Automating the multimodal analysis of musculoskeletal imaging in the presence of hip implants

Marta Bianca Maria Ranzini

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

Research Degree: Medical Imaging [RRDMEISING01]

University College London

November 17, 2020

I, Marta Bianca Maria Ranzini, 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.

A Beppe.

Mi piace pensare che ti sei fatto dare un passaggio
dalla cometa che passava di qui in quei giorni.

Abstract

In patients treated with hip arthroplasty, the muscular condition and presence of inflammatory reactions are assessed using magnetic resonance imaging (MRI). As MRI lacks contrast for bony structures, computed tomography (CT) is preferred for clinical evaluation of bone tissue and orthopaedic surgical planning. Combining the complementary information of MRI and CT could improve current clinical practice for diagnosis, monitoring and treatment planning. In particular, the different contrast of these modalities could help better quantify the presence of fatty infiltration to characterise muscular condition after hip replacement.

In this thesis, I developed automated processing tools for the joint analysis of CT and MR images of patients with hip implants. In order to combine the multimodal information, a novel nonlinear registration algorithm was introduced, which imposes rigidity constraints on bony structures to ensure realistic deformation. I implemented and thoroughly validated a fully automated framework for the multimodal segmentation of healthy and pathological musculoskeletal structures, as well as implants. This framework combines the proposed registration algorithm with tailored image quality enhancement techniques and a multi-atlas-based segmentation approach, providing robustness against the large population anatomical variability and the presence of noise and artefacts in the images.

The automation of muscle segmentation enabled the derivation of a measure of fatty infiltration, the Intramuscular Fat Fraction, useful to characterise the presence of muscle atrophy. The proposed imaging biomarker was shown to

strongly correlate with the atrophy radiological score currently used in clinical practice.

Finally, a preliminary work on multimodal metal artefact reduction, using an unsupervised deep learning strategy, showed promise for improving the post-processing of CT and MR images heavily corrupted by metal artefact.

This work represents a step forward towards the automation of image analysis in hip arthroplasty, supporting and quantitatively informing the decision-making process about patient's management.

Impact Statement

This work developed automated processing tools for the analysis of Computed Tomography (CT) and Magnetic Resonance (MR) images of patients with hip implants. It showed the potential benefit of combining multimodal information to obtain patient-specific anatomy and quantitative assessment of muscular condition after hip replacement.

From a clinical perspective, this work represents a step forward towards personalised treatment in the orthopaedic field. First, by providing the automated segmentation of volumetric images, it reduces the burden of manual segmentation for musculoskeletal radiologists. As a result, the volumetric information these images provide can be fully exploited both in terms of volumetric rendering of the patient's anatomy and in terms of deriving quantitative metrics from the regions of interest. Such metrics could help better characterise the clinical condition of the patient, for instance in relation to a healthy population. In addition, a three-dimensional visualisation of all musculoskeletal structures could inform orthopaedic surgeons on the need for revision surgery, the choice of surgical approach and the extent of muscle damage associated with the implant. It also opens up new clinical research questions. For instance, could the proposed imaging biomarkers for muscle atrophy be employed in a longitudinal study to support the monitoring of patients with hip implants? And also, are the benefits of a multimodal analysis counterbalancing the costs of acquiring multiple scans? Such questions could be addressed by the development of clinical studies which take advantage of the tools here developed.

Further methodological research could also stem from the presented work.

The advance of deep learning technologies is revolutionising healthcare and specifically the medical imaging field. However, these approaches often rely on the availability of large annotated datasets, which are difficult and labour-intensive to construct in the medical field. The proposed framework could facilitate the application of deep learning in the musculoskeletal imaging field by speeding up the generation of such datasets, as it would provide a means to automatically segment regions of interest and thus generate annotated data. Moreover, this work represents a baseline for the combination of CT and MR information in the case of heavily corrupted images, and fosters further research for the synthesis of missing modalities: can we artificially generate a CT from the MRI of a patient and vice versa? This question is particularly relevant in hip imaging, since the availability of synthetic CT images would spare gonad exposure to ionising radiation associated with CT scanning.

In conclusion, most of the work discussed in this thesis has potential to improve the clinical workflow in hip replacement patient's management, by supporting clinicians with patient-specific information and by promoting further research towards a clearer understanding of implant failure mechanisms.

Acknowledgements

First and foremost, I would like to thank my supervisor Dr Marc Modat, for constantly supporting me both professionally and personally throughout the PhD journey, and for being a person that really cares for others. I am also grateful to Prof Sébastien Ourselin for giving me this great opportunity, and for constantly making sure I was happy at work. I would like to sincerely thank Prof Alister Hart and Dr Johann Henckel for their insights on the clinical aspects of the project, for giving me the opportunity to bring some of my work to the patients and for their valuable advice during the PhD.
Special thanks also go to all my co-authors, in particular to Prof Tom Vercauteren, Dr Michael Ebner, David Drobny, Irme Groothuis and Kerstin Kläser for their responsiveness and for being a pleasure to work with. A very special mention is for Dr Erika Molteni, who really has been a role model to me and who taught me so much about research - and about life.

Despite the many challenges, this journey brought me to meet some of the most brilliant minds and the most precious friends. I would like to thank all the Cool Kids for always been fun and supportive, and all the friends from my CDT year, especially Bianca, Nooshin, Irina, Maria, Danny and Ruaridh. I am also thankful to my friends Vale, Fra, Gabri, Pol and Ste, for always being there for me, even when I don't deserve it.
I owe my deepest gratitude to Maddy and to my large, messy but wonderful family with all aunts, uncles and cousins, whose support meant the world to me. Above all, I would like to thank my parents for keeping strong, for being so patient with me and for teaching me resilience.

Lastly, the greatest thank you is for Carlo: this journey made sense only because I shared it with you. Thank you for being with me and guiding me through the good and bad times.

Vorrei innanzitutto ringraziare il mio relatore Dr Marc Modat, per il suo supporto costante sia professionalmente che personalmente, e per essere una persona che davvero tiene agli altri. Ringrazio anche il Prof Sébastien Ourselin per avermi dato questa opportunità e per il continuo interesse al mio benessere. Vorrei ringraziare sinceramente il Prof Alister Hart e il Dr Johann Henckel per il loro contributo clinico al progetto, per avermi dato l'opportunità di presentare parte del mio lavoro ai pazienti e per il loro sostegno durante il PhD. Un ringraziamento speciale va a tutti i miei co-autori, in particolar modo al Prof Tom Vercauteren, al Dr Michael Ebner, a David Drobny, Irme Groothuis e Kerstin Kläser, con cui è stato un piacere lavorare. Vorrei ringraziare specialmente la Dr.ssa Erika Molteni, che è stata per me un modello da seguire, mi ha insegnato tanto sulla ricerca - e sulla vita.

Nonostante tutte le sfide, questo percorso mi ha permesso di conoscere menti brillanti e amici molto preziosi. Vorrei ringraziare tutti i Cool Kids per essere sempre divertenti e di grande aiuto; vorrei ringraziare anche gli amici del CDT, specialmente Bianca, Nooshin, Irina, Maria, Danny e Rory. Un enorme grazie ai miei amici Vale, Fra, Gabri, Pol e Ste, per esserci sempre per me, anche quando non lo merito.

Un grazie profondissimo va alla Maddy e alla mia gigante, incasinata ma meravigliosa famiglia, con tutti gli zii, le zie, i cugini, il cui sostegno è stato fondamentale. Soprattutto, vorrei ringraziare i miei genitori, per aver tenuto duro, per essere stati cosí pazienti con me e per avermi insegnato a stringere i denti.

Infine, il grazie più grande va a Carlo: questo percorso ha avuto senso solo perché l'ho condiviso con te. Grazie per essere al mio fianco e per guidarmi in tutti i momenti, bello o brutti che siano.

Contents

List of Figures

List of Tables

List of Abbreviations

2D	Two-dimensional
3D	Three-dimensional
4D	Four-dimensional
AAM	Active Appearance Model
ADN	Artefact Disentanglement Network
AE	Absolute Error
ANOVA	Analysis of Variance
ASM	Active Shape Model
BE	Bending Energy
CNN	Convolutional Neural Network
CSA	Cross-Sectional Area
CT	Computed Tomography
DS	Dice Score
EM	Expectation-Maximisation
FF	Fat Fraction
FFD	Free-Form Deformation
FOV	Field-of-View
GAN	Generative Adversarial network
GMAX	Gluteus Maximus
GMED	Gluteus Medius
GMIN	Gluteus Minimus
GMM	Gaussian Mixture Model
GT	Ground Truth
HRA	Hip Resurfacing Arthroplasty
HU	Hounsfield Unit
IMAT	Intermuscular Adipose Tissue
IFF	Intramuscular Fat Fraction
IoU	Intersection over Union
LNCC	Locally Normalised Cross Correlation
LOOCV	Leave-One-Out Cross Validation
LTS	Least Trimmed Squares
MADN	Multimodal Artefact Disentanglement

	Network
MAR	Metal Artefact Reduction
MARS	Metal Artefact Reduction Sequence
MHRA	Medicines and Healthcare products Regulatory Agency
MoM	Metal-on-Metal
MR	Magnetic Resonance
MRF	Markov Random Field
MRI	Magnetic Resonance Imaging
MSK	Musculoskeletal
NMAR	Normalised Metal Artefact Reduction
NMI	Normalised Mutual Information
NN	Neural Network
PCA	Principal Component Analysis
PET	Positron Emission Tomography
RMAR	Refined Metal Artefact Reduction
ROI	Region of Interest
SAM	Statistical Appearance Model
SAT	Subcutaneous Adipose Tissue
SNR	Signal-to-Noise Ratio
SRR	Super-Resolution Reconstruction
SSM	Statistical Shape Model
STEPS	Similarity and Truth Estimation for Propagated Segmentations
TE	Time to Echo
TFL	Tensor Fasciae Latae
THR	Total Hip Replacement
TRE	Target Registration Error
TR	Repetition Time
TSE	Turbo Spin-Echo
US	Ultrasound
UTE	Ultra-short Time-Echo
vs	versus
ZTE	Zero-Time-Echo

Chapter 1

Medical imaging for hip arthroplasty

Medical imaging is increasingly becoming a powerful tool to help in diagnosis, treatment and management of patients in all clinical fields, thanks to the rapid development of novel imaging techniques and automated analysis tools. However, in the musculoskeletal (MSK) field, advanced imaging analysis methods struggle to translate to clinical practice, due to either lack of generalisability or difficulty in tailoring musculoskeletal models to patient-specific anatomy. As a result, current decisions for patient management in the MSK field often rely on clinical evaluation along with only a qualitative assessment of medical imaging examinations.

As they mostly limit the range of motion, MSK conditions have a strong impact on life quality and on the ability to perform basic daily tasks. Taking into account the increase of both life expectancy and prevalence of risk factors such as obesity and sedentary lifestyle, MSK problems are now recognized as a primary burden on both individuals and health systems. The World Health Organization, for instance, indicates the development of technologies to facilitate monitoring and delivering of clinical services in MSK health as a research priority [1]. In the UK, MSK disorders represent the leading cause of pain and disability. In November 2017 the government launched a *Muscu-*

loskeletal Diseases Profile tool[1], with the aim of summarising key MSK data to promote further understanding and improvement of the health service. They later showed that in 2018 up to 17% of the population in England reported a long term MSK problem. Considering the UK data only, 15-20% of general practitioner consultations relate to MSK problems [2]. Therefore automated methods for MSK medical image analysis are needed to aid prevention, diagnosis and treatment in clinical practice, leading to advancements in the service provided.

In this scenario, this project focuses on improving the clinical workflow for medical image analysis in the context of hip arthroplasty surgery.

1.1 Hip arthroplasty

Hip arthroplasty is the second most common surgical procedure after hernia repair [3], and consists of replacing a pathological hip joint with a prosthesis. This joint is composed of the femoral head and the acetabulum (a round cavity in the pelvis), and it is surrounded by several ligaments and robust muscles. Being one of the largest joints in the human body, it provides both stability and great flexibility, thanks to its *ball-and-socket* configuration, where the femoral head (the ball) moves freely into the acetabulum (the socket). The principal cause of hip joint degeneration is osteoarthritis (91.9% of diagnoses in the UK [4]), which determines pain and stiffness. According to the *Musculoskeletal Disease Profile*, 10.9% of the population aged 45 and over are affected by hip osteoarthritis, making this a major public health issue.

To restore the normal functioning of the joint, a hip arthroplasty procedure called total hip replacement (THR) is typically performed: both the acetabulum and the femoral head are replaced by a modular implant inserted into the femur through a stem (Fig. 1.1). In younger patients, a more bone-preserving alternative procedure called hip resurfacing arthroplasty (HRA) is sometimes preferred, where a resurfacing cup is applied to reshape the femoral head

[1] `https://fingertips.phe.org.uk/profile/msk`

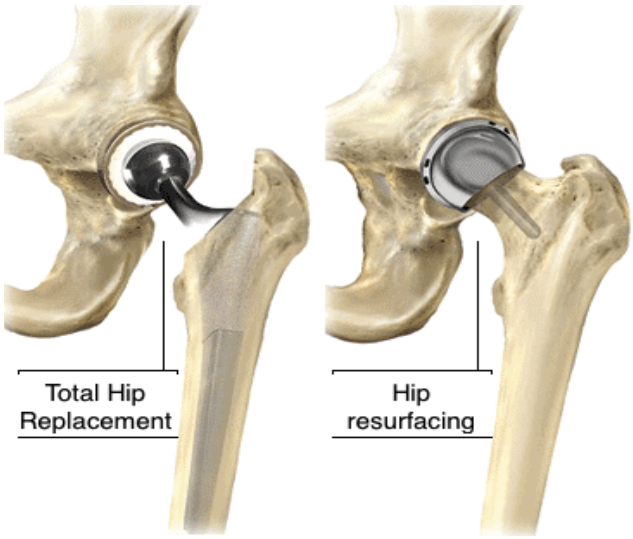


Figure 1.1: Main types of hip arthroplasty procedures.
Image from `http://www.mydepuyhiprecall.com/2013/05/the-depuy-hip-litigation-guide.html`.

only [5]. In the UK, the National Joint Registry recorded almost 1 million hip arthroplasty procedures since 2003, with about 92000 performed in 2017 alone [4]. The reported age range of the patients spanned between 7 to 105 years (median age: 69), and higher prevalence was noted among women, who represented approximately 60% of the cases. Overall this type of interventions reports successful outcomes in alleviating symptoms and improving life quality, with a decrease from 92.0% before surgery to only 17.8% of patients indicating moderate or severe pain after surgery [6]. However, commonly used implants are associated with a non negligible failure rate - 7.27% at 14 years from primary surgery according to the UK National Joint Registry [4] - and consequent need to replace the implant in revision surgery. Aseptic loosening, fracture, dislocation, adverse reaction to particulate debris and infection are the most frequent indications to justify revision. Interestingly, revision rates are found to increase with lower age at primary surgery. For instance, male patients below 55 are associated to a risk of revision surgery of 16% at 14 years from the primary one, compared to only 2.5% for men aged 75 and above. Overall, the failure rate increases up to 19-22% when Metal-on-Metal (MoM) implants have been utilised in primary surgery, this type of prosthesis being

linked to adverse inflammatory reactions and muscular wastage [7, 8]. The former induce either solid or cystic inflammatory masses in the periprosthetic tissue, called *pseudotumours*, and can determine nerves or tendons compression and consequent pain [9, 10]. Muscular wastage or atrophy manifests as a reduction of muscular mass, both as volume shrinkage and as increased fatty infiltration, which weakens the muscle. Hip abductor muscles are most concerned by atrophy, with prevalence ranging between 22% and 90% of the cases depending on the muscle [8].

Given the established risk associated with MoM implants, their use has been dramatically reduced after 2012, and formal guidelines for the management of implanted patients have been released and updated by the Medicines and Healthcare products Regulatory Agency (MHRA) in a Medical Device Alert [11] since then. Their recommendations include regular blood metal ion level tests to quantify metal debris (in particular to measure the concentration of Cobalt and Chromium ions), as well as the use of Magnetic Resonance Imaging (MRI) or Ultrasounds (US) to assess soft tissue conditions.

Medical imaging plays therefore an important role in assisting and monitoring patients with hip implants. Moreover, understanding the mechanisms of implant failure is an active research area: medical imaging, coupled with failed implant retrieval analysis [10], can shed new light on these mechanisms and strongly impact the surgical outcomes.

1.2 Medical imaging for hip arthroplasty

Medical imaging is involved throughout the hip arthroplasty clinical workflow for different tasks (Fig. 1.2). Both in pre-operative and post-operative settings, radiography and computed tomography (CT) imaging are the routine modalities to assess the bone stock, to plan surgery and to verify delivery, given their optimal contrast for osseous structures [12]. Compared to radiography, the three-dimensional (3D) nature of CT acquisition makes it most suitable to assess the correct positioning of the implant, by measuring the acetabu-

lar inclination and version angle with respect to the anterior pelvic plane. In the case of MoM patients, CT-measured implant position already proved to correlate with increased metal ions levels in blood [13]. In particular, cobalt and chromium levels were found to sharply rise with cup inclination greater than 45° and with insufficient version ($< 20°$). CT also provides clearer visualisation of prosthesis dislocation or fracture, and it facilitates the detection of osteolysis [14]. These features combined with its fast and relatively cheap acquisition make CT a useful diagnostic tool for bone and implant visualisation. However, its use is associated with radiation exposure to the gonads and consequent higher risk of developing cancer. This risk limits its repeated applicability for monitoring purposes, especially for younger women in fertile age [15].

For soft tissue visualisation Magnetic Resonance Imaging (MRI) and Ultrasounds (US) are generally preferred if not explicitly required as follow-up measures [11]. Thanks to its outstanding soft tissue contrast, MRI already proved to be the most suitable modality to assess muscular conditions on suspicion of inflammation or muscle atrophy [16]. Due to the presence of metal in most of the implants, the Metal Artefact Reduction Sequence (MARS) MRI is frequently adopted to reduce the susceptibility artefacts around the implant and consequently improve the image quality. The MARS MRI is a modification of the Turbo Spin-Echo (TSE) sequence, where an intensified slice-select gradient and a smaller flip angle are coupled with a broader bandwidth to enhance the imaging gradients over the susceptibility-induced ones, however at the expense of a reduced Signal-to-Noise Ratio (SNR) [17]. Although the effect of the metal artefact is reduced through this sequence, the implant still remains completely obscured, making any evaluation on its position or on the bone-implant interface impossible on the MR image [14]. In general, MR requires long acquisition times and lacks signal for cortical bone, which hampers accurate identification of bone boundaries, especially in the joint space. Bone MR sequences such as Zero-Time-Echo (ZTE) MRI are a promising solution to this problem [18], but

at present they are still at research stage, most of these sequences requiring high field scanners ($B_0 > 3$T), and have not been tested in presence of implants yet. Nonetheless, MRI is increasingly becoming a popular modality in MSK and specifically in hip arthroplasty due to its excellent soft tissue contrast and its non-ionising nature.

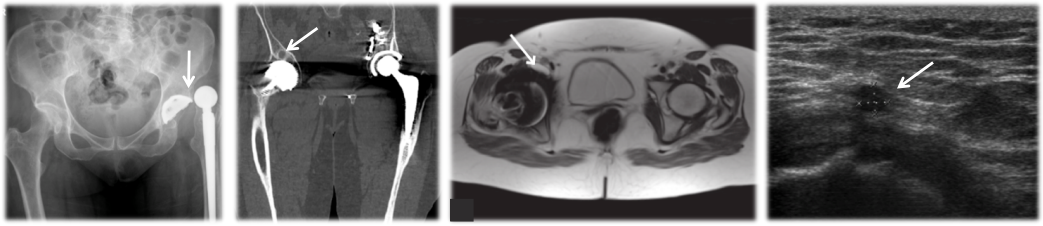


Figure 1.2: Examples of medical imaging for hip arthroplasty. White arrows indicate regions of clinical interest. (a) X-ray image[1] showing a dislocated implant. (b) CT image[1] reporting osteolysis of pelvic bones. (c) MARS MR image[2] presenting a fluid-like soft-tissue lesion. (d) US image[3] showing the presence of a pseudotumour.

1.3 Current challenges in MSK imaging

Despite the relevant role of medical imaging for the diagnosis, treatment and monitoring of patients in orthopaedics, the development and application of advanced automated imaging analysis techniques are lagging behind other clinical fields. The main challenges in MSK are the lack of standardised acquisition protocols, the large inter-subject variability of the structures of interest, and their typically extended size, often requiring large and highly variable imaging Field-of-View (FOV) or patient's pose across different acquisitions [20, 21].

In the specific case of hip arthroplasty, one additional major challenge is the presence of strong metal-artefact induced noise in the images. The artefacts greatly degrade the image quality, often covering the anatomical features of interest. In CT, the higher X-ray attenuation of the metallic prosthesis with respect to biological tissue determines photon starvation at the signal detectors. This results in darker and brighter streaks (star artefacts) crossing the full field-

[1]Image from `http://www.complexhipsurgery.com/metal-debris-disease`
[2]Image from Robinson *et al.*[14]
[3]Image from Matharu *et al.*[19]

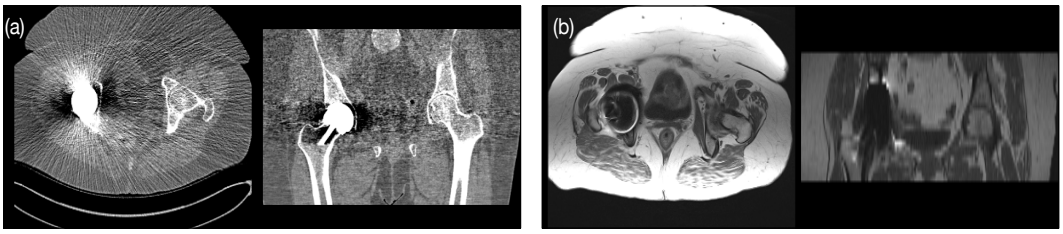


Figure 1.3: Appearance of metal-artefact-induced noise in a CT image (a) and in a MARS MR image (b).

of-view in the axial plane, also affecting regions far away from the implants (Fig. 1.3(a)). Being widespread, the artefacts hamper the clear identification of different organ boundaries, especially between different muscles, potentially reducing the diagnostic confidence [22]. By contrast, in MRI the metal-induced artefact appears as a dark irregular shadow more localised around the implant. The shadow is originated by the absolute lack of detectable signal from the metal object, combined with susceptibility variations between metal and biological tissue that produce local modification of the static magnetic field, often leading to frequency shift, further signal loss and geometrical image distortion [23]. Therefore, the exact position of the implant, as well as its interface with the bone stock are not visible in MRI, and the muscles closer to the prosthesis might be partially shaded or distorted (Fig. 1.3(b)). Moreover, fat-suppression or fat-enhancement MRI sequences such as DIXON might not be effective in the presence of metal, as they rely on the ability to identify the narrow frequency separation between fat signal and water signal [23]. Thus, the quantification of fatty infiltration within muscular structures might be hindered as well. Combined together, these aspects strongly limit the generalisability and applicability of existing methods for image analysis in hip arthroplasty. As a result, current decisions for patient post-surgical management often rely only on separate, subjective and qualitative assessments of imaging data. For instance, the evaluation of muscle atrophy in clinical practice is based on visual slice-by-slice inspection and qualitative scoring of the total fatty infiltration percentage (e.g. less than 20% or more than 70% in the whole muscular volume) [24]. This type of assessment does not account for the three-dimensional

nature of the anatomy, discarding the volumetric information embedded in 3D imaging data.

Finally, as mentioned CT and MRI yield complementary information - bones and implant the former, soft tissue optimal visualisation the latter. However, they are assessed independently from each other and seldom integrated for improved and more complete clinical picture of all MSK structures jointly and relatively to the implant.

1.4 Research hypotheses and objectives

In this context, this project is based on the research hypothesis that the current clinical workflow for hip arthroplasty could be enhanced by improving the processing, use and analysis of the images clinically available. It is hypothesised that a unique framework combining the complementary information of CT and MRI could support the diagnosis and treatment of implant failure, as well as monitoring of well-functioning implants. In particular, such framework could have clinical value for imaging biomarkers extraction and surgical planning. It would help better delineate implants, muscular and skeletal structures concurrently and their relative spatial localisation, towards a more accurate definition and visualisation of patient-specific anatomy. This would benefit the customisation of surgical planning to minimise the damage to healthy MSK tissue and potentially result in longer-lasting implants after revision, especially relevant for younger patients [4]. Moreover, a more accurate muscle damage analysis could take advantage of the richer muscular information from MRI and the standardisation of CT intensities to better quantify fatty infiltrations, together with the clearer implant localisation the latter provides. This could help identify the muscles at greater risk of developing atrophy and therefore improve implant failure assessment.

In the direction of improving the clinical workflow for hip arthroplasty, the aims of my project are twofold:

1. To develop tools for the combination of CT and MRI complementary

MSK information, employing state-of-art and novel approaches that could ease and automate the image analysis and that could be translated to real clinical settings.

2. To develop methods for the automatic extraction of quantitative imaging biomarkers to help the characterisation of pain origin and the assessment of implant failure.

To this end, I will employ images that have been collected retrospectively for routine clinical practice, and will develop tailored information processing tools in the fields of image processing, image registration (*i.e.* the process of aligning same anatomical structures in different images) and segmentation (*i.e.* the delineation and classification of regions of interest).

1.4.1 Contributions

The first main contribution of this work is the development of a novel composite framework for the automated joint segmentation of pelvic CT and MR images in the presence of hip Metal-on-Metal implants. The proposed pipeline is designed to handle low-quality clinical images, characterised by highly anisotropic resolution and strongly affected by metal artefact induced-noise. One key aspect of the pipeline is the introduction of a novel non-linear diffeomorphic registration algorithm that preserves the local rigidity of bones, allowing for a robust intra-subject multi-modality alignment. The automated segmentation is inclusive of relevant bones, muscles and implants, and it is obtained through a multi-channel multi-atlas segmentation propagation approach. This latter combines information from both CT and MRI and employs a new atlas of implanted hips I built for this purpose, providing robustness against the large population variability of MSK structures. The derived fused segmentation can be overlaid on both imaging modalities, highlighting also those structures which the single modalities lack contrast for (e.g. the different abductor muscles in CT, the bones and the implants in the MR). It can also be employed for patient-specific volume rendering, which helps the 3D visualisation of the

relative spatial positions of the considered anatomical structures.
The obtained segmentation provides an automated means for definition of regions of interest, from where quantitative imaging biomarkers can be extracted. Towards this aim, the second contribution of this work is the derivation of a novel automated quantification of fatty infiltration, the Intramuscular Fat Fraction (IFF), which measures the percentage of fat in muscular volumes. The proposed derivation takes advantage of both imaging modalities and shows potential as a quantitative imaging biomarker of muscle atrophy.
Finally, I introduced a novel data-driven approach to metal artefact reduction to further improve the image quality in both CT and MRI and facilitate downstream analysis. As the appearance of the artefact differs in the two modalities, the proposed method makes use of the contextual information to simultaneously correct missing data in both CT and MRI.

1.5 Thesis organisation

This thesis is organised as follows. Chapter 2 presents a review of the literature, discussing the current state-of-the-art approaches for medical image information processing in the MSK field. Special focus is kept on current quantification methods of muscle atrophy, pelvic CT and MRI registration and segmentation approaches, as well as techniques for metal artefact reduction in these modalities. After presenting the dataset used for this thesis project in Chapter 3, I will introduce in Chapter 4 a novel registration algorithm I proposed to improve the alignment of CT and MR images, which allows for non-linear deformation of soft tissue while preserving the rigidity of bony structures. This chapter also presents the experiments to assess the accuracy and the robustness of the proposed algorithm. Chapter 5 introduces the developed pipeline for the joint automated segmentation of clinical CT and MRI from hip arthroplasty patients. A validation analysis is reported in comparison with single modality approaches, and the effects of the inaccuracies from the previously discussed registration algorithm are also analysed. In Chapter 6, I will introduce a novel

automated method for quantification of fatty infiltration using both CT and MRI, which is benchmarked against the current clinical evaluation of muscle atrophy. The last contribution of this thesis is reported in Chapter 7, where a deep learning multimodal framework for metal artefact reduction is presented, and the benefits of fusing the CT and MRI information for this task are analysed. Finally in Chapter 8 a summary of the obtained results and future trends for medical imaging in hip arthroplasty are discussed.

Chapter 2

Medical image analysis in MSK

Medical image analysis refers to all techniques for correcting, manipulating and improving images after acquisition in order to extract meaningful information and facilitate their interpretation. Among these techniques, registration and segmentation tools play a fundamental role in image processing pipelines. The former allow to identify anatomical correspondences between different images, which therefore guarantees reliable comparisons in both longitudinal or multimodal studies - where different time points acquisitions or different imaging modality scans from the same subject are compared - and in cross-sectional studies - where inter-subject comparisons are performed. Segmentation techniques help delineate specific regions of interest (ROIs) within the images, which are needed for further quantitative and/or qualitative analysis (e.g. volume or shape analysis, texture analysis). This often results in the derivation of summative metrics, called imaging biomarkers, that are relevant for the characterisation of a physiological or a pathological condition within the ROI.
Despite the efforts of the medical imaging research community towards the automation of image processing and quantitative biomarkers extraction, the full translation of such tools to clinical practice is still far from being reached. In the MSK field, the main bottleneck is the lack of robust and generalisable segmentation techniques for automated ROI definition [20]. As a result, though cumbersome and time-consuming, manual ROI definition is often considered as the only reliable solution for imaging biomarker extractions.

In this chapter, I will firstly present the current imaging biomarkers used in the clinical literature for the assessment of muscle wastage in hip arthroplasty. Then, I will discuss the problem of automated ROI definition in MSK, and state-of-the-art segmentation and registration methods attempts to address it. Finally, I will focus on the problem of metal artefact induced noise in medical imaging for hip arthroplasty and most recent approaches to reduce its impact on medical image analyses.

2.1 Imaging biomarkers for hip arthroplasty

An imaging biomarker is defined as an "anatomic, physiologic, biochemical, or molecular parameter" that can be detected or measured from a medical image to determine the presence or the severity of a pathology [25].
In the context of hip arthroplasty, an important factor to establishing implant failure is the development of muscle atrophy. Atrophy manifests as severe muscle wastage, and can be assessed on MR images in terms of muscular volume shrinkage and by quantification of the fatty infiltration within the muscle (intramuscular fat) [26]. Such quantities can therefore represent useful imaging biomarkers to help diagnosis and patient monitoring.

2.1.1 Current clinical assessment

The quantification of muscle atrophy from MRI is currently performed using a single-slice scoring system, which provides a visual estimation of the percentage of fat infiltration into the muscular volume. The most commonly adopted is the Pfirrmann grading system [24], which consists of assessing axial T1-weighted MR images at one-third and two-thirds of the distance between the greater trochanter tip and the iliac crest (see Figure 2.1 for anatomical reference). A similar approach was introduced by Bal and Lowe [27]: visual scores are assigned for muscular volume and fatty change with respect to the asymptomatic hip side. The classification description of the two scoring scales is reported for reference in Table 2.1, showing the qualitative and subjective nature of such systems.

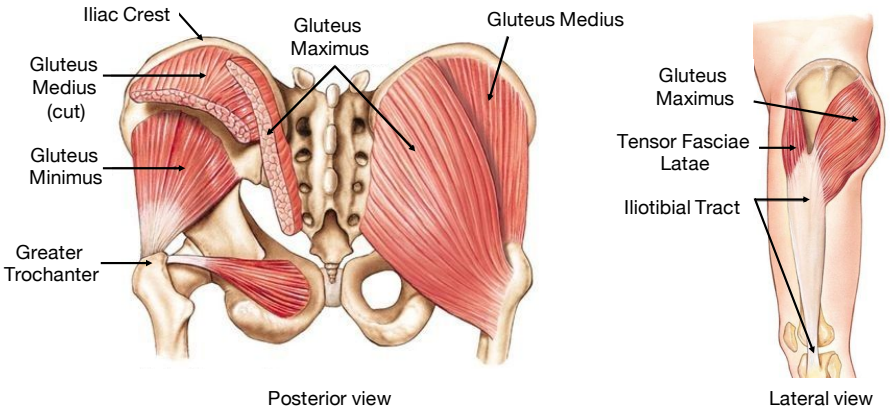


Figure 2.1: Anatomical representation of the abductor muscles considered in this work. The greater trochanter and the iliac crest are also shown as typical skeletal references for the assessment of muscle atrophy [24]. Image adapted from `https://anatomyinfo.com/gluteal-muscles/`.

Table 2.1: Comparison of the currently used radiological scores to rate muscle atrophy in terms of size changes and fatty infiltration.

	Pfirrmann score [24]	**Bal & Lowe score** (as reported by [26])
Grade 0	No intramuscular fat	Normal
Grade 1	Some fat streaks present	Not exceeding 30% decrease in muscle size
Grade 2	Evident presence of fat, but less than muscle tissue	Decrease muscle size and 30%-70% fatty infilitration
Grade 3	Same amount of fat and muscle	>80% muscle size decrease, >70% fatty infiltration
Grade 4	More fat than muscle	-

These and similar scores are applied in other muscular structures as well, however they are strongly rater-dependent and discard all the volumetric information that the image provides. For instance, in the case of the roteator cuff muscle, Vidt *et al.* [28] showed that such 2D assessment is not representative of the respective 3-dimensional measures. An exemplar case for gluteus maximus classification is shown in Figure 2.2, demonstrating the subjectivity of such scoring systems, characterised by large inter- and intra-rater variability. A more objective volumetric quantification of atrophy is therefore required to

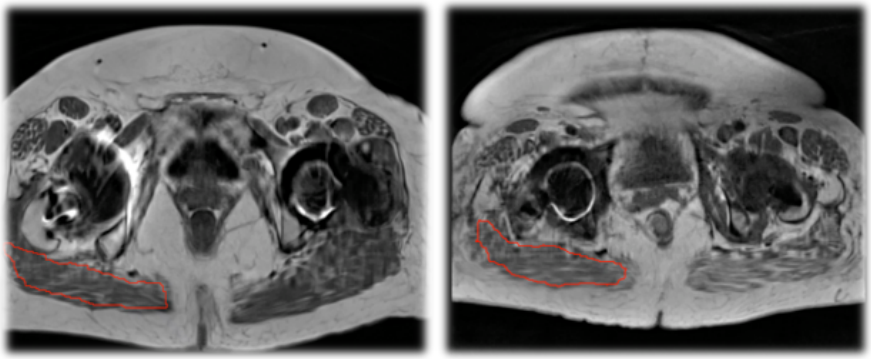


Figure 2.2: Example of muscle atrophy scoring for the gluteus maximus, delineated by the red contour. Although the pattern of fatty infiltration appears similar in both cases, the same rater classified subject (a) as grade 3 ($>$ 70% fatty infiltration and reduction in muscle size) and subject (b) as grade 2 (30$-$70% fatty infiltration and reduction in size) according to Bal and Lowe scale [27].

drive the characterisation of muscular condition and help the surgeons in the management of implanted patients.

2.1.2 Quantitative imaging biomarkers

The availability of methods to segment a whole muscle as ROI allows to derive volumetric quantitative biomarkers. Most methods to quantify muscular wastage have been focussing on two measures: the gross muscle volume and the amount of intramuscular fat. Apart from where explicitly stated, all the methods presented next for muscle image biomarkers extraction are based on manual segmentation of the gross muscular volume to define the ROI.

Muscular volume assessment

The quantitative assessment of muscular volume has already proved its relevance in many MSK diseases such as neuromuscular disorders, muscular dystrophies, inflammatory myopathies, sport injuries and sarcopenia [29]. In its most basic form, it is simply computed as the voxel volume in millimetres times the number of voxels within a specific ROI. A surrogate metric that is commonly used in clinical routine is the Cross-Sectional Area (CSA), measured as the muscle area at one or few specific 2D slices of the whole volume. While being a fast and easy measure, it is scarcely reproducible and not representa-

tive of the full volume [30].

Actual 3D measurements from full-volume muscle segmentation were reported by Grimaldi *et al.* [31, 32] in the context of hip joint pathology analysis. With respect to their control cohort, in the case cohort they identified significant volume reduction between the symptomatic and the asymptomatic hip side for the gluteus maximus (GMAX), the gluteus medius (GMED) and the piriformis muscles. A statistically non-significant trend for volume shrinkage of the gluteus minimus (GMIN) was found as well, while the tensor fasciae latae (TFL) showed no significant asymmetry. Similar analysis and conclusions were obtained by Skorupska *et al.* [30] in the study of the low back with leg pain, where more than 50% of the diseased subjects presented smaller volume for the GMAX, the GMIN and the piriformis.

While the cited work required manual segmentation of the muscular volume, fully automated segmentation and quantification was proposed by Karlsson *et al.* [29] and by LeTroter *et al.* [33], both employing an atlas-based approach to separate different muscles (see Sec. 2.2 for a more detailed description of the segmentation methods). The former took advantage of a specific MRI acquisition, the two-point Dixon chemical-shift protocol, to acquire water-fat separated MR images and correct the muscle volumes for fat infiltration, reporting good agreement with the volumes obtained from the manual ground truth segmentation. The method was validated on healthy subjects only and thus no clinical relevance of the muscular volume in pathological conditions was reported. The latter compared a semi-automated single-atlas and a fully-automated multi-atlas approaches to obtain volume estimation of the quadratus femoris in healthy subjects. The volume was computed by the truncated cone formula, which assumes an inter-slices conical shape of the muscles [34]. Interestingly, in the case of longitudinal analysis, the authors reported a better agreement between the volume from manual segmentation and the semi-automated single-atlas approach, compared to the multi-atlas one. They suggested that, although time-consuming, performing manual segmentation on

the first time-point and then use it to automatically segment subsequent time-points should be preferred to a fully-automated method.

Fat infiltration

As already introduced in Sec. 2.1.1, current measures of fatty infiltrations are qualitative visual estimations at few characteristic slices [24, 35]. However, severe fatty muscular degeneration is associated with poor functional outcome [35], leading to a need for defining standardised ROI selection and more accurate intramuscular fat quantifications [36, 37].

Most state-of-the-art methods are based on tailored MR acquisition with enhanced contrast for fat signal, such as chemical-shift MR sequences [38, 39]. Given a manually selected ROI, the fat infiltration is quantified as fat fraction (FF):

$$FF = \frac{S_f}{S_f + S_w} \times 100 \tag{2.1}$$

where S_f and S_w are the total signal from fat and the total signal from water in the considered volume. An exemplar image of fat segmentation for FF estimation is shown in Figure 2.3. The relevance of such estimation of the intramuscular fat was recently demonstrated in a longitudinal clinical study reported by Morrow *et al.* [39], where FF was found significantly correlated with reduced thigh and calf muscle strength and with other clinical assessment scores in patients with neuromuscular diseases. However, this type of MR acquisition is limited by longer scan time and by increased noise and confounding factors [38]. Moreover, it is very sensitive to magnetic field inhomogeneities, and therefore unsuitable in the presence of metal implants [23].

An Ultra-short Time Echo (UTE) sequence was employed by Ugarte *et al.* [40], who introduced a novel automated segmentation of fat within a manually delineated ROI. This method relies on a fuzzy c-means clustering of the intensities, with the addition of spatial connectedness constraints to enforce a filament-like structure to the regions classified as fat. While providing a volumetric map of the fatty infiltration, no clear quantification is however proposed nor benchmarked for pathological conditions. An automated fat fraction quan-

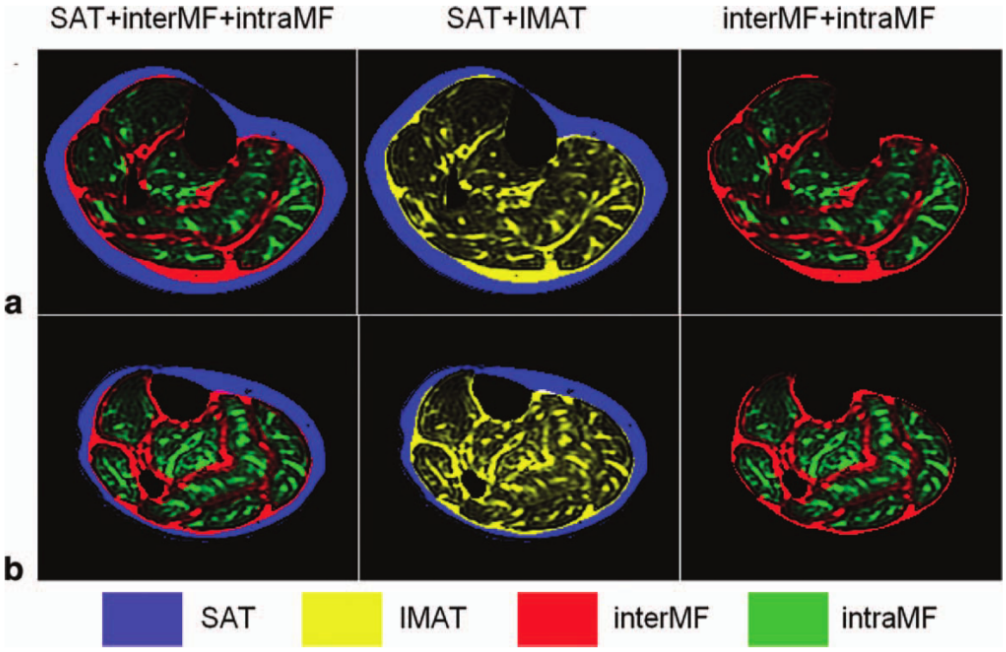


Figure 2.3: Segmentation of fat in MR derived from water-fat separated imaging of the calf musculature. The subcutaneous adipose tissue (SAT), the inter- and intra-muscular fat are highlighted. The intermuscular adipose tissue (IMAT) is simply the union of inter- and intra-muscular fat. Image from Karampinos *et al.* [38].

tification derived from standard T1-weighted MR images was proposed by Lareau-Trudel *et al.* [41] for patients with facioscapulohumeral muscular dystrophy type 1. A k-means clustering algorithm was first applied to classify each voxel as background, adipose tissue or muscle tissue. An active-contour segmentation algorithm was then applied to separate the subcutaneous adipose tissue from the muscle boundary. By combining this with the previous classification the percentage of fat within muscular volume could be obtained. They found signficant correlation between their automated fat fraction measure and a 4-point visual scale of fatty infiltration, and even higher correlation when the fat fraction was converted to a logarithmic scale. While fully automated, they still reported a substantial failure rate, with 20% of the cases requiring manual adjustments in the segmentation. More recently, Gadermayr *et al.* [37] compared different automated and semi-automated methods to define the ROI,

and from each of them they calculated the fat fraction as:

$$FF_{\gamma}(s) = \int_{\gamma}^{s_{max}} h(s)ds \tag{2.2}$$

where $h(s)$ is the normalised intensity histogram of volume s, with highest intensity being s_{max}, while γ is a manually selected threshold to separate muscular and fat intensities. They also introduced a metric to quantify the error on FF estimation derived from the automated ROI as the maximum L1 distance from manual ROI-derived FF, for a specified threshold γ.

Finally, although most fat fraction quantification methods have been developed from MRI, it is worth mentioning a CT-based analysis proposed by Momose *et al.* [42]. They used volume and mean CT radiodensity information of manually segmented Gluteus Medius as a measure of fatty degeneration in patients with hip osteoarthritis. The 3D volumetric assessment was once again shown to more accurately reflect muscle strength compared to CSA measures. The CT radiodensity was found significantly lower in the pathological hip side when compared to the contralateral healthy side. More relevant, it significantly correlated with a reduction of hip abductor strength, clinically measured with a dynamometer.

2.1.3 Current challenges and open questions

As presented in the previous section, a variety of definitions and techniques have been investigated, but none has yet defined a standardised and robust approach for the estimation of muscle atrophy. To the best of my knowledge, no method explicitly tackles the issue of imaging biomarkers quantification in the presence of implants, which heavily degrade the image quality. Moreover, only few works directly mention and address the problem of partial volume effects, *i.e.* the presence of piled-up signal due to the limited imaging resolution and potential tissue heterogeneity within the voxel. Such effects could bias the voxel classification into fat or muscular tissue and thus make the biomarker estimation more difficult. However, the main obstacle still remains

the automation of the processing pipelines and of the biomarker extractions. In fact, most methods still require manual input from external users, either as definition of ROI or threshold selection.

In the musculoskeletal field, the application of automated image analysis tools is hampered by the exceptional shape variability of musculoskeletal structures within the population. Together with the inconsistency and variety of image acquisition protocols, this poses strong limitations to the generalisability of existing approaches, and therefore their customisation to subject-specific cases or to pathological conditions [20, 21]. In addition, these factors curb the availability of sufficiently large annotated data sets to be used in supervised machine learning or deep learning approaches, which are showing very promising results in other medical imaging applications. Finally, a further complication is the presence of metal objects in the field of view, as in the case of hip replacement. The induced artefacts greatly vary in appearance depending on the image modality and the shape of the implants. By corrupting the signal intensity from numerous voxels, either localised around the implant or spread across the FOV, the presence of metal further hampers the ability to discriminate between tissue types and thus reliably estimate the muscle atrophy.

The following sections review the current state-of-the-art strategies to address the described problems.

Summary box: Challenges in imaging biomarkers for hip replacement

- Based on qualitative, rater-dependent visual scores
- Lack of standardisation of quantitative muscle atrophy estimation
- Lack of robust and automated segmentation techniques
- Lack of robust and automated image analysis tools

2.2 Segmentation of musculoskeletal anatomy

In medical imaging, segmentation is the process of classifying and delineating boundaries of semantically connected regions such as different biological structures. As a result, a new same-size image is usually produced, where each pixel/voxel stores a label indicating the class of the respective pixel/voxel in the original image (Fig. 2.4). Common applications of segmentation include the localisation and monitoring of tumours or any other pathological region, longitudinal analysis of organ volume, shape and texture modifications, as well as delineation of target regions for treatment or surgical planning [43].

The most basic approach to perform image segmentation is by manual delineation of contours. In case of medical applications, this is a heavily time-consuming task, subject to intra- and inter-rater variability. Nonetheless, this

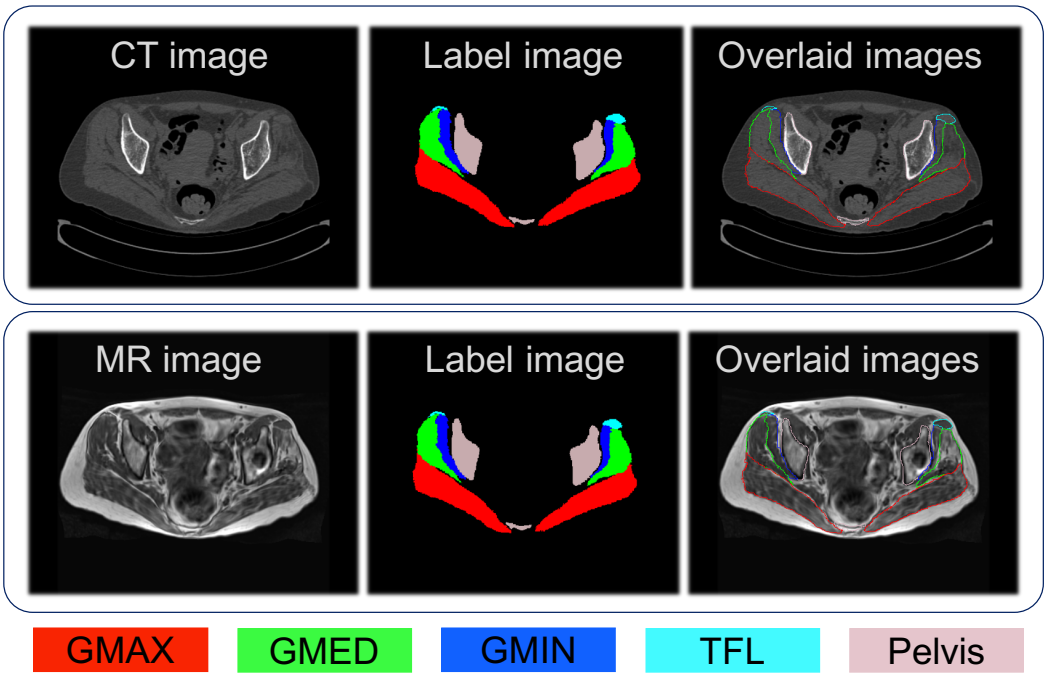


Figure 2.4: Same-subject CT and MR images and respective manual segmentation (label image) of gluteus maximus (GMAX), gluteus medius (GMED), gluteus minimus (GMIN), tensor fasciae latae (TFL) and pelvic bones. The third column shows the segmentation overlaid to the respective grey-level image. While muscular and skeletal tissue is easily distinguishable, the correct boundaries between different muscles or between different bones are harder to define due to the lower intra-class intensity variability.

technique benefits from human expertise in recognising relevant objects, hence it is often considered as the ground truth. In the past thirty years, several methods have been elaborated to achieve automatic or semi-automatic segmentation of medical images, their applicability being strictly connected to the image modality and the type of segmented structures. The main algorithmic approaches relevant for MSK can be categorised in the following groups:

- *Intensity-thresholding approaches* make use of filtering techniques on the voxel intensity values. They can be applied stand-alone or combined with other basic image processing tools, such as smoothing and/or morphological operations, in a sequence of steps tailored for specific applications.

- *Iterative clustering* refers to unsupervised classification techniques aiming at simultaneoulsy maximising the intra-class similarity while minimising the inter-class one. For the purpose of image segmentation, they are applied to the voxel intensity values and they are effective on imaging modalities where there is a clear and reproducible relationship between tissue type and voxel intensity (e.g. in CT). An examples is the k-means clustering algorithm: given the number of expected classes and an initialisation of the average intensity per class, it iteratively assigns each voxel to the class whose average is closest and then updates the average value. The operation is repeated until no voxel classification is modified anymore. An extension of the k-means algorithm is the fuzzy c-means, where each voxel is not associated with a single class but with a membership value (between 0 and 1) per class, representing its distance to the average of that class. This makes the classification more robust to outliers and implicitly considers partial volume effects, but is still affected by noise and does not encompass any shape information. If assumptions can be made on the intensity distribution of the images, model-based clustering can also be developed. A Gaussian Mixture Model (GMM) is often chosen in medical imaging for this task using an Expectation-Maximisation (EM) framework to obtain the maximum likelihood esti-

mates of the mean and standard deviation per class. Anatomical priors, as well as noise modelling and Markov Random Field regularisation can be encompassed within the EM framework to tailor the segmentation algorithm to specific tasks [44]. However, such approaches would fail if the intensity distributions of different classes present extensive overlapping.

- *Active shape models (ASM)* indicate the process of applying a local search algorithm to fit a shape model to an image, in order to locate and segment the structure of interest. Statistical shape models (SSM) are typically used as models to be fit by the ASM approach. SSM are population-derived models most suited for structures with characteristic shape and location. Given a training dataset, they are built by (1) localising a set of easily identifiable landmarks along the shape contour in each image, (2) rigidly registering the shapes into the same reference frame, (3) using principal component analysis (PCA) to derive a shape model defined as the mean shape and its variations modes. When applied to a new image with the ASM, these models provide a priori knowledge on the shape to segment in order to identify the same object within the image [45]. The SSM can be combined with texture information to learn correlations between shape and texture (statistical appearance models or SAM). In this case, after warping each training image to the mean shape, a texture vector is extracted from each of them, an eigen-model for texture is built using PCA, and is then combined with the shape model. The process of fitting this model to a new image is called *active appearance model* (AAM), where not only the shape but also the texture is taken into account to obtain the automated segmentation of the structure of interest [46]. AAMs take greater advantage of the available information compared to ASMs, but are computationally expensive and strongly reliant on a good initialisation.

- *Atlas-based methods* are based on the same ASM concept of applying prior information to a new image. The prior knowledge in this case de-

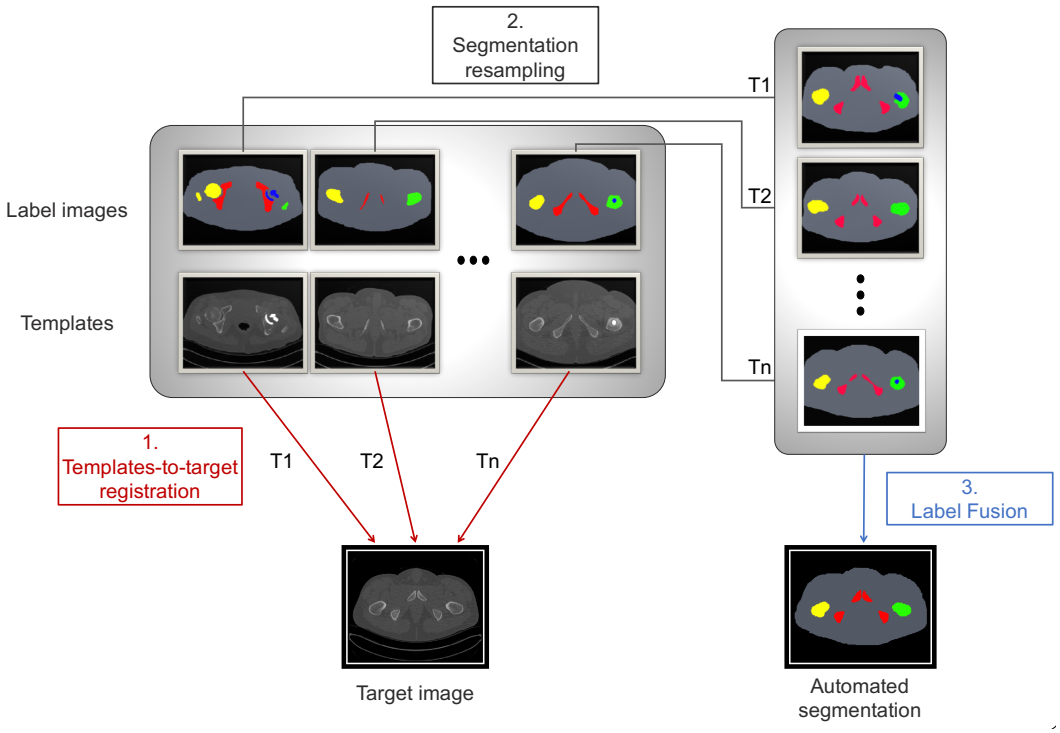


Figure 2.5: Schematic representation of the multi-atlas automated segmentation approach. First, all the template images are registered to the target. The optimised transformations T1, ..., Tn are used to resample the respective label images to the target space (segmentation propagation). The final automated segmentation is estimated through a label fusion technique.

rives from one or more atlas images, which are sets of intensity image (template) and its known segmentation (label image) [47]. Differently from ASM approaches, the final segmentation is obtained by propagating the label images into the target image space. More specifically, first all the available atlases are aligned to the target using image registration. The same alignment is applied to their respective label images so as to remap them onto the target image (segmentation propagation). Each propagated segmentation represents therefore a candidate segmentation for the target. A final consensus is then defined using a label fusion technique (*e.g.* majority voting, STAPLE [48], STEPS [49]). A schematic representation of the multi-atlas automated segmentation is shown in Figure 2.5. Compared to SSM, atlas-based methods are more generalisable across subjects and across modalities, as they do not rely

on the presence of specific shape features, but only on the possibility to achieve a reasonably accurate registration. However, the need to perform multiple registrations is also a drawback in terms of computational cost.

- *Region-based methods* separate homogenous regions by identifying boundaries in the image. If edges are clearly visible in the image, such methods are typically more robust to noise than pure intensity-based ones as they can encompass spatial connectivity information. However, they often require some manual initialisation, which they are very sensitive to. Region-growing methods such as watershed belong to this category, together with contour modelling, either parametric (active contours or snakes) or non-parametric (level sets). Active contours [50] are parametric curves optimised by minimisation of an energy functional. They are typically physics-based models that balance boundary evolution to image edges with specific shape constraints (e.g. continuity, boundary smoothness). Level sets [51] are a contour evolution method based on implicit surfaces. Differently from active contours, they are also able to model self-intersection and changes in topology.

- *Graph-cut segmentation* considers each voxel in an image as the node of a graph, connected to the neighbouring voxel by an edge (n-link). Each node is also connected by so-called t-link edges to different terminal nodes, designating the different possible classes. Both n-links and t-links are associated with nonnegative weights such that the segmentation problem becomes the identification of the best graph cut minimising the sum of the weights of the severed edges [52]. In practice, this is equivalent to the minimisation of a cost function including regional properties, boundary terms and possibly hard constraints.

- *Deep learning segmentation* refers to a broad range of fully data-driven algorithms that make use of large training data to learn the features best representing the structure under study. A deep learning model is

expressed as a neural network (NN), a collection of layers connected to each other to construct highly-dimensional non-linear functions. The neural network is then trained to map the inputs to a desired output using an optimisation scheme. Supervised learning is the most common setting, where the training set is composed of grayscale images and their respective label images (typically manual segmentations). In this case the neural network is optimised to correctly predict the label images, given the grayscale as input. If trained correctly, the model is able to generalise to unseen images and thus automatically predict their segmentation. In medical imaging, convolutional neural networks (CNNs) are the most widely used model, as they employ convolutional filters to extract image features in a tractable way on such high-dimensional data as medical images [53]. While being very powerful modelling techniques, they require large amounts of reliable labelled training data, which is often not readily available in medical imaging applications. A few approaches already exist in the literature, which aim at obviating the need for large labelled training sets. Data augmentation is the most common approach, and consists in artificially injecting input data variability during training through randomly transforming the available data (e.g. by applying random rotations/flipping along the image axes or by adding random noise)[54]. This approach typically favours the generalisation to unseen data as it prevents the network from simply memorising the training set, but is limited by the variability the random transforms are able to capture. Alternatively, a patch-based sampling scheme can be adopted, when the segmentation problem is applied to large 3D volumes. This refers to the technique of randomly sampling portions (patches) of the volume, instead of feeding the full volume at once. Similarly, 2.5D deep learning techniques use 2D CNNs but with 3D input, where the input usually includes a few neighbouring slices or multi-view slices. Both these approaches can be used to enlarge the number of training samples while

keeping the model parameters to a tractable number, however they might struggle to capture the global context medical images carry [53]. More sophisticated techniques include transfer learning - where a deep learning model pre-trained on a different task (e.g. natural image classification) is slightly modified and fine-tuned on the target task (e.g. medical image segmentation) [55] - and active learning - where the network output is iteratively presented to a human expert who corrects any mistake or annotates scribbles which are fed to the network to improve the segmentation [56]. Finally, semi-supervised learning represents a promising and growing field towards the compensation of missing labelled data. During training, these approaches leverage both labelled and unlabelled data to improve the generalisability of the learnt model. Typically, the model is first partially trained using only the labelled data, then pseudo-labels are generated with this model also for the unlabelled data. The model is then retrained with all the available labels, typically by taking into account the uncertainty associated with each label. The process can also be iterated to further improve the performance [57].

Task-specific hybrid approaches combining the discussed categories are becoming more and more popular in the field. For instance, SSM or atlas-based methods can be used to initialise a segmentation further refined by a graph-cut algorithm; non deep-learning methods and manual refinements can be adopted to generate the large training datasets needed to train the deep learning models.

A few challenges still characterise the application of the described techinques in the musculoskeletal field. As already mentioned, the main bottleneck is the large inter-subject variability in both shape and texture, especially of muscles. Moreover, most of MSK structures are relatively wide, requiring either very large image FOV or methods able to cope with object occlusion, which happens when the object of interest is partially outside the FOV or hidden by other elements in the scene. Other challenges involve the small intra-class ap-

Table 2.2: Typical intensity values in CT imaging measuring the radiodensity of tissue in Hounsfield Units (HU): $HU = 1000 \times (\mu_{tissue} - \mu_{water})/(\mu_{water} - \mu_{air})$, with μ indicating the X-ray attenuation coefficient [59, 60].

Substance	CT intensity (HU)
Water	0
Air	-1000
Fat	[-100, -50]
Soft Tissue	[+30, +300]
Bone	[+300, +3000]
Heavy metals	>+3000

pearance variability: while it is reasonably simple to distinguish muscular and skeletal structures, it is generally harder to define clear boundaries between different adjacent muscles or adjacent bones. An example is shown in Figure 2.4, reporting both a CT and an MR segmented scans.

2.2.1 Bone segmentation

As introduced in Section 1.2, CT is the preferred modality for bone visualisation. In MR, the highly rigid structure of cortical bone induces a fast-decaying signal that is often not detectable, causing a lack of information in the reconstructed image in correspondance to cortical areas. As a result, the cortical bone appears as black, while trabecular bone tissue has a gray-to-white appearance, close in intensity to fat [58]. By contrast, in CT the signal intensity at each voxel is directly proportional to the X-ray attenuation coefficient characteristic of the tissue presented in the volume. Typical ranges of image intensities are reported in Table 2.2. Being highly attenuating with respect to other tissue types, bones produce areas of high contrast in the image, which makes them more easily distinguishable. Hence most of the work presented in the literature for hip joint segmentation has been developed for this latter imaging modality.

Segmentation from CT images

Given the well-modeled relationship between image intensity and tissue type, the easiest and most common form of bone segmentation in CT is thresholding: voxels associated with intensities within a certain range of values (typically +300/+3000 HU [60]) are classified as bone tissue. This technique is effective for cortical bone, as its intensities are well separated from other tissues ($HU > 1800 - 1900$), but it fails in identifying the trabecular bone and bone marrow, as their intensities are closer to other tissue types ($300 \leq HU \leq 400$). This worsens in presence of pathological conditions such as osteoporosis, where the bone boundary becomes very narrow [61]. Simple thresholding also does not permit to distinguish and classify the different types of bones.

A large variety of more sophisticated automated methods has been presented in the literature for the segmentation of the hip joint, focusing mostly on the correct identification of the acetabular space. Four main categories can be identified [62, 63]:

- *Intensity-thresholding approaches.* One of the most recent attempts was reported by Cheng *et al.* [64], who proposed an iterative adaptive thresholding classification, based on Bayes decision rule, with the application of a valley-emphasising filter to enhance contrast in the acetabular space. A coarse-to-fine approach was presented by Guo *et al.* [65], where morphological operations and an EM bone boundary definition strategy are coupled with a 3D surface voxel tracing algorithm to further refine the segmentation. While typically fast, these methods tend to fail in case of severely diseased hips, as bone boundaries are less sharp or less well-defined [65].

- *Active shape models (ASM).* Examples of this approach in the pelvic domain are reported by Seim *et al.* [66], who focused only on the pelvic bone, and by Keinmueller *et al.* [67], who extended the model to the whole articulation by including also the proximal femur. Similarly, Yokota *et al.* [68] implemented a hierarchical SSM approach to firstly

segment the pelvis and the distal femur and then refining the segmentation of the femoral head near the acetabular space. This latter method proved to be robust for diseased hip joints as well.

- *Atlas-based methods.* For hip segmentation, a single-atlas approach was presented by Pettersson *et al.* [69], while multi-atlas frameworks were proposed by Whitmarsh *et al.* [61] and more recently by Besler *et al.* [70]. The use of atlases naturally allows for the independent labelling of each and every bone, making this method more robust to cases with narrow acetabular space, provided that an accurate atlas-to-target registration can be obtained.

- *Graph-cut segmentation.* A fully automated application of this method to the hip joint segmentation was proposed by Krcah *et al.* [71], which does not require any prior shape model but only exploits intensity information as prior. Alternatively, Pauchard *et al.* [63] presented an interactive graph-cut framework for femur segmentation, where the user provides initial quick scribbles of the foreground and background to be included as hard constraint in the graph-cut and can also perform manual editing on the result iteratively.

In the context of combining multiple techniques, work has been proposed by Chu and collaborators in two separate articles. In their first work [62], they combine a robust landmark-based initialisation with a multi-atlas segmentation approach, which is further improved by fitting an articulated SSM to preserve the joint structure. In the second one [72], the multi-atlas based segmentation is employed to constrain a multi-label graph-cut algorithm. An example of hybrid method with SSM and graph-cut is offered by Huang *et al.* [73] for femur segmentation. To avoid the need of population models, region-growing methods have been combined with patient-specific optimal thresholding [74] or other variational approaches such as phase-field fracture models [75] to better separate femur and pelvic bones. They have the advantage of fairly reduced

computational cost with respect to atlas-based methods, though their generalisability to pathological hips together with their robustness to noise still need to be proven.

More recently, deep learning approaches have been introduced for bone segmentation in CT. Some examples on whole-body CT showed high segmentation accuracy but focussed on segmenting the skeletal components as a whole, without differentiating individual bones [76, 77]. Lindgren Belal *et al.* [78] proposed a cascade of CNNs to automatically segment 49 different bones, including the pelvic bone. They used whole-body data, which would be computationally intractable for a deep learning model. To overcome this problem, a first CNN was trained to locate relevant anatomical landmarks, and each landmark was then fed to a second CNN together with a corresponding patch of the CT image for automated segmentation. Similar work was proposed by Schnider *et al.* [79], who compared different 3D U-Net models for segmenting 126 different bone structures (including pelvic bones and femora). In both works, however, no segmentation accuracy is reported for the pelvis or the femora. An automated segmentation of the proximal femur was very recently proposed by Zhao *et al.* [80], who obtained highly accurate results on 3D CT images cropped around the femoral head using a 3D V-Net architecture. They also showed significant correlation ($R^2 = 0.9956$) when comparing the volumes from the automated and the manual segmentations. Interestingly, the ground truth to train this model was built using thresholding and a contour detection algorithm, followed by manual refinements, on the training data points.

Altogether, despite this large variety of approaches, global thresholding coupled with manual editing still remains the most commonly applied segmentation method for bones in CT [81].

Segmentation from MR images

Despite the limitations of this modality for bone visualisation, the use of MR is still appealing for its lack of ionising radiation exposure and for its relatively good contrast for cartilaginous tissue.

Methods similar to the ones described for the segmentation in CT images are often applied. Schmid *et al.* [82] combined a SSM initialisation of a physics-based deformable contour evolution which accounts for the most probable modes of variations and a Markov-Random-Field regularisation. This was extended by Chandra *et al.* [83], with the introduction of clinically relevant weighting maps into the statistical shape model, to focus and improve the segmentation of 3T MRI in the most critical areas of the acetabular space. More recently, Damopoulos *et al.* [84] proposed the use of a random forest classifier to obtain an initial probabilistic segmentation of the femur in radial MR scans, which is then used to drive the registration of a template mesh to the target image. However, such methods often require a manual detection of landmarks to initialise the model fitting, and most commonly rely on shape models built from a healthy population, whose applicability to pathological scenarios is still unclear.

Xia *et al.* [85] developed a comparison between a CT-to-MR multi-atlas approach - where the segmentation from template CT images are propagated to the T2-weighted MR space - and an ASM method built on a CT-derived statistical shape model of the hip joint. Both methods proved to be robust and accurate, with the higher computational burden of the multi-atlas approach being the only major difference. However, the use of a 3T MR scanner on healthy volunteers might have facilitated the achievement of good segmentation results in the MR.

Due to the different appearance of cortical and trabecular bone in MRI, region-growing methods are less effective for bone segmentation, as they tend to converge to the trabecular bone only or leak into soft tissue where the cortical bone is very thin [58]. To overcome this issue, Arezoomand *et al.* [58] developed an atlas-based semi-automatic initialisation of an active contour segmentation of 3T MRI scans.

Deep learning methods have been recently coming into play for this task in MSK. An example on knee MRI is given by Liu *et al.* [86], where an encoding-

decoding CNN is trained to predict bone and cartilage segmentation of the joint, subsequently refined through a marching cube and simplex deformable model to extract a final mesh of the segmented structures. T2-weighted MR images acquired on a 3T scanner were employed in this case as well. Promising results for femur segmentation in T1-weighted images were reported by Zeng *et al.* [55] using a similar 3D U-net architecture. To overcome the small size of the training dataset, they took advantage of a transfer learning approach from a network trained to classify sport images. A comparison of 2D and 3D CNN architectures for proximal femur segmentation in high-resolution T1-weighted MR images from a 3T scanner was presented by Deniz *et al.* [87]. They reported very high segmentation accuracy (Dice similarity = 0.95 ± 0.02) and showed that 3D CNN models and the use of dilated convolutions boosted the segmentation performance. Such high accuracy and the computational gain at testing are clear advantages of well-trained deep learning models over other techniques. Still, however effective and popular, the lack of large annotated training datasets is currently the main holding-back factor for such deep learning approaches to flourish in hip imaging.

2.2.2 Muscle segmentation

A variety of approaches can be found in the literature for the segmentation of muscles and other soft tissue in both CT and MR images. While fat is characterised by a specific range of values on the HU scale and can be easily segmented in CT images through thresholding, muscle intensity distribution overlaps with that of other organs and tissue types, [-29, 150] HU, hence clear boundaries between neighbouring muscles and/or organs are hard to delineate [88, 89]. On the other hand, MRI provides remarkable contrast for soft tissue structures, but its wider application in orthopaedics is currently limited by its high costs and by the lack of standardised acquisition protocols. Together with the considerable shape and texture variability of muscular structures, this hampers the development of standardised and accurate segmentation techniques which could aid the image analysis [20].

Segmentation from CT images

Few attempts have been reported in the literature for the segmentation of muscles in CT, mostly focussing on chest and abdominal area for the quantification of body composition and fatty or lean tissue distribution. Different approaches have been proposed either based on fitting statistical shape and appearance models (Chung *et al.* [90]), on combination of atlas-based and active contour models (Zhang *et al.* [88]), or on deep learning segmentation (Lee *et al.* [89]). However, all of these methods are applied to the whole musculature, thus being inadequate to differentiate the individual muscles specifically.

Regarding individual muscles, the segmentation of the psoas major in CT was addressed by Kamiya *et al.* [91] by means of a mathematical shape model, defined from a set of landmarks and muscle centerlines, which *per contra* is not trivially extendable to pathological cases. Atlas-based approaches are currently the most appropriate solution for simultaneous segmentation and classification of several different muscles, as they allow to include prior geometrical knowledge about the shapes and the relative locations of the considered structures. Wei *et al.* [92] employed a single-atlas segmentation propagation approach for the paraspinal muscle, where the atlas is affinely registered to the target, followed by a local contour optimisation to improve the convergence to the correct muscle location, and further refined by active contour segmentation. Interesting work was presented by Otake *et al.* [93] to estimate patient-specific muscle fiber orientation from CT: the segmentation of gluteus maximus and medius is firstly derived with a hierarchical multi-atlas approach, and a geometric muscle fibers template is fit to a CT-derived structural tensor vector field within the gross muscle volume. However, the success of this approach relies on sufficient intramuscular fat within the muscle to allow for the fibers to be detected.

At the time of writing, the most relevant results are reported by Yokota *et al.* [94]. In order to segment 19 skeletal muscles in the hip and thigh region, they introduced a hierarchical multi-atlas approach in three steps shown in Fig. 2.6. First, the skin and the bones of the target image are automatically

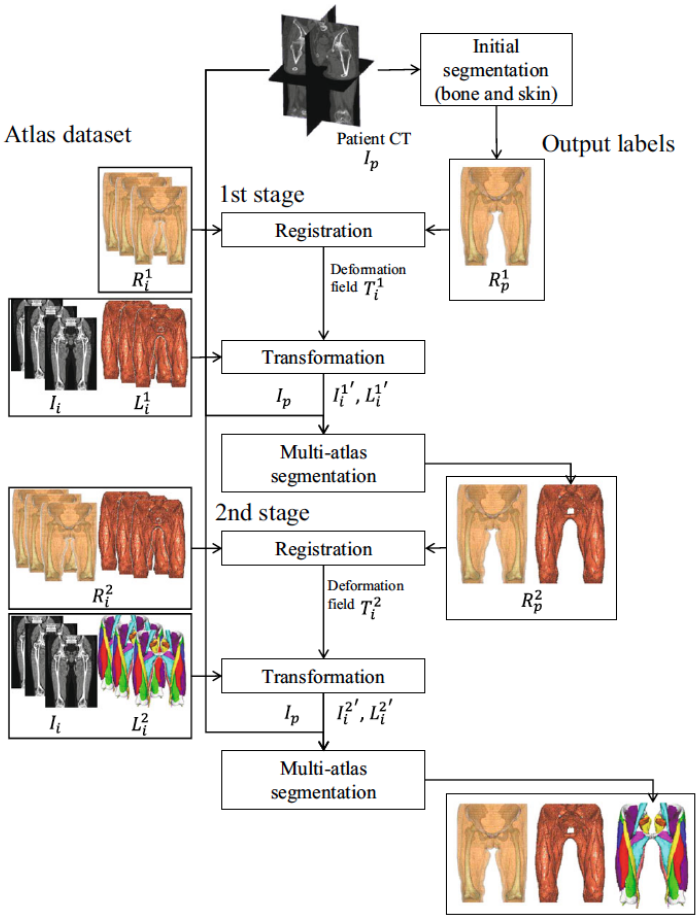


Figure 2.6: Hierarchical multi-atlas approach for automated muscle segmentation proposed by Yokota *et al.*. Three iterations of multi-atlas segmentation propagation and label fusion are performed, first to segment bone and skin, second for the full musculature and finally for 19 individual muscles classification. (Image from Yokota *et al.* [94]).

segmented and a label-based registration is used to align all the templates to the target image (spatial normalisation); then, the whole muscular tissue is segmented using the spatially normalised atlases and a new label-based registration of the whole muscles is used to update the spatial normalisation; a final iteration of atlas-based segmentation propagation is used to obtain the separate muscles labelling. In a cross-validation study with 20 female patients, they reported an average Dice score [95] from all muscles of about 80%, showing improvement against no use of hierarchisation. Interestingly, they demonstrated how a 2D assessment of segmentation accuracy led to either under- or overestimation on at least 5 muscles, when compared to full volumetric measures.

A major drawback of Yokota's method is the high computational cost, due to the multiple iterations of atlases-to-target registrations. The same group very recently tackled this issue using a deep learning framework, which is slow at training but computationally very efficient at testing on unseen data [96]. They proposed a 2D Bayesian U-Net, where Monte Carlo dropout is applied at inference to estimate the segmentation uncertainty. On the same 19 muscles, this method improved the Dice Score to 89%, and it also showed potential for an active learning framework, where high uncertainty regions are presented to a human rater for reassessment and manual correction. Although the authors mentioned that all their subjects had total hip replacement on one hip side, in both studies results are shown only for the contralateral (non-implanted) hip side. It is thus unclear what was the perfomance on the implanted side and what was the impact of metal artefacts.

Segmentation from MR images

The interest in muscle segmentation in MRI initially arose from the possibility of developing musculoskeletal models for biomechanical dynamic simulations. The first attempts were therefore limited to an adult average anatomy, with no customisation to subject-specific geometry [21]. An example is the 3D finite element model developed by Blemker *et al.* [97] derived from manual segmentation of hip muscles. As the accuracy of biomechanical simulations was shown to benefit from patient-customisation [98], more research efforts have been focussing on automated muscle segmentation from MRI, also in connection with localisation of tendon insertions.

Although graph-based modeling or active contour initialisation are found in the literature [41, 99, 100], similarly to CT, atlas-based methods are still the most commonly used for muscle segmentation in MRI, either stand-alone or in combination with other refinement techniques. In fact, edge-based methods tend to fail in determining separate muscle boundaries, especially when heavy fat infiltration is present, as they cannot distinguish between intramuscular and intermuscular fat without a shape model [37].

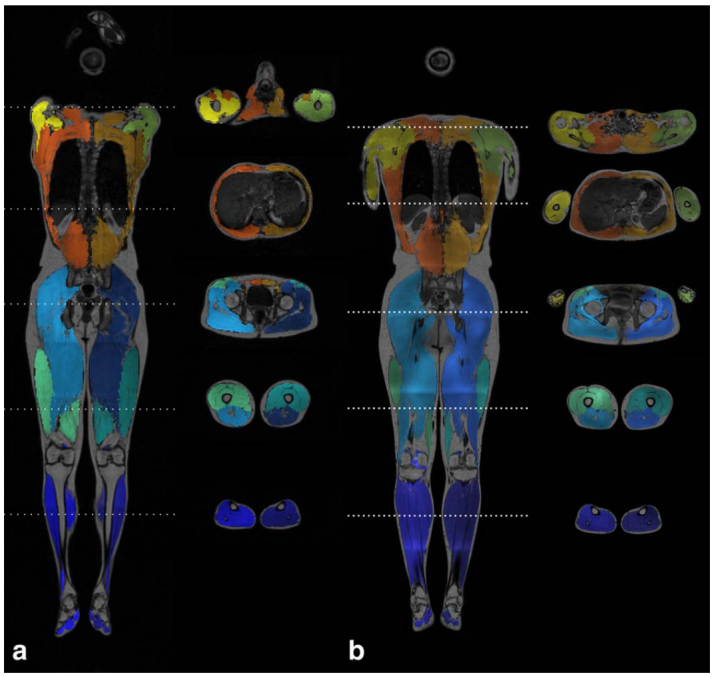


Figure 2.7: Example of atlas-based automated muscle segmentation from water-fat separated MRI. Results from a 1.5T (a) and a 3T (b) atlas datasets are shown. (Image from Karlsson *et al.* [29]).

Similar atlas-based approaches are discussed by Jurcak *et al.* [101] and by Prescott *et al.* [102], the former exploiting a probability atlas to initialise a geodesic active contour algorithm to segment the quadratus lomborum muscle, the latter selecting the template most similar to the target image to initialise a level-set segmentation of four quadriceps muscles. A full-body 10 muscle groups segmentation was achieved by Karlsson *et al.* [29], taking advantage of Dixon acquisition with water-fat separated MR images for enhanced contrast of muscular boundaries (Fig. 2.7). The multi-atlas based automated segmentation was validated on 3T MR images, using both a 1.5T and a 3T atlas datasets. For the same group of muscles, Ahmad *et al.* [103] proposed a semi-automatic algorithm based on patient-specific atlas: the manual segmentation of a single 2D slice (ideally the mid-FOV one) is used as an atlas by registering and propagating its segmentation to neighbouring slices; the process is iterated by updating the atlas image with the newly segmented slices and moving therefore along the full volume. On the same idea of propagating manually

annotated slices, Ogier *et al.* [104] presented a framework that combines both an interpolation and a propagation schemes from the inferior and superior manually segmented slices. The former registers the references slices together and linearly interpolates between their segmentation to obtain the intermediate slices contours, while the second performs pairwise registration and label propagation of each slice with the nearest reference one. A deep learning for thigh muscle and fat segmentation was proposed by Yao *et al.* [105], integrating a network for the detection of edges and regions with a traditional active contour refinement and a classification network. Still within the thigh area, Mesbah *et al.* [106] presented the advantages of using fat- and water-separated MRI images to obtain an accurate segmentation of different muscular groups and the intramuscular fat in healthy and pathological subjects. Taking advantage of the multimodal information, they employed a Markov Gibbs Random Field model which included atlas-based prior shape knowledge, labelling spatial coherence and fat and muscle intensity differences. They interestingly compared their approach with a CNN segmentation model (DeepMedic [107]), demonstrating comparable performances and arguing that their approach is more computationally efficient and would not require re-training if new atlases are available.

Most of the discussed methods are adapted only for healthy conditions and often require manual interaction. In the case of pathological muscles, no method is able to accomodate for metal-artefact corrupted images. For MoM hip arthroplasty patients, to the best of my knowledge the only automated segmentation method of hip abductor muscles in MRI was presented by Klemt *et al.* [108] in previous work from my group. They proposed a multi-atlas based segmentation propagation approach based on a robust Least Trimmed Square registration framework, which proved to achieve high accuracy segmentation for both healthy and severely atrophied muscles.

2.2.3 Joint skeletal and muscular segmentation

Examples of combined musculoskeletal segmentation can hardly be found in the literature for hip anatomy. A first attempt, still within the context of segmentation for biomechanical simulations, was presented by Gilles *et al.* [109], who implemented a multi-resolution simplex mesh model of musculoskeletal structures that can be registered to a MR image in a semi-automated way. A biomechanical modelling technique was proposed by Kohout *et al.* [110], based on a surface mesh generic model that can be fit to EOS imaging data by alignment of three landmarks. While such models are suited for biomechanical simulations, they do not really fit locally to the anatomy of the patient, hence they cannot be used to evaluate imaging biomarkers or generically assess pathological conditions.

Only very recently a few more works on joint skeletal and muscular segmentation have been published, concurrently to the development and progression of this thesis project. The methods proposed by the Nara Institute of Science and Technology group [94, 96], which were discussed in section 2.2.2, do report the ability to segment also the pelvic bones and the femora together with 19 muscles. However, no segmentation performance is reported for the skeletal structures, as they were not the main focus of the work. For the purpose of body composition analysis, Hemke *et al.* [111] trained a 2D U-Net to automatically segment CT slices into subcutaneous adipose tissue, muscle, bone and other generic tissue. A segmentation of the intramuscular fat was then derived through morphological erosion of the muscle mask and subsequent thresholding of the CT image at -30 HU within the muscle area. The model is however not able to differentiate separate muscles or bones, and it was also debatably trained and tested to segment a single standardized CT slice, with no clear assessment of the whole muscle volume. A more robust deep learning approach, called Attention-Pyramid network (APNet), has been recently introduced by Liang *et al.* [112] for the segmentation of 50 pelvis anatomical structures on MR images. They combined spatial pyramid pooling to capture model local

context with an attention mechanism to preserve the global context, reporting a mean Jaccard index of 80.27% on the test image and improving over standard deep learning architectures.

A last very relevant development from the Nara group concerns the segmentation of CT images of patients with THR also on the implanted hip sides [113]. Using simulated data, they trained two U-Nets: the first one refines a standardized metal artefact reduction technique called Normalized Metal Artefact Reduction (NMAR, see Sec. 2.4); the second performs the segmentation of 19 hip muscles, femur, pelvis and sacrum. The training data was generated by impainting implant data onto pre-operative CT images, forward projecting the images in the sinogram space, merging them and reconstructing the corrupted image with filtered back projection to produce the artefact. The MAR network and the segmentation network were then tested on 3 post-operative CTs with manual segmentation of gluteus maximus and medius. An example of their segmentation result is reported in Fig. 2.8. It shows how an automated segmentation technique developed on healthy population (i.e. non corrupted images) performs poorly on cases with artefacts, thus demonstrating the need for segmentation methods more robust to metal-induced noise. Although very promising, their method is based on a supervised approach for the metal artefact reduction. As no real CT image can be obtained with metal but without artefact, they need to rely on simulated data where the artefact is artificially generated. Thus, the performance of the U-Net is highly dependent on how realistic this simulated data is (see Sec. 2.4 for more details). As it was tested on 3 subjects only, further validation would also be desirable to assess the accuracy and generalisability of their approach.

2.2.4 Summary of current methods limitations

A number of limitations still characterises the currently available approaches for musculoskeletal segmentation. First, the majority of the methods focuses either on bony or muscular structures, with no means of fusing them to better visualise their spatial relationship. In the past ten years, most of the combined

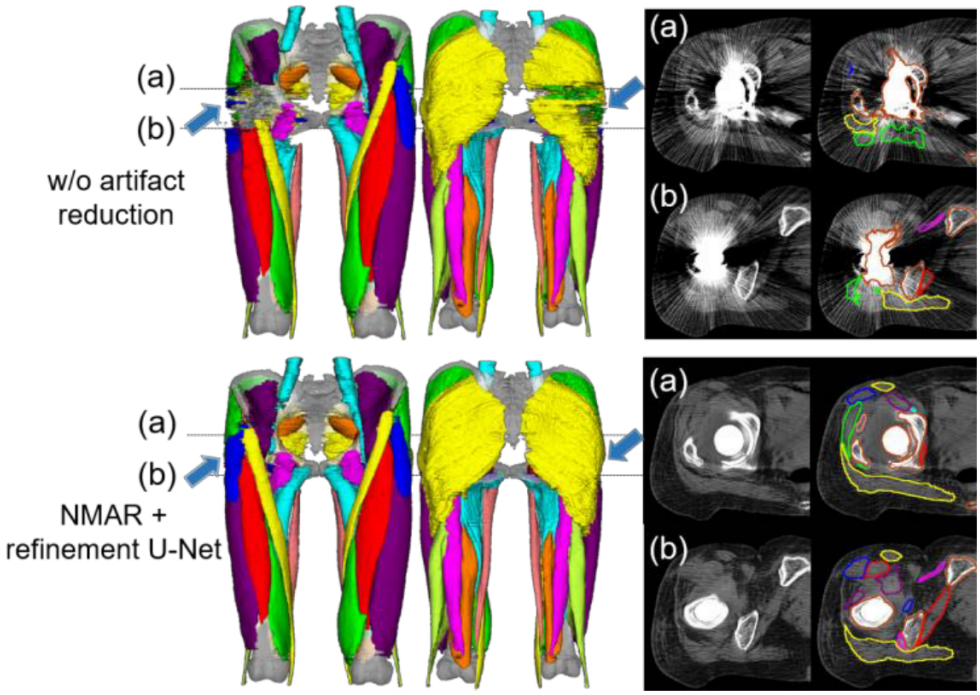


Figure 2.8: Example of automated segmentation of CT images including hip implants using the pipeline proposed by Sakamoto *et al.* [113]. The top row displays the performance of a U-Net based segmentation when no explicit compensation is applied for the metal artefact. The 3D rendering of the segmentation is reported together with two exemplar axial slices, also overlaid with their segmentation result. The bottom row shows the outcome of their proposed approach, where a refined metal artefact reduction method is applied before the segmentation (Image from Sakamoto *et al.* [113]).

musculoskeletal segmentation techniques were developed for the purpose of generic biomechanical simulations, not tailored to patient-specific anatomy. However, recent literature has shown that such models would benefit from customisation to the subject, thus a volumetric model derived from patient's imaging would be more accurate for these tasks [98].

In general, most of the discussed image-based methods are built on prior geometrical knowledge from healthy subjects. This consequently limits the physiological and pathological variability they are able to encompass and account for. This problem becomes even more relevant in presence of strong artefacts in the image, hampering the robustness of the model-fitting procedure, such as in the case of MoM hip arthroplasty patients. With the exception

of Klemt *et al.* [108] and Sakamoto *et al.* [113], no method has reported reliable performance in presence of hip implants within the image FOV.

My work sets in this context, addressing the automated segmentation and analysis of musculoskeletal anatomy from low-quality and heavily artefacted hip imaging data. In particular and differently from all the methods currently available in the literature, I propose the combination of CT and MRI information, taking advantage of both modalities to enhance both the inter-tissue and the intra-tissue variability. The effective fusion of multimodal information requires accurate registration, thus the state-of-the-art processing techniques for this task will now be discussed.

Summary box: Challenges in MSK segmentation

- Little work on combined muscular and skeletal segmentation
- Models including prior shape knowledge are built on healthy population
- Lack of robust methods to deal with metal artefact induced noise
- Lack of standardisation of scanning protocols
- Large shape variability in population
- Typically large inter-tissue variability (e.g. easy differentiation between bone and muscle), but small intra-tissue variability (e.g. difficult differentiation among different muscles).

2.3 Registration of pelvis CT and MRI

In medical imaging, registration indicates the process of establishing a one-to-one mapping between the coordinates in one space and those in another, in order to find corresponding anatomical or functional points in two or more images [114, 115]. A general registration algorithm comprises a transformation model, defining the geometric transformation between images, and a cost function, typically including a similarity measure to quantify the alignment of the images and a penalty term to promote desired properties on the transformation model (e.g. smoothness, regularisation of volumetric expansion and contraction). An optimisation scheme is adopted to iteratively update the transformation model parameters in order to minimise the cost function (Figure 2.9). Different registration algorithms are substantially characterised by different choices of transformation model, cost function and optimisation scheme. In particular, the complexity of the transformation model defines the degree of alignment achievable: simple rigid or affine models provide global alignment (same transformation is applied to all voxels, parametrised by 6 to 12 degrees of freedom), while nonlinear transformation models allow for more localised deformations, as the number of degrees of freedom becomes of the same order of magnitude as the number of voxels.

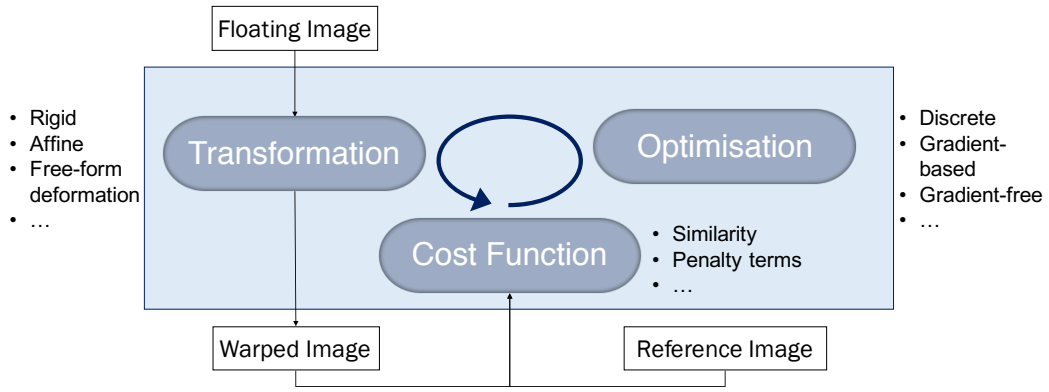


Figure 2.9: General scheme for image registration algorithms. A transformation model is chosen to warp a floating image and align it to a reference image. The model parameters are iteratively updated through an optimisation scheme in order to minimise a cost function, which typically measures the quality of the alignment between the reference and the warped images. Image adapted from Modat [116]

.

Registration algorithms can be designed to align images acquired with different modalities (multimodal registration). This is particularly useful in comparing and fusing the information derived from each modality. As mentioned in Sec. 1.2, CT provides excellent contrast for bones and implant, while soft tissue anatomy is better visualised on MRI. Hence, CT-MRI registration is an extremely helpful tool in musculoskeletal imaging to combine their complementary information. This is indeed an active research area, especially in the field of pelvic radiotherapy treatment planning, where CT provides the tissue attenuation information to calibrate the treatment while MRI allows for more accurate contouring of cancerogenous tissue (e.g. for prostate, rectal or cervical cancer). In this context, the first methods to perform multimodal registration were manual or semi-automatic, mostly based on the manual annotation of landmarks and on the identification of the best affine transformation to match these points [117]. Kerkhof *et al.* [118] proposed a bone-driven registration by simply thresholding the CT and extracting a bone segmentation to which the MR is registered by maximisation of mutual information [119]. A different structure-driven nonlinear approach was proposed by Rivest-Hénault *et al.* [120], who included a further term in the registration cost function to quantify the similarity between tissue contours delineated in both modalities. Similarly, Noorda *et al.* [121] presented a multimodal framework in which, after initial global alignment through rigid and affine registration, they employed a B-spline parametrisation for nonlinear registration and tailored the cost function to combine information from the image intensity distributions, from the gradient images and from the bone segmentation. An interesting learning-based registration method has been introduced by Cao *et al.*, which makes use of image synthesis to avoid a direct multimodal approach: a random forest regression is trained to synthesize the CT from the MR and vice versa; then each modality is registered to its synthetic same-modality counterpart and the two registration results are finally fused to provide a final deformation field for the multimodal registration [122].

Being tailored to specific datasets and most often to a specific organ or region of interest within the FOV, the discussed registration methods would not achieve a satisfactory alignment on extended FOV as the ones typically used to assess musculoskeletal structures. Among the challenges that have not been addressed yet is the low-quality and highly anisotropic resolution of MR images routinely acquired in MSK, bound to the short acquisition time that is usually available. The difference in the high in-plane and very low out-of-plane resolution (up to a factor 10) strongly affects the ability of intensity-based registration algorithms to achieve accurate anatomical correspondences along the three imaging planes. Moreover, in the specific case of hip arthroplasty, the presence of metal implants in the imaging field introduces artefacts and induced-noise that deteriorate the image quality in both CT and MR. However, the main problem affecting the CT-MR registration in the pelvic domain is related to differences of subject's pose within the two scanners. While a global rigid or affine transformation would be insufficient to cope with different stretching or compression of soft tissue, a nonlinear transformation would introduce unrealistic shape deformations in rigid structures such as bones and implants. The applied transformation should therefore allow for low-frequency nonlinear deformations of soft tissues, but also preserve the shape of rigid structures.

2.3.1 Nonlinear registration with rigidity constraint

The problem of nonlinear registration with locally rigid behaviour has wide application not only for musculoskeletal but also for other anatomical structures. One of the first examples in the literature tackles this issue for breast imaging [123], where discrepancies of tumour volume measurements were observed before and after nonlinear registration of contrast-enhanced MR to its baseline. To solve this, the authors proposed a free-form deformation (FFD) registration framework where control points within specific masks - i.e. segmented breast lesions - were coupled and bound to have the same displacement. Such a solution would therefore model a rigid transformation defined by translation

only and no rotation within each mask. To provide a more general framework to this problem, a novel transformation model called polyrigid or polyaffine was introduced in 2005 by Arsigny *et al.* [124]. The main idea behind this approach is to define some fuzzy regions associated with independent rigid or affine transformations, and obtain the full-FOV deformation field by smooth interpolation of all these local components. In particular, they propose a velocity field parametrisation that weighs the contributions of the locally rigid regions, from which the displacement field can be derived through integration, ensuring diffeomorphic transformations. An application of this formulation to the MSK field was presented by Seiler *et al.* [125], who included a polyaffine regularisation within the log-demons registration framework for femur allograft selection in the treatment of bone tumours.

Extending the idea of deriving a diffeomorphic nonlinear transformation from a set of rigid local ones, Commowick *et al.* [126] proposed a two-step strategy that computes spatially sparse local rigid transformations using a block-matching approach without the need for initialisation or prior knowlegde of the rigid parts. A dense velocity field is then interpolated from the rigid matrices in the logarithmic domain.

A widely applied solution was suggested by Staring *et al.* [127] and consists of introducing a penalty term in the registration cost function that would promote a rigid behaviour in user-defined masks. Given a deformation field $\boldsymbol{\phi}$ parametrising the transformation between a reference image I_R and a floating one I_F, the optimisation problem to be solved by the registration algorithm is formulated as:

$$\max_{\boldsymbol{\phi}}[\mathcal{S}(\boldsymbol{\phi}; I_R, I_F) - \alpha\mathcal{P}_{rigid}(\boldsymbol{\phi})] \tag{2.3}$$

where $\mathcal{S}$ represents a measure of similarity between the reference and the floating image after transformation and $\mathcal{P}$ is the penalty term, whose contribution is weighted by the hyperparameter α. In Staring's formulation, this penalty term includes three components (penalisation of non-null second order derivatives of $\boldsymbol{\phi}$, promotion of orthonormality and of incompressibility) and applies only

for the displacement associated to voxels within rigid masks. This approach was extended for spine CT-MRI registration by Reaungamornrat *et al.* [128], who added a constraint to impose injectivity to the transformation model. In the context of CT synthesis for radiotherapy treatment planning, Dowling *et al.* [129] generated a well-aligned CT-MRI training dataset using the structure-guided nonrigid registration proposed by Rivest-Hénault *et al.* [120]. However, their algorithm relies on availability of accurate contours of the same anatomical structure in both modalities, which are matched in the registration through the addition of a soft constraint to the cost function. While being effective, all these methods need to find a trade-off between the terms in the cost function, particularly between the similarity measure and the rigidity constraint. Hence, deviations from a strictly rigid transformation might still be allowed, and a careful selection of the rigidity penalty term weight α must be carried out to establish the right balance in the cost function. For these reasons, such methods are often referred to as "soft constraints".

Opposite to this is the concept of "hard constraint", indicating that the constraint conditions have to be met exactly. Haber *et al.* [130] presented a mathematical formulation of the image registration problem where a nonlinear transformation model can be enforced to be strictly rigid in specified areas. Using the same formalism as above, the optimisation problem is the following:

$$\begin{aligned} &\max_{\boldsymbol{\phi}}[\mathcal{S}(\boldsymbol{\phi}; I_R, I_F)] \\ &\text{subject to } \boldsymbol{\phi}(\boldsymbol{x}) - R_j(\boldsymbol{x}) = 0 \quad \forall \boldsymbol{x} \in \text{Mask}_j \end{aligned} \tag{2.4}$$

where R_j indicates a rigid or affine transformation specific to the j-th rigid mask. In their work, the authors report a proof of concept on 2D examples only, and do not explicitly consider the integration of such formulation into a broader diffeomorphic registration framework. Nonetheless, the use of such image-based "hard constraints" should be preferred as it limits the optimisation only to biologically and physically plausible transformations, which preserve the volume and the shape of rigid structures.

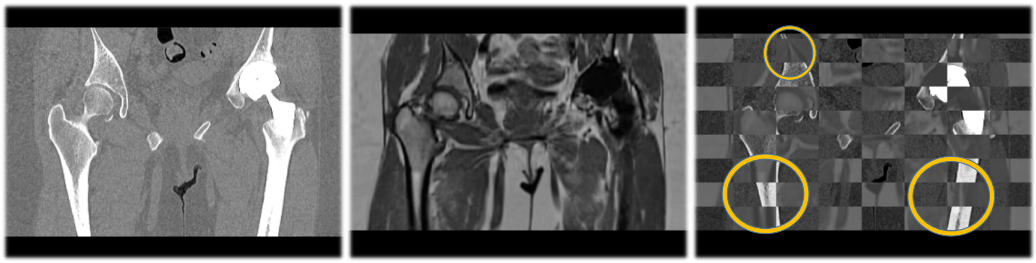


Figure 2.10: Example of checkerboard visualisation (on the right) to assess the registration results. The reference CT image (left) and the affinely registered MR (centre) from the same subject are shown. The yellow circles in the checkerboard illustrate areas of clear misalignment between the CT and the MR.

2.3.2 Validation methods

Given the wide variety of registration algorithms available, it is always important to define a validation protocol that evaluates the accuracy of the achieved alignment and the robustness of the results [115]. Unfortunately, registration is an ill-posed problem and there is no ground truth to compare with. The first and the simplest evaluation is based on visual assessment by overlaying the reference and the warped floating image and visually inspecting their alignment. In this sense, checkerboard visualisation can help identify discontinuities in boundaries and misalignments (Fig. 2.10). However, it is a subjective and qualitative evaluation, which is often insufficient to assess registration performance on 3D images.

The gold standard approach consists of using firmly fixed markers positioned before the image acquisition and visible on all the considered modalities. The markers can be used to estimate the registration error as the residual distance between their pairs (homologous points). In most cases, such an approach is invasive and unfeasible on routinely clinically acquired data. More common is instead the post-acquisition localisation of specific anatomical landmarks in the image FOV that can be used as homologous points. The Target Registration Error (TRE) is defined as the distance between corresponding landmarks after application of the optimised transformation, and it is often indicated as a measure of registration accuracy [131].

If contours of regions of interest are available in both images, the similarity

between the segmented structures can also be quantified as a measure of accuracy. A very common metric to check the segmentations overlap is the Dice Score [95]:

$$DS = 2 \times \frac{||S_R \cap S_W||}{||S_R|| + ||S_W||} \tag{2.5}$$

where S_R and S_W are respectively the segmentation of the reference and of warped images.

Another validation option consists in the synthesis of a ground truth deformation and in testing how well the registration algorithm recovers the transformation. This is however limited by the ability of simulating realistic deformations. Finally, the increase of image similarity could be checked as an indicator, although no conclusion should be drawn only based on this as the registration algorithm is generally designed to maximise such measure.

The accuracy is not the only desirable property of a registration algorithm. The biological plausibility of the transformation is also a fundamental aspect to be considered. When registering two images of the same patient, the structural topology has to be preserved in order to prevent from unrealistic folding of the tissue. This is monitored by computing the Jacobian determinant $|J|$ of the transformation at each voxel, which indicates the amount of deformation applied at that point: $|J| > 1$ indicates volumetric expansion; $|J| = 1$ indicates volume preservation (*e.g.* in the case of rigid transformation); $0 < |J| < 1$ indicates volume compression; finally, $|J| < 0$ indicates topology breaking and should therefore be avoided. As a result, diffeomorphic transformation models are usually preferred as they naturally guarantee the preservation of topology. As discussed in the previous paragraph, the biological plausibility of the transformation concerns also the need for rigid behaviour within rigid structures, which is fundamental in the musculoskeletal field. Finally, inverse-consistency is also sought to guarantee that the transformation is invariant to the choice of reference or floating image.

In conclusion, the validation of registration algorithms cannot rely on a single accuracy measure, but has to include different comparisons and analyses to

assess the performance and the presence of all the desired features in the transformation.

Summary box: Challenges in pelvis MR-CT registration

- Need to cope with different patient's pose within the scanners
- Robustness to metal artefact induced noise
- Preservation of the rigidity of skeletal structures while locally deforming soft tissue
- Accuracy throughout large field-of-view
- Robustness to anisotropic resolution of imaging data
- Extensive validation

2.4 Metal Artefact Reduction

Metallic implants are one of the main causes of image quality degradation in medical imaging. In patients with hip replacement, the size of the metallic prosthesis makes the artefacts even more severe and extended [132], hampering the diagnostic interpretation of the images in the most clinically relevant areas, i.e. close to the implant. As a result, the introduction of successful metal artefact reduction (MAR) techniques in hip replacement imaging is of great importance and thus is an active field of research.

2.4.1 MAR in CT images

In Computed Tomography, the presence of metal objects in the FOV causes the corruption or incompleteness of projection data, producing bright and dark streaks that radiate from the metal source throughout the reconstructed image. The signal corruption comes from the physics of X-ray absorption. In the presence of a polychromatic X-ray beam, the higher attenuation coefficient of metal induces greater absorption of low-energy photons and consequent "beam hardening" - the process by which the average beam energy increases. Because of the higher energy, the dominant photon-matter interaction mechanism is no longer transmission but Compton scattering, causing part of the photons to be detected at different angles than the ones expected from transmission. Thus, the assumptions of transmissive model which the image reconstruction algorithms are based on are not met anymore, and the estimation of the attenuation coefficient along each projection may be altered. In addition, the higher absorption reduces the photon flux to the detectors, thus increasing the statistical noise of the measures. Finally, the sudden and sharp change in attenuation coefficient at the edges between metal and biological tissues detemines non-linear partial volume effects, which affect the accuracy of the image reconstruction.

Non-learning-based approaches

Numerous approaches have been proposed in the literature for MAR in CT and can be grouped in acquisition improvements, sinogram completion, iterative reconstruction and image post-processing methods [132].

Regarding the acquisition improvements, the most noteworthy development is the introduction of dual-energy CT. This technique combines two polychromatic beam acquisitions to synthesise a virtually monoenergetic image, thus reducing the impact of beam hardening effects. Phantom studies demonstrated a substantial 74% reduction of the metal artefact from hip implants, although at the expenses of a slight decrease of the contrast-to-noise ratio [133].

Sinogram completion methods refer to the synthesis of missing or corrupted projection data. In the sinogram space, the artefacts are localised around the metal trace, so theoretically they can be more easily compensated for [134]. The typical algorithm consists of segmenting the metal trace in the sinogram space, replacing the corrupted data and finally reconstructing the CT image (e.g. through filtered back projection). Such approaches work best on the original raw data, however they can also be applied to virtual sinograms: the reconstructed image is first forward projected into the sinogram space (virtual sinogram), the corrupted data is corrected and then the image is reconstructed back to its original space. To this category belongs the current state-of-the-art MAR approach on reconstructed CT, namely the normalized MAR (NMAR) [135]. The algorithm, reported in Fig. 2.11, initially reconstructs the image and generates a metal-only image and a prior image (including bone, soft tissue and air) using intensity thresholding. These are both forward projected and the original sinogram is normalised with the sinogram of the prior image. Interpolation from neighouring pixels is performed over the metal trace, the sinogram is denormalised and the final corrected image is reconstructed with back-projection. While correcting for the metal artefact, the interpolation step of NMAR and similar sinogram completion techniques typically introduces new interpolation errors and loss of details close to the metal. In

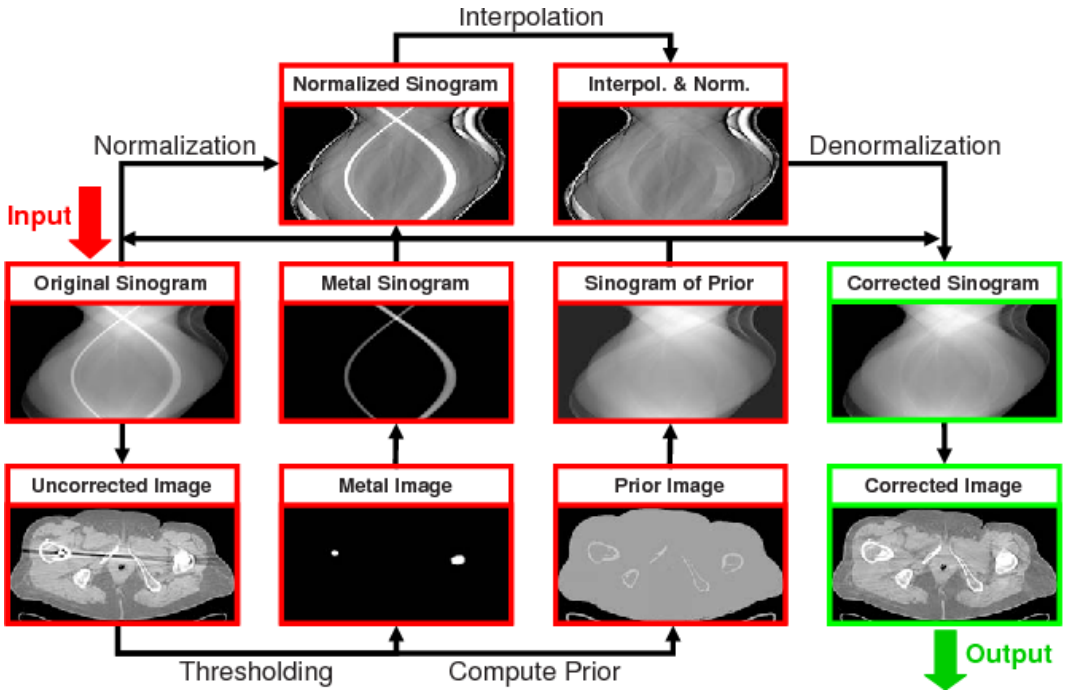


Figure 2.11: Normalized Metal Artifact Reduction (NMAR) algorithm. Intensity thresholding is initially applied to generate a metal-only image and a prior image (including bone, soft tissue and air). All the images are forward projected and the prior sinogram is used to normalise the original one. The metal trace is then replaced with interpolation from neighbouring intensities and the corrected sinogram is denormalised and back-projected to the image domain. (Image from Meyer *et al.* [135])

.

an attempt to address these limitations, Treece [134] introduced an iterative sinogram correction algorithm called Refined MAR (RMAR), where the prior is not obtained for the whole image at once but it is iteratively estimated independently for each projection angle, by comparing with uncorrupted areas, and by using a bitonic filter for the interpolation. The RMAR algorithm was developed for reconstructed images and, although slower than the more traditional NMAR, demonstrated a reduction of secondary interpolation artefacts and better details preservation. However, such methods rely on the quality of the image-derived priors, and residual metal artefact or severe photon starvation effects might still be present in the corrected image. Finally, in the absence of actual raw data from the scanner, the need to forward-project onto the sinogram space and back-project into the image space typically requires

making assumptions on the scanner and acquisition protocol (often not available). It also causes accumulation of interpolation errors.

Iterative reconstruction algorithms define the MAR as an optimisation problem: starting from an initial reconstructed image estimate, they compute its virtual sinogram and compare with the real sinogram, and update the reconstructed image so as to minimize the difference between the virtual and the real sinogram. In the optimisation, the metal trace is either ignored, downweighted or corrected with tailored physical model assumptions or prior information. However, they tend to be very slow and more importantly they require access to the real raw data, often not accessible on commercial CT scanners [132].

Finally, a few examples of MAR as post-processing on reconstructed CT images are present in the literature. They typically use filtering and morphological operations and can in some cases incorporate shape prior information to reduce the streaks throughout the FOV (e.g. Naranjo *et al.* [136]). Despite the clear advantage of not requiring the projection data, they have not been shown as effective as sinogram completion techniques, especially in more severe artefacts such as in the pelvis.

Deep learning approaches

Traditional physics-based or iterative reconstruction methods are now being challenged by novel deep neural network approaches, which are data-driven and less dependent on physical model assumptions. However, most methods [137, 138] are trained in a supervised fashion, relying on pairs of images with artefact and their respective ground truth without artefact. During training, deep learning networks learn how to regress the denoised image from the corrupted image by comparing their output to the ground truth. With metal artefacts, real ground truth images (with metal objects in the field-of-view, but no artefacts) cannot be acquired. Thus, such supervised methods need to employ surrogate ground truths, such as pre- and post-operative paired data or simulated images.

An example of correction directly in the image domain is the work proposed by Zhu *et al.* [139], where a 2D fully convolutional network is trained first on a simulated anthropomorphic head phantom with dental implants, and then on a real cylindrical phantom. In their work, testing is always performed on the same type of data as training, so the generalisability of this approch to real clinical data is not demonstrated. Similarly, Huang *et al.* [140] proposed a patch-based CNN learning the residuals between a corrupted image and the artefact-free one, and showed qualitative clearer reconstruction on 15 real clinical images compared to standard sinogram completion with linear interpolation. A supervised learning in the sinogram space was instead proposed by Park *et al.* [141], who however claimed efficacy on primary beam hardening effects only, due to the limitations of the training data synthesisation process. More promising strategies combine deep learning with traditional sinogram completion algorithms, in particular with NMAR. Gjesteby *et al.* [142] initially proposed the use of a CNN as a sinogram refinement step after application of NMAR. They further extended their approach to a patch-based two-branch CNN to enforce residual learning: one branch receives as input the NMAR corrected patch, while the other processes a corresponding "artefact-only" image, obtained by subtracting the original corrupted image from the same image after low-pass filter; the two branches are finally merged to produce an artefact-free patch. They also applied a perceptual loss to train their network: using a pretrained feature extractor, the corrected and the ground truth (artefact-free) images are compared in a feature space instead of the traditional image space. Such modifications proved to increase the performance and reduce over-smoothing effects on simulated data [138]. A different approach was proposed by Zhang *et al.* [143], where a patch-based 2D CNN is adopted to fuse different existing sinogram-based MAR techinques, concatenated and used as multi-channel input. The CNN output is then segmented into bone, water and air with a k-means clustering algorithm to produce a prior image, which can be finally applied for the original sinogram interpolation.

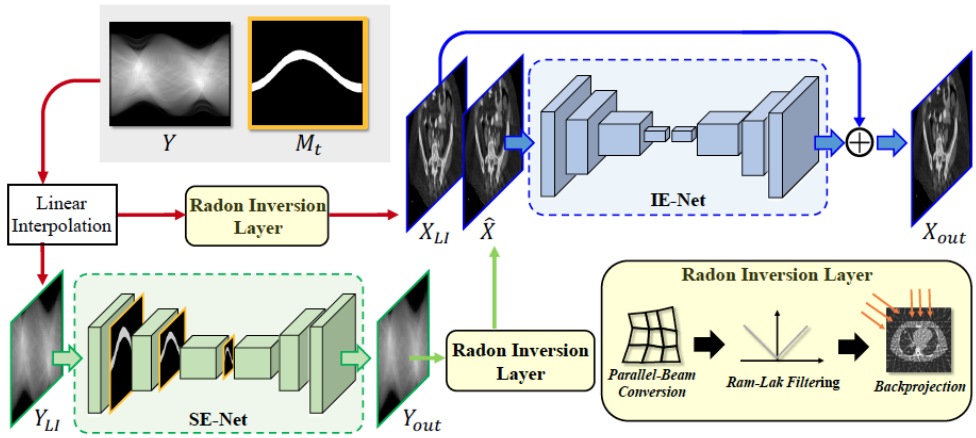


Figure 2.12: Dual domain network (DuDoNet) for metal artefact reduction in CT. The method combines a CNN to correct the sinogram (SE-NET) and a CNN to correct the reconstructed image (IE-Net), connected by a differentiable Radon inversion layer. (Image from Lin *et al.* [144]).

While the methods discussed so far operate either in the image or in the projection domain, Lin *et al.* [144] recently proposed to combine the two domains. They argue that correcting in the image domain is hampered by the structured and non-local nature of the artefact, and on the other hand sinogram interpolation methods introduce secondary artefacts. They thus suggest to jointly train two CNNs - one in the image domain (IE-Net), the other in the sinogram domain (SE-NET) - which are connected by a novel reconstruction layer (Radon inversion layer). Their combined network Dual Domain Network (DuDoNet) is shown in Fig. 2.12. By having a differentiable reconstruction layer, they are able to backpropagate information from the image-domain to the sinogram-domain, avoiding the introduction of secondary artefacts. This is enforced with a training loss including a sinogram recovery loss (corrected sinogram vs. artefact-free sinogram), an image recovery loss (corrected image vs. artefact-free image) and a Radon consistency loss (sinogram-corrected image before the image network vs artefact-free image).

Generative Adversarial Networks for MAR in CT

A deep learning model that has become very popular recently is the Generative Adversarial Network (GAN) [145]. The basic GANs are composed of two

elements: a generator and a discriminator. The former is meant to generate samples from the same distribution as the training data, while the discriminator aims at learning to differentiate between real samples and generated (fake) ones. During training, the generator uses the discriminator classification to learn to produce more realistic samples, and thus tries to fool the discriminator into classifying fake samples as real. At the same time, the discriminator keeps trying to improve its differentiation ability. A schematic representation of the GAN principles is shown in Figure 2.13. GANs find their theoretical basis in game theory, where the generator and the discriminator act as two players trying to maximise their output while not having control on the actions of the other player. More formally, let us consider the generator as a function $G_{\theta_G} : X \to Y$ mapping between a domain X and the domain of the training samples Y, and dependent on the parameters θ_G. The discriminator can be modelled instead as a function $D_{\theta_D} : Y \to \{0, 1\}$, associating each input to the probablity of being real or fake, and dependent on the parameters θ_D. The discriminator is in effect a binary classifier, thus its loss function can be defined as a cross-entropy loss:

$$\mathcal{L}_D(\theta_D, \theta_G) = -\frac{1}{2}\mathbb{E}_{y \sim p_{data}} \log D_{\theta_D}(y) - \frac{1}{2}\mathbb{E}_x \log(1 - D_{\theta_D}(G_{\theta_G}(x))) \tag{2.6}$$

where $\mathbb{E}$ indicates the expectation value over the respective distribution. Considering a zero-sum formulation for the game, the loss function for the generator can be defined as:

$$\mathcal{L}_{GAN}(\theta_D, \theta_G) = -\mathcal{L}_D(\theta_D, \theta_G) \tag{2.7}$$

The optimal solution for the joint estimation of all the model parameters is:

$$\theta_D^*, \theta_G^* = \arg\min_{\theta_G} \max_{\theta_D} \mathcal{L}_{GAN}(\theta_D, \theta_G) \tag{2.8}$$

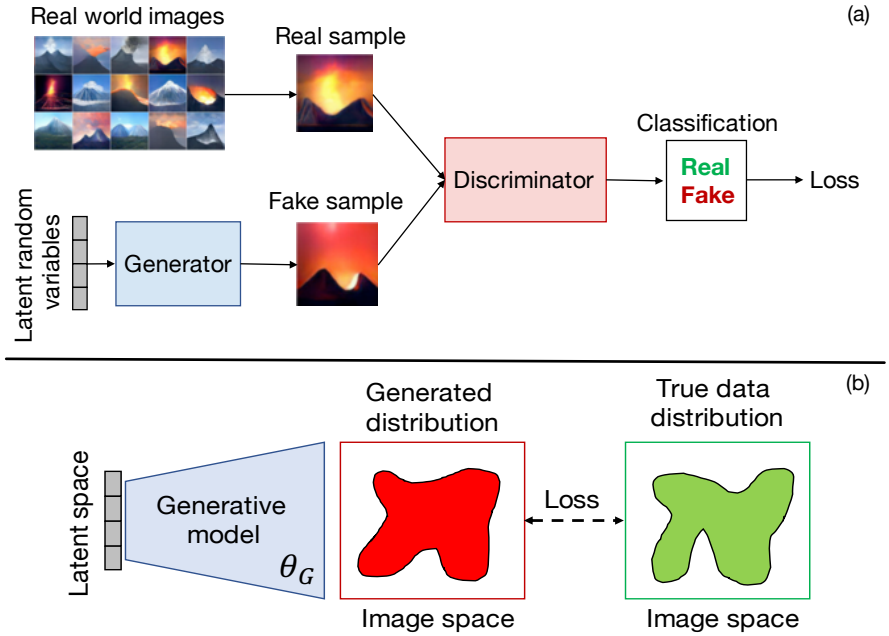


Figure 2.13: Generative adversarial networks. (a) Schematic representation, (b) Working principle: the generator implicitly learns the probability distribution of the real data.

This formulation is also called *minimax* game and its solution is given by a Nash equilibrium, meaning that the optimal parameter θ_D^* is a local maximum of $\mathcal{L}_D$ and jointly θ_G^* is a local minimum of $\mathcal{L}_G$ [145].

The GAN formulation is an unsupervised generative model that implicity learns the probability distribution of the training set. As a result, by sampling from this learnt distribution, new samples can be produced. This generative property is one of the main strengths of GANs, as it allows to still generate realistic images without the need of paired data, *i.e.* without the need of a ground truth. As a matter of fact, GANs have found most common applications in image-to-image translation, style transfer and image synthesis, such as mimicking art styles (e.g. Van Gogh or Monet) from real photographies, or generating realistic pictures of non-existing people. Compared to other deep learning methods (e.g. autoencoders or variational autoencoders), the generators have been regarded as producing qualitatively less blurred, more realistic and more diverse samples [145, 146]. These properties make GAN approaches

very appealing in a problem such as metal artefact reduction, since in CT this can be thought both as a denoising and a style transfer problem. However, GANs come with a series of non trivial drawbacks, the most important being the instability of training. The convergence of the model to the Nash equilibrium is difficult to reach, and requires training monitoring and balancing of the different adversarial components [146]. For instance, if the discriminator becomes too accurate and its loss converges, no information can be propagated back to the generator to improve the fake samples (vanishing gradient problem). Moreover, GANs typically require larger training datasets than supervised settings: the more data are seen at training, the better the real data distribution can be modelled and thus the more realistic samples are produced. The training instability and the need for big dataset make GANs often unsuccessful for 3D images, confining their application in the medical imaging field only to 2D slices or patch-based approaches. Least but not last in importance, GANs have been demonstrated to hallucinate features in the generated samples and introduce geometrical distortions [147]. This is clearly undesirable for the medical imaging applications, since any modification or domain transfer generated through GANs need to preserve the anatomical content of the images.

Because of these difficulties, the application of GANs to the MAR problem in CT has been limited to a supervised setting, where the generator output is still compared to the ground truth, and the adversarial training works mostly as a regulariser [137, 148]. In this regard, the work by Wang *et al.* [137] for MAR for cochlear implants is exemplar. They introduced a conditional GAN (cGAN) trained on paired pre-operative CT and post-operative CT images of the ear. The original GAN loss function is modified to jointly minimise the L1 distance between the generated artefact-corrected post-operative CT and the ground truth pre-operative CT:

$$\mathcal{L} = \min_{\theta_G} \max_{\theta_D} \mathcal{L}_{GAN}(\theta_D, \theta_G) + \lambda L_1(\theta_G) \tag{2.9}$$

where $L_1(\theta_G) = ||y - G_{\theta_G}(x)||_1 = \sum_{i=1}^{n} |y - G_{\theta_G}(x)|$, with y indicating the pre-operative ground truth and $G_{\theta_G}(x)$ the generated artefact-corrected CT from the post-operative input x. The network was trained on 90 CT pairs from which 2D slices were extracted and preprocessed with a piecewise linear intensity normalisation, to further enhance the image contrast. The authors reported a significant reduction in segmentation error of the intra-cochlear structures from their post-operative CT test set, operating directly in the image domain (as opposed to sinogram space). They also acknowledged that, while having a 3D network would be theoretically better for volumetric anatomical modelling, they were not able to obtain improved performance in 3D.

Still, the most appealing characteristics of GAN approaches is the ability to work in an unsupervised setting, where no ground truth is required. Inspired by current developments in image-to-image translation, Liao *et al.* [149] recently proposed a valid alternative to the supervised setting in the case of MAR. They introduced the Artefact Disentanglement Network (ADN), an unsupervised adversarial training scheme to disentangle the artefact from the anatomy appearance in CT images, showing state-of-the-art performance on both synthetic and real data (Figure 2.14). Given two unpaired images - one with metal artefact, the other without, but from different subjects, the ADN is trained to: (1) remove the artefact from the first input and (2) add the same artefact on the artefact-free image. By simultaneously trying to remove and synthesise artefact, with the aid of the discriminators, the network learns to produce realistic artefact-free images. In addition, starting from an artefact-free CT, the output of the artefact synthesis branch is forward passed through the artefact removal one so as to reproduce the original input CT. This "cycle" consistency promotes the geometrical preservation of the image anatomical content, counteracting the geometrical distortion problems of traditional GANs. Liao's work defines a relevant starting point for the multimodal metal artefact reduction technique presented in this thesis. Thus, further details about the Artifact Disentanglement Network will be provided in Chapter 6.

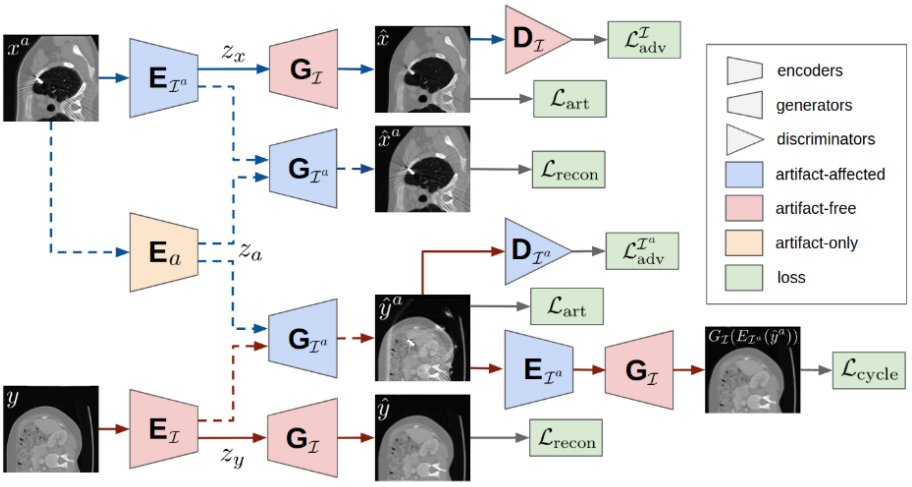


Figure 2.14: Overview of the Artifact Disentanglement Network proposed by Liao *et al.* [149] for metal artefact reduction in CT. The network is composed of: three encoders E, mapping the input images to a lower dimensional representation; two decoders acting as generators G and mapping the encoded lower dimensional representation back to the image domain, with or without the artefact; two discriminators D, classifying images (with or without the artefact) as either real or fake. Corrupted images undergo both self-reconstruction and artefact removal. Artefact-free images undergo self-reconstruction, artefact synthesis and subsequent artefact removal for cycle consistency.

MR-driven MAR in CT

A last set of approaches still in its infancy is the use of MRI as prior information to correct the CT. Research in this area has focussed only on metal artefact from dental implants, and typically in the setting of trimodality (CT/MR/PET) acquisitions for radiation therapy. A pilot study on two subjects was presented by Anderla *et al.* [150] and extended by Delso *et al.* [151]. Similarly to MR-based PET attenuation correction methods, they proposed to derive a pseudo-CT from the MR image (based on simple thresholding segmentation of the MRI) and replace the corrupted voxels in the real CT with the pseudo-CT values. They however acknowledged that this fails in case of MRI signal void, thus it is not able to reduce the artefacts in cortical bone regions.

A similar approach was proposed by Park *et al.* [152] but in a more general setting without the need of multimodality scanner. Their method is based on CT-MR intra-subject registration, and subsequent replacement of corrupted

CT pixels with HU values from an artefact-free region. The MRI data is used to estimate the transformation between artefact-free slices and corrupted ones, which is then applied to the respective CT slices to replace the corrupted pixels. While conceptually simple, the method relies on the assumption that artefact-free slices are close enough to the corrupted ones to provide an accurate registration. Thus, its application to large hip implants with extended corrupted FOV is not straightforward. Lastly, Nielsen *et al.* [153] more formally modeled the CT artefact reduction as a CT value regression from corresponding MR patches. They proposed to learn the joint probability distribution of artefact-free CT and MR intensities using Gaussian kernel density estimation, and model the artefacts as zero mean Gaussian noise on real CT values. Bayes' rule can then be applied to infer the corrected CT value from corrupted ones. Also, they interestingly employ an expectation-maximisation framework to estimate per-patient model parameters, making the correction subject-specific and independent on the MR acquisition. Although the methodology is sound and mathematically robust, at present its validation is limited to qualitative comparisons and lack of thorough testing of the modelling assumptions.

The discussed works are all limited to dental implants or surgical clips, whose artefacts in the MRI space are generally more localised and less detrimental than the larger hip implants. Also, they are mostly simulations or small cohort studies, thus a more thorough validation is needed to demonstrate their effectiveness and scalability. On a broader perspective, these methods have been shown to prevent the introduction of secondary artefacts typical of sinogram interpolation techniques. On the other hand, they require at least two modalities and are very sensitive to accurate registration between MR and CT. Nonetheless, they all show value in combining multimodal information if available, indicating a promising direction of investigaton.

2.4.2 MAR in MR images

In Magnetic Resonance Imaging, metal objects induce local magnetic field inhomogeneities that cause intensity and geometrical distortions in the recon-

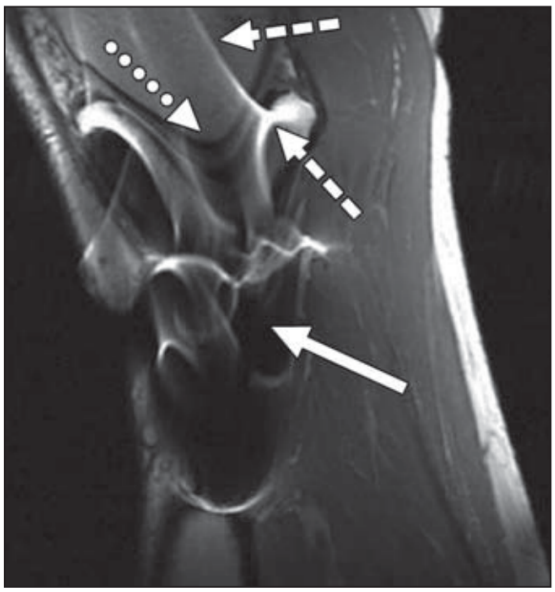


Figure 2.15: Knee MRI scan affected by metal-induced artefacts due to the presence of stainless steel screws. Darker areas caused by signal loss are indicated by the solid arrow, and pile-up effects by the dashed arrow. The dotted arrow indicates an area of geometrical distortion. Image from Hargreaves *et al.* [23].

structed image. Due to the highly rigid structure or metal, no signal is received from the implant itself, which thus appears dark in the image. Moreover, the sudden change of magnetic susceptibility between metal and biological tissue induces severe variations in the static magnetic field dependent on the implant shape, size and type. This determines a change of the characteristic resonance frequency, leading to significant signal loss and incorrect spatial information encoding. As illustrated in Figure 2.15, these susceptibility artefacts typically appear as blackened areas at and close to the implant, partially shadowing the neighbouring structures (due to signal loss), together with bright spikes or geometrical distortion due to spatial displacement of the signal and accumulation in neighbouring regions (pile-up). Finally, being based on the chemical shift between water and fat resonance frequency, fat suppression protocols tend to fail in the presence of metal, because of the metal-induced frequency shift [23].

In MRI research, efforts have focused mostly on image acquisition improvements: tailored MR sequences such as MARS [17], MAVRIC [154] or SEMAC [155] have proven effective in reducing the extension of the shad-

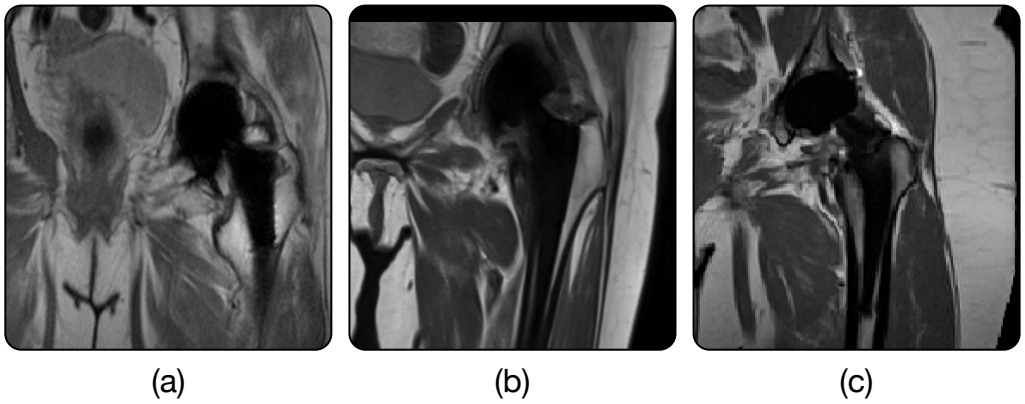


Figure 2.16: Examples of metal artefact reduction sequences in pelvic MRI. (a) MAVRIC (General Electric, from `http://newsroom.gehealthcare.com/mavric-sl-imaging-mr-conditional-implants-to-spot-adverse-reactions-fast/`). (b) SEMAC (Siemens Healthineers, from `https://www.siemens-healthineers.com/en-us/magnetic-resonance-imaging/clinical-specialities/msk-imaging`). (c) MARS acquisition [17], used in this thesis project .

owing, but cannot completely eliminate it, making the clear visualisation of the implant in MRI impossible (Fig. 2.16). To the best of my knowledge, no post-acquisition method has been proposed in the literature to reduce the artefact in reconstructed MR images. The only relevant work is a phantom study on metal artefacts from dental implants [156], which incorporates CT information by registering the two images, segmenting the phantom teeth from the CT and inpainting them on the MR image. However, the main focus of this work was the evaluation of the artefact extension with different MRI sequences and only limited description on how the artefact was compensated for in both modalities was provided. Although limited to a phantom study and thus not proven effective in any real clinical application, this preliminary work shows the potential of using CT information to recover the MRI signal and it opens up new directions of investigation for metal artefact reduction in MRI.

Summary box: Challenges in Metal Artefact Reduction

- Physics-based or iterative MAR techniques in CT still have residual artefacts and can introduce secondary artefacts
- Deep learning strategies for MAR in CT require either artefact-free ground truth images and/or large amount of training data
- MAR in MR only relies on tailored acquisition protocols that reduce the shadowing effect but cannot fully restore the signal
- No post-reconstrution MAR techniques in MR have been proven effective in clinical data

2.5 Summary

The review of the current literature highlighted a lack of effective image analysis methods for combining and segmenting CT and MRI information in the presence of hip implants. The main bottlenecks are the large variability both of the anatomical structures under study and of the acquisition protocols currently used, which often limit the generalisability of the existing approaches. Moreover, most methods applied to the pelvic anatomy do not consider the case of hip implants, which heavily affect the quality of the images and introduce strong artefacts. However, the fusion of skeletal and muscular information using a multimodal approach could help better visualise the patient-specific anatomy and support the extraction of quantitative and clinically relevant imaging biomarkers. The following chapters in this thesis will explore this potential and will illustrate automated processing pipelines specifically developed for multimodal imaging in the presence of hip implants.

Chapter 3

Data description and preprocessing

The dataset used in this thesis project comprises CT and MR images from 87 subjects that have been treated with MoM hip arthroplasty. Data was selected retrospectively, based on availability of images from both modalities, among patients that were referred to Charing Cross Hospital (London, UK) for unexplained hip pain between 2006 and 2012. Summary demographic statistics are reported in Table 3.1. All the MR images were acquired on a Siemens MAGNETON Avanto 1.5T scanner, with the Metal Artefact Reduction Sequence (MARS) presented in Sabah *et al.* [26]. This standard protocol includes the collection of two T1-weighted Turbo Spin Echo highly anisotropic images: an axial acquisition (TE = 8 ms, TR = 509 ms, typical imaging resolution = $0.78 \times 0.78 \times 7.02$ mm^3) and a coronal acquisition (TE = 7.1 ms, TR = 627 ms, typical imaging resolution = $1.25 \times 1.25 \times 6.00$ mm^3). Most of the CT acquisitions were performed on a Siemens SOMATOM Sensation 16 scanner, but 8 cases acquired on a Siemens SOMATOM Definition AS+ machine. Tube voltage varied in the range [80, 120] kVp.

Bal and Lowe radiological score [27] of gluteus medius atrophy was available for 35 subjects, while Cobalt and Chromium blood concentration information was retrieved for another 35 subjects (only 16 cases had both pieces of information).

	Females	**Males**	**Total**
Number of subjects	54	33	87
Unilateral cases	36	26	62
Bilateral cases	18	7	25
Mean [Range] age	53.74 [23, 74]	56.88[35, 70]	54.93 [23, 74]
MRI within 1 month from CT	42	21	63
MRI within 6 months from CT	4	2	6
MRI within 12 months from CT	3	2	5
MRI within 24 months from CT	4	6	10
MRI within 37 months from CT	1	2	3

Table 3.1: Summary of the dataset demographic statistics. Unilateral cases refer to patients with only one implanted side, while bilateral cases to subjects with implants on both hip joints. The mean age at the time of CT acquisition is also shown, together with the number of cases grouped according to the time difference between the CT and the MR acquisition.

Among the available CT-MRI pairs, data from 11 subjects were randomly selected for manual segmentation in order to build the template dataset for the automated segmentation tool proposed in this work. The generation of the template dataset is described in Sec. 3.2, after the introduction of the generic preprocessing steps implemented for all the dataset in Sec. 3.1.

3.1 Image quality enhancement

The images in the described dataset were acquired in routine clinical setting, and thus with standard clinical quality (as opposed to research image quality). Specifically, no Metal Artefact Reduction technique was applied on the CT data, while the MR acquisition was limited by scanning time, resulting in a series of multi-planar 2D acquisitions with highly anisotropic resolution (up to a factor of 10 difference between the in-plane and out-of-plane resolution). In addition, for both modalities it was not possible to access the raw data, but only the reconstructed images in DICOM format. Thus, a series of preprocessing steps have been implemented in this work to enhance the quality of clinically acquired data.

CT preprocessing

The reconstructed CT images have initially been corrected for metal artefact using the wxDICOM software[1], based on the Refined Metal Artefact Reduction method proposed by Treece [134] and described in Sec. 2.4.1. After conversion to the easier-to-handle NIfTI format, a cubic interpolation scheme was utilised to resample the corrected CT to isotropic resolution of $1 \times 1 \times 1\text{mm}^3$. Finally, thresholding and morphological operations were applied to obtain a body mask and exclude signal from the CT couch. A summary of the processing steps is shown in Figure 3.1. These corrected CT images have been used for all the experiments discussed in this thesis, with the exception of Chapter 7, where a novel method for multimodal metal artefact correction is introduced.

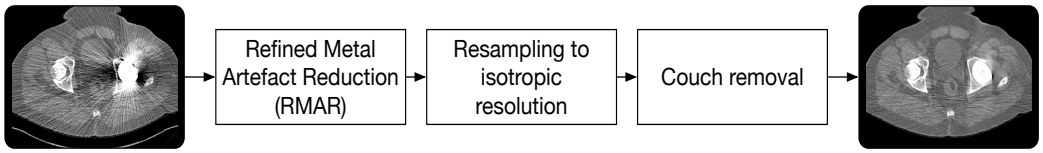


Figure 3.1: Processing pipeline for image quality enhancement in CT.

MRI preprocessing

For each subject, the axial and the coronal MRI acquisitions were first converted from DICOM to NIfTI format. Because of the typically large scanned FOV, the images appeared heavily affected by bias field inhomogeneities. Hence, a global N4 bias field correction [157] was first applied to each acquisition (using SimpleITK). A residual bias field effect was however still visible, especially in the subcutaneous adipose tissue. To compensate for it, I automatically generated masks of the adipose tissue regions, using the Expectation-Maximisation segmentation algorithm proposed by Van Leemput *et al.* [44]. These masks were utilised to estimate the residual bias field and apply a compensating correction on the whole FOV, still using the N4 algorithm. After this correction, the intensity of the two MRI acquisitions were normalised using a histogram matching technique that aligns specified quantile values of the two

[1] `https://mi.eng.cam.ac.uk/Main/GMT_wxDicom`

intensity histograms and rescales all the intensities accordingly (SimpleITK HistogramMatchingImageFilter). The two views were finally combined into a single volume at 1×1×1 mm^3 resolution using the super-resolution reconstruction (SRR) algorithm developed by Ebner *et al.* [158]. This method generates a single high-resolution 3D volume from a series of orthogonal MRI stacks of low-resolution 2D slices. The use of a SRR algorithm has several advantages. Firstly, it allows us to compensate for the highly anisotropic resolution of the acquired data, which would adversely affect the subsequent registration to the CT image; secondly, it compensates for the lack of 3D MRI acquisition, as 2D multislice sequences are usually preferred in clinical routine due to scanning time constraints. The MRI quality enhancement pipeline is shown in Figure 3.2, while an example of the sole effect of the SRR is reported in Figure 3.3.

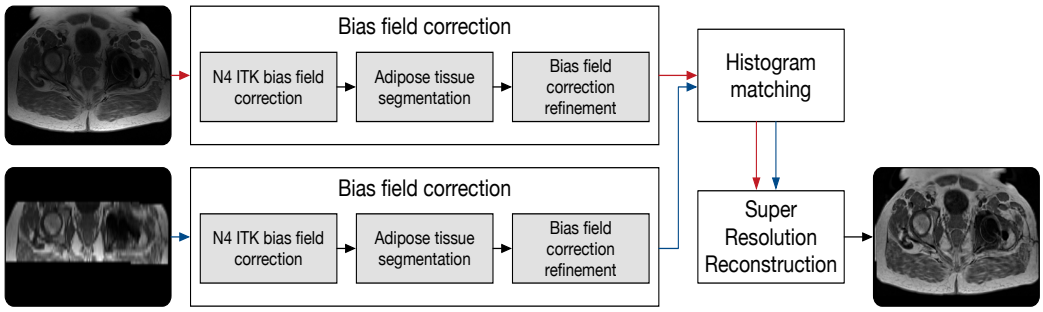


Figure 3.2: Processing pipeline for image quality enhancement in MRI.

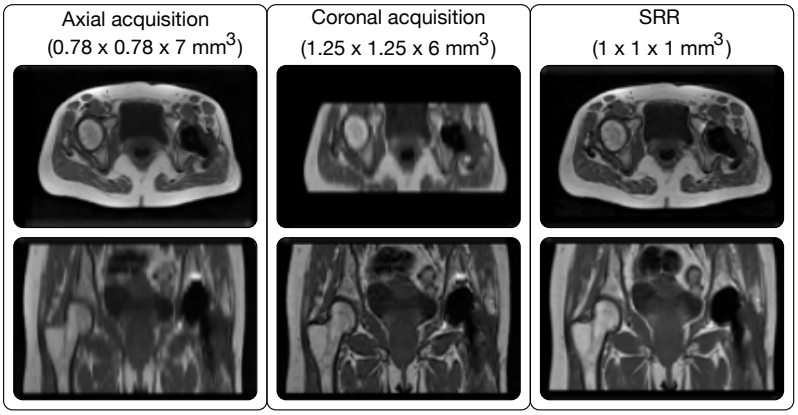


Figure 3.3: Example of the input axial and coronal MRI acquisitions and the respective super-resolution reconstructed (SRR) MRI, showing how low-resolution information is combined to maintain high resolution in both imaging planes.

3.2 Template dataset generation

The work presented in this thesis is based on the availability of a series of template images with associated manual labelling of structures of interest. The manual segmentation was needed both for the actual automated segmentation pipeline and for quantifying the performance of the proposed tools. As no such dataset is publicly available in the context of hip replacement, one of the contributions of this thesis project is the generation of a template dataset of paired CT-MR pelvic images and associated musculoskeletal segmentation.

The template dataset consists of 11 subjects (10 unilateral, 1 bilateral), whose MR images and CT were acquired on the same day. For each subject, the CT and MR images were preprocessed for image quality enhancement as discussed in the previous section (Sec. 3.1).

Given their higher contrast for bony structures, the CT images were used to manually delineate[2] the pelvic bones, the femora and the implant(s). Manual segmentation of Gluteus Maximus (GMAX), Gluteus Medius (GMED), Gluteus Minimus (GMIN) and Tensor Fasciae Latae (TFL) was performed on the MR image after super-resolution reconstruction was applied. These muscles are part of the abductor group and were deemed as the most relevant for clinical assessment by the clinical collaborators of this project.

The two modalities were then registered into the same reference frame - full details on the intra-subject registration method will be introduced in Chapter 4 - and the respective bones and muscles manual segmentation masks were merged into a single multi-label segmentation image. I finally performed further manual refinement, in order to guarantee non-overlapping regions between the segmented structures.

The template dataset was then organised into two subsets: the implanted hip sides (12 sides) and the non-implanted ones (10). For the sake of simplicity, I will refer to these latter as the healthy sides, although I acknowledge that the absence of the implant does not imply absence of pathology nor implant-

[2]Manual segmentations were performed using ITKSnap tools. `www.itksnap.org`

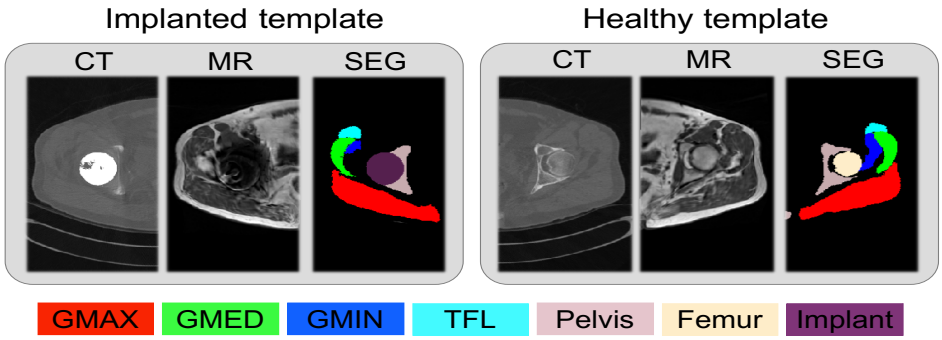


Figure 3.4: Examples from the template sets, composed of CT, the respective non-linearly registered MRI (see Chapter 4 for algorithmic details), and their fused segmentation of hip joint bones and abductor muscles. An implanted (left) and a healthy (right) template sides are shown. A single slice from the 3D volume is displayed for illustration purposes.

induced artefact. In order to separate and group all the implanted and the healthy hip sides, each template set of CT, MR and label images was split along the CT-derived sagittal axis of symmetry, and reoriented according to the presence or absence of an implant. For this task, I developed a symmetry-detection algorithm that, in analogy with the mechanics of rigid body, computes the inertia tensor of the CT image using its intensities as mass values. Specifically, to balance the high intensity values of the implant typically present only on one side, the inertia tensor is obtained from the input CT after rigidly registering and averaging it with its left-right flipped image. The mid-sagittal plane of symmetry can then be extracted by the inertia tensor. Simple thresholding at 3000 HU is then exploited to assess the presence of implants in each hip side of the original CT. All implanted sides are then oriented to be on the left and the healthy sides on the right.

Within each subset, all the templates were rigidly aligned into their mid-space through the robust group-wise registration framework proposed by Klemt *et al.* [108]. Only the header orientation matrices were updated to avoid intensity resampling.

Figure 3.4 shows two examples from template sets, one for the healthy and one for the implanted hip sides.

3.3 Landmarks annotation

The eleven subjects selected for the template dataset generation were also manually annotated with landmarks. The annotation was performed in the original CT and MR image spaces, i.e. before the intra-subject registration but after the application of the quality enhancement preprocessing. For this task, I designed a protocol for the selection of imaging landmarks in CT and MRI of patients with hip implants, which was reviewed by an expert musculoskeletal radiologist. I selected five anatomical landmarks on each hip side (total of ten landmarks per subject), of which three are located on skeletal and two on muscular structures. The landmarks were chosen according to three criteria: (1) being easily identifiable in both imaging modalities; (2) being spread across the full field-of-view; (3) being located on the structures I aim to segment automatically. In addition, the selected landmarks are not gender related, not age related and very little susceptible to normal variants. Below is the detailed description of each landmark, together with an example for each of them.

(1) Greater trochanter

On the **coronal** view I search for the top tip of the greater trochanter in the middle slice of appearance of the femur. In the described dataset, the slice number increases as posterior to anterior. I consider as first slice the one where the first hyper-intense spot appears in the middle of Gluteus Maximus, while the last slice is the one where the femoral head is lastly visible before disappearing. Given the position in millimetres of these two slices, the middle position is computed as (initial position + final position)/2. I consider the closest visible slice to this mid-point.

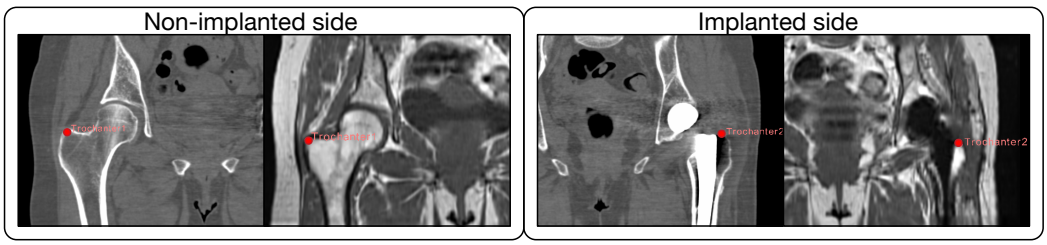


Figure 3.5: Greater Trochanter landmark

(2) Tensor Fasciae Latae (TFL)

On the **axial** view I search for the most anterior and medial tip of the TFL in the middle slice of appearance. As for the femur, I search for the first and the last slice of appearance of the TFL. The first (inferior) slice is the one where the muscle firstly emerges from the tendon (i.e. where the tendon seems to enlarge and it is not just a thin stripe anymore). The last (superior) slice is the one where the TFL gets completely covered by the Gluteus Medius and Minimus (afterwards the top tip of the TFL is not visible anymore). Using these slices, I compute the mid position in millimetres and in the respective slice I choose the most anterior and most medial tip.

Figure 3.6: Tensor Fasciae Latae landmark

(3) Pelvic Brim

On the **axial** view, I select a point on the pelvic brim. This corresponds to the anterior-medial corner of the pelvic bone just above the femoral head. Moving inferior to superior, I select the first slice where the femoral head is not visible any more (not even as brighter "shadow" in the middle of the pelvic bone) and I choose the anterior-medial corner.

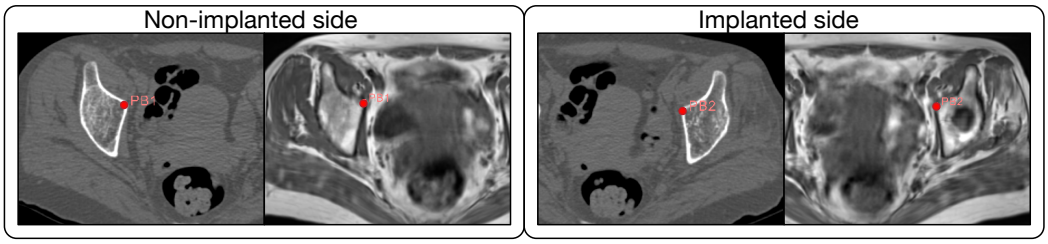


Figure 3.7: Pelvic Brim landmark

(4) Gluteus Maximus (GMAX) Sacrum origin

On the **axial** view, I select the point where GMAX inserts onto the coccyx. Moving inferior to superior, I select the first slice where this is clearly visible and in particular I choose the corner of the GMAX inserting onto the bone.

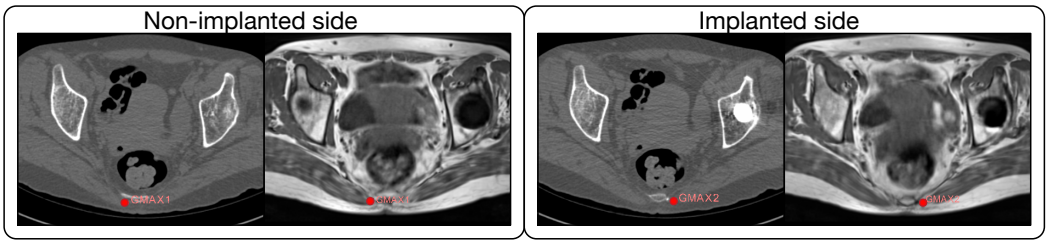


Figure 3.8: Landmark at the GMAX Sacrum origin

(5) Ischium

On the **axial** view, moving inferior to superior I choose the first slice where the pelvic bone splits into the ischium and the symphysis pubis. In this slice, I select the most external vertex of the ischium.

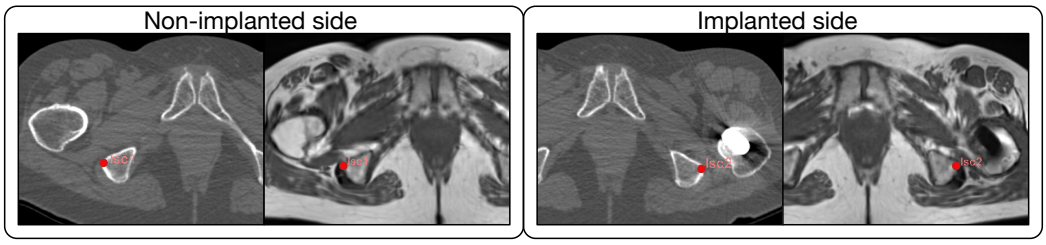


Figure 3.9: Ischium landmark

As the identification of clear landmarks on extended and generally homogeneous structures such as muscles is not trivial, to account for choice bias I performed the manual selection twice for each image at different times. The average reproducibility error across all the 11 subjects is reported in Fig. 3.10.

This protocol was designed for the purpose of characterising the registration error on the structures of interest throughout the FOV. I thus acknowledge that the described landmarks are not directly linked to clinical relevance.

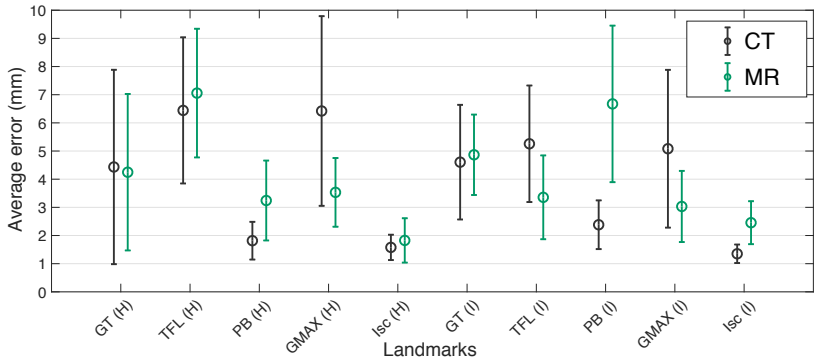


Figure 3.10: Manual selection reproducibility error for the 10 landmarks in CT and in MR images (mean and 95% confidence interval are reported). Landmarks: Greater Trochanter (GT), anterior tip of the Tensor Fasciae Latae (TFL), Pelvic Brim (PB), Gluteus Maximus sacrum origin (GMAX) and Ischium (Isc). (H) indicates the healthy hip side, (I) the implanted one.

Chapter 4

Multimodal registration with rigidity constraints

Motivation

Accurate alignment of CT and MRI is needed to combine their multimodal information.

Contribution

A novel nonlinear intra-subject registration algorithm is introduced, characterised by a diffeomorphic transformation model constrained to a strictly rigid behaviour on bones and implants.

The main aim of this work is the combination of information from different imaging modalities to ease the assessment of both muscular and skeletal structures in patients with hip replacement. Towards this aim, intra-subject image registration is a crucial step for merging information, as it allows to align anatomically corresponding points from different images. However, many challenges characterise this task.

First, intensity-based registration is heavily affected by noise and artefacts in the images. The image quality enhancement steps described in the previous chapter were designed to help with this problem, in particular by applying a metal artefact reduction technique on the CT data [134] and by combining

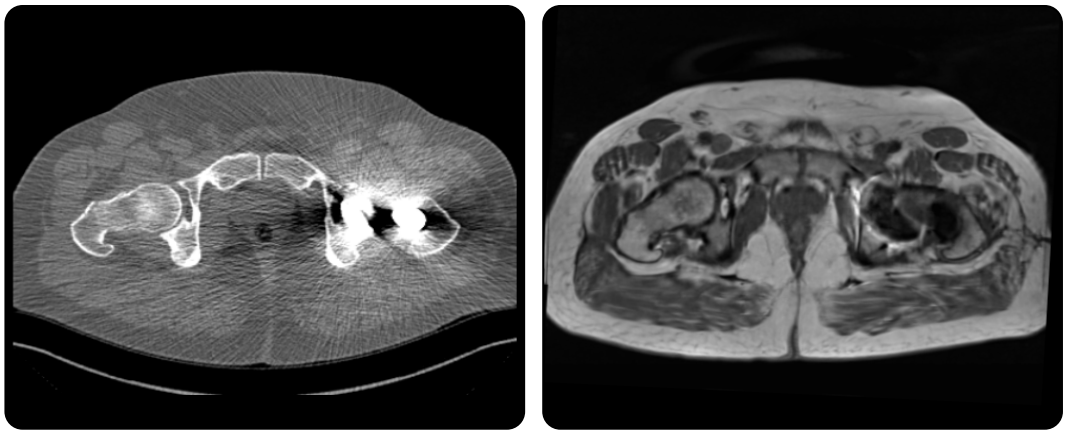


Figure 4.1: Example of different patient's pose in the CT scanner (left) and in the MRI (right). The difference in the couch shape determines difference in bone position and soft tissue deformations that cannot be recovered by simple global registration

multiple MR acquisitions into a single isotropic SRR MR [158]. Moreover, the images are characterised by large variability of the anatomical structures as well as typically large imaging FOV, thus the robustness of registration algorithms could be hampered by these factors.

Yet, the main challenge is the difference in the patient's pose within the two scanners. As illustrated in Figure 4.1, the CT scanners used for this project were characterised by a round couch, while the MR scanner had a flat one. This results in different pelvic positions in the two scanners, with varying amount of soft tissue compression, which cannot be recovered by a rigid or affine registration. On the other hand, a fully nonlinear registration would allow for unrealistic deformations in rigid structures such as bones and implants. A desirable transformation model for this task should therefore accomodate both a nonlinear deformation of soft tissue and preservation of rigidity in incompressible structures. State-of-the-art approaches for this problem rely on "soft constraints", *i.e.* regularisation terms added to the cost function which penalise deviation from rigid behaviour in defined areas [127, 128] or that enforce the alignment of corresponding masks from the two images [129]. These approaches, however, need to find a balance between the terms in the cost

function, which might still cause deviations from a strictly rigid transformation and therefore requires a careful selection of the rigidity penalty term weight. A "hard constraint" solution such as that proposed by Haber *et al.* [130] should be preferred, as it enforces a strict fullfilment of the rigidity constraints in the defined areas.

This chapter introduces a novel multimodal nonlinear registration framework able to compensate for a different patient's pose within the CT and the MR scanners while preserving the rigidity of bony structures. I will first present the mathematical formulation of the proposed method as a "hard constraint", its inclusion in a diffeomorphic framework and provide the details of the algorithmic implementation. The performance of the proposed approach will be assessed on the available template dataset, from which I will discuss its advantages and limitations. The work described here was initially presented at the 5^{th} Workshop of Computational Methods and Clinical Applications in Musculoskeletal Imaging (MSKI) held in conjuction with MICCAI 2017 [159], and then published in the Computer Methods and Programs in Biomedicine journal [160].

4.1 Transformation model

In order to prevent implausible deformation of bones, I introduced an intensity-based nonlinear registration framework that applies a strictly rigid transformation to all the voxels within specified masks. This method extends the mathematical formulation of hard rigid constraints presented by Haber *et al.* [130] to the 3D case and is embedded into a diffeomorphic framework.

Let's first consider the original formulation from Haber *et al.* [130]. Given a reference image $R : X \to R$, defined in the reference space $X \subseteq \mathbb{R}^3$, and a floating image $F : Y \to F$, defined in the floating space $Y \in \mathbb{R}^3$, the registration algorithm aims at optimising a transformation model $\boldsymbol{\phi} : X \to Y$ which is enforced to be rigid within specific areas defined by a set of masks M, with

$M_j \subset X \; \forall M_j \in M$, and nonlinear elsewhere. Therefore, given the cost function

$$\mathcal{C}(R, F; \boldsymbol{\phi}) = (1-\lambda)\, \mathcal{S}(F(\boldsymbol{\phi}(X)), R(X)) \; - \lambda Reg(\boldsymbol{\phi}), \tag{4.1}$$

where $\mathcal{S}$ is a measure of similarity between the reference and the warped floating image and Reg a regularisation term weighted by a coefficient λ, the optimisation problem is formulated as:

$$\max_{\phi} \mathcal{C}(R, F; \boldsymbol{\phi}) \quad \text{subject to } \boldsymbol{\phi}(\mathbf{x}) - \mathcal{R}_j(\mathbf{x}) = \mathbf{0} \quad \forall \mathbf{x} \in M_j, \, \forall M_j \in M \tag{4.2}$$

with $\mathcal{R}_j : \mathbb{R}^3 \to \mathbb{R}^3$ being a rigid transformation applied to all voxels $\boldsymbol{x}$ within the j-th mask. Using this formulation, the rigid constraint in the bony structures is embedded directly into the transformation model $\boldsymbol{\phi}$ as a "hard constraint". This differs from currently proposed approaches based on "soft constraints", which simply penalise large deviations from a rigid behaviour within specified regions, e.g. [127, 128] (Sec. 2.3.1, Eq. 2.3).

As introduced in Sec. 2.3.2, the preservation of rigidity in incompressible structures is not the only desirable characteristics of the transformation model. Topology preservation and inverse-consistency are also necessary to guarantee the biological plausibility of any analysis based on image registration. These properties are naturally provided by diffeomorphic transformation models, being differentiable one-to-one mappings. For this reason, in this work we take advantage of a stationary velocity field formulation for the transformation model, originally proposed by Arsigny *et al.* [161] and developed it into a registration framework by Ashburner [162]. Assuming a constant velocity field $\boldsymbol{v}$, the deformation field is obtained by integrating the velocity field over time. This can be computed by the Euler integration method, which breaks the integration path into many small time-steps h: $\boldsymbol{\phi}^{(t+h)} = \boldsymbol{\phi}^{(t)} + h\boldsymbol{v}(\boldsymbol{\phi}^{(t)}) = (\boldsymbol{x} + h\boldsymbol{v}) \circ \boldsymbol{\phi}^{(t)}$ [162]. Starting with an identity transformation at time $t = 0$, the velocity field is divided into n steps and the final deformation is given by a series of compositions of each small time step:

$$\begin{aligned}
\boldsymbol{\phi}^{(1/n)} &= \boldsymbol{x} + \frac{\boldsymbol{v}(\boldsymbol{x})}{n} \\
\boldsymbol{\phi}^{(2/n)} &= \boldsymbol{\phi}^{(1/n)} \circ \boldsymbol{\phi}^{(1/n)} \\
\boldsymbol{\phi}^{(3/n)} &= \boldsymbol{\phi}^{(1/n)} \circ \boldsymbol{\phi}^{(2/n)} \\
&... = ... \\
\boldsymbol{\phi}^{(n/n)} &= \boldsymbol{\phi}^{(1/n)} \circ \boldsymbol{\phi}^{((n-1)/n)}
\end{aligned} \tag{4.3}$$

The computation can be simplified by choosing n as a power of 2 and by applying a scaling-and-squaring exponentiation [161]: an initial scaling is applied to the velocity field to represent a small deformation and then the field is iteratively composed with itself to approximate the integration (squaring). For instance, choosing $n = 8$ leads to:

$$\begin{aligned}
\boldsymbol{\phi}^{(1/8)} &= \boldsymbol{x} + \frac{\boldsymbol{v}(\boldsymbol{x})}{8} \\
\boldsymbol{\phi}^{(1/4)} &= \boldsymbol{\phi}^{(1/8)} \circ \boldsymbol{\phi}^{(1/8)} \\
\boldsymbol{\phi}^{(1/2)} &= \boldsymbol{\phi}^{(1/4)} \circ \boldsymbol{\phi}^{(1/4)} \\
\boldsymbol{\phi}^{(1)} &= \boldsymbol{\phi}^{(1/2)} \circ \boldsymbol{\phi}^{(1/2)}
\end{aligned} \tag{4.4}$$

So long as the small step is a diffeomorphic deformation, the final composed deformation field will preserve diffeomorphicity as composition of two diffeomorphisms is a diffeomorphism itself (in practice, small deviations might still be possible due to numerical approximations). The use of stationary velocity fields directly enables the explicit estimation of the inverse transformation. The velocity field is indeed a member of the Lie algebra, and the deformation field can be considered as the exponentiation of the velocity field: $\boldsymbol{\phi}^{(1)} = \mathrm{Exp}(\boldsymbol{v})$ [162]. Thanks to the properties of exponential maps, we have that $\boldsymbol{\phi}^{(-1)} = \mathrm{Exp}(\boldsymbol{v})^{-1} = \mathrm{Exp}(-\boldsymbol{v})$. Thus, the inverse transformation can directly be approximated with scaling-and-squaring of $-\boldsymbol{v}$ from the floating space identity deformation.

The stationary velocity field model provides us with a sound mathemat-

ical framework to easily incorporate rigid behaviour in specified areas while preserving the desirable geometrical properties of diffeomorphisms. The rigidity constraints expressed in Equation 4.2 can indeed be applied to the velocity field $\boldsymbol{v}$ after the initial scaling. The integration to the final deformation is then approximated through the composition of this small field by itself several times (squaring step). As the scaled deformation is set to rigid, due to the properties of composition the final deformation is maintained rigid within the desired areas. In other words, with reference to Equation 4.4, the final deformation $\boldsymbol{\phi}^{(1)}$ will be guaranteed rigid if $\boldsymbol{\phi}^{(1/8)}$ is set to be rigid.

4.2 Registration algorithm

The proposed transformation model has been embedded in the `NiftyReg` framework, an open-source package developed by Modat *et al.* containing utilities for medical image registration[1]. This package uses the symmetric block-matching algorithm [163, 164] for rigid and affine registration, while the nonlinear registration algorithm is based on the Free-Form Deformation (FFD) parametrisation using cubic B-splines [165, 166]. A symmetric and diffeomorphic nonlinear registration is also implemented with a stationary velocity field model parametrised with cubic B-splines basis [167].

4.2.1 Parametrisation with cubic B-splines

In `NiftyReg`, the parametrisation of the stationary velocity field takes advantage of the efficient formulation of the Free-Form Deformation algorithm [166]. The transformation $\boldsymbol{\phi}$ is not optimised at each and every voxel but only over a lattice of control points $\{\mu\}$ overlaid on the reference image. The displacement at each voxel is then derived by cubic B-spline interpolation:

$$\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x} + \sum_{l=0}^{3}\sum_{m=0}^{3}\sum_{n=0}^{3} B_l\left(\frac{x}{\delta_x} - \lfloor\frac{x}{\delta_x}\rfloor\right) B_m\left(\frac{y}{\delta_y} - \lfloor\frac{y}{\delta_y}\rfloor\right) B_n\left(\frac{z}{\delta_z} - \lfloor\frac{z}{\delta_z}\rfloor\right) \mu_{i+l,j+m,k+n} \tag{4.5}$$

[1] https://github.com/KCL-BMEIS/NiftyReg

where i, j, k are the indices of the first considered control point, and $\delta_x, \delta_y, \delta_z$ denote the control point spacing along each direction. B represents the approximated cubic B-spline basis functions:

$$\begin{aligned} B_0(u) &= (1-u)^3/6 \\ B_1(u) &= (3u^3 - 6u^2 + 4)/6 \\ B_2(u) &= (-3u^3 + 3u^2 + 3u + 4)/6 \\ B_3(u) &= u^3/6 \end{aligned} \tag{4.6}$$

Cubic B-splines are characterised by a local support of 4 control point width, meaning that each control point influences a region of size $4\delta_x \times 4\delta_y \times 4\delta_z$ centred in the control point itself. As a result, the spacing of the control points defines the amount of local detail of the deformation. A multi-resolution coarse-to-fine (also named pyramidal) approach can be used to favour faster computation, to optimise the capture range and to avoid local minima thanks to the smoothing derived from downsampling: starting with a large spacing, low frequency components of the transformation are optimised and by iteratively reducing the spacing higher frequency components are estimated while the transformation gets refined more locally. Such approach can easily be applied to the cubic B-splines parametrisation, as the deformation can be refined and preserved identical when halving the spacing of the control point grid [116].

When using the standard FFD algorithm, the control point positions directly parametrise the transformation. In the stationary velocity field model, instead, the control points are used to define where the velocity field is estimated. The final transformation is obtained through exponentiation, approximated with the scaling-and-squaring approach (Eq. 4.4). The exponentiation is performed only at the control points, not at the voxel level, using a cubic spline interpolation. The dense deformation field at each voxel is finally interpolated from the control point grid deformation.

As mentioned, this parametrisation also provides us with both the direct and inverse transformation, which are optimised simultaneously. By default, the

optimisation is performed through a conjugate gradient scheme in `NiftyReg`. At each iteration of the optimisation process, the gradient of the cost function $\mathcal{C}(R, F; \boldsymbol{\phi}) = (1-\lambda)\, \mathcal{S}(F(\boldsymbol{\phi}(X)), R(X)) - \lambda Reg(\boldsymbol{\phi})$ is computed at each control point and a line search along the direction of the gradient is performed to identify a local maximum of the cost function. The gradient at the current maximum is added to the estimate of the control point transformation parameters, and the inverse update is performed similarly for the backward transformation. The forward and backward velocity fields are averaged in the reference and the floating space to enforce symmetry on the transformation model. Finally, the scaling-and-squaring exponentiation is applied to obtain the current displacement field at the control points. This algorithm, reported in Algorithm 1 for the forward model, is iterated until convergence of each pyramidal level or until the maximum number of iterations is reached.

Algorithm 1 Control point parametrisation update (forward transformation)

Compute the gradient, G, of the current cost function value for each control point μ:

$$G(\mu) = \nabla \mathcal{C}(R, F; \mu), \qquad \forall \mu \in \{\mu\}$$

Perform a line search along the direction of G
for each step in line search **do**
  Update the current control point parameters:

$$\mu \leftarrow \mu + G(\mu), \qquad \forall \mu \in \{\mu\}$$

  if stationary velocity field parametrisation **then**
    Symmetrise with backward transformation

$$\mu \leftarrow 0.5(\mu - \mu^{backward}), \qquad \forall \mu \in \{\mu\}$$

  end if
end for
if stationary velocity field parametrisation **then**
  Scaling-and-squaring exponentiation (n steps)
end if

4.2.2 Introduction of rigid constraints

In order to introduce the rigidity constraints, I modified the optimisation scheme to update the transformation parameters as in Algorithm 2.

Algorithm 2 Control point parametrisation update with rigid constraints

Compute the gradient, G, of the current cost function value for each control point μ:

$$G(\mu) = \nabla \mathcal{C}(R, F; \mu), \qquad \forall \mu \in \{\mu\}$$

Perform a line search along the direction of G:
for each step in line search **do**
  Update the current control point parameters:

$$\mu \leftarrow \mu + G(\mu), \qquad \forall \mu \in \{\mu\}$$

  if stationary velocity field parametrisation **then**
    Symmetrise with backward transformation

$$\mu \leftarrow 0.5(\mu - \mu^{backward}), \qquad \forall \mu \in \{\mu\}$$

    Down-scale the control point parameters

$$\mu \leftarrow \mu / 2^n, \qquad \forall \mu \in \{\mu\}$$

  end if
  for each mask M_j **do**
    Define the subset $\{\mu\}_j$ of the control points within M_j
    Least Trimmed Square regression of the rigid transformation $\mathcal{R}_j$:

$$\mathcal{R}_j = \mathrm{LTS}(\{\mu\}_j)$$

    Update the parameters with the estimated rigid displacement:

$$\mu \leftarrow \mathcal{R}_j(\mu) \qquad \forall \mu \in \{\mu\}_j$$

  end for
  if stationary velocity field parametrisation **then**
    Up-scale control point parameters

$$\mu \leftarrow \mu * 2^n, \qquad \forall \mu \in \{\mu\}$$

  end if
end for
if stationary velocity field parametrisation **then**
  Scaling-and-squaring exponentiation (n steps)
end if

Specifically, at each iteration the gradient of the cost function is used to update the new parameters of each control point as described in the previous section. Within each mask associated with a rigid region, I extract the parameters of all the control points and use a Least Trimmed Square (LTS) regression to robustly estimate the rigid transformation that best fits the displacement associated with each control point, discarding 50% of points with highest residuals as outliers. The obtained rigid transformation is subsequently applied to update the current estimate of the parameters associated with each control point within the mask (i.e. the displacement in a FFD transformation model or the velocity field in a stationary velocity field one). Of note, before any optimisation is performed, the rigid masks are resampled to the control point grid space and dilated by 1. This is necessary to compensate for the local support of B-splines: by this dilation, the effect of the rigidity constraint is extended slightly beyond the initial rigid area, thus making sure that the impact of non-linear transformation from control points at the border of the rigid area is negligible within the original rigid mask itself. In case of a pyramidal coarse-to-fine approach, this operation is repeated at the beginning of the optimisation for each pyramidal level.

The proposed approach provides a formulation for the gradient computation for both rigid and non-rigid areas that can be easily incorporated into a gradient-based optimisation scheme, and can be used in both a velocity field parametrisation and standard Free-Form Deformation one.

4.3 Application to pelvic CT-MRI registration

The proposed registration algorithm was tested on the clinical application relevant for this project, *i.e.* the intra-subject registration of CT and MRI of patients with hip implants. The registration pipeline includes three steps: (1) generation of bone masks to localise areas where rigidity should be enforced; (2) initial affine registration; (3) nonlinear registration with rigidity constraints.

Bone masks generation: The described transformation model requires masks for the anatomical regions to be kept rigid - bones and implants in this application. In particular, these structures can be better segmented from CT images, thus the rigidity masks are extracted from this modality. As manual contouring is labour-consuming, I propose to take advantage of the template datasets to automatically derive segmentation masks of bones and implants, using a multi-atlas segmentation propagation pipeline as the one described in Figure 2.5. The target CT image is first split along the sagittal axis of symmetry, with the same algorithm described in Sec. 3.2. For each hip side, template CT images are registered to the target CT image via affine registration [164] followed by free-form nonlinear registration [166] with default parameters in `NiftyReg`; the template label images are then propagated with the respective estimated transformation and final consensus is obtained using the STEPS label fusion algorithm [49]. The two hip sides are finally combined back together into the full FOV. The pipeline for automated bone masks extraction was implemented in NiPype [168].

Affine registration step: As a first step, an initial global alignment is achieved by affinely registering the SRR MRI to the CT using the symmetric block-matching algorithm available in `NiftyReg` [164]. The SRR MRI is then resampled into the CT space using the estimated affine transformation and a cubic interpolation scheme.

Nonlinear registration step: A more refined local alignment is then achieved through the non-linear registration step, where rigidity is enforced on areas specified by binary masks. For this thesis' application, the Normalised Mutual Information (NMI) was used as a measure of similarity, being the most commonly adopted solution for multimodal registration. This is an entropy-based similarity measure introduced by Studholm *et al.* [169] which is maximised when the reference and the warped floating images have the most information in common, and thus when they are aligned. The bending energy (BE) was selected as a regularisation term [166] in order to promote smooth deformations.

The final cost function for this application can therefore be written as:

$$\mathcal{C}(R, F; \boldsymbol{\phi}) = (1-\lambda)\, NMI(F(\boldsymbol{\phi}(X)), R(X)) \, - \lambda BE(\boldsymbol{\phi}), \tag{4.7}$$

Previous work [170] also tested the efficacy of Locally Normalised Cross Correlation (LNCC) as a measure of similarity for intra-subject registration, but it was found performing worse than NMI for this application. Indeed, NMI is generally more robust to noise, which is heavily affecting the CT and MR images of patients with hip implants. To further reduce the impact of noise, during the registration intensity clipping was applied to both reference CT and floating MRI at the 1^{st} and 99^{th} percentiles of their intensity distributions.
In order to ensure smooth transitions in the deformation field, the proposed algorithm is iterated on a five-level course-to-fine pyramidal approach, using a final control point spacing of 5 mm. As the images are fairly large, this choice is suitable to sufficiently capture both low and high frequency changes. To account for the local support of the cubic B-spline parametrisation, the rigid masks are dilated by 1 control point at each level of the resolution pyramid.

4.4 Validation and experiments

The registration pipeline described in the previous section was tested on the template subjects, due to the availability of landmark annotation as well as manual segmentation to perform quantitative assessment. A series of experiments were designed to determine: (1) the influence of the rigidity constraints on the registration accuracy; (2) the impact of automated bone mask generation process on the registration; (3) potential failure in recovering femoral rotation.

4.4.1 Effect of the rigidity constraints

In order to assess the impact of the rigidity constraints on the registration accuracy, I compared the CT and SRR MRI registration for each template subject using the proposed rigidity constraints on bones and implants or fully

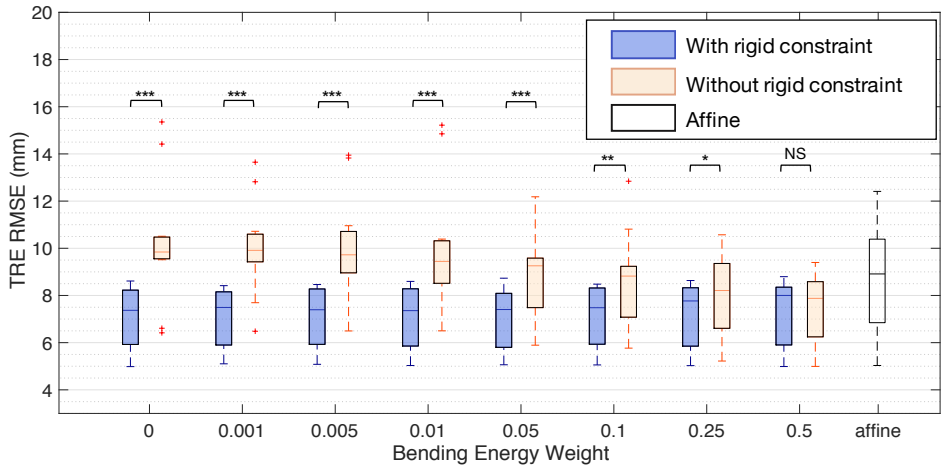
non-linear registration. The transformation was optimised for varying regularisation weights λ, so as to idenfity the most appropriate parameter selection for the available dataset. The manual segmentation masks of pelvis, left femur, right femur and implants were used to localise the rigid areas.

The quantification of the registration accuracy was obtained by computing the Target Registration Error (TRE) on 10 anatomical landmarks - 5 for each hip side - manually selected in both skeletal (3 landmarks) and muscular (2 landmarks) structures. More details about the landmarks annotation process was reported in Sec. 3.3. Given a transformation $\boldsymbol{\phi}$, the TRE for a specific landmark i was calculated in a symmetric form as:

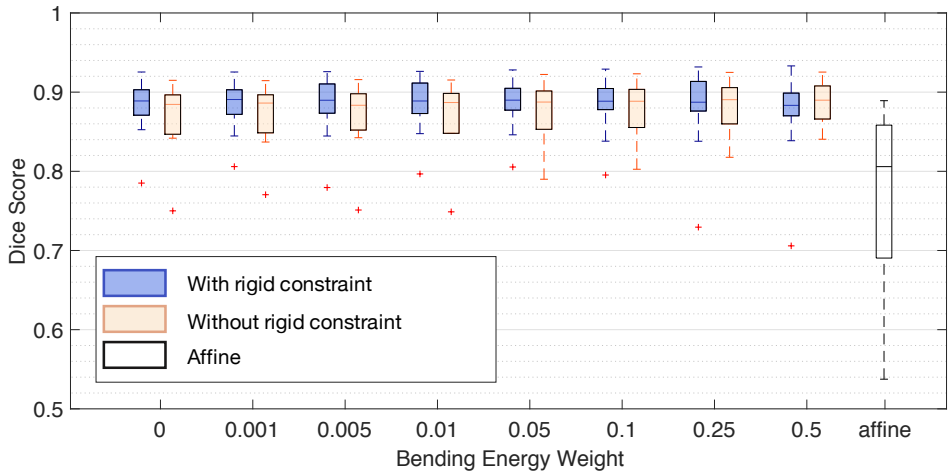
$$\mathrm{TRE}_i = \frac{1}{2}(\|\boldsymbol{x_i} - \boldsymbol{\phi}^{-1}(\boldsymbol{y_i})\|_{\ell^2} + \|\boldsymbol{\phi}(\boldsymbol{x_i}) - \boldsymbol{y_i}\|_{\ell^2}) \tag{4.8}$$

where $\boldsymbol{x_i}$ is the landmark position in the CT space and $\boldsymbol{y_i}$ is the corresponding landmark position in the MRI.
For each landmark and each subject I computed the average of the TRE values from the different manually selected landmark sets. For each subject and for each registration approach, I reported the root mean square error (RMSE) of the TRE values across the ten landmarks. Figure 4.2(a) shows the RMSE TRE distributions at varying bending energy weights for the registration with and without rigid constraints. Overall, the rigidly-constrained registration algorithm not only provides clinically plausible deformations, but it also outperforms the standard non-linear one in effectively reducing the TRE RMSE, producing therefore a more accurate alignment at the landmark locations. I also observed a reduced sensitivity to the choice of the regularisation parameter. Moreover, the introduced algorithmic steps for the rigidity constraint did not impact the total computation time (on the same machine, the rigidly constrained non-linear registration required 93 minutes on average, compared to 96 minutes for the fully non-linear approach).

As a further comparison, I additionally manually delineated the gluteus medius of healthy hip sides on the template CT images and computed the



(a)



(b)

Figure 4.2: (a) Comparison of TRE RMSE values obtained from the rigidly-constrained non-linear registration and the standard fully non-linear one with varying weights of the bending energy regularisation term. The TRE RMSE for the affine registration is reported as well. Wilcoxon signed rank test was performed between pairs with same registration parameters, and statistically significant differences are reported (NS indicates non significant, * indicates p-value $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$). (b) Dice Score for gluteus medius segmentation overlap between registered CT and MRI (only the healthy hip sides are considered). Wilcoxon signed rank test between pairs with same registration parameters did not highlight any significant difference (all $p > 0.05$).

Dice score between the registered CT segmentation and MRI based segmentation. In particular, to obtain a symmetric estimation I computed the average between the Dice score in the CT space (by warping the manual MRI segmentation with the inverse transformation) and the Dice score in the MRI

space (by warping the manual CT segmentation with the direct transformation). In Fig. 4.2(b), the distributions of the Dice score values are reported with and without the use of the rigid constraint using the same set of λ values. In agreement with the previous result, higher Dice scores are obtained with the proposed method, with similar improved performances for regularisation weights lower than 0.1, although non statistically significant.

Both experiments have been used to select an optimal bending energy weight for further analysis. I selected $\lambda = 0.001$, being the value that yields the best TRE RMSE, while still providing high degree of overlap for the segmented gluteus medius muscles (median Dice score [minimum, maximum] = 0.89 [0.81, 0.93]).

A visual comparison of the registration results for one case is presented in Fig 4.3. At equal regularisation ($\lambda = 0.001$), the Jacobian determinant maps show that the use of the rigid constraints enforces a volume-preserving deformation within bones and implant, ensuring anatomical plausibility of the applied transformation. For a fair comparison, I also reported the outcome of the standard non-linear registration with its optimal regularisation weight ($\lambda = 0.5$). While it promotes smoother transformations, it does not provide sufficient local deformation in the soft tissue to compensate for the different patient's position, especially on the implanted side. Also, the Jacobian determinant map still shows volume variations in the bony area, which are instead avoided by design in the proposed method.

4.4.2 Effect of automated bone masks

The proposed registration pipeline takes advantage of an automated segmentation method to define the masks for the rigidity constraints. The following experiment was designed to test the robustness of the registration to the use of automated masks. Indeed, an inaccurate bone mask generation would result in incorrect application of the rigidity constraint, and thus it would adversely influence the intra-subject registration and all the subsequent analysis.

As an initial check, I quantified the segmentation accuracy by computing the

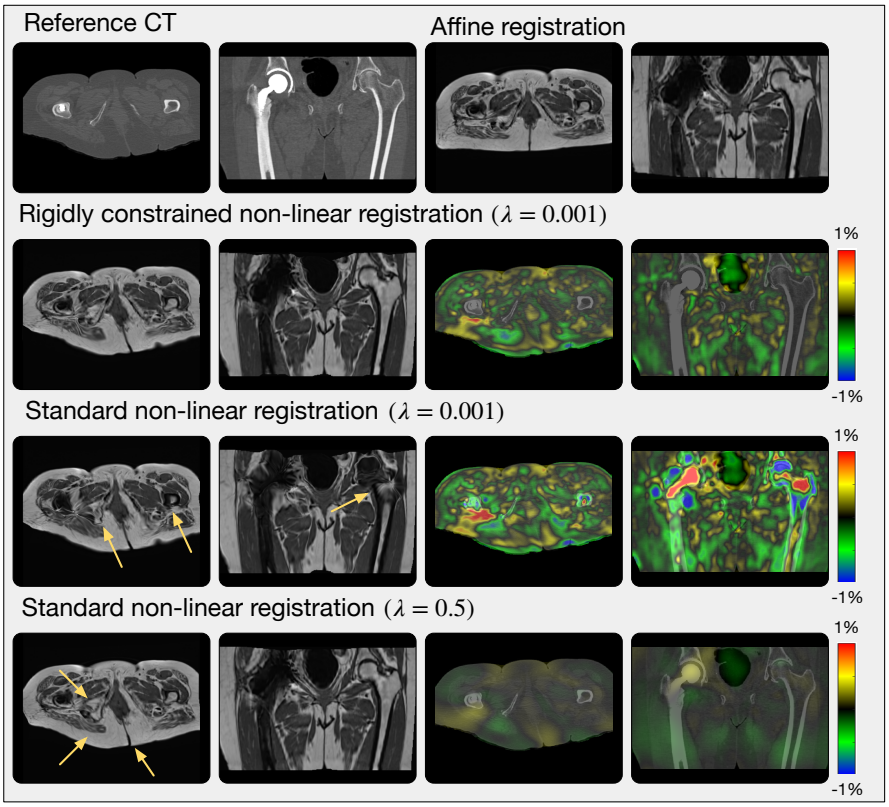


Figure 4.3: Example of registration results. The central axial and coronal views for the reference CT, the affinely registered MRI, the non-linearly registered MRI with and without the use of the rigidity constraints are reported. The coloured insets show the Jacobian determinant maps (in percentage of volume change) for each non-linear deformation. Results are displayed with the rigidly-constrained registration optimal regularisation weight ($\lambda = 0.001$) to showcase the sole effect of the rigidity constraints. The result of the optimal fully non-linear regularisation weight ($\lambda = 0.5$) is also shown. Yellow arrows point at areas where the standard non-linear registration fails to recover a good alignment. Differently from this latter, the proposed rigidly constrained non-linear registration allows for more localised deformation and better soft tissue alignment while preserving the volume and shape of bones and implant.

Dice score between the automated and the manual bone masks. Table 4.1 presents and median and the range of Dice score values for the pelvic bone, the healthy femora, the implanted femora and the implants. The agreement between the manual and the automated bone masks is satisfactory, although not perfect. This is especially true in the pelvic bone, which is the most variable and the most affected by FOV variations among the considered structures. To measure the impact of the segmentation inaccuracy on the intra-subject regis-

	Median [Minimum, Maximum] Dice Score
Pelvis	0.91 [0.59, 0.93]
Femur (Healthy)	0.96 [0.95, 0.97]
Femur (Implanted)	0.92 [0.84, 0.95]
Implant	0.93 [0.83, 0.95]

Table 4.1: Median [Minimum, Maximum] Dice Score values between the automated bone masks and their manual ground truth for the 11 subjects in the template dataset.

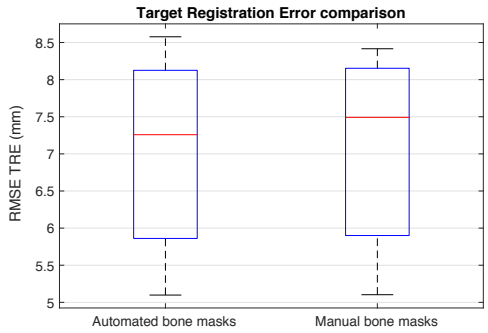


Figure 4.4: Root Mean Square Error of Target Registration Error computed over 10 landmarks for the 11 template subjects. Comparison between intra-subject CT-MRI registration with manual bone masks and automated ones (λ= 0.001).

tration, I compared the Target Registration Error (TRE) between registrations with either automated or manual bone masks. The boxplot graph reported in Figure 4.4 shows this comparison. A very slight increase in the TRE is observed with the automated bone masks, however no significant difference was identified (Wilcoxon signed rank test, 5% confidence level). Although I acknowledge that this analysis is limited only to the 11 subjects whose ground truth was available, this experiment shows that the intra-subject registration is robust against segmentation errors due to the use of automated bone masks.

4.4.3 Rotation recovery analysis

The last experiment aims at identifying failure modes of the incorporation of rigidity constraints. As shown also by the TRE analysis at varying λ (Fig. 4.2) and by the example in Fig. 4.3, the use of the rigidity constraints naturally acts as regulariser of the applied deformation. However, the deformation field

can be highly discontinuous at the joint between the pelvis and the femur, for instance in the case of internal/external rotation of the femur. Moreover, the proposed transformation parametrisation does not model sliding effects.

To try and quantify the ability of our model to correctly recover the transformation for the rigid structures, I compared the estimated rigid transformations with those obtained from simple rigid registrations focused on one bone at a time. The implant was discarded for this analysis, as it was not possible to estimate a rigid transformation focussing on the implant only due to the metal artefact in the MRI. For each of the 11 template subjects, the MRI-to-CT rigid transformations were computed separately for each bone with the symmetric block-matching algorithm implemented in `NiftyReg` by masking the considered structure in the reference space. Given the estimated transformation matrix

$$T = \begin{bmatrix} T_{11} & T_{12} & T_{13} & T_{14} \\ T_{21} & T_{22} & T_{23} & T_{24} \\ T_{31} & T_{32} & T_{33} & T_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.9}$$

and assuming the order of the rotation is around the $x-$axis first, then y, then z, the three Euler angles were derived as:

$$\begin{aligned} \theta_x &= \tan^{-1}(T_{32}, T_{33}) \\ \theta_y &= \tan^{-1}(-T_{31}, \sqrt{T_{11}^2 + T_{21}^2}) \\ \theta_z &= \tan^{-1}(T_{21}, T_{11}) \end{aligned} \tag{4.10}$$

where $\tan^{-1}$ indicates the four quadrant arctangent.

Similarly, the rigid transformation was extracted for each mask from the deformation field obtained with the rigidly-constrained non-linear registration algorithm. The Euler angles were computed as in Equation 4.10 and compared to the rigid-only ones.

The results for the pelvis, the non-implanted femora (healthy) and the implanted femora are presented in Fig. 4.5. For the latter structure, three subjects had to be excluded due to failure of optimising the rigid transformation focussed on the femur only. A general tendency to underestimate the rotation

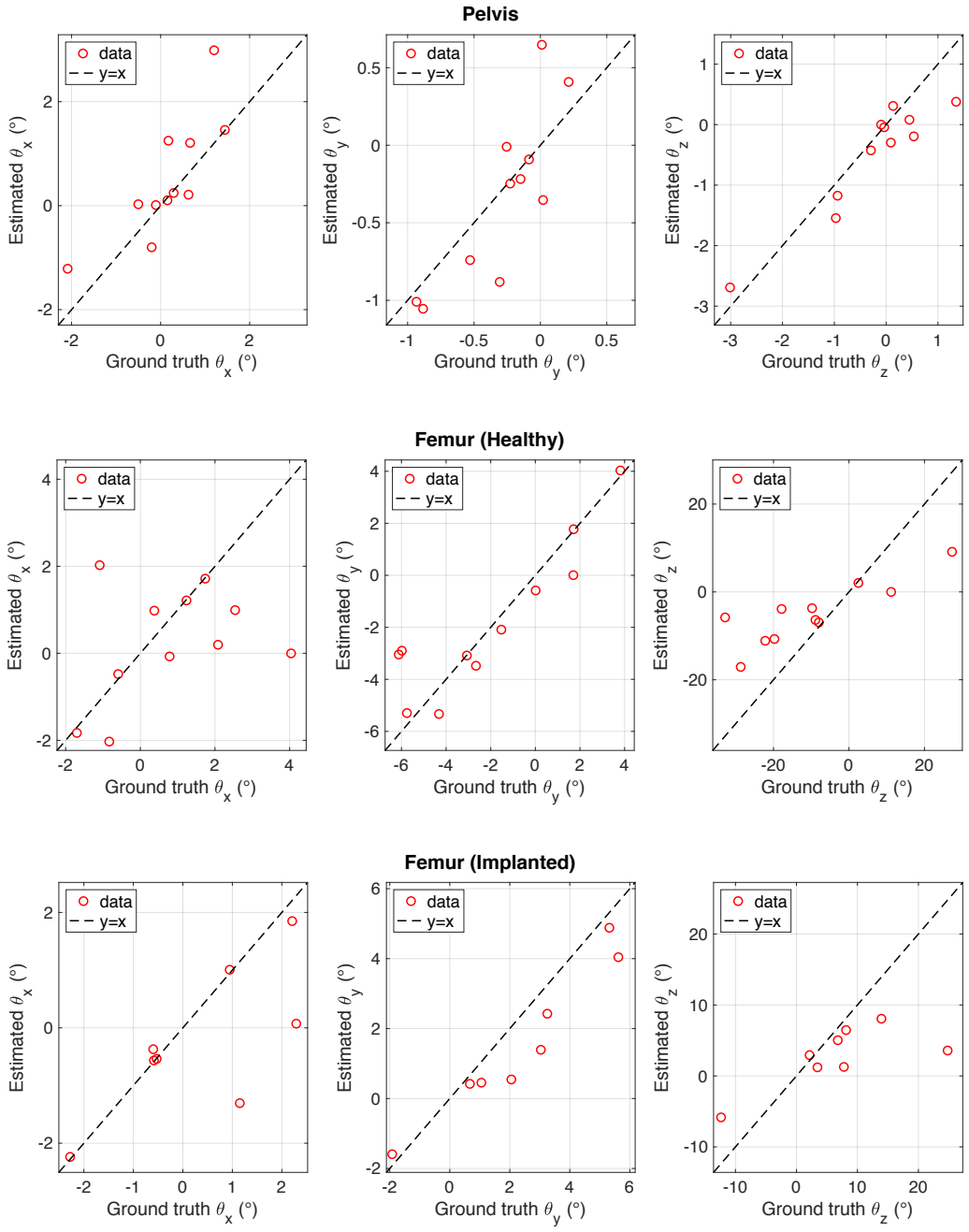


Figure 4.5: Rotation recovery analysis, comparing the ground truth angles and the angles obtained with rigidly-constrained non-linear registration algorithm. The bisector is displayed in black for reference.

is visible from these results, especially for θ_z in both healthy and implanted femora, which represents flexion/extension of the leg. This is expected as the non-linear registration algorithm optimises a transformation that aligns both the bones and the surrounding soft tissue, and it is regularised for smooth and diffeomorphic transitions over a control point grid with 5 mm spacing. It is therefore likely that the optimisation converges to local minima that might not exactly correspond the global minimum for the specific structure.

The absolute error for each angle was computed as $AE(\theta_{GT}, \theta_{rec}) = |\theta_{GT} - \theta_{rec}|$, with θ_{GT} being the Euler angle from the single-bone registration considered as the ground truth, and θ_{rec} the Euler angle recovered from the proposed rigidly-constrained non-linear registration algorithm. The median, minimum and maximum absolute errors are reported in Table 4.2, together with the respective ground truth angle ranges for reference.

		Pelvis	**Femur (H)**	**Femur (I)**	**Overall**
θ_x	GT	0.50 [0.11, 2.09]	1.24 [0.38, 4.04]	1.05 [0.53, 2.29]	0.89 [0.11, 4.04]
	AE	0.53 [0.01, 1.79]	0.86 [0.02, 4.04]	0.14 [0.01, 2.45]	0.47 [0.01, 4.04]
θ_y	GT	0.22 [0.01, 0.93]	3.07 [0.02, 6.12]	2.54 [0.67, 5.62]	1.61 [0.01, 6.12]
	AE	0.20 [0.01, 0.64]	0.60 [0.02, 3.08]	0.71 [0.25, 1.64]	0.45 [0.01, 3.08]
θ_z	GT	0.44 [0.04, 3.01]	17.85 [2.49, 32.81]	7.96 [2.15, 24.79]	7.28 [0.04, 32.81]
	AE	0.32 [0.01, 0.97]	11.05 [0.44, 27.00]	4.05 [0.79, 21.21]	1.71 [0.01, 27.00]
All θ	GT	0.29 [0.01, 3.01]	3.07 [0.02, 32.81]	2.28 [0.53, 24.79]	1.48 [0.01, 32.81]
	AE	0.24 [0.01, 1.79]	1.18 [0.02, 27.00]	1.17 [0.01, 21.21]	**0.58 [0.01, 27.00]**

Table 4.2: Rotation angle recovery: median [minimum, maximum] absolute error (AE) in degrees between the Euler angles from the single-bone rigid transformation and from the rigidly-constrainted non-linear deformation field. The median [minimum, maximum] ground truth angles (GT) are also reported for reference. H and I indicate the healthy and implanted hip sides respectively.

On the available template dataset, the absolute error appeared systematically higher for θ_z in both femora, which can be explained by the large range of associated ground truth angles. Overall, the median absolute error was found to be 0.58° although it raised up to 27°, especially along the flexion/extension axis, as also noticed from Fig. 4.5. The trend of the absolute error with respect

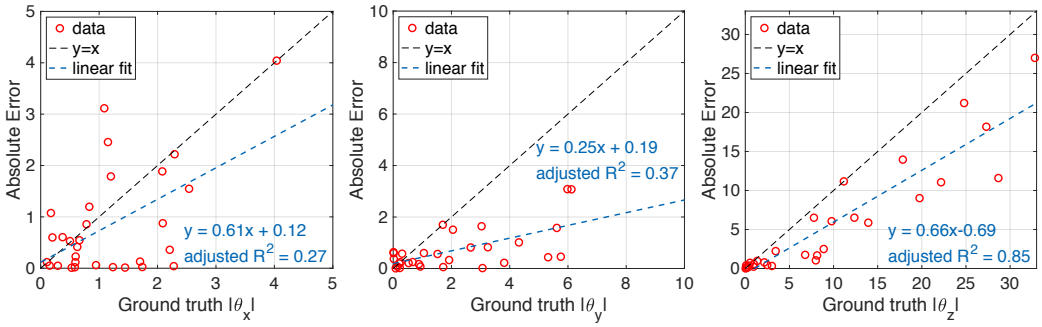
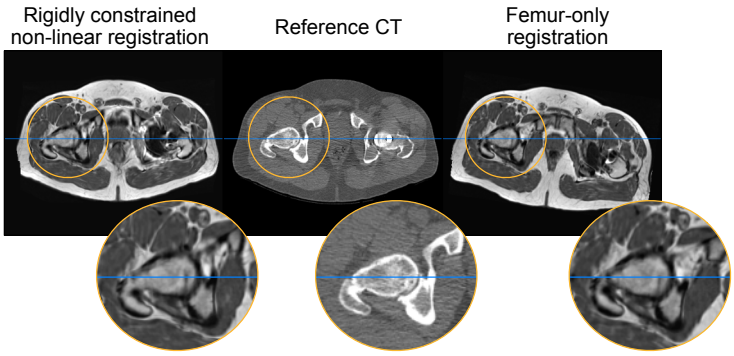


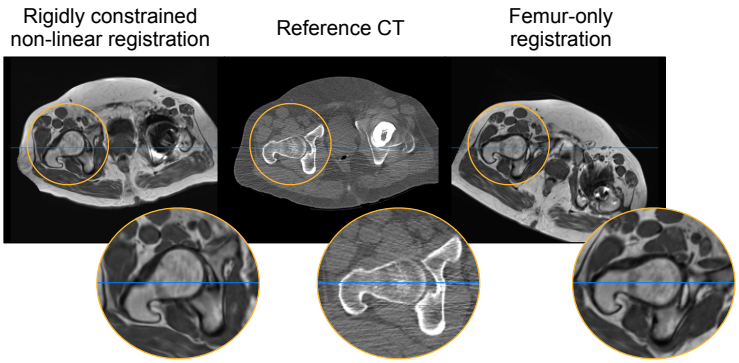
Figure 4.6: Absolute error versus ground truth Euler angles. The bisector is shown in black for reference, as well as the trend line (in light blue) obtained from ordinary least squares fit. Note that the display range varies across the different angles.

to the magnitude of the ground truth angle is shown in Fig. 4.6. Especially for θ_z, a linear trend was observed between the angle and its associated absolute error. This trend is particularly problematic for large rotation angles, where an error bigger than 50% can have a substantial impact on the final alignment. However, in our dataset the median rotation angles were 0.29°, 3.07° and 2.28° for the pelvis, the healthy femora and the implanted femora respectively, with 90% of the angles below 12°. For the sake of illustrating the impact of rotation errors, Figure 4.7 shows an example of angle recovery for the healthy femur with an error close to the femora median absolute error, while Figure 4.8 shows the case reporting the worst error. In the context of this application for hip imaging, the error on the angle recovery has none or little impact on the registration for small rotations, with satisfactory alignment on the femur as well as on the surrounding soft tissue. For large rotations of the femur, its position is not well recovered, although a reasonably good alignment is achieved throughout the FOV. Two factors might have influenced this result: first, the optimisation converged to a local minimum and no further improvement was possible; second, such large rotation might have induced folding in the surrounding tissue, which is however impeded by design in the stationary velocity field model. The algorithm thus found a trade-off between the allowed rotation and the preservation of diffeomorphism.



	θ_x	θ_y	θ_z
GT	0.78	-3.07	-8.88
Rec	-0.07	-3.08	-6.38
AE	0.85	0.01	2.50

Figure 4.7: Example of rotation recovery performance on healthy femur: case with errors similar to the femora median absolute error. Figure (left): the reference CT image, rigidly-constrained non-linear registered MR image and the femur-focussed rigidly registered MR. Table (right): ground truth (GT) angles from the femur-only registration, recovered (Rec) angles from the rigidly-constrained non-linear registration and the respective absolute error (AE). Despite a 2.50° error on one of the angles, the femur alignment is still satisfactory.



	θ_x	θ_y	θ_z
GT	-1.09	-6.12	-32.81
Rec	2.03	-3.05	-5.81
AE	3.12	3.07	27.00

Figure 4.8: Example of rotation recovery performance on healthy femur: case with-worst error. Figure (left): the reference CT image, rigidly-constrained non-linear registered MR image and the femur-focussed rigidly registered MR. Table (right): ground truth (GT) angles from the femur-only registration, recovered (Rec) angles from the rigidly-constrained non-linear registration and the respective absolute error (AE). The inset shows a poor alignment on the femur of the non-linear registraiton, probably due to convergence to a local minimum, as other anatomical structures appear well-aligned.

Because of skewed distributions, the sign-test was used to compare the ground truth and the recovered angles. No significant differences were found in any of the three angles and in any of the anatomical structures (significance threshold at 0.05), although no robust conclusion can be drawn from this test due to the very small sample size.

4.5 Discussion

The combination of multimodal information requires accurate alignment of the different images, achievable through intra-subject image registration. In order to compensate for differences in patient's position within the scanners, I developed a novel CT-MRI intra-subject registration algorithm that allows for local deformation of the soft tissue while preserving the rigidity of bone structures. The introduction of rigidity constraints is crucial to obtain clinical trust, as it guarantees fidelity in the applied anatomical deformations. Indeed, differently from other approaches such as Staring *et al.* [127] and Reaungamornrat *et al.* [128], the hard-constraint formulation limits the optimisation only to strictly rigid transformations within rigid structures, preserving their shape and volume. As the hard constraints have been embedded in a diffeomorphic transformation model, the proposed method also guarantees the preservation of topology and avoids unrealistic tissue folding.

For the pelvic images registration task, this registration algorithm proved to be more robust to the choice of regularisation parameters and more accurate than standard same-parametrisation non-linear registration (Sec. 4.4.1). The addition of further algorithmic steps to optimise the rigid transformations did not impact on the overall computation time, as the rigidity masks actively regularise the transformation and favour convergence. Moreover, the algorithm showed robustness against segmentation inaccuracies for the rigid areas that might arise from the use of automated mask extraction (Sec. 4.4.2).

This framework was developed in the context of intra-subject pelvic CT and MRI registration, but its implementation is anatomy-agnostic and generic: it can be applied to any anatomy requiring rigidity in specific areas, so long as masks of the rigid areas are available. In addition, its incorporation in a cubic B-spline parametrisation makes it flexible and suitable for both Free-Form Deformation model as well as its diffeomorphic extension with stationary velocity field.

Despite the demonstrated improvements on intra-subject registration ac-

curacy and transformation plausibility, the proposed method still presents some limitations. Firstly, the non-linear step is initialised with a global registration. If this registration fails or is not sufficiently accurate, the non-linear step would not be able to compensate for global misalignment due to limited capture range and local optima. Secondly, the current parametrisation does not model sliding effects, which can cause large discontinuities in the deformation field within the acetabular space when large femoral rotation occurs between the patient's position in the MR and in the CT scanners. When compared to rigid registration focussed on individual bones only (Sec. 4.4.3), the proposed technique introduces critical errors for hip application on large rotations ($> 10°$), although they seem to represent only a tiny percentage of the observed rotation angles (templates dataset median rotation angle is about $1.5°$).

While not directly observed in this application, the order of optimisation of the rigid transformations might have an impact on the final result, especially if the rigid structures are close to one another (e.g. spine application). Further tests on a different dataset would thus be needed to assess the importance of the rigid masks order.

I finally acknowledge that the validation set is limited in sample size, as ground truth segmentation and landmark annotation were available for eleven subjects only. Although the template subjects greatly vary in anatomy, implant type and extent of metal artefacts, the generalisability of this registration algorithm should be further tested on a larger dataset.

Chapter 5

Automated multimodal segmentation

Motivation

Image segmentation is needed for quantitative and qualitative analysis, but it is a cumbersome and time-consuming task. In MSK, the development of automated segmentation tools is hindered by the large variability of musculoskeletal structures and the presence of metal artefact induced noise.

Contribution

A fully automated pipeline for the joint segmentation of CT and MR images of patients with hip implants is presented. It provides skeletal, muscular and implant segmentation, and it is robust to the low quality of clinical data.

Image segmentation allows for the delineation of regions of interests (ROIs) for quantitative or qualitative analyses relevant for the clinical interpretation of the images themselves. Recent advances in image processing have enabled the automation of medical image segmentation, mitigating the burden of manual annotation. However, automating the segmentation of musculoskeletal structures is challenging because of the large anatomical variability in the

population, the lack of standardised scanning protocols, and the small intra-tissue variability which hampers the separation of different bones or different muscles. The presence of metal implants further complicates the segmentation, due to the introduction of metal artefacts and noise.

Most methods proposed in the scientific literature focus on the segmentation of either skeletal or muscular structures. In both cases, atlas-based segmentation approaches and statistical shape models are the most common choice, as they take advantage of prior anatomical knowledge to infer the segmentation of new images [33, 62, 68, 72, 82, 83, 85, 103].

Very little work has been proposed to combine the segmentation of all relevant musculoskeletal structures of the hip joint. One of the most remarkable examples is the work by Yokota *et al.* [94], where a hierarchical multi-atlas approach is employed to obtain automatic segmentation of pelvis, femur and 19 muscles in CT images (see also Fig. 2.6). Their atlas dataset consisted of 20 subjects (augmented to 40 templates by mirroring with respect to the sagittal plane), and the approach was tested with a Leave-One-Out Cross Validation study, reporting an average Dice score of 80.21±6.13 % on all muscles. More recently, Liang *et al.* [112] proposed a deep learning framework called Attention-Pyramid Network (APNet), which automatically segments 50 anatomical structures on pelvic MR images. The network was trained on 240 slices from 10 subjects and tested on 20 slices of a single patient, resulting in an average Intersection-over-Union (IoU) of 80.27%. Finally, quantification of segmentation accuracy of musculoskeletal structures in the pelvis was also reported by Hiasa *et al.* [171], whose focus was however the MR-to-CT synthesis using a CycleGAN approach. They trained a two-channel 2D U-Net for segmentation of four musculoskeletal structures using 20 labeled CT images and respective synthetic MRI. At testing, they automatically segmented 10 real MR images coupled with their synthetic CT. Their approach shows promise for effectively incorporating multimodal information for both synthesis and segmentation.

The discussed methods were all developed and tested on healthy hip sides, with little or no metal artefact corruption. Only very recently Sakamoto *et al.* [113] presented an automated segmentation for patients with hip implants. Their method combines two U-Nets to first refine the metal artefact correction and then obtain the segmentation of 19 muscles and hip bones. However, their analysis is limited by the use of images with simulated metal artefact and by testing on only 3 real post-operative images.

In order to address the need for robust image segmentation of musculoskeletal anatomy for patients with hip replacement, I developed a fully automated pipeline to jointly process CT and MRI of the same patient and combine their multimodal information. The proposed approach generates the segmentation of the pelvic bones, the implants and the abductor muscles which are at greatest risk of developing atrophy or hypertrophy after hip arthroplasty. To achieve robustness and high accuracy, this pipeline combines: (1) tailored pre-processing to reduce the impact of metal artefact as well as to improve the quality of MRI data; (2) diffeomorphic nonlinear registration algorithm respective of local rigidity of the bones; (3) a multi-channel multi-atlas segmentation propagation approach, which accomodates for the large population variability of the considered anatomical structures.

This chapter first presents the automated segmentation pipeline, with particular focus on the multi-channel multi-altas segmentation propagation step. The validation of the method will then be discussed and the results will be compared with current state-of-the-art approaches. The work here described was originally presented at the 5^{th} Workshop of Computational Methods and Clinical Applications in Musculoskeletal Imaging (MSKI) held in conjuction with MICCAI 2017 [159], and then published in the Computer Methods and Programs in Biomedicine journal [160].

5.1 Fully automated pipeline for multimodal segmentation

The proposed segmentation framework was developed for the processing of CT and MRI data of patients with hip implants. It considers three inputs: one CT image, and two anisotropic MR acquisitions at different high-resolution planes (as they are often acquired in clinical practice for faster scanning time). The inputs are processed in three sequential blocks of steps: image quality enhancement, intra-subject MRI-CT image registration with rigidity constraints on bone, and multi-atlas based multimodal segmentation.

A schematic representation of this pipeline is shown in Fig. 5.1. Taking advantage of the NiPype framework for the implementation [168], this pipeline makes use and extends image processing utilities from `NiftyReg`[1], `NiftySeg`[2], `FSL`[3] and `SimpleITK`[4].

Image quality enhancement . The purpose of the first processing block is to enhance the quality of the input images, to deal with routinely acquired clinical data. The steps presented here were specifically implemented for the available data types (*i.e.* one CT image and two anisotropic MR T1-weighted acquisitions). However, the modular formulation of the pipeline allows for tailoring these steps to different data types. For instance, it can be adapted to manage only partial data - e.g only one MR image available - or extended to include different acquisitions - e.g. other sequences MR images.

As previously discussed in Sec. 3.1, the CT image is initially corrected for metal artefact [134] and resampled to $1 \times 1 \times 1$ mm^3 resolution. For the MR images, the axial and the coronal acquisitions are corrected for bias field effects, their intensity distributions are matched by means of histogram normalisation and finally they are fused into a single volume at $1 \times 1 \times 1$ mm^3 resolution with the Super Resolution Reconstruction algorithm proposed by Ebner *et al.* [158].

[1] https://github.com/KCL-BMEIS/NiftyReg
[2] https://github.com/KCL-BMEIS/NiftySeg
[3] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki
[4] http://www.simpleitk.org/

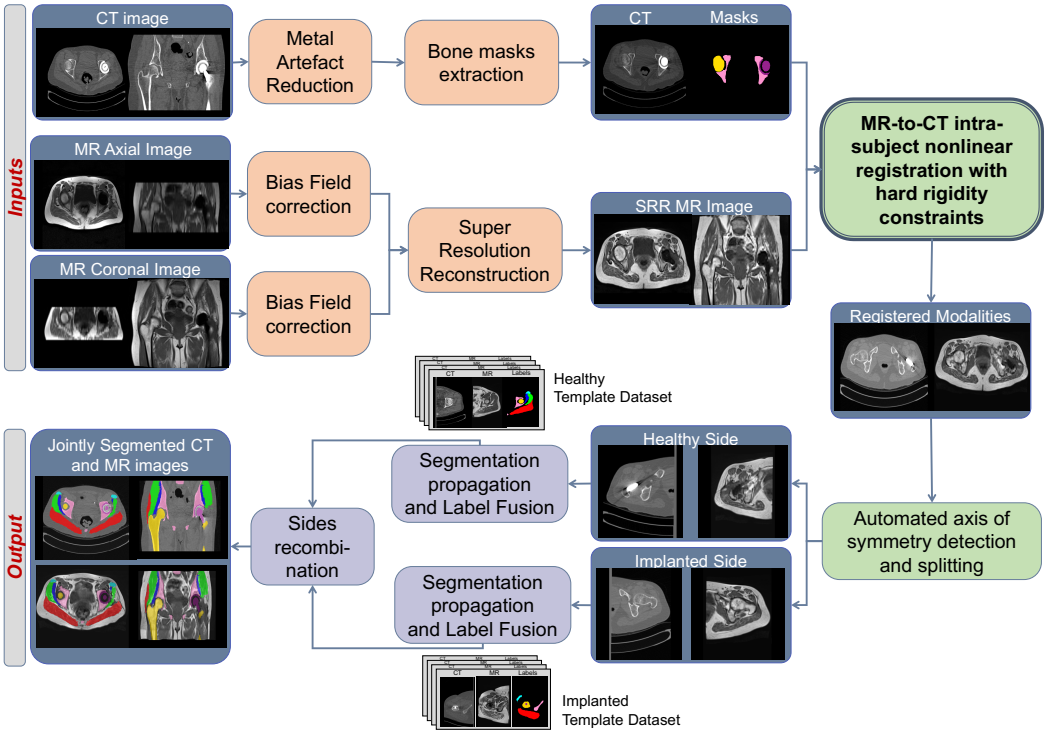


Figure 5.1: Proposed framework for joint automated segmentation of CT and MR pelvic images. The three blocks composing the pipeline are highlighted: image quality enhancement (orange), where the two modalities are firstly processed independently; our novel intra-subject multimodal registration (green), where the proposed rigidly constrained non-linear registration provides alignment of the CT and MRI while guaranteeing a rigid behaviour in bones; multi-atlas based automated segmentation (purple), where the joint CT and MRI segmentation of the two hip sides are separately obtained and then recombined in the full FOV.

Intra-subject registration with rigidity constraints. The second block of the pipeline aims at aligning the CT and the SRR MRI of the same subject. This requires a multimodal non-linear registration to compensate for different patient's pose within the two scanners while respecting the rigidity of bony structures. The steps in this block follow the same presented in Sec. 4.3. First, pelvis, femora and implant masks are automatically derived from the CT image using the multi-atlas segmentation propagation strategy described in Sec. 4.3. Then, an initial global alignment is achieved by affinely registering the SRR MRI to the CT using a symmetric block-matching algorithm [164]. A more refined local alignment is then achieved through the non-linear registration step with rigidity constraints on the extracted masks. After the registration,

the images are split along the axis of symmetry as automatically detected from the CT image (Sec. 3.2) and reoriented to have implanted hips as left hip side, and healthy hips as right side - in line with the reorientation of the template subsets.

Multi-atlas segmentation propagation . In the last block, the separated multimodal hip sides are automatically segmented using multi-channel multi-atlas segmentation propagation and label fusion, taking advantage of the template dataset described in Sec. 3.2. The sides and their obtained labels are finally reoriented and recombined to their original space. The obtained segmentation labels the pelvis, the femora, the implants, GMAX, GMED, GMIN and TFL, and it can be overlaid on both CT and MRI. A more detailed explanation of the segmentation pipeline is provided in the next section.

5.2 Multi-atlas segmentation approach

The final block of the proposed pipeline uses the template datasets in a multi-atlas segmentation propagation and label fusion framework to estimate the final segmentation of each hip side.

Segmentation propagation. Given the selected dataset (either healthy or implanted) for the considered hip side, the templates CT-MR images are registered to the target CT-MRI (stacked into a 4D volume) through the following steps:

1. *Rigid registration*: All the templates are rigidly aligned to the target independently, using symmetric block-matching [164].

2. *LTS average of rigid transformations*: The optimised rigid matrices are averaged in the log-Euclidean space, discarding 50% of them as outliers through least trimmed square regression. Since all the templates are co-registered to their mid-space, this guarantees robustness against potentially failed template registrations.

3. *Affine registration*: All the templates are affinely aligned to the target.

Instead of initialising the registration with the previously computed rigid matrix for each template, the LTS rigid average is employed for all of them.

4. *LTS average of affine transformations*: Similarly to the rigid case, a LTS average affine transformation is computed.

5. *Non-linear registration*: Using the LTS average affine matrix as initialisation, each template is non-linearly registered to the target using a multi-channel approach that aligns simultaneously both imaging modalities. Both CT and MRI contribute equally to the cost function. The optimisation will thus converge to a final transformation that balances the alignment of the CTs and the alignment of the MR images. In this approach, we make use of a standard Free-Form Deformation registration algorithm, parametrised through cubic B-splines [166]. A bending energy penalty term is also added to the cost function to regularise the optimised deformation field. Using T_i to refer to the i-th template image and R for the reference (target) image to segment, the total cost function is:

$$\mathcal{C}(R,T_i) = (1-\lambda)\left(\frac{1}{2}\mathcal{S}(R^{CT},T_i^{CT}) + \frac{1}{2}\mathcal{S}(R^{MR},T_i^{MR})\right) + \lambda \mathrm{BE} \quad (5.1)$$

where $\mathcal{S}$ indicates a similarity measure, BE the bending energy term, weighted by the hyper-parameter λ.

6. *Segmentation propagation*: The final estimated non-linear transformations are deployed to resample the label image of the respective templates onto the target space, using a nearest-neighbour interpolation scheme.

Label fusion. As for the bone-mask creation, a final consensus is obtained from the candidate segmentations by means of the STEPS label fusion algorithm [49]. This technique considers the propagated segmentations as potential classifiers, and locally ranks them based on the similarity between the target

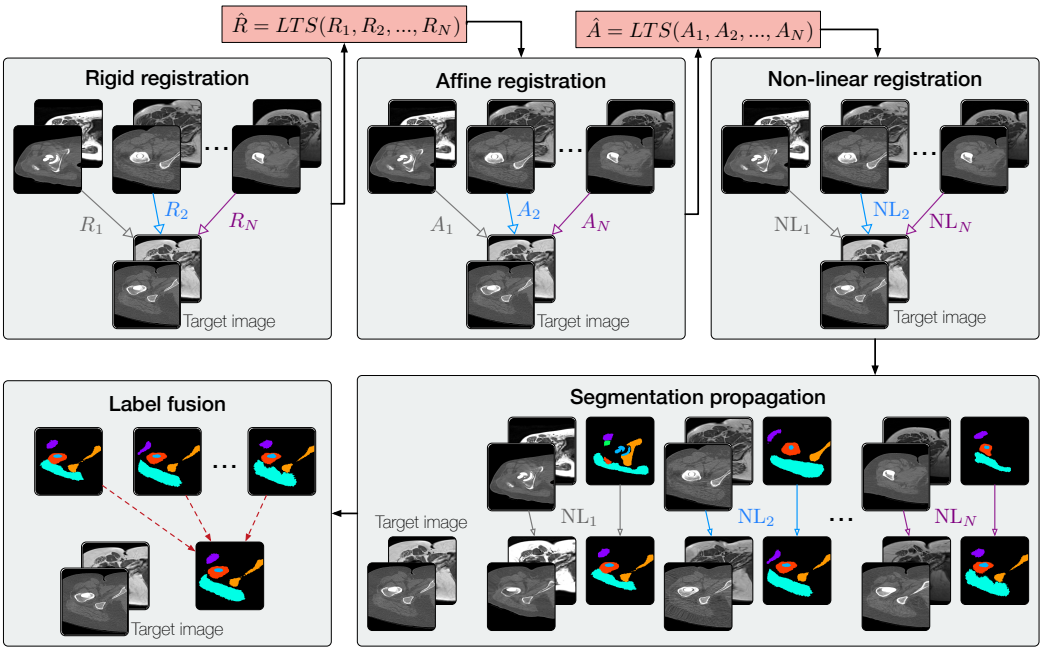


Figure 5.2: Multi-channel multi-atlas segmentation propagation and label fusion pipeline for automated segmentation of registered CT and MRI. Each template subject is aligned to the target image with rigid registration (R) followed by affine registration (A) and non-linear one (NL). Each step is initialised by the least-trimmed square (LTS) average of all template transformations from the previous step. The segmentations are then resampled using the estimated transformations and fused into a final consensus.

image and the respective registered template. The similarity is measured on a local patch with locally normalised cross correlation (LNCC) and spatial consistency is favoured by the use of a Markov Random Field (MRF) regularisation. For this application, STEPS was modified to use a multi-channel version of LNCC to rank the templates, defined as the sum of LNCC values from each channel. Two main hyper-parameters require tuning: the weight β of the MRF regularisation, and the number N of the top ranked classifiers to retain in the final majority voting to estimate the consensus label in each voxel. A diagram of the segmentation pipeline is displayed in Fig. 5.2.

This procedure is performed for both hip sides separately. The two sides are finally reoriented back and recombined to their original full FOV, providing a final multi-label image that can be overlaid on both CT and MRI, highlighting all the segmented musculoskeletal structures.

5.3 Pipeline validation

This section describes a series of experiments that have been performed to validate the proposed multimodal automated segmentation pipeline. It focusses mostly on quantifying the performance of the multi-channel multi-atlas segmentation propagation scheme introduced in the previous section. The reader is referred to Sec. 4.4 for a detailed performance analysis of the intra-subject registration block.

All the experiments described below made use of the 11 template subjects (10 unilateral cases and 1 bilateral), being the only cases with available manual ground truth segmentation of bones and muscles. The main validation experiment, described in Sec. 5.3.1, is a Leave-One-Out Cross Validation (LOOCV) study that aimed at: (1) identifying the optimal hyper-parameters for the automated segmentation, and (2) evaluating the pipeline performance with the optimal set of parameters. The generalisability of the proposed approach was tested further with the nested LOOCV study presented in Sec. 5.3.2, assessing the performance of the pipeline on an hold-out set. Finally, continuing the analysis presented in Sec. 4.4.2, the robustness of the full pipeline was tested against the use of automated bone mask generation, and the results are discussed in Sec. 5.3.3.

5.3.1 Leave-One-Out Cross Validation

To assess the performance of the automated segmentation pipeline, I designed a LOOCV experiment on the template datasets. Given the healthy (N=10 hip sides) or implanted (N=12 hip sides) dataset, I performed the multi-atlas based automated segmentation step for each template image using the remaining $N-1$ templates and varying the segmentation propagation and the label fusion parameters. The obtained segmentations were then compared to the ground truth by computing the Dice Score for each label.

This test was performed with three different settings: (1) using only the CT images; (2) using only the MR images; and (3) using the registered 4D CT-MR images. No muscle segmentations were available for the CT, and similarly no

		Segmentation propagation		Label fusion	
		Similarity	λ	N	β
CT	Healthy	LNCC	0.001	3	2
	Implanted	NMI	0.01	3	2
MR	Healthy	NMI	0.001	3	3
	Implanted	LNCC	0.1	5	2
Multimodal	Healthy	LNCC	0.01	3	2
	Implanted	LNCC	0.1	3	3

Table 5.1: Optimal hyper-parameters of the multi-atlas automated segmentation selected from each LOOCV experiment setting. λ is the bending energy weight in the template registration cost function, while N is the number of top ranked classifiers to retain and β is the MRF weight for the STEPS algorithm.

bones and implant labelling were available on the MRI, thus only the available labels were considered in the single-modality experiments. For each setting, an hyper-parameter search was performed to tune the multi-atlas automated segmentation parameters. In particular, for the segmentation propagation I tested the use of NMI or LNCC as similarity measure, as well as tuned the bending energy weight ($\lambda = [0.001, 0.01, 0.1]$). For the label fusion, also based on previous analysis [170], I fixed the standard deviation of the Gaussian kernel for the LNCC computation to $\sigma = 5$ voxels, while I performed tests for the number of top ranked classifiers to retain in the majority voting ($N = [3, 5]$) and for the MRF weight ($\beta = [2, 3]$). The optimal set of multi-atlas-based automated segmentation parameters was selected as the one maximising the lowest Dice Score across subjects and across labels; when different sets of parameters had lowest values within 0.05 difference, I chose the one with the highest median Dice Score. The result of the selection is reported in Table 5.1. The quantitative and qualitative analyses that follow were obtained with this selection of parameters.

Visual examples of the performance of the proposed multimodal multi-atlas automated segmentation pipeline are shown in Fig. 5.3. Table 5.2 displays instead the median Dice score values for bones, muscles and implants

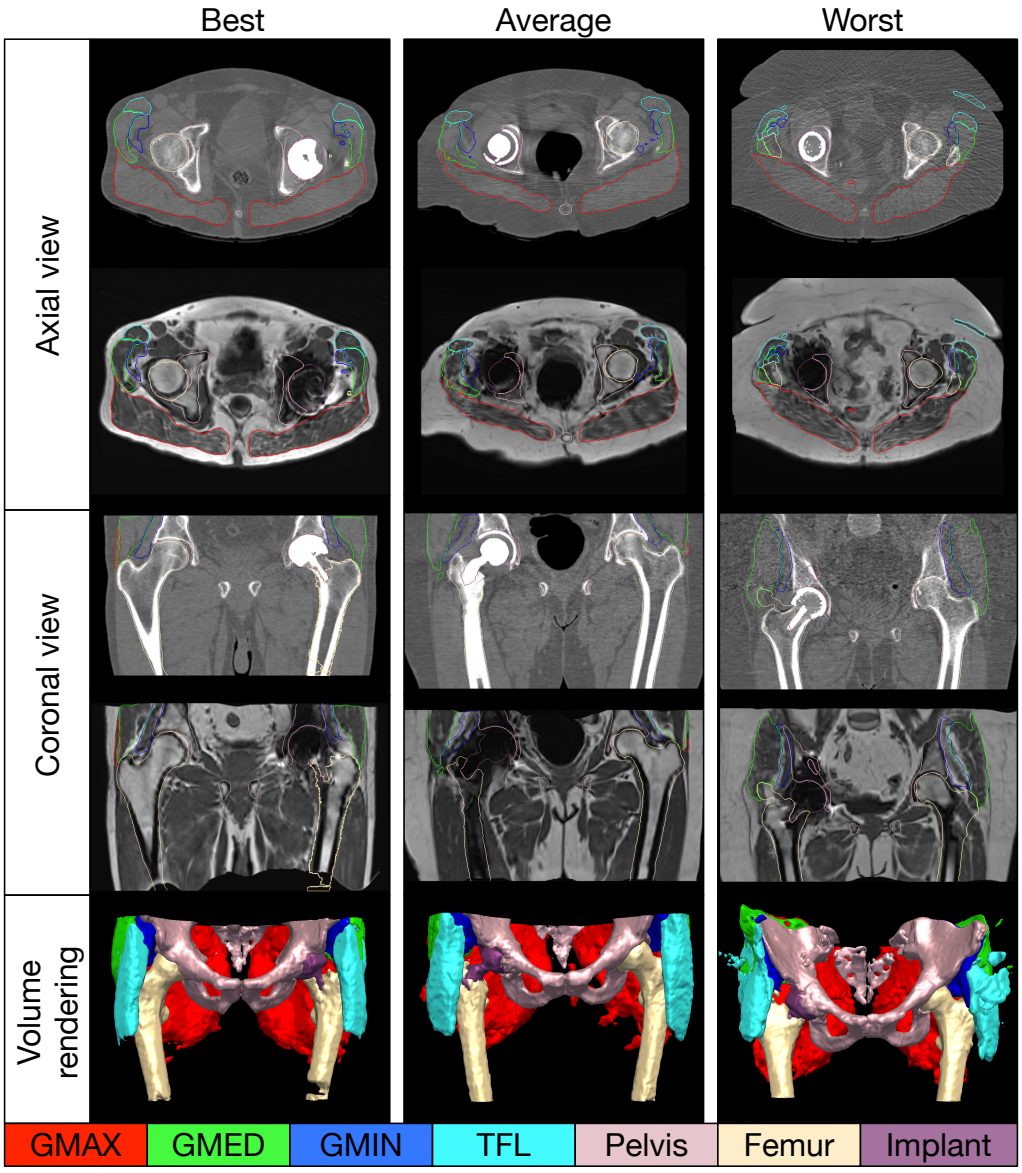


Figure 5.3: Examples of automated segmentation results obtained with the proposed multimodal pipeline. Left, middle and right columns correspond to the subjects with best, average and worst mean Dice score across the segmented structures. They also demonstrate the large variability of musculoskeletal shapes, implant types, patient's position and degree of artefacts present in this type of images. For the first case the pipeline achieved plausible segmentation throughout the field of view, while in the second case the muscular structures are less well-defined. In the third case, instead, the strong metal artefact in the MR image affected the templates-to-target registration and the consequent segmentation accuracy.

obtained with the best set of parameters for each experiment setting. A more detailed per-label comparison is reported in Fig. 5.4. Regarding the considered structures, the bones and the implant are overall better segmented with

Leave-One-Out-Cross-Validation for model parameter selection

		Healthy side		
	CT	*MR*	*Multimodal*	*p-value*
Bones	0.95 [0.72, 0.97]	N.A.	0.93 [0.75, 0.96]	0.0064*
Muscles	N.A.	0.84 [0.59, 0.93]	0.85 [0.66, 0.92]	0.0806
		Implanted side		
	CT	*MR*	*Multimodal*	*p-value*
Bones	0.91 [0.70, 0.94]	N.A.	0.89 [0.69, 0.91]	0.0002*
Muscles	N.A.	0.83 [0.60, 0.93]	0.81 [0.61, 0.92]	0.1051
Implant	0.92 [0.81, 0.95]	N.A.	0.91 [0.84, 0.95]	0.8501

Table 5.2: Median Dice Score values and 95% percentile ranges for bones, implant and muscles: comparison between single- and multi-modality results. Wilcoxon signed rank test was performed to test the null hypothesis of same distribution for the multi-modality- and the respective single-modality-derived Dice Scores (obtained p-values are reported and starred are the cases of rejection of the null hypothesis with 5% significance level). N.A. indicates cases where the manual segmentation was not available.

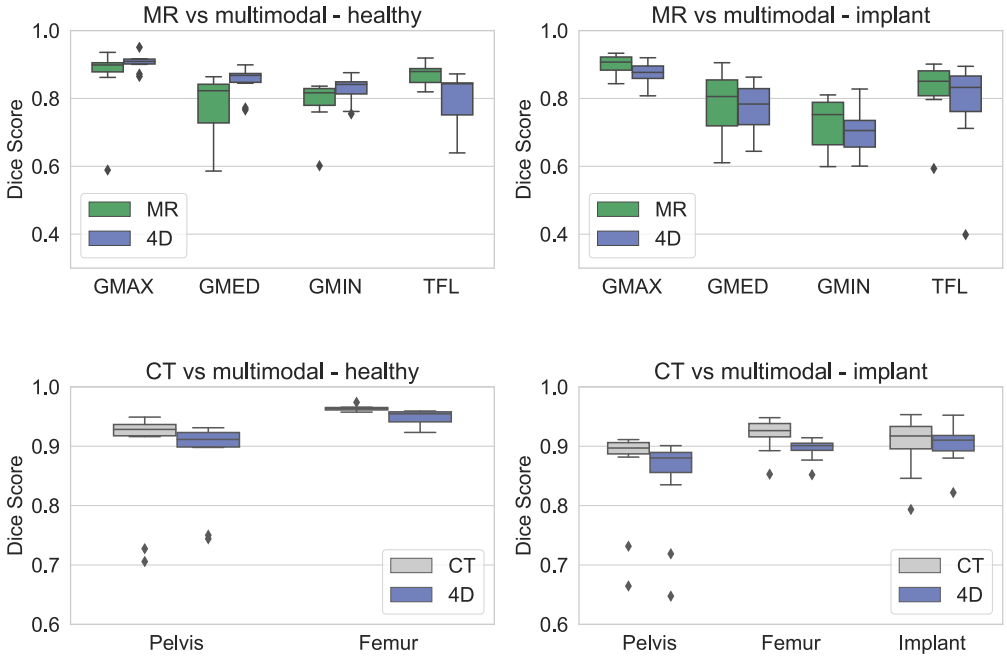


Figure 5.4: Boxplots illustrating the Dice score distribution obtained for the best set of parameters in the LOOCV experiments. The multimodal approach was compared with the CT for bones and implants, and to MRI for muscles. Results are separated in healthy and implanted hip sides.

respect to muscles, given their lower shape and texture variability. To compare the single- vs the multi-modality performance, I tested the hypothesis of same underlying distributions for Dice Score values with the Wilcoxon signed rank test (5% significance level). Significant difference was found only for the segmentation of the bony structures, although the sample size is too small to draw any definite conclusion. The implanted side provided slightly lower values with the multimodal pipeline. This side is more affected by residual metal artefact in CT and lack of metal intensity information in the MRI, which hamper both the intra- and inter-subject registration. Moreover, in the multimodal experiment the multi-atlas segmentation parameters are selected based on all the musculoskeletal structures, finding a trade-off between the segmentation accuracy of bones, muscles and implant. This differs from the single-modality cases, where the parameters are selected only on bones and implant (CT) or muscles (MRI), therefore being more performant on such structures.

Although the quantitative results do not show a significant improvement over the single modality approach, the proposed multimodal framework is able to provide a consistent and unified solution to the segmentations of both CT and MRI. This approach guarantees no overlap between the segmented ROIs, which cannot be ensured by the use of independent approaches for muscular segmentation on the MRI and bone or implant on the CT. On the template set, I verified that on average 2% of the voxels labeled as muscle on the MRI overlapped with CT-labeled bone voxels in the manual segmentations.
Moreover, without the use of a registration framework able to combine the two modalities while maintaining their biological plausibility, the joint segmentation of muscles, bones and implants would be more challenging and less accurate on the single modality. To test this hypothesis, I adopted the same LOOCV framework to obtain an automated segmentation of the Gluteus Medius on CT images, exploiting the available manual segmentations of the healthy side as already described in Sec. 4.4.1. The comparison of the Dice Score values from CT, MRI and combined multimodal framework is reported

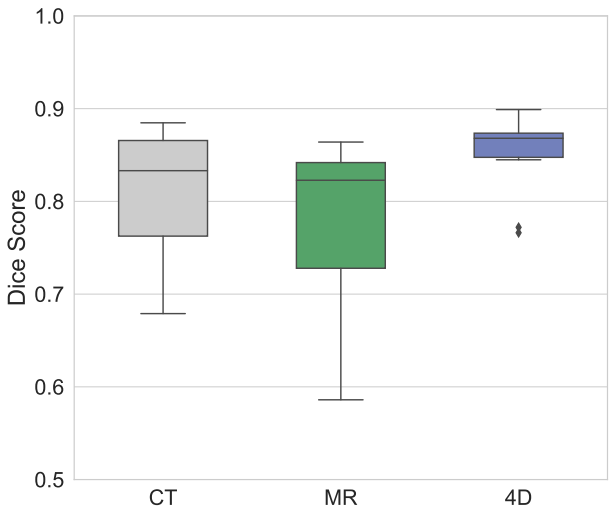


Figure 5.5: Dice Score values for healthy-side GMED automated segmentation using only CT, only MR or our 4D CT-MR framework. The Dice Score is computed against the manual segmentation on the CT space for the CT framework, on the MR space for the MR framework, and on the MR space then registered to the CT space for the multimodal framework.

in Fig. 5.5. A visual example is also shown in Fig. 5.6, reporting the automated segmentation of GMED and the ground truth. While the multimodal approach provides a reasonable result, the single modality results are less satisfactory: the CT-only automated segmentation is not able to clearly recover the boundary between fat and muscle; in the MRI-only, parts of the femur are wrongly classified as muscle.

For the sake of illustration, Figs. 5.7 and 5.8 report a similar example on the GMIN muscle, being the closest to the hip joint. Due to lack of manual ground truth on the CT, these figures only compare the segmentation obtained from MRI and the multimodal framework, for the healthy and the implanted hip sides. The Dice score values for GMIN resulting from the LOOCV experiment are reported in the summative Figure 5.4. This example shows that on the healthy side the use of the CT in the multimodal framework allows for better recovering of the muscle insertion on the greater trochanter, while on the implanted side it prevents bone tissue to be wrongly labelled as muscle.

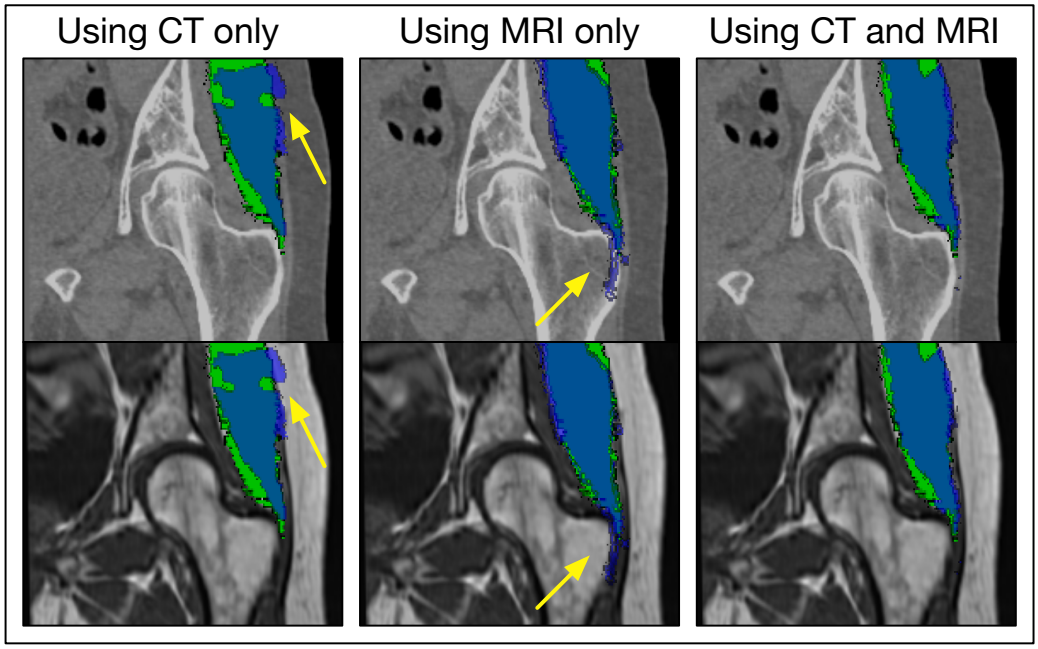


Figure 5.6: An example of GMED automated segmentation from the LOOCV experiments. The automated segmentation (blue) is overlaid onto the manual ground truth (green). Although the first two columns show results for the single-modality experiments, I reported both images for clearer visualisation of failure. For this case, Dice Score values are: 0.68 (CT only), 0.72 (MR only), 0.77 (multimodal approach).

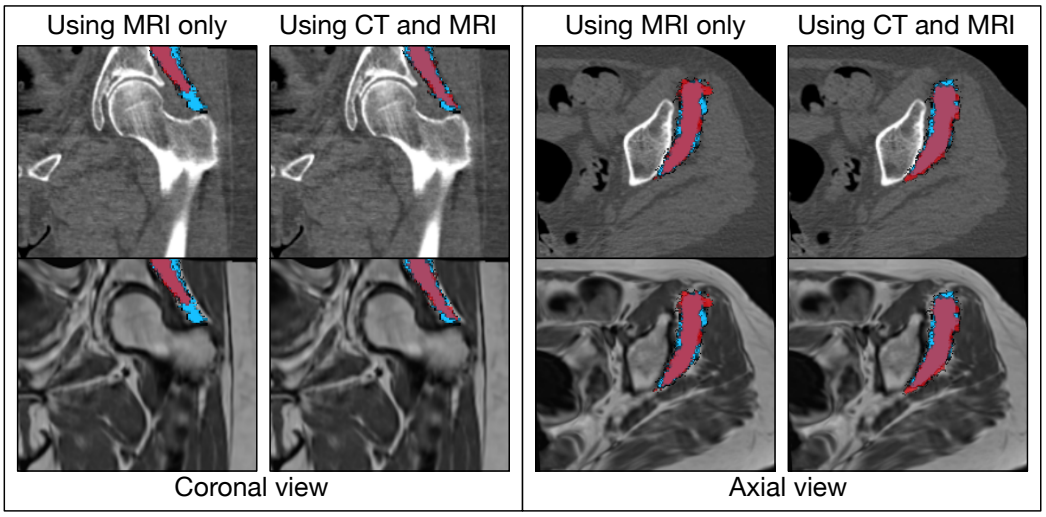


Figure 5.7: An example of GMIN automated segmentation from the LOOCV experiments (healthy hip side). The automated segmentation (red) is overlaid onto the manual ground truth (light blue). A coronal and an axial view of the same subject are reported. Dice Score values are: 0.60 (MRI only), 0.76 (multimodal approach). No manual segmentation was available in the CT space for GMIN, so CT only results cannot be shown.

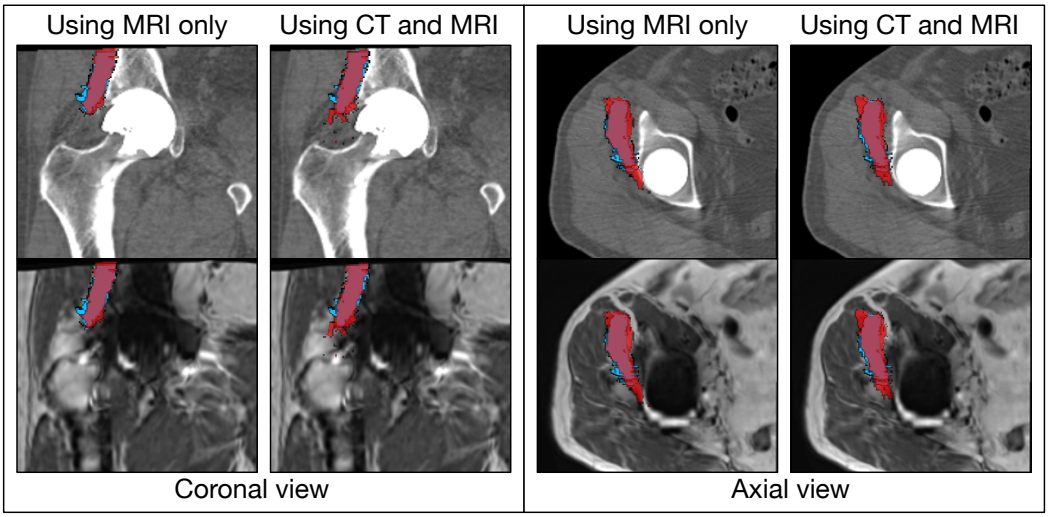


Figure 5.8: An example of GMIN automated segmentation from the LOOCV experiments (implanted hip side). The automated segmentation (red) is overlaid onto the manual ground truth (light blue). A coronal and an axial view of the same subject are reported. Dice Score values are: 0.81 (MRI only), 0.78 (multimodal approach). No manual segmentation was available in the CT space for GMIN, so CT only results cannot be shown.

5.3.2 Generalisability analysis

As a further test of robustness of the proposed segmentation pipeline, I extended the LOOCV experiment to a nested version, in order to verify the generalisability of the proposed approach to unseen data. In particular, I used N-1 subjects to perform the parameter selection in the same LOOCV fashion as described in the previous paragraph, and then I tested the performance of the model with the selected parameters on the hold-out subject. By iterating this over all the template cases, I obtained the summary Dice score statistics reported in Table 5.3. These results are in line with the full dataset LOOCV analysis, which represents the optimal performance, showing the robustness of this approach on a more realistic setting. The median Dice Scores are unchanged or within 1% difference. A very slight reduction of the Dice score range in some of the classes (e.g. muscles in the implanted hip sides) is imputable to the use of less templates for the parameter selection. Due to the small number of templates currently available and because of the large population variability, removing a template subject that highly contributes to representing this

Nested Leave-One-Out-Cross-Validation for Generalisability				
Healthy side				
	CT	*MR*	*Multimodal*	*p-value*
Bones	0.94 [0.65, 0.97]	N.A.	0.94 [0.75, 0.96]	0.0513
Muscles	N.A.	0.85 [0.65, 0.92]	0.86 [0.71, 0.92]	0.2789
Implanted side				
	CT	*MR*	*Multimodal*	*p-value*
Bones	0.91 [0.64, 0.95]	N.A.	0.88 [0.64, 0.92]	0.00002*
Muscles	N.A.	0.83 [0.60, 0.93]	0.80 [0.58, 0.91]	0.0254*
Implant	0.92 [0.70, 0.96]	N.A.	0.91 [0.82, 0.94]	0.1294

Table 5.3: Median Dice Score values and 95% percentile ranges for bones, implant and muscles obtained from the nested LOOCV (p-values for 5% significance level of Wilcoxon signed rank test between monomodal and multimodal comparisons are shown). N.A. indicates that manual segmentation was not available.

variability might bias the parameter selection towards other types of subjects (*e.g.* removing a subject with particularly high distribution of adipose tissue might lead to poor segmentation of patients with more robust constitution). This more adversely impacts the multimodal approach especially in the more variable muscular structures, as the inter-subject registration needs to balance the alignment of both CT and MRI. Overall, however, this experiment did not show a significant drop in performance when tested on the hold-out set, so it can be concluded that the proposed approach is able to generalise to unseen data.

5.3.3 Effect of automated bone masks

A final set of experiments aimed at extending the analysis on propagation of segmentation errors due to the use of the automated bone masks extraction (Sec. 4.4.2). To verify whether any error would propagate to the end of the pipeline, the multi-atlas segmentation propagation was performed on both CT-MRI pairs registered with manual bone masks and CT-MRI registered with automated masks. A LOOCV scheme was used in this case as well both in the automated mask generation and in the multi-channel multi-atlas segmentation propagation to avoid the introduction of biases.

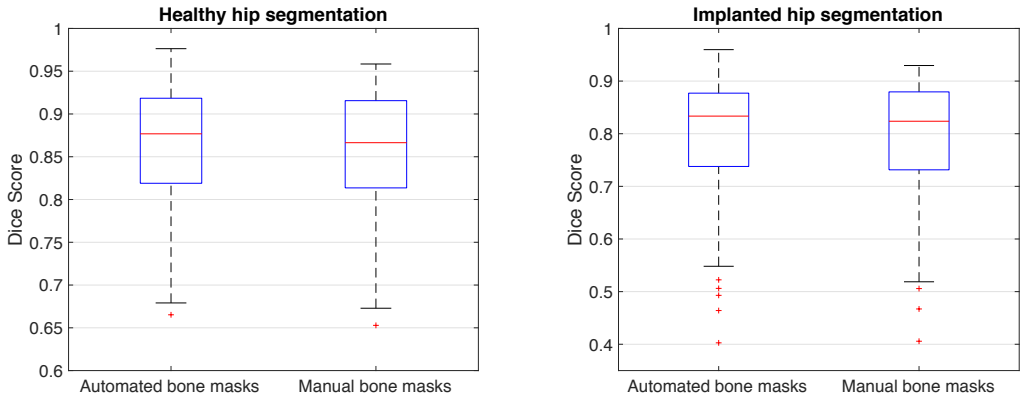


Figure 5.9: Dice Score for all the structures obtained after intra-subject registration with either automated or manual bone masks.

Figure 5.9 shows the resulting Dice Score distributions for all the structures of interest (bones, muscles and implants). No statistically significant difference was found (Wilcoxon signed rank test, 5% confidence level), proving the robustness of the full segmentation pipeline against the use of automated bone mask definition for the intra-subject CT-MRI registration.

5.4 Discussion

An automated processing framework to register and jointly segment same-subject pelvic CT and MR images was developed to facilitate the definition of regions of interest for hip arthroplasty assessment. This framework allows to highlight both muscular and skeletal structures on all available modalities. The implant segmentation, for example, can be overlaid on the MR image, where the metal artefact completely shadows it. Visualising the spatial position of the implant with respect to muscular structures could help better determine the muscles at greater risk of developing atrophy or the presence of other inflammatory lesions. Also, by reducing the burden of manually selecting regions of interest through an automated segmentation scheme, 3D rendering of patient-specific anatomy as well as volumetric imaging biomarkers can be more easily derived.

The proposed pipeline was developed taking into consideration the large population variability and the sources of image artefact and noise, which ham-

per the generalisability of automated image analysis in MSK.

The image quality enhancement block makes it suitable for routinely acquired clinical data, heavily corrupted by noise and artefacts. The use of a tailored bias field correction scheme and of the super resolution algorithm helps obtain high quality MR images with isotropic resolution, without the need for longer volumetric MR acquisitions. The introduction of a Metal Artefact Reduction (MAR) technique for the CT also facilitates the registration tasks, and thus promotes accurate final segmentation.

As already discussed in Chapter 4, the intra-subject registration with rigidity constraints enables accurate alignment of the CT and MRI, while guaranteeing the anatomical fidelity in the applied deformations, crucial to obtain clinical trust.

Should the intra-subject registration succeed, the subsequent steps of the pipeline are not expected to fail but rather to be non-optimal in the worst case. This is due to the robustness of the multi-channel multi-atlas segmentation block. First, a robust affine scheme is used for inter-subject registration, which allows up to 50% of the pairwise affine registrations to fail without compromising the final outcome. Second, the fusion algorithm automatically selects the template images that appear to be the most similar to the target image and thus the best registered. This second aspect also enables the pipeline to deal with several non-linear inter-subject registration failures. In the Leave-One-Out Cross Validation experiment, the presented multi-channel multi-atlas segmentation approach proved to achieve good accuracy with median Dice Score of 0.90 for skeletal and 0.84 for muscular structures. The implanted sides overall reported a slightly lower accuracy, but it is reasonable to expect an improvement on this performance following further advances in CT-MR registration quality, such as the introduction of the sliding modelling, and increased template dataset size.

The multimodal approach introduced a small reduction in performance when compared to the use of CT for bone segmentation and MRI for muscle

segmentation (Table 5.2). However, the experiment on the GMED showed that there is added value in the use of multimodal information to segment structures which the single modality lacks constrast for (*e.g.* muscles in CT). The lack of contrast for muscular structures in CT images hampers the templates manual delineation process, introducing larger variability in the segmentations propagated to a target space and therefore hindering the accuracy of the final consensus. In the MRI space, the manual segmentation of the templates is more reliable, but the qualitative nature of this modality makes the inter-subject registrations more difficult in the multi-atlas approach. The proposed 4D multimodal framework takes advantage of both modalities, as the CT helps regularise the templates-to-target registration while exploiting the more trustworthy manual segmentation from the MR, leading therefore to more robust results.

5.4.1 Comparisons with the literature

A comparison with four of the most recent and most similar works was carried out to further evaluate the performance of the proposed framework in relation to the state-of-the-art. It is worth noting that only approximate and inconclusive interpretations of this comparison can be drawn, as the methods have been trained and tested with different datasets, varying in size, quality and presence/absence of metal artefact induced noise.

Results on the healthy hip sides were compared to: Yokota *et al.* [94], who proposed a hierarchical multi-atlas approach for MSK segmentation in CT images, and thus is the most similar in terms of goal and methodological approach; Hiasa *et al.* [171], who presented a multimodal framework for CT-MR image synthesis, followed by a 2D UNet for joint segmentation of MSK structures in MR-synthetic CT pairs, and thus it is the most similar in the use of multimodal information; and Liang *et al.* [112], who used a multi-resolution CNN to segment pelvic MR images. The performance metrics reported in their original articles are compared to the proposed approach in Table 5.4. The proposed approach outperformed the multimodal segmentation network

presented by Hiasa *et al.* [171] in all the compared structures, though I acknowledge that their use of synthetically generated data for training might not be representative of performance on real data. Yokota *et al.* [94] reported slightly higher Dice Score for GMAX, GMED and TFL (still within one standard deviation), while improved performance was observed on the smaller GMIN, typically characterised by higher shape variability. Similar performance on bone segmentation was observed in comparison to Liang *et al.* [112], although they reported consistently higher accuracy in muscle segmentation. However, results were provided for a single subject only, so it is difficult to assess the validity of this comparison. In addition, the values displayed here for Liang *et al.* [112], are the average between the left and right hip sides, which are instead reported in the original paper. When observing their original values, the left hip side appeared systematically worse than the right hip side (*e.g.* IoU = 93.65/84.70, IoU = 95.28/77.84, IoU = 92.82/58.30 for right/left GMAX, GMED and GMIN respectively), and no clear explanation of this behaviour is provided.

Likewise, a comparison with the cascade of U-Nets proposed by Sakamoto *et al.* [113] for the segmentation of metal-artefact corrupted CTs is reported in Table 5.5. Similar accuracy was obtained in the segmentation of GMAX and slightly worse for GMED, although again the small test dataset size invalidates any conclusion.

5.4.2 Conclusions

The presented segmentation framework represents a promising tool for the automation of image analysis in hip arthroplasty. The combination of multi-modal information proved its efficacy in providing a robust and unified solution for the segmentation of both skeletal and musculare structures, as well as implants.

The block structure of the proposed framework makes it flexible and readily adjustable to various types of input data. Despite relying on a dataset of registered CT-MRI atlases, the pipeline could be applied to a single modality and

	Dice score		
	Yokota *et al.* [94] (CT only)	Hiasa *et al.* [96] (MR+synthetic CT)	Proposed (MR + CT)
Nr test cases	20	10	10
Pelvis	-	0.808 ± 0.036	**0.882** ± 0.072
Femur	-	0.883 ± 0.029	**0.949** ± 0.013
GMAX	**0.921**	-	0.906 ± 0.024
GMED	**0.875**	0.804 ± 0.040	0.850 ± 0.045
GMIN	0.697	0.669 ± 0.054	**0.826** ± 0.040
TFL	**0.807**	-	0.797 ± 0.088

	Intersection over Union (%)	
	Liang *et al.* [112] (MR only)	Proposed (MR + CT)
Nr test cases	1	10
Pelvis	**79.92**	79.53 ± 10.12
Femur	90.17	**90.24** ± 2.15
GMAX	**89.18**	82.87 ± 3.82
GMED	**86.56**	74.13 ± 6.20
GMIN	**75.56**	70.53 ± 5.32
TFL	**84.19**	67.01 ± 10.93

Table 5.4: Comparison with literature results. Literature methods were tested on non implanted patients, hence only "healthy" hip sides were considered for the proposed method. Mean values are reported with their standard deviation (if available). In bold is the highest value for each structure.

	Dice score	
	Sakamoto *et al.* [113] (CT only)	Proposed (MR + CT)
Nr test cases	3	12
GMAX	**0.91**	0.89 ± 0.03
GMED	**0.88**	0.81 ± 0.07

Table 5.5: Comparison with literature results presented by Sakamoto *et al.* [113] for the segmentation of muscles in patients with hip implants. Results are average of both implanted and contralateral hip sides. In bold is the highest value for each structure.

still provide the whole musculoskeletal range of structures while requiring only little modification. However, further analysis should be performed to test the robustness of the pipeline when used on single image or on other acquisition types.

This approach showed accuracy comparable to other state-of-the-art methods, although further adjustments could have a positive impact in improving its performance. For instance, variability in anatomical and implant structures could be better captured by increasing the number of template datasets. Moreover, a more thorough hyper-parameter search could further optimise the segmentation propagation and label fusion steps. Finally, and most importantly, the only active measures for reducing the impact of the metal artefact in the images are presently the application of RMAR [134] to CT, and the use of NMI in the intra-subject registration, being a similarity measure robust to noise. Better approaches to metal artefact reduction in both CT and MRI could be beneficial and favour higher segmentation accuracy, as they would reduce the impact of noise in the several registration steps required by this approach.

Nonetheless, the proposed pipeline is a valuable tool towards the automation of ROIs definition, from which clinical biomarkers could be derived to inform the clinicians about skeletal and muscular conditions around the implant. This potential application of the segmentation framework will be explored in the next chapter.

Chapter 6

Imaging biomarkers for muscle atrophy

Motivation

Development of muscle atrophy is a sign of implant failure, but current clinical assessment relies on subjective and qualitative visual scoring systems.

Contribution

Introduction of a multimodal imaging biomarker, the Intramuscular Fat Fraction, for automated quantification of fatty infiltration and muscle atrophy assessment.

The interpretation of musculoskeletal images and the assessment of implant failure could be facilitated by the identification of reliable and quantitative imaging biomarkers. In the context of muscle atrophy, two quantities are of main interest to define muscular condition: the muscle gross volume and the quantification of fat infiltration within the muscle [29]. Current clinical practice is based on a visual and qualitative assessment of few 2D slices, which is therefore strongly subjective and does not account for the whole muscular volume. A few works have proposed 3D measures of volume [30–32] and fat infiltration [38–40, 42]. They however rely on the time-consuming manual segmentation of the whole muscle, making it difficult to scale up for clinical

deployment.

The segmentation pipeline proposed in the previous chapter provides a fully automated means for (1) patient-specific musculoskeletal anatomy visualisation and volumetric rendering, and (2) delineation of regions of interest (ROI) for the extraction of clinically relevant information. By overcoming the need for manual segmentation, it enables the automated derivation of 3D metrics to quantify muscle atrophy.

This chapter reports a series of studies to evaluate the feasibility of automatically extracting imaging biomarkers of muscle atrophy from the automated segmentation. Two measures will be analysed and compared against current clinical assessment: the volume percentage difference - measuring lateral-contralateral volume asymmetry - and the Intramuscular Fat Fraction - measuring the percentage of fatty infiltration in the muscles. Finally, a preliminary study of the association between muscle atrophy and implant metal wear will be presented.

6.1 Volume asymmetry

Volume measurements are the first and most basic quantitative assessment that can be extracted from 3D muscle segmentation, being simply the product of the number of segmented voxels and the voxel volume. A straightforward potential measure of muscle atrophy can be derived from volumes by quantifying the muscle volume asymmetry between the painful and the contralateral hip side. Such an approach has already been reported in the literature for the characterisation of hip osteoarthritis [31, 32], showing significant GMAX and GMED asymmetries in patients with advanced pathology (19.7% and 12% volume reduction respectively), and a statistically non-significant shrinkage trend for GMIN as well (around 8% for both mild and advanced pathology). Similar analysis was performed by Takao *et al.* [172], who reported a mean volume difference of 18.6±7.1% in 13 of 17 muscles pre-operatively from 20 patients with hip osteoarthritis. No post-operative analysis was however reported.

In line with the current literature, I thus estimated the volume asymmetry in the available dataset to verify its potential use as a biomarker for muscle atrophy. Volume asymmetry was defined in terms of volume percentage difference between the implanted hip side and the contralateral one:

$$V_{\text{diff}} = 100 \times \frac{V_{\text{contralateral}} - V_{\text{implanted}}}{V_{\text{contralateral}}} \tag{6.1}$$

In case of bilateral prostheses, the side reported as painful in the radiological report was considered as "implanted" and the non-painful side as contralateral. In order to establish the accuracy of the automated segmentation for this task, I firstly compared the derivation of the volume percentage difference from manual and automated segmentations on the template dataset. Then, I estimated the volume asymmetry in the whole dataset and compared results across muscles. Finally, I compared it with the visual radiological score of muscle atrophy to verify its ability to differentiate subjects according to current clinical classification. The results of these three analyses are discussed in the next sections.

6.1.1 Automated vs manual segmentation comparison

As one of the main goals of this work is the automation of image segmentation and biomarkers extraction, I firstly evaluated the ability to accurately recover the volume percentage difference from the automated segmentation.
To this aim, I considered the eleven template subjects and their respective manual segmentation or automated segmentation as obtained from the LOOCV experiment (described in Sec. 5.3.1). The volume percentage difference was computed from either the manual and the automated segmentation according to Eq. 6.1 for the three abductor muscles. Their comparison is shown in Fig. 6.1.

Overall, this analysis resulted in a mean absolute error of 15.3 ± 11.8% between the volume difference from manual and automated segmentation (9.5 ± 6.4% for GMAX, 18.9 ± 11.9% for GMED, 17.5 ± 13.6% for GMIN). How-

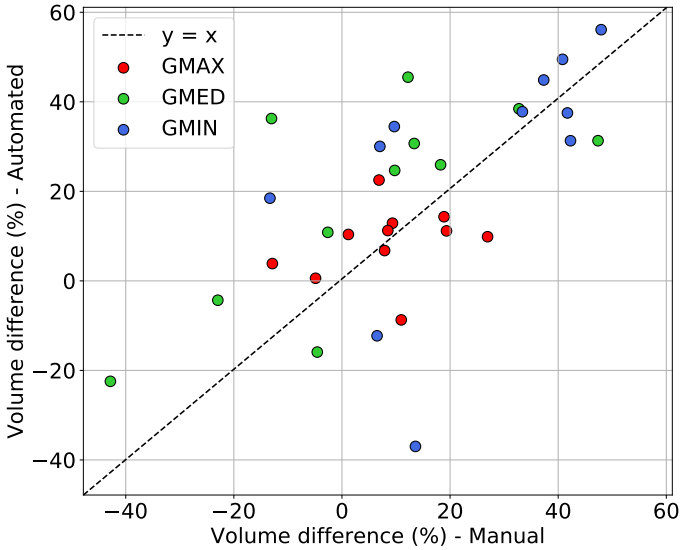


Figure 6.1: Comparison between painful-contralateral volume percentage difference estimated from automated vs. manual muscle segmentation. The identity line is displayed in black for reference.

ever, Wilcoxon signed-rank test with 5% significance level did not highlight significant difference between the two distributions.

6.1.2 Volume asymmetry in the available dataset

The previous analysis allows us to quantify the potential error introduced by the use of automated segmentation to extract volume measures. Keeping this in mind, I extended the volume asymmetry analysis to the remaining data for which the manual segmentation was not available, with the aim of evaluating the volume asymmetry in the available population sample.

The dataset . For this analysis, I considered the full dataset described in Chapter 3, excluding only the subjects used to build the template dataset and the subjects with bilateral implants whose painful side information was not available. In total, 69 subjects were thus automatically processed with the pipeline described in Chapter 5 and the painful-contralateral volume percentage difference was derived from the automated segmentation. Demographic information for this subset of the data is reported in Table 6.1. No substantial change is noted in the demographics distribution of this subset with respect to the full dataset.

Volume asymmetry analysis

	Females	Males	Total
Number of subjects	43	26	69
Unilateral cases	30	22	52
Bilateral cases	13	4	17
Mean [Range] age	53.74 [23, 74]	57.88 [35, 70]	55.30 [23, 74]
MRI within 1 month from CT	32	15	47
MRI within 6 months from CT	4	2	6
MRI within 12 months from CT	2	1	3
MRI within 24 months from CT	4	6	10
MRI within 37 months from CT	1	2	3

Table 6.1: Summary of the dataset demographic statistic for the subset employed in the volume asymmetry analysis.

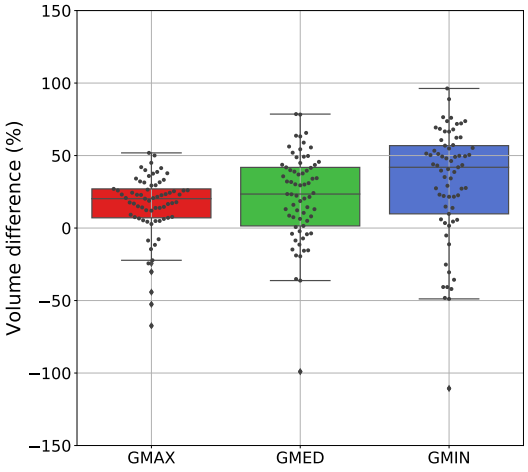


Figure 6.2: Distribution of painful-contralateral volume percentage difference for the three abductor muscles in the dataset.

	Median [2.5, 97.5]% range
GMAX	20.25 [-46.71, 46.50]
GMED	23.45 [-55.06, 69.42]
GMIN	41.92 [-67.39, 80.20]

Table 6.2: Median and 95-percentile range of painful-contralateral volume percentage difference for the three abductor muscles.

Results . The distributions of the volume percentage difference for the three abductor muscles are shown in Figure 6.2, with the corresponding median and 95 percentile range reported in Table 6.2. Unsurprisingly, these figures show a trend towards higher volume shrinkage the closer the muscle is to the implant, with GMIN reporting the highest volume asymmetry. The difference in muscle size also reflects into higher variability of volume difference for GMED and GMIN compared to the larger GMAX.

Although a direct comparison is not possible, the values obtained from this analysis seem compatible with the results reported by Takao *et al.* [172]. They estimated an average volume shrinkage of 18.6% between painful and contralateral side *pre-operatively*, while no post-operative estimation is available for comparison. Qualitatively, the increase in volume shrinkage observed in this dataset is not surprising, since these patients have been referred to Charing Cross Hospital for potential failing implant and need for revision surgery.

Nonetheless, considering the fairly large error associated with the proposed automated estimate of the volume difference, further validation is needed to assess the clinical use of this measure as a biomarker for muscle atrophy.

6.1.3 Comparison with radiological score

In order to verify the validity of the volume percentage difference as a useful indicator for muscle atrophy, I benchmarked this candidate biomarker against the current clinical evaluation. In particular, I considered a subset of cases with associated Bal and Lowe scoring [27] of Gluteus Medius and analysed its association with the proposed volume percentage difference.

The dataset . Within the dataset, 27 subjects had associated muscle atrophy grading for GMED (excluding also cases belonging to the template dataset). The grading had been performed by an expert radiologist according to the Bal and Lowe visual scoring system [27]: grade 0 corresponds to no atrophy, grade 1 is less than 30% change in muscle, grade 2 between 30% and 70%, and grade 3 more than 70% fatty change. Such radiological scores were available only for the implanted hip sides, and only the symptomatic side was considered in case of bilateral hip replacement. Summary demographic statistics are reported in Table 6.3. Although smaller, this subset preserves most of the full dataset characteristics: 2:1 female-male splitting, unilateral cases more than double the bilateral ones, and similar distribution across time difference between MR and CT. A slight increase in the average age and range is noted, due to the lack of clinical score for youngest subjects.

Radiological score correlation analysis			
	Females	**Males**	**Total**
Number of subjects	18	9	27
Unilateral cases	11	8	19
Bilateral cases	7	1	8
Mean [Range] age	56.61 [35, 74]	60.22 [45, 69]	57.81 [35, 74]
MRI within 1 month from CT	14	5	19
MRI within 6 months from CT	2	0	2
MRI within 12 months from CT	0	0	0
MRI within 24 months from CT	1	4	5
MRI within 37 months from CT	1	0	1

Table 6.3: Demographics of the patients whom Bal and Lowe radiological score for muscle atrophy is available.

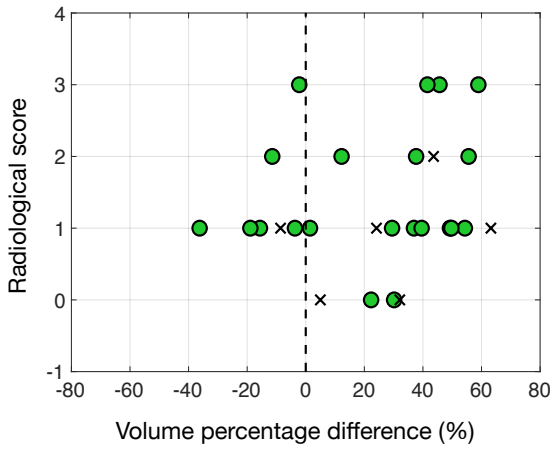


Figure 6.3: Painful-contralateral GMED volume difference against radiological score for muscle atrophy. Cases whose CT and MRI were acquired with more than 6-month difference are reported as crosses.

Results . The association between the four groups of the Bal and Lowe scoring system and the volume percentage difference was tested with an Analysis of Variance (ANOVA) study. The distribution of the volume percentage difference values grouped per score are shown in Figure 6.3. As seen from the figure, the volume percentage difference does not allow for a clear separation of classes. The one-way ANOVA supported this conclusion, indicating no significant difference in the means of the four groups ($p > 0.05$). The Spearman's rank-order correlation coefficient ρ_s, computed between the volume percentage

difference and the associated radiological score, further confirmed the lack of significant correlation ($\rho_s = 0.188$, p-value$= 0.347$).

6.1.4 Limitations

The analyses presented in this section suggest that the volume percentage difference might not be a reliable biomarker for muscle atrophy. Its current definition is characterised by several limitations. First, the value of the volume percentage difference could be completely biased by the image FOV and by the patient's position within the scanner: being the muscles quite extended and not always fully imaged within the FOV, a difference in volume might stem from asymmetric coverage of the left and right hip sides due to patient's pose, rather than actual volume shrinkage. In other words, the same muscle might be truncated at slightly different heights on the painful and on the contralateral side. A standardisation of the FOV coverage at acquisition would thus be required to make sure the whole muscle volumes are included. In addition, this biomarker definition requires the identification of a non-implanted or at least non-painful hip side to be used as a reference. This might not be trivial or applicable in patients with bilateral hip implants.

Assuming the volume percentage difference from the manual segmentation as a ground truth estimation, the error analysis shows a discrepancy between the manual and the automated estimation up to the same order of magnitude of the actual painful-contralateral differences to be captured. This suggests that the error introduced by the automated method might further bias the biomarker calculation, and thus make the quantification of the actual asymmetry not trustworthy. Moreover, Takao *et al.* [172] reports significant asymmetry already pre-operatively, adding further difficulties in disentangling the effect of the implant in the development of muscle wastage.

Taking all these considerations into account, at present the proposed volume percentage difference is not suitable as a clinical biomarker. Perhaps, a measure of volume shrinkage between pre- and post-surgery in a longitudinal analysis could be a potential future direction of investigation.

6.2 Intramuscular Fat Fraction

The development of muscle atrophy is characterised not only by a reduction of volume but also by increased fatty infiltration within the muscle region (intramuscular fat). A standardised and quantitative assessment of fatty degeneration could thus provide a relevant biomarker for muscle wastage.

The estimation of fatty infiltration has traditionally been derived from water-fat separated MR imaging, using chemical-shift acquisition protocols [38, 39], and requiring manual annotation of the muscle region to consider. Such protocols are however very sensitive to magnetic field inhomogeneities and thus less reliable for patients with metal implants [23]. Automated muscle-fat segmentation methods in T1-MR imaging have been introduced as well [37, 41], still reporting the need for manual intervention in either defining the ROI or selecting an appropriate thresholding to classify voxels as muscle or fat. The use of CT radiodensity to measure the percentage of fat was also proposed in the literature, showing correlation to muscle strength in patients with hip osteoarthritis [42]. Previous studies however rely on manual segmentation of muscles in CT, as the lack of intra-class contrast hampers the automated segmentation of individual muscles.

In this work, given the availability of same-subject CT and MR images, I propose an estimation of the intramuscular fat fraction which takes advantage of multimodal information to characterise fatty infiltration in patients with hip implants. The quantitative nature of CT and its clear separation between fat and muscle intensities favour the automated tissue classification within regions of interest [89]. This can be combined with the proposed automated segmentation pipeline to obtain accurate MRI-driven muscle boundary delineation also in the CT space, even in the presence of strong artefact induced noise. In this way, a quantification of fatty infiltration can be fully automatically derived from the input CT and MR images.

In order to differentiate intramuscular fat and lean tissue within each muscle ROI, the output of the presented segmentation pipeline needs to be further

post-processed. Each muscle label is extracted independently from the multi-label segmentation output, and considered as a mask of the ROI for the fatty infiltration estimation. The mask is eroded by three voxels to ignore intermuscular fat at the volume boundary and potential segmentation inaccuracies [38]. Within each eroded ROI mask, the CT image is segmented by thresholding at -30 HU [89], classifying each voxel as either adipose tissue or lean muscle. From this classification, the percentage of fat tissue in the muscle gross volume can be computed as Intramuscular Fat Fraction (IFF):

$$\mathrm{IFF} = \frac{V_{\mathrm{fat}}}{V_{\mathrm{lean}} + V_{\mathrm{fat}}} \tag{6.2}$$

where the volume V of fat/lean muscle is the total number of voxels classified as fat/lean muscle multiplied by the voxel volume. In other words, MRI helps define the boundaries of the different muscles, i.e. the different ROIs, while CT helps differentiate fat and lean muscle tissue within each ROI.

Similarly to the analysis performed for the volume percentage difference, the next paragraphs will firstly illustrate the quantification of error introduced by the use of automated segmentation, and then the correlation of the proposed IFF with the radiological score for muscle atrophy.

6.2.1 Automated vs manual segmentation comparison

The first experiments aim at quantifying the robustness of the approach against the use of automated segmentation. The eleven template subjects were used for this analysis, as they were the only cases with available manual segmentation. As automated segmentation, I employed the results of the LOOCV experiments (Sec. 5.3.1). For both segmentations, the IFF was automatically computed according to Eq. 6.2, after the processing steps of erosion and CT thresholding. Figure 6.4 plots the IFF obtained from the automated vs. manual segmentation, the latter being considered as the ground truth value. Results are grouped according to the presence/absence of hip implants (implanted/healthy hip side). The mean absolute errors for each abductor muscle

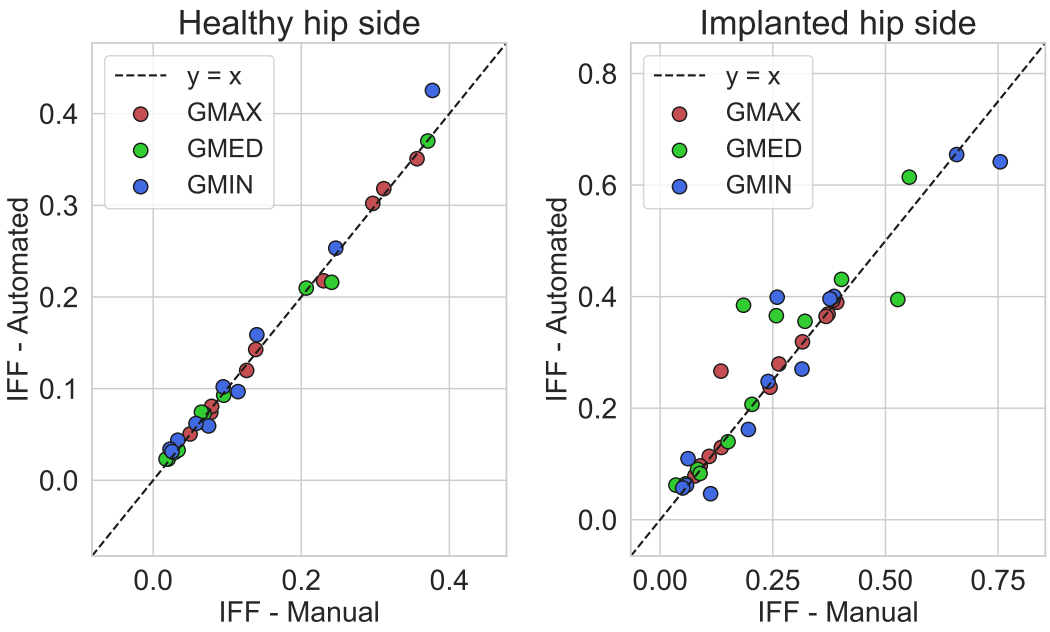


Figure 6.4: Comparison between intramuscular fat fraction (IFF) estimated from manual segmentation of muscles and IFF from automated segmentation as proposed in this work. Healthy hip side results on the templates dataset are shown on the left, implanted hip sides on the right. The identity line is displayed in black for reference.

	Mean Absolute Error	
	Healthy	Implanted
GMAX	0.005 ± 0.003	0.016 ± 0.035
GMED	0.006 ± 0.007	0.052 ± 0.060
GMIN	0.015 ± 0.012	0.042 ± 0.043
Total	0.008 ± 0.009	0.037 ± 0.049

Table 6.4: Mean absolute error and associated standard deviation between the IFF from manual and from automated segmentation on the 11 template subjects.

are summarised in Table 6.4. No significant difference was found for any of the muscles when comparing the automated and manual estimation (Wilcoxon signed rank test with significance level $\alpha = 0.05$). The GMAX reported the smallest error in both healthy and implanted hip sides, while GMED and GMIN had a slightly higher error, especially on the implanted side. This might be linked to the smaller size of these muscles (and thus more impacted by a constant erosion factor of 3). Most likely, this is due to the proximity

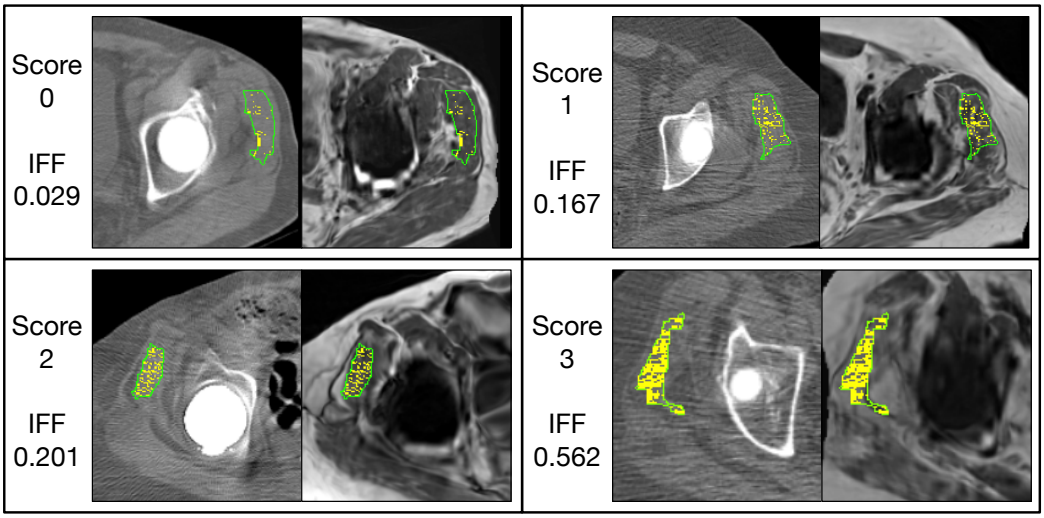


Figure 6.5: Examples of registered CT and MRI for each radiological score, and the automated fat segmentation from the proposed method. The eroded gross muscle boundary is reported in green. The fat segmentation is indicatedin yellow. The derived IFF is also reported, showing an increase of fat infiltration in agreement with the radiological score.

with the implant, and hence to the less accurate segmentation (already observed in the LOOCV experiment, Fig. 5.4). Overall, the estimation of the IFF from the automated segmentation appears fairly robust to segmentation inaccuracies, with most errors within 5-10% of the corresponding IFF value.

6.2.2 Comparison with radiological scores

The proposed automated quantification of IFF was benchmarked against the current radiological scores for muscle atrophy. The same 27-subject dataset previously described (Sec. 6.1.3) was used for this analysis as well. The CT and MR images of these subjects were automatically segmented and the IFF for the pathological GMED was then derived. The association between the computed IFF and the gold-standard radiological score was tested using the one-way ANOVA F-test. Examples for each radiological score of the obtained fat segmentation and the respective IFF values are reported in Fig. 6.5.

For the sake of comparison, I estimated the IFF from the MR images, i.e. discarding the extra information that the CT could provide. In this case, I manually identified a threshold for each subject to segment fat and lean muscle

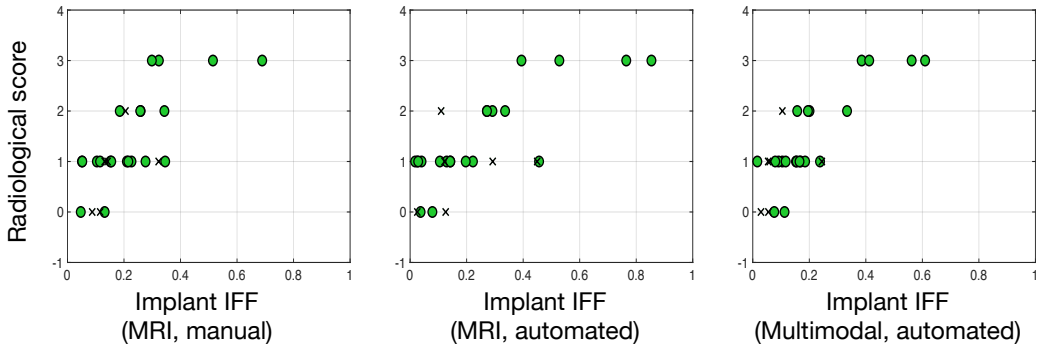


Figure 6.6: Intramuscular Fat Fraction of GMED (implanted side), estimated with different thresholding methods and plotted against the corresponding visual radiological score. Cases whose CT and MRI were acquired with more than 6-month difference are reported as crosses.

within the GMED mask and I computed the IFF from this segmentation. The same analysis was performed with an automated thresholding on MRI. The automated threshold was estimated by: (1) manually thresholding the GMED-masked MRI for the templates; (2) computing for each template,the percentile of the GMED intensity distribution the threshold corresponds to; (3) deriving the average percentile across templates. On this dataset, this corresponded to the 41st percentile of the full range of intensities in the masked MRI. The GMED muscles of the 27 subjects whose radiological score was available were then automatically segmented into fat and lean muscle using this percentile threshold, and the IFF was derived.

Figure 6.6 displays the association of the IFF values and the radiological scores for the three analysed cases. The results of the performed statistical analysis are reported in Table 6.5. These results show better inter-class separation when IFF is computed from the CT image, suggesting a stronger association between this IFF estimation method and the radiological score. The Spearman's rank-order correlation analysis also shows stronger significant correlation with the radiological score when the IFF is computed using the proposed multimodal approach. A fairly clear separation trend is also visible in the manual fat segmentation from MRI, which is however cumbersome and time-consuming. The use of an automated percentile thresholding in MRI

	MRI manual	**MRI automated**	**Multimodal automated (proposed)**
p-values for pair-wise comparisons with two-sample t-test			
0 vs 1	4.71E-01	5.68E-01	4.67E-01
0 vs 2	1.25E-01	2.02E-01	7.26E-02
0 vs 3	1.91E-04*	3.54E-05*	2.49E-07*
1 vs 2	5.27E-01	6.29E-01	3.34E-01
1 vs 3	3.14E-04*	2.59E-05*	8.91E-08*
2 vs 3	2.50E-02*	2.38E-03*	3.28E-05*
p-values for all groups comparison with One-way ANOVA			
	1.48E-04*	1.86E-05*	7.14E-08*
Spearman's rank-order correlation			
ρ_s	0.733	0.750	0.815
p-value	1.36E-05*	6.62E-06*	2.29E-07*

Table 6.5: Statistical analysis comparing the radiological scores and the IFF values obtained manually from MRI, automatically from MRI, and automatically with the proposed multimodal approach. The statistical significance threshold was set to 0.05 for all tests (significant values are indicated by *). The first block of the table reports the p-values for the pair-wise comparison between the different classes, grouped according to the radiological score. T-test was used, coupled with the Tukey's honestly significant difference procedure for multiple comparisons correction. The second block reports the p-values obtained with the One-way ANOVA F-test comparing all group means. The last block reports the Spearman's rank-order correlation coefficient ρ_s and its associated p-value.

does not improve the class separation and the choice of the threshold might not generalise well to other data. On the contrary, the thresholding in the CT domain can be directly applied to any other CT image without loss of generalisability. It can be noticed that the group with radiological score equal to 1 is the most overlapping with other groups. This result is not unexpected, as this analysis does not account for the inter- nor intra-rater variability of this visual score. In fact, class 1 represents the intermediate situation where no clear atrophy pattern can be distinguished in the image. Further analysis will therefore be needed to better represent this intermediate class.

6.2.3 Limitations

The presented analysis of Intramuscular Fat Fraction as an imaging biomarker for muscle atrophy is currently limited by the following aspects.

Sample size. The IFF was benchmarked on 27 subjects and only for the gluteus medius. Despite the encouraging results, a more thorough validation is required to assess its clinical relevance. This should include the estimation of intra- and inter-rater variability of the current visual scoring system used as gold standard, and extension of the approach to the other abductor muscles. Confounding factors such as gender and unilateral/bilateral implant were analysed and not found significantly relevant in this specific subset of patients. However, a deeper investigation of other factors such as time from primary surgery or presence of comorbidities would be useful to better characterise the biomarker on a larger cohort.

CT-MRI time difference. A large time gap between the acquisition of the CT and the MRI would jeopardize the reliability of the comparison between the multimodal IFF estimation and the MRI radiological score, as muscle change could have occured in between. However, clinical literature shows that significant gluteus medius atrophy change appears after a mean interval of 11 months between scans [8]. In the available dataset only six cases had a time difference greater than 6 months, and except for one case with grade 2 atrophy, all the others were associated with no or very little fatty infiltration. For this reason I included all the available cases in the analysis, in order to maximise the sample size. A very similar trend was indeed found even when the aforementioned six cases were excluded.

Partial volume effects. The effect of partial volume or any residual metal artefact hyper- or hypo-intensities was not explicitly modeled in the proposed estimation of IFF. Attempts to address this limitation were performed with a Gaussian Mixture Model (GMM) segmentation approach instead of simple thresholding to separate the lean muscle and fat classes within a ROI. The GMM would allow the modelling of partial volume effects by providing fuzzy

membership of each voxel to each class. However, when analysed in relationship with the radiological scores, no significant separation between groups was obtained.

Despite these limitations, the proposed IFF showed a strong association with the radiological score in the available dataset, and represents a suitable candidate for the automated quantification of muscle atrophy. To further assess its potential use as an imaging biomarker, I therefore investigated its relationship with another relevant biomarker for implant failure: elevated blood metal ion levels.

6.3 Muscle atrophy and blood metal ion levels

The follow-up of patients with MoM hip implants includes cross-sectional imaging as well as monitoring of whole blood metal ion levels, as indicated by the MHRA [11]. In particular, increased concentration of Cobalt ([Co]) and Chromium ([Cr]) in blood might indicate the presence of metal wear and potential risk of soft tissue reaction. Current indications require patients with metal ion levels higher than 7 ppb (7 ng/mL) having closer follow-up and cross-sectional imaging. However, there is no agreed ion levels threshold to automatically instruct revision surgery, and patient-management decisions should combine imaging findings, blood metal levels as well as patient's factors and implant type.

Metal wear determines local tissue reactions, associated with increased inflammatory response as the immune system develops hypersensitivity to metal antigen components [9]. While the link to development of pseudotumours is fairly well-understood, at present there is no evidence of relationship between increased metal wear and the development of muscle atrophy. Only one study in the literature addressed this issue. They found no associaton between gluteal muscle atrophy - assessed with the Bal and Lowe score - and Co or Cr ion levels in blood [173], in a cohort of 263 patients with unilateral ASR XL total hip replacement (posterior surgical approach). In line with our previous

findings, they reported atrophy on GMED and GMIN to be more common than in GMAX. However, only moderate-to-severe GMIN atrophy was found significantly more prevalent in patients with [Cr] exceeding 5 ppb than in patients with low levels.

While the study design is robust and is based on a very homogeneous cohort of subjects, the use of a crude visual scoring system such as Bal and Lowe might make the analysis scarcely sensitive to atrophy variations. Also, the blood metal ion levels were used only to separate the subjects into two classes: cases with elevated concentration (either [Cr] or [Co] > 5 ppb) or low.

Under the hypothesis that the radiological score might be too crude of a measure to reveal association between muscle atrophy and metal wear, I performed a correlation study between the IFF and the blood metal ion levels in the available dataset. The goal is to analyse the relationship between muscle atrophy and metal wear and test whether the use of continuous measures might better capture the presence of an association. The results, described in the next paragraphs, are to be considered only an exploratory analysis, given the small sample size and the heterogeneity in implant types and surgical approaches.

6.3.1 Correlation analysis

Dataset. In this project's cohort, the blood metal ion levels information was available for 35 subjects. However, 4 subjects had been previously selected as templates, and another 4 had to be discarded due to lack of painful side information. In total, 27 subjects were retained for this study and the demographics are reported in Table 6.6. As per previous experiments, the subjects were automatically segmented with the proposed pipeline and the IFF was derived according to the procedure in Sec. 6.2.

Results. The Pearson's correlation coefficient was calculated between the IFF of each muscle and the concentration of Cobalt and Chromium in blood. As shown in Table 6.7, all muscles reported significant correlation with the blood metal ion levels, with GMED showing the strongest (and most significant) association. Cobalt and Chromium levels are strongly correlated with each

Blood metal ion levels analysis			
	Females	**Males**	**Total**
Number of subjects	20	7	27
Unilateral cases	13	7	20
Bilateral cases	7	0	7
Mean [Range] age	55.10 [30, 74]	57.00 [42, 69]	55.59 [30, 74]
MRI within 1 month from CT	18	4	22
MRI within 6 months from CT	1	1	2
MRI within 12 months from CT	0	0	0
MRI within 24 months from CT	1	1	2
MRI within 37 months from CT	0	1	1

Table 6.6: Demographics of patients entering the association analysis between the intramuscular fat fraction and the blood metal ion levels.

	GMAX		**GMED**		**GMIN**	
	ρ	p-value	ρ	p-value	ρ	p-value
Co	0.498	0.008	0.574	0.002	0.439	0.022
Cr	0.440	0.022	0.464	0.015	0.393	0.043

Table 6.7: Pearson's correlation coefficient ρ between the intramuscular fat fraction and the concentration of Cobalt (Co) and Chromium (Cr) in blood. All coefficients are statistically significant at 0.05 significance level.

other ($\rho = 0.882$, with p-value $= 5.34 \times 10^{-11}$). As all these comparisons are performed on the same subjects, this analysis suggests that the Cobalt concentration might be slightly more strongly associated with the estimated IFF.

Available demographic and clinical information was inspected to uncover the presence of confounding factors. Visually, no evident separations or clusters were identified based on sex or implanted hip side (left, right or bilateral). This was confirmed with a one-way ANOVA analysis to compare the IFF or the blood metal ion concentrations across groups (i.e. female vs male, left vs right, right vs bilateral, left vs bilateral). In all cases no significant difference was found (5% significance level, corrected for multiple comparisons). Conversely, the age of the patient was found significantly correlated with IFF in GMED

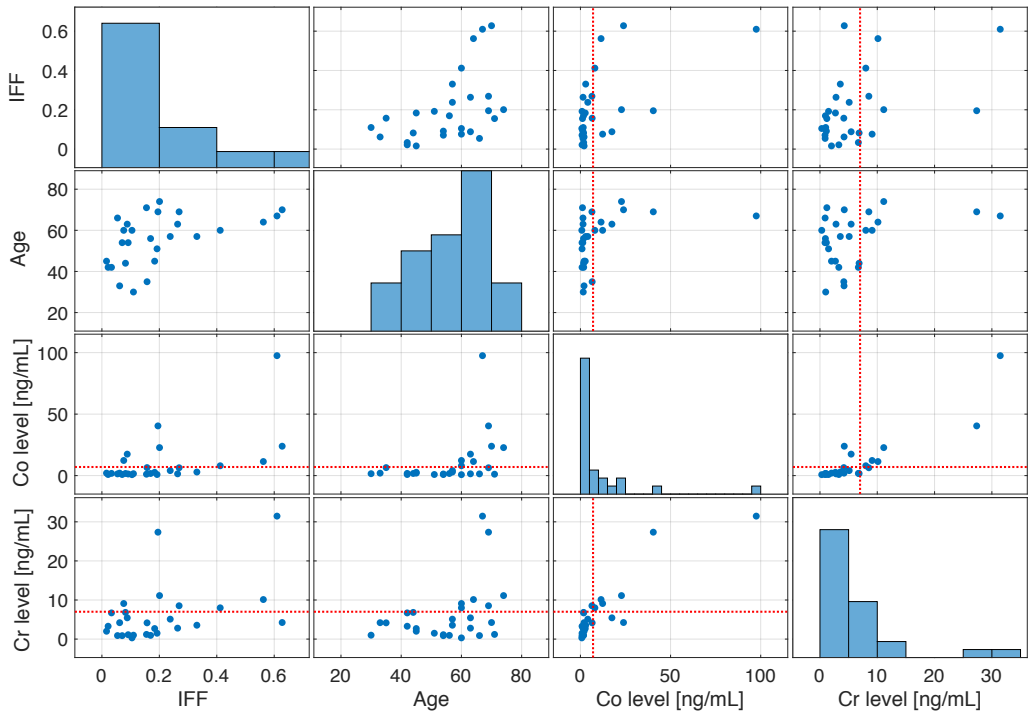


Figure 6.7: Matrix plot showing the relationship between GMED IFF, age of the patient, Cobalt and Chromium blood concentration levels. Red dotted lines show the 7 ppb threshold currently used as indication for patient's stricter follow-up.

($\rho = 0.490$, p-value $= 0.009$) and GMIN ($\rho = 0.461$, p-value $= 0.016$). Similar correlation coefficient was found between age and Cobalt blood concentration ($\rho = 0.399$, p-value $= 0.039$), and between age and Chromium blood concentration ($\rho = 0.373$, p-value $= 0.056$), but only the former reached signficance. A trend towards higher IFF was indeed noted for older patients, and this matched also with increased blood metal ion levels. A matrix plot is shown in Figure 6.7 to illustrate the trends captured by the correlation analysis, associating IFF for GMED, Cobalt and Chromium blood concentrations and the age of the patients. GMAX and GMIN demonstrated very similar distributions, so they are not reported here.

Overall, it was observed that the only two subjects with very low Co and Cr levels ($<$ 1ppb) reported low IFF in all three muscles (< 0.16). One case had highest IFF in the GMAX and the other in GMED. At the other end of the spectrum, the subjects with both [Co] and [Cr] above the 7 ppb thresh-

	GMAX	GMED	GMIN
[Co], [Cr] $<$ 1 ppb (2)	0.106	0.088	0.039
[Co], [Cr] $<$ 7 ppb (16)	0.164	0.135	0.179
[Co] $\geq$ 7 ppb or [Cr] $\geq$ 7 ppb (3)	0.155	0.328	0.265
[Co], [Cr] $\geq$ 7 ppb (6)	0.230	0.343	0.378

Table 6.8: Average Intramuscular Fat Fraction per muscle according to concentration ranges of Cobalt and Chromium levels. The number of cases per each group is indicated in round brackets in the first column.

old consistently showed high IFF on all three muscles, with highest IFF in GMIN (4 cases) and GMED (2 cases). The intermediate cases ([Co], [Cr] $<$ 7 ppb) showed greater variability: 9 cases had highest IFF in GMAX, 2 cases in GMED and 5 cases in GMIN. Finally, the three cases with only one of the metal ion levels above the 7 ppb threshold showed increased IFF with respect to the intermediate cases (1 case with highest IFF in GMAX, 1 in GMED and 1 in GMIN). These results thus indicate that, while GMAX demonstrates some degree of atrophy across all level groups, the highest concentrations of Cobalt and Chromium are associated with GMED and GMIN as the most atrophied muscles. The average IFF per group, shown in Table 6.8, further confirms the observed trend of increasing IFF with increased blood metal ion levels, particularly for GMED and GMIN. As these muscles are closest to the implant, these findings might suggest a potential association between implant wear and muscle atrophy.

6.3.2 Limitations

This analysis mostly aimed at providing descriptive information of the observed relationships between IFF and blood metal ion levels. These results should be considered only a preliminary exploratory phase and would require further investigation. In addition to the limitations of the IFF calculation discussed in Sec. 6.2.3, this study is hampered by the large heterogeneity of the sample and the current lack of information to characterise it. Differently from the study proposed by Reito *et al.* [173], the analysed cohort presents a variety of

hip implants (eithr THR and HRA), a variety of surgical approaches and the presence of both unilateral and bilateral cases. Moreover, the time difference between the imaging and the blood tests could not be retrieved, although it is reasonable to assume they were acquired as part of the same screening program, under MHRA guidance. Finally, only a very small subset of patients had both radiological score and blood metal ion levels available. This information would help replicate the analysis of Reito *et al.* [173] and support/discard the hypothesis of IFF being more sensitive than the radiological score in capturing atrophy variations.

Nonetheless, the observed trends and correlations prompt further clinical investigation to better understand whether there might be a link between metal wear and muscle atrophy, and to more clearly analyse the role of confounding factors, e.g. considering how physically active the patients are.

6.4 Discussion

Current assessment of muscle atrophy in hip arthroplasty is based on a crude visual scoring system applied to MR images. Such a system is qualitative, often subjective and reliant on only few 2D slices instead of considering the whole muscular volume. The lack of automated techniques for ROI delineation is the main bottleneck in extracting volumetric measures to assess muscular conditions. In order to address this problem, I tested the feasibility of deriving imaging biomarkers for muscle atrophy from the automated segmentation pipeline proposed in this thesis.

I first considered a quantification of volume asymmetry based on the volume percentage difference between the lateral (painful) and contralateral hip side. This measure showed to suffer from an error comparable in magnitude to the effect to capture. Also, in a cohort of 27 subjects, it was unable to differentiate cases according to their radiological score classification. For these reasons, the proposed measure for volume asymmetry was deemed unsuitable as a muscle atrophy biomarker.

A second measure, the Intramuscular Fat Fraction, was then derived to characterise the fat infiltration in muscles. The proposed IFF estimation takes advantage of the multimodality information to automatically and more robustly classify lean muscle and fat tissue. Whilst retaining the 3D nature of muscles, this approach avoids the need for manual ROI selection, which is currently the most common practice for 3D fatty infiltration estimation [38, 40]. Differently from MR-based approaches, the use of a standardised range of intensities makes the automated classification into lean muscle and fat straightforward from the CT. The proposed measure can be estimated for each hip side independently, and proved to be less sensitive to gross muscle segmentation inaccuracies. Notably, it also showed strong association with the radiological score assessing GMED muscle atrophy.

While promising, the presented analysis is still constrained by the limited number of cases available for each radiological score. It would be of interest to extend the IFF analysis to the other gluteal muscles, as well as to include an estimation of the inter-rater variability of the current visual scoring system, in order to assess the reliability of the clinical benchmark. It would also be important to increase the dataset to better analyse the role of confounding factors, such as type of implant, time from primary surgery or the level of physical activity of the patients. Nonetheless, the proposed multimodal muscle atrophy estimation is advantageous as it combines the robustness of MRI-driven gross muscle volume segmentation with a consistent and reproducible fat quantification from CT.

Finally, a preliminary exploratory analysis investigated the relationship between implant metal wear - as measured by blood concentration of Cobalt and Chromium - and muscle atrophy - as measured by the IFF. The two biomarkers were found significantly correlated. In particular, GMED and GMIN reported the highest IFF in subjects with blood ion levels above the MHRA-indicated threshold of 7 ppb. No effect of gender or implanted hip side was observed in the distribution of the biomarkers, while age was found

positively correlated with both IFF and the Cobalt blood metal ion levels. Given the small sample size and its vast heterogeneity, no clinical conclusion can be drawn on the association between metal wear and muscle atrophy. However, these results may justify further investigation of this relationship to improve our current understanding of implant failure and its impact on muscular conditions.

Chapter 7

Multimodal metal artefact reduction

Motivation

Metallic implants induce strong artefacts in hip CT and MR images, hampering their diagnostic interpretation in the most clinically relevant regions.

Contribution

Introduction of a novel CT-MR metal artefact reduction technique, which takes advantage of multimodal information using an automated, data-driven, unsupervised approach.

Metallic implants are one of the main causes for image quality degradation in medical imaging of patients with hip implants. In Computed Tomography, the metal's higher attenuation coefficient causes signal corruption, resulting in bright and dark streaks that irradiate from the metal source throughout the reconstructed image. In Magnetic Resonance Imaging, metal objects induce local magnetic field inhomogeneities that cause intensity and geometrical distortions in the reconstructed image. These susceptibility artefacts typically appear as blackened areas close to the implant, partially shadowing the neighbouring structures. Metal artefacts are more pronounced close to the implant, thus reducing the diagnostic information provided by the images. This motivates research efforts towards more effective metal artefact reduction (MAR)

techniques in both CT and MRI.

Numerous approaches have been proposed in the literature for MAR in CT [132]. Traditional physics-based or iterative reconstruction methods are now being challenged by novel deep neural network approaches, which are data-driven and less dependent on physical model assumptions. However, most methods [137, 138] are trained in a supervised fashion, relying on either pre- and post-operative paired data (not always available) or simulations (not realistic enough). A solution to the supervised training setting was recently proposed by Liao *et al.* [149]. They introduced an unsupervised adversarial training scheme to disentangle the artefact from the anatomy appearance in CT images, showing state-of-the-art performance on both synthetic and real data. In MRI research, efforts have focused mostly on image acquisition improvements: Tailored MR sequences such as MARS [17] or SEMAC [155] have proven effective in reducing the extension of the shadowing, but cannot completely eliminate it, preventing the visualisation of the implant in MRI.

In this Chapter, I introduce a novel unsupervised deep learning MAR method for *jointly* correcting same-subject CT and MR hip images. The proposed Multimodal Artefact Disentanglement Network (MADN) extends the unsupervised approach of Liao *et al.* [149] by introducing a similarity loss that induces the network to learn shared information between CT and MRI content. As a result, the CT correction takes advantage of the sharper contrast of MRI throughout the field of view, while the MRI correction is helped by implant localisation information from the CT. As the appearance of the artefact is different in CT and MRI, it is hypothesised that making use of their contextual complementary information would help better correct for the artefact in both modalities.

This work represents a preliminary analysis of the feasibility and potential benefits of a multimodal metal artefact reduction approach, and was accepted for presentation at the IEEE International Symposium on Biomedical Imaging 2020 [174].

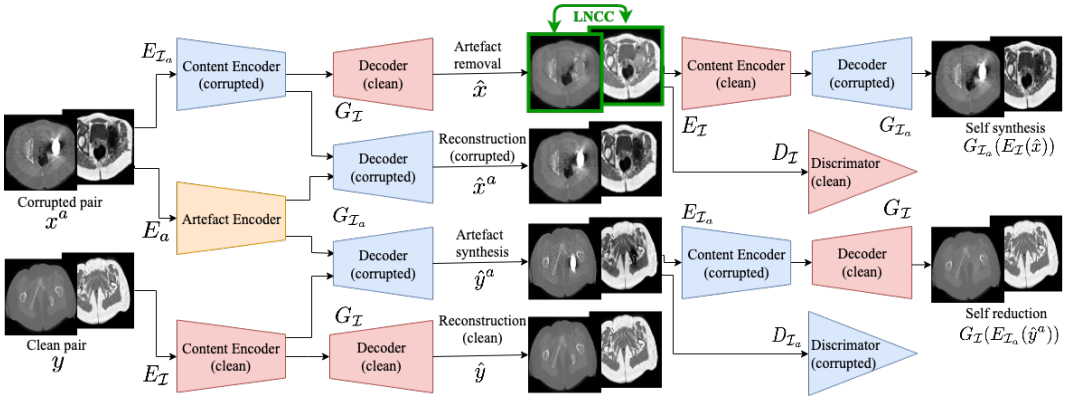


Figure 7.1: Schematic representation of MADN. A multichannel ADN architecture [149] was integrated with a novel similarity loss term to adapt it to the multimodal scenario.

7.1 From ADN to MADN

This work builds upon the Artefact Disentanglement Network (ADN) recently proposed by Liao *et al.* [149] to perform unsupervised MAR on CT images. The ADN uses two sets of unpaired images, one including metal artefact corrupted CTs and one with clean non-corrupted images. The key idea is to use encoder-decoder networks coupled with adversarial training to learn a latent representation of the data where the artefact is disentangled from the anatomical content. This separation allows to reconstruct the corrupted images using only the latent content representation, therefore removing the artefact. It also allows to reconstruct the denoised images with the latent artefact representation and thus synthesising corrupted images.

Let $\mathcal{I}$ and $\mathcal{I}_a$ be the domains of clean and corrupted images respectively. The network architecture (Fig. 7.1) is composed as follows: Three encoders ($\mathrm{E}_{\mathcal{I}} : \mathcal{I} \to \mathcal{C}$, $\mathrm{E}_{\mathcal{I}_a} : \mathcal{I}_a \to \mathcal{C}$, $\mathrm{E}_a : \mathcal{I}_a \to \mathcal{A}$) map the input images to either the content $\mathcal{C}$ or the artefact $\mathcal{A}$ latent spaces; two decoders map the latent space back to the image domain and work as generators ($\mathrm{G}_{\mathcal{I}} : \mathcal{C} \to \mathcal{I}$, $\mathrm{G}_{\mathcal{I}_a} : \mathcal{C} \times \mathcal{A} \to \mathcal{I}_a$); finally two discriminators ($\mathrm{D}_{\mathcal{I}}$, $\mathrm{D}_{\mathcal{I}_a}$) define whether an input is real (i.e. coming from the real distribution of $\mathcal{I}$ or $\mathcal{I}_a$ respectively) or fake (i.e. synthetically generated by the decoders $\mathrm{G}_{\mathcal{I}}$ and $\mathrm{G}_{\mathcal{I}_a}$ respectively). We refer the reader to the original manuscript [149] for the layer-by-layer outline of the network.

Given a corrupted image $x^a \in \mathcal{I}_a$ and a clean image $y \in \mathcal{I}$, we can define their encoding as $c^a = \mathrm{E}_{\mathcal{I}_a}(x^a), a = \mathrm{E}_a(x^a)$ and $c = \mathrm{E}_{\mathcal{I}}(y)$. Indicating by$\hat{}$the decoded images, from x^a we obtain a reconstructed corrupted image $\hat{x}^a = G_{\mathcal{I}_a}(c^a, a)$ and its corrected version $\hat{x} = G_{\mathcal{I}}(c^a)$. Similarly, $\hat{y}^a = G_{\mathcal{I}_a}(c, a)$ is the synthetically corrupted image from the clean input and $\hat{y} = G_{\mathcal{I}}(c)$ is the reconstructed clean image. To guarantee the expected outputs, the network is trained to minimise the following total loss function:

$$\mathcal{L}_{tot} = \lambda^{\mathcal{I}}_{adv}\mathcal{L}^{\mathcal{I}}_{adv} + \lambda^{\mathcal{I}_a}_{adv}\mathcal{L}^{\mathcal{I}_a}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{sr}\mathcal{L}_{sr} + \lambda_{art}\mathcal{L}_{art} \tag{7.1}$$

The first two terms are the traditional adversarial losses that promote a realistic generation of clean and corrupted images from $G_{\mathcal{I}}$ and $G_{\mathcal{I}_a}$:

$$\begin{aligned} \mathcal{L}^{\mathcal{I}}_{adv} &= \mathbb{E}_{\mathcal{I}}[\log D_{\mathcal{I}}(y)] + \mathbb{E}_{\mathcal{I}_a}[1 - \log D_{\mathcal{I}}(\hat{x})] \\ \mathcal{L}^{\mathcal{I}_a}_{adv} &= \mathbb{E}_{\mathcal{I}_a}[\log D_{\mathcal{I}_a}(x)] + \mathbb{E}_{\mathcal{I},\mathcal{I}_a}[1 - \log D_{\mathcal{I}_a}(\hat{y}^a)] \end{aligned} \tag{7.2}$$

The reconstruction loss guarantees that same-branch encoding-decoding correctly reconstructs the input, thus ensuring the preservation of patient's anatomy:

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathcal{I},\mathcal{I}_a}[||\hat{x}^a - x^a||_1 + ||\hat{y} - y||_1] \tag{7.3}$$

The self-reduction loss promotes cycle consistency within the cycle "clean - corrupted - clean":

$$\mathcal{L}_{sr} = \mathbb{E}_{\mathcal{I},\mathcal{I}_a}[||G_{\mathcal{I}}(E_{\mathcal{I}_a}(\hat{y}^a)) - y||_1] \tag{7.4}$$

Finally, the artefact consistency loss enforces that the artefact removed through the denoising path is the same added from the artefact-synthesis path, *de facto* training E_a to encode the artefact only:

$$\mathcal{L}_{art} = \mathbb{E}_{\mathcal{I},\mathcal{I}_a}[||(x^a - \hat{x}) - (\hat{y}^a - y)||_1] \tag{7.5}$$

Each loss term is weighted by the respective hyper-parameter λ.

7.1.1 Multimodal Artefact Disentanglement Network

In this work, the ADN is extended to a multimodal case, using two-channel inputs x^a and y, with CT image as first channel and respective registered MRI as second channel. The network learns to correct for the artefact on both modalities simultaneously, using multimodal information to encode the anatomical content of the images. To further enforce this sharing of information between the modalities, a loss term is introduced to maximise the similarity of the artefact-corrected images. This is motivated by the idea that two different images of the same object appear less similar if corrupted by noise or artefacts, especially when the artefacts present with different patterns in the two images. Conversely, the two images should look more similar if artefact-free. The aim is thus to improve the artefact reduction for both modalities by maximising the similarity between the output channels: firstly, the high-frequency and full-field-of-view nature of the artefact in the CT could be corrected through comparison with artefact-free MRI regions; secondly, the implant lack of signal in MRI could be compensated by the CT information, and better reconstruction should be achieved. Locally Normalised Cross Correlation (LNCC) was chosen as a measure of similarity, as it is suitable for multimodal comparison and it can be efficiently incorporated onto a neural network framework thanks to its convolution formulation [175]. The new similarity loss term is thus defined as:

$$\mathcal{L}_{sim} = 1 - \mathbb{E}_{\mathcal{I}_a}[|LNCC(\hat{x}_{CT}, \hat{x}_{MRI})|] \tag{7.6}$$

In addition, we also consider a self-synthesis consistency loss for the cycle "corrupted - clean - corrupted", that constitutes a full cycle loss together with the self-reduction loss:

$$\mathcal{L}_{cycle} = \mathcal{L}_{sr} + \mathbb{E}_{\mathcal{I}_a}[||G_{\mathcal{I}_a}(E_{\mathcal{I}}(\hat{x}), a) - x^a||_1] \tag{7.7}$$

Experimentally, this helps obtain sharper output images, especially with small training set size. The final total loss for training the MADN architecture is thus

$$\mathcal{L}_{tot} = \lambda^{\mathcal{I}}_{adv}\mathcal{L}^{\mathcal{I}}_{adv} + \lambda^{\mathcal{I}_a}_{adv}\mathcal{L}^{\mathcal{I}_a}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cycle}\mathcal{L}_{cycle} + \lambda_{art}\mathcal{L}_{art} + \lambda_{sim}\mathcal{L}_{sim} \quad (7.8)$$

7.2 Experimental setup

Dataset preparation . The dataset employed to train and test the MADN architecture included 65 3D CT-T1 MARS MRI pairs from subjects with metal hip implants, and 63 CT-T1 MRI clean pairs, with no metal artefacts.

The clean pairs were obtained from a dataset of whole-body CT and MRI acquisition protocol. The CTs were acquired on a Discovery 710 GE scanner (Tube voltage 140 kVp) with voxel size 1.37×1.37 ×3.27 mm^3. The MR images were acquired on a 3T Siemens Biograph mMR PET/MR scanner (TE=1.23ms, TR=4.02ms and flip angle=10°) with voxel size 0.67×0.67×5 mm^3. Pre-processing included bias field correction followed by fusion between stages using a percentile-based intensity harmonisation. Each CT-MRI pair was aligned with non-linear registration using a cubic b-spline free-form deformation algorithm [166]. An affine groupwise registration was then performed to align all the pairs into the same space, and all images were subsequently cropped at the pelvic region.

The corrupted pairs were selected from the dataset originally described in Chapter 3. With the exception of the 11 cases with manual ground truth segmentation, all the other cases were selected randomly to approximately match the number of 2D slices present in the clean dataset. Similarly to the clean dataset, intra-subject alignment was performed with non-linear registration using a cubic b-spline free-form deformation algorithm. They were then affinely remapped to the groupwise space generated from the clean dataset. Note that the CT images were initially corrected with the Refined Metal Artefact Re-

duction (RMAR) [134] for more accurate transformation estimation. However, the uncorrected images were utilised in all the experiments subsequently described.

11 pairs of corrupted CT-MRI were associated with manual segmentation of four muscles - Gluteus Maximus (GMAX), Gluteus Medius (GMED), Gluteus Minimus (GMIN) and Tensor Fasciae Latae (TFL). These subjects were thus left out from training and used as test set. As the architecture is trained on 2D slices, 12 randomly selected subjects from the clean CT-MRI pairs were also excluded from the training, to balance the number of training slices from both artefacted and artefact-free datasets. Overall, 54 CT-MRI pairs with artefacts (corresponding to 2159 2D slices) and 51 CT-MRI pairs without artefacts were used for training (corresponding to 2193 slices).

Evaluation strategy . Due to the lack of a ground truth, two different evaluation strategies were developed. To quantify the effect of the MAR on CT, I computed the standard deviation of the intensities (σ_{CT}) within the muscle regions. The presence of metal artefact induces noise even farther from the implant, causing fluctuations of the intensities from their true value, and therefore higher standard deviation. This analysis was performed before correction (No MAR) and after correction with: (RMAR CT) a conventional MAR algorithm [134]; (ADN CT) correction using an ADN model trained on CT only; (Multichannel ADN) correction using an ADN two-channel model trained on CT and MRI; (MADN) the proposed correction based on two-channel ADN model with LNCC similarity loss. For the MRI, a segmentation propagation experiment was performed: each test MRI was registered to all others using an intensity-based free-form deformation algorithm [166], their manual segmentation was propagated with the estimated transformation and compared with the manual ground truth using the Dice score. In addition to No MAR, Multichannel ADN and MADN, for this task I also trained an ADN model using MR only (ADN MR).

Implementation details . The architecture was developed in PyTorch, based on the original ADN implementation [149]. A detailed representation of the different components of the architecture is given in Fig. 7.2 (discriminator architecture), in Fig. 7.3 (elementary building blocks of the encoder-decoder paths, and in Fig. 7.4 (encoder and decoder diagrams).

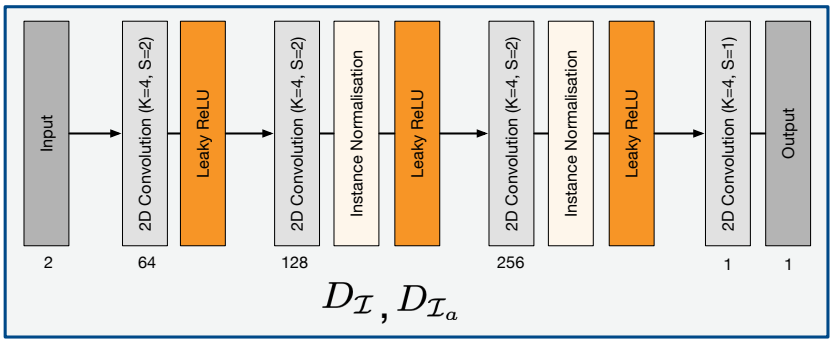


Figure 7.2: Both discriminators in the MADN have the same architecture, alternating 2D convolutional layers (kernel size KxK and stride S) with Instance Normalisation and Leaky ReLU.

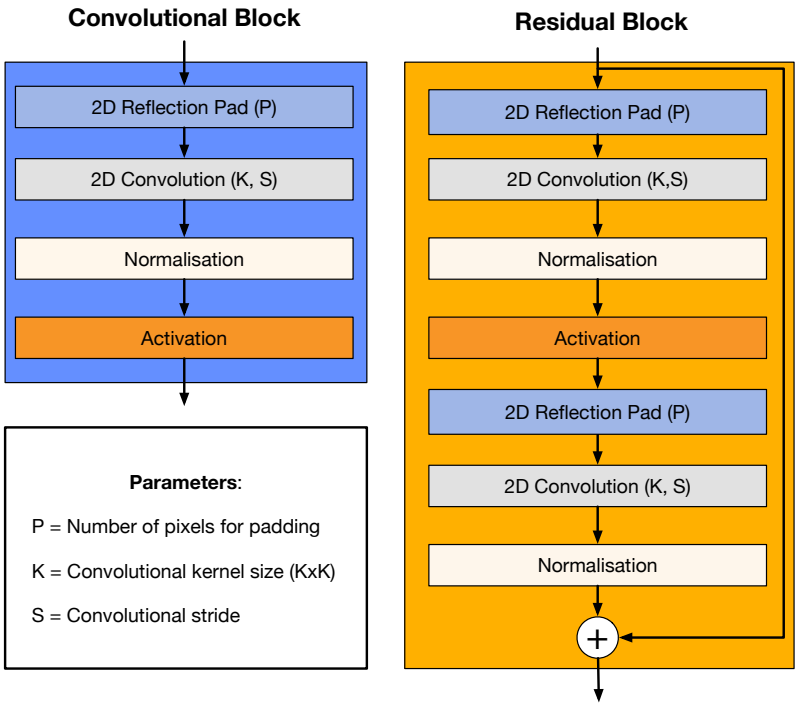


Figure 7.3: Convolutional and Residual building blocks used in the encoders and decoders of the MADN architecture.

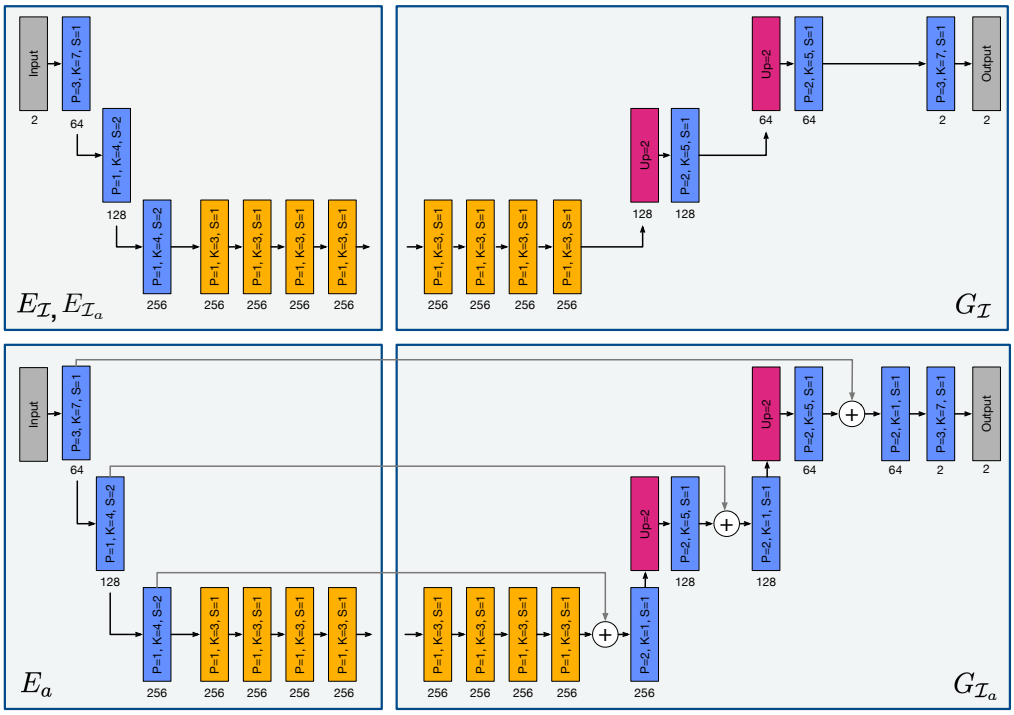


Figure 7.4: MADN encoders and decoders. Recalling Fig. 7.3, the blue blocks are convolutional blocks and the yellow ones are residual blocks. The number of feature channels is reported below each block. The upsampling layers (magenta) are nearest-neighbour upsampling operations by the factor indicated in the box. The three encoders $E_{\mathcal{I}}$, $E_{\mathcal{I}_a}$, E_a (blocks on the left) use Instance Normalisation and ReLU activation function in both convolutional and residual blocks. The two decoders $G_{\mathcal{I}}$, $G_{\mathcal{I}_a}$ have similar architecture, but the latter concatenates (+) artefact features from the artefact encoder E_a. In both decoders, Instance Normalisation is used in the residual blocks and Layer Normalisation in the convolutional blocks. Leaky ReLU is used throughout the decoders, but for the last convolution where the hyperbolic tangent is used instead.

All data-driven models were trained for 10 epochs on 2D slices with ADAM optimiser and learning rate $= 10^{-5}$. The loss weights were set to $\lambda^{\mathcal{I}}_{adv} = \lambda^{\mathcal{I}_a}_{adv} = 1.0$, $\lambda_{cycle} = \lambda_{rs} = \lambda_{rec} = \lambda_{art} = 20.0$. For the proposed MADN, we set $\lambda_{sim} = 1.0$, and LNCC estimated through a Gaussian kernel with $\sigma = 5$. A random flipping data augmentation strategy was employed during training.

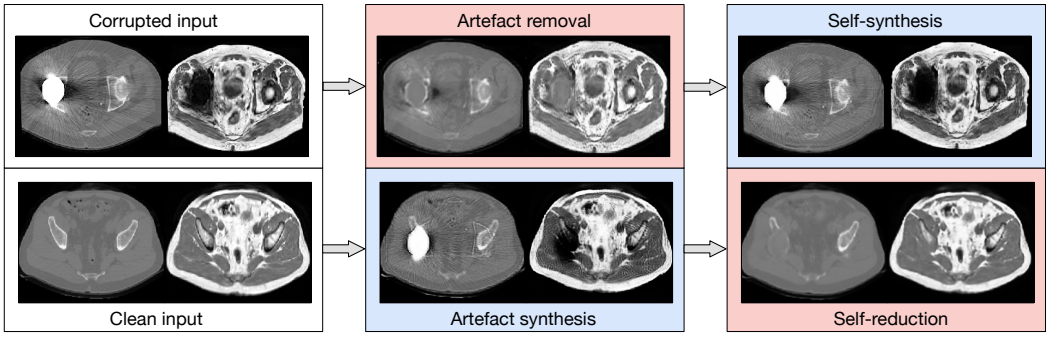


Figure 7.5: Example of MADN outcomes during training (9-th epoch). One pair of corrupted and one pair of clean images (from different subjects) are simultaneously input to the network. Top row: the corrupted pair undergoes artefact removal and subsequent artefact synthesis to produce the original input. Bottom row: the clean pair is corrupted with the same artefact removed from the corrupted pair (artefact synthesis) and then denoised again to produce the original input (self-reduction). In reference with Fig. 7.1, pink indicates artefact removal and light blue artefact synthesis.

7.3 Results

For the sake of illustrating the working principle of MADN, Fig. 7.5 shows the intermediate outputs of the architecture during training. In particular, it can be noticed how the artefact removed from the corrupted pair is transferred to the clean pair in the artefact synthesis. Vice versa, the artefact removal produces images that are more similar to the clean pair in texture. The self-synthesis and self-reduction steps further promote the disentanglement of artefact and anatomical content, by generating images that reproduce the respective inputs.

The results obtained after inference on the 11 unseen testing data are summarised in Figs. 7.6, 7.7, 7.8 and Table 7.1. Figure 7.6 presents a visual comparison of all the tested MAR methods on some exemplar test cases. On the CT, the proposed MADN approach is the most effective in reducing the streaks artefacts throughout the full field of view, as the correction is also driven by non-corrupted MRI corresponding areas. This reduction is also demonstrated by the decrease in σ_{CT} within the muscular tissue (Fig. 7.7), on either the implanted and non-implanted hip sides. The bilateral case (example

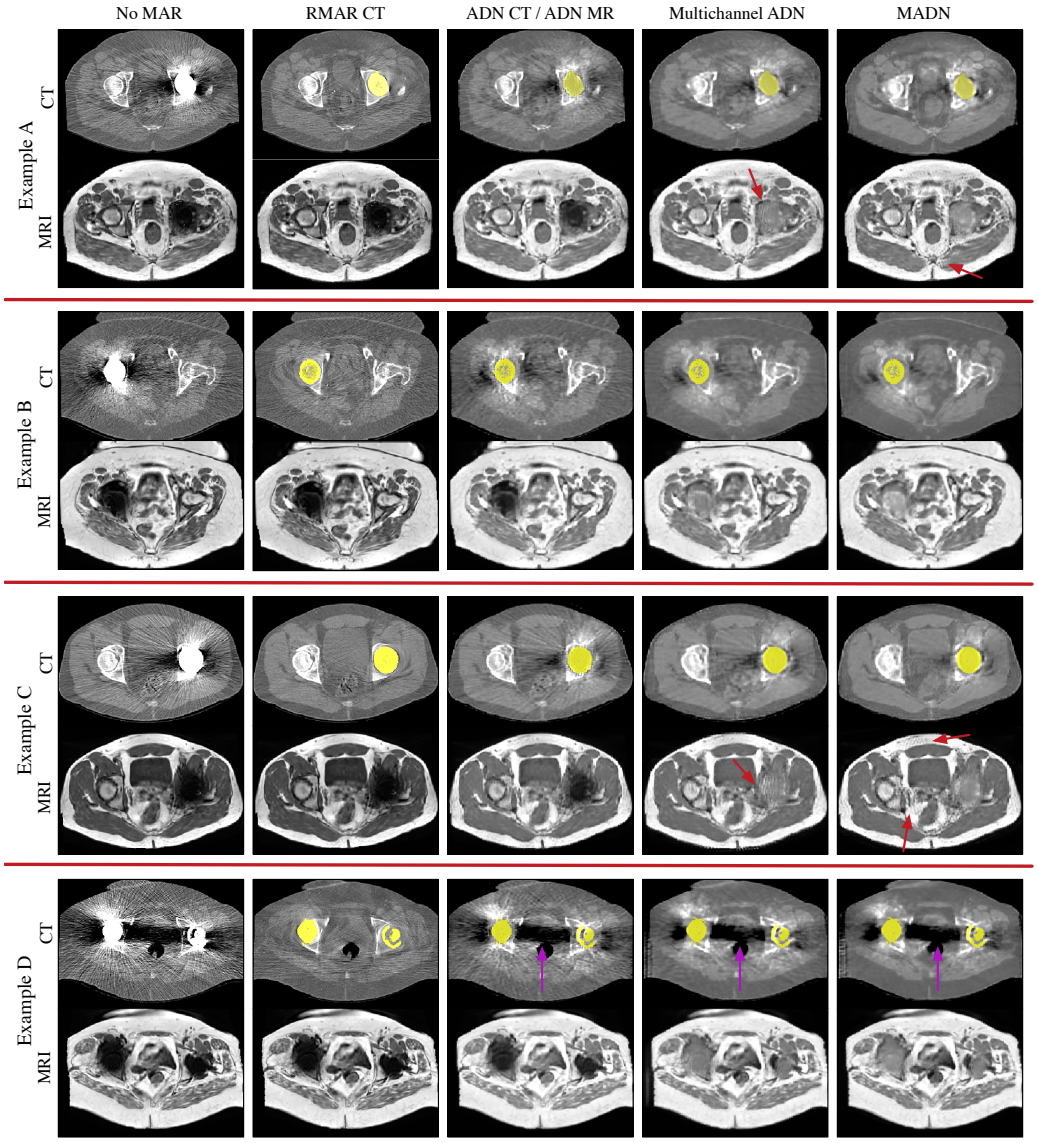
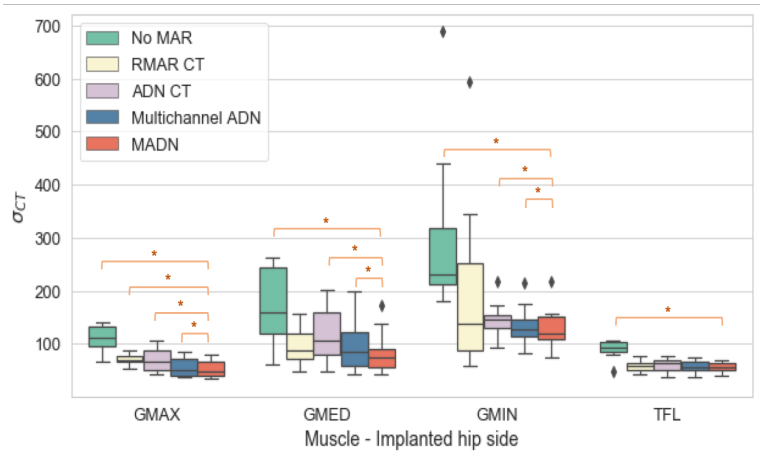


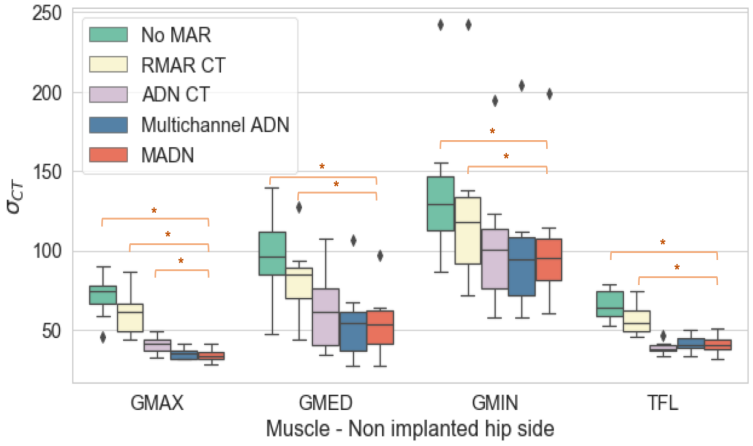
Figure 7.6: Visual comparison of MAR methods on CT and MRI. The implant is highlighted in yellow in all the corrected CT images. No overlay is applied to the MRI to better display the effect of the MAR. Red arrows indicate deconvolution checkerboard artefacts, while purple ones the residual photon starvation artefact which learning methods are unable to compensate for.

D in Fig. 7.6) shows however that all learning methods fail to compensate for the strong photon starvation effect and to recover the signal within the two implants. In this case, the RMAR is the most effective, although not fully able to eliminate all the streak artefacts.

On the MRI, Fig. 7.6 shows that training with MR images only (ADN



(a)



(b)

Figure 7.7: Standard deviation of CT intensity values within specific muscles. Cases significantly different from MADN are indicated by * (two-tailed paired t-test with $p < 0.05$). (a) Implanted hip side. (b) Non implanted hip side.

MR) is not sufficient to learn an embedding of the artefact and therefore correct for it. The multichannel ADN and MADN approaches instead identify the corrupted area correctly and attempt to regress the signal in it. However, the shape of the implant or the surrounding tissue is still not fully recovered. Furthermore, as indicated by the red arrows, the learning methods might introduce checkerboad artefacts which are typical of suboptimal upsampling strategies in the CNN [176]. The quantitative experiments reported in Table 7.1 and Fig. 7.8 show that MADN provides slightly better alignment for GMAX and TFL, but it performs worse on GMIN. It is worth noticing that the manual segmen-

	GMAX	GMED	GMIN	TFL
No MAR	0.76±0.15	**0.54**±0.17	**0.37**±0.18	0.41±0.27
ADN MR	0.76±0.18	0.53±0.16	0.34±0.19	0.44±0.26
Multichannel ADN	**0.77**±0.17	0.53±0.16	0.31±0.19	0.46±0.27
MADN	**0.77**±0.16	**0.54**±0.15	0.32±0.19	**0.50**±0.26

Table 7.1: Mean and standard deviation of Dice score for MR-to-MR inter-subject segmentation propagation task.

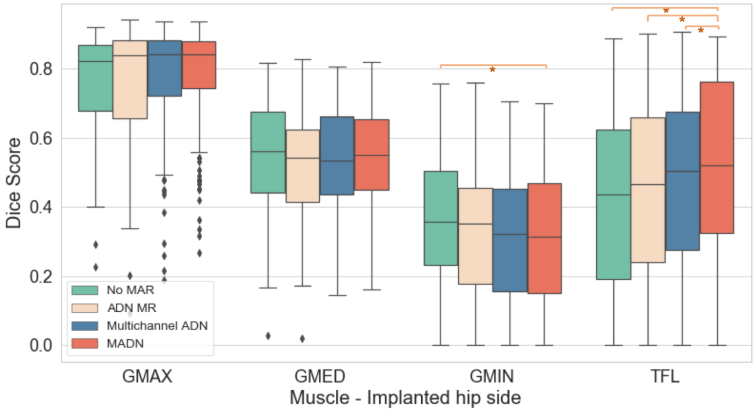


Figure 7.8: Dice score obtained from MR-to-MR inter-subject registration and segmentation propagation. Cases significantly different from MADN are indicated by *(two-tailed paired t-test with $p < 0.05$).

tations were performed on the non-corrected MRI, where GMIN is the most affected by the artefact. This makes it challenging to determine whether such result is due to less accurate registration or unreliable ground truth. Further analysis is thus needed to better quantify the MAR impact on MRI.

7.4 Discussion

The work presented in this Chapter is a preliminary study on the potential of a multimodal data-driven approach for metal artefact reduction. The idea of a multimodal approach was motivated by the different appearance of the artefact and thus by the hypothesis that one modality could compensate for the missing/corrupted information in the other. A data-driven unsupervised learning approach was chosen to compensate for the lack of paired corrupted-clean data. This problem is indeed characterised by the lack of a clear ground truth, due to the difficulty to obtain images with the implant but without the

associated artefacts.

The proposed approach showed some improvements over standard MAR and other unsupervised learning strategies on CT images with unilateral implant. Sharing information with the respective MRI reduced the presence of residual streak artefacts throughout the FOV, as also quantitatively shown by the intensity standard deviation analysis (Fig. 7.7). The MRI is however the modality that would benefit the most from an effective multimodal MAR technique, since existing methods are mostly applied at acquisition and thus limited by the physics of MRI. For this modality, the proposed MADN still proved suboptimal, as the reconstructed signal in the implant area is little informative compared to the uncorrected image. It is nonetheless encouraging that the network is able to correctly recognise the corrupted region and that this is clearly linked to the information sharing with CT. In fact, no effect was noted in the output after artefact correction when training the ADN only with MRI (column 3 in Fig. 7.6). This seems to suggest that the MADN architecture might have the capability to further improve the metal artefact reduction in MRI as well, but the model capacity was perhaps not adequately dimensioned to the task at hand. One of the main limitations of the proposed analysis is in fact the need for a more thorough architecture design. The ADN architecture was not modified with respect to the original network [149], keeping the same amount of parameters for either single- or multi-modality tests. Further tailoring of the architecture might also help identify a more suitable upsampling strategy, to eliminate the aforementioned checkerboard artefacts. In addition, a more thorough hyperparameter search needs to be conducted, ideally based on the identification of appropriate performance metrics evaluated on a validation set. Finally, different multimodal similarity measures could also be implemented in place of LNCC (e.g. Cross Correlation or Normalised Mutual Information) to test their impact on the reconstruction.

The generalisability of the proposed approach requires more extensive validation, as the test set is currently limited in size due to lack of ground truth

segmentation for quantitative analysis. The generalisability assessment also includes a substantial analysis of the impact of CT-MRI registration errors. Empirically, it was observed that, in the presence of large misalignments, an excessive weight on the similarity measure loss term might induce geometrical deformations in one of the two modalities, which compensate for the misalignment. For this reason, cases with largely inaccurate registration had precautionarily been removed from the training set, as the primary objective of the study was solely the MAR. It does however require further investigation and potentially opens up further research questions on the possibility of a data-driven approach for simultaneous registration and metal artefact reduction. This aspect is also relevant for the potential integration of this MAR technique on the automated segmentation pipeline presented in Chapter 5. The MAR requires accurate intra-subject registration; vice versa the intensity-based intra-subject registration would be improved by higher-quality data. To compensate for this cyclic relationship, an iterative approach could be implemented in the pipeline, alternating the MAR and the intra-subject registration towards a refinement of both.

In conclusion, the qualitative examples and quantitative results on the CT suggest that the use of the multimodal approach for MAR could be beneficial, as it combines different information to learn a better embedding of the anatomy and of the artefact. Future work will investigate further improvements on the MRI correction and therefore better assess whether the benefits of a multimodal approach could overweight the need for multiple modalities.

Chapter 8

Conclusions

8.1 Summary

This thesis has presented the development of automated medical image analysis tools to improve the clinical workflow in orthopaedics, and specifically in the context of hip arthroplasty. It proposed solutions to address the need for combining the complementary skeletal and muscular information provided by CT and MRI, towards patient's specific anatomical modelling and muscle atrophy quantification.

To achieve the desired goals, a novel intra-subject registration algorithm was presented in Chapter 4. The proposed method guarantees the preservation of bone rigidity while allowing for local soft tissue deformations. Thus, it is able to compensate for patient's pose variations in CT and MRI, while still applying anatomically plausible deformations. This is achieved by incorporating hard rigidity constraints into a stationary velocity field transformation model, yielding also a desirable diffeomorphic behaviour. The algorithm was tested on the intra-subject pelvic CT-MRI registration task, showing a significant reduction of registration error compared to a fully non-linear transformation. It also favoured robustness against the choice of the registration parameters and against inaccuracies of the rigid region segmentation masks.

The registration algorithm was embedded into a composite pipeline for the joint automated segmentation of CT and MRI in the presence of hip implants.

The framework, introduced in Chapter 5, was built on three main blocks: (1) image quality enhancement, to improve the quality of routinely acquired clinical data; (2) CT-MR intra-subject registration, to provide robust alignment of anatomically corresponding structures while preserving the biological plausibility of the deformations; (3) multi-atlas based segmentation, to automatically derive musculoskeletal and implant segmentation while taking advantage of population variability information. Although tested using only a small sample, the proposed approach performed comparably to automated segmentation on a single modality, with the advantage of unifying muscular, skeletal and implant segmentation. The obtained segmentation resulted in a median Dice score of 0.90 for skeletal and 0.84 for muscular structures. This segmentation accuracy is in line with state-of-the-art approaches [94, 112, 113, 171], which mostly dealt with non-implanted hips and thus were not impacted by metal artefacts.

The availability of an automated segmentation scheme reduces the burden of manually tracing regions of interest. This favours both the visualisation of patient-specific anatomy through 3D rendering of the segmentation, as well as the extraction of quantitative biomarkers from the ROIs. This latter aspect was investigated in Chapter 6, with particular focus on the quantification of muscle atrophy. To this end, I proposed an imaging biomarker that calculates the percentage of fat in the muscular volume, called Intramuscular Fat Fraction. This biomarker takes advantage of accurate ROI definition based on the multimodal segmentation, and exploits the quantitative nature of CT intensities, which makes it reproducible and easy to compute. For GMED, this biomarker showed strong association with the radiological scoring system currently used in the clinical assessment of muscle atrophy. It was also found to correlate with blood metal ion levels, another indicator of implant failure [11]. However, the small sample size of the available cohort currently limits the strength of any conclusion from this analysis, thus further investigation on a larger scale should be conducted to assess its actual clinical relevance.

Although tailored to the hip arthroplasty application, the discussed registration, segmentation and biomarker extraction approaches are still partially affected by the presence of metal artefacts. Residual streak artefacts in the CT may bias the fat/lean muscle segmentation and thus the derived IFF. The distortions and lack of signal caused by the susceptibility artefacts in MRI mostly impact the registration and the segmentation steps. A preliminary attempt to address this limitation was presented in Chapter 7. A novel multimodal metal artefact reduction was proposed, which uses an unsupervised data-driven strategy to learn a latent representation of the artefact and disentangle it from the anatomical content. This approach proved effective on the CT, and showed its ability to exploit the multimodal information. However, it requires further development to adequately correct the MR image. If proven beneficial for both modalities, this technique could be integrated in the automated segmentation framework and potentially lead to more accurate segmentation and biomarker extraction.

All the contributions presented in this thesis are currently limited by too little generalisability analysis, due to lack of data with associated ground truth. A more thorough validation should therefore be performed to more deeply evaluate the potential clinical benefits of a multimodal analysis against the associated imaging costs. Nonetheless, these tools showed potential value towards the automation of image analysis. They could thus support and inform the decision-making process about patient's management, improving the clinical services in MSK health in accordance with the indications of the World Health Organization [1].

8.2 Future developments

Motivated by the current limitations of this work, two main directions could be explored in future work: (1) the adoption and further development of deep learning approaches to build upon the proposed image analysis pipeline, (2) the assessment of further clinical applications of this pipeline in biomarker

extractions and surgical planning/outcome measurement.

8.2.1 Deep learning in MSK

Deep learning is gaining popularity in the scientific community due to its outstanding performance in classification, segmentation and regression tasks. At present, its use in MSK is limited to monomodal imaging applications, mostly in the detection and segmentation of vertebrae or knee cartilage segmentation [53]. It can be anticipated that novel deep learning segmentation methods will outperform atlas-based approaches in MSK as well, both in accuracy and speed. However, the success of most deep learning algorithms still depends on the availability of large annotated datasets that can be used for training. The proposed multi-atlas based segmentation pipeline can thus play an important part in helping the construction of such datasets.

Further methodological developments could also arise from the presented work. In particular, I envisage two potentially relevant advancements that take advantage of data-driven modelling.

Integrating image registration and MAR. Deep learning approaches have the capacity of learning shared latent representations of the anatomical structure under study from multiple imaging modalities [177]. The preliminary work presented in Chapter 7 demonstrated this ability towards a multimodal approach for metal artefact reduction and paves the way for further research in this direction. In particular, a strong interconnection between multimodal MAR and intra-subject registration was observed when mis-alignment between the input CT and MRI was present: the network would spatially deform one of the two to maximise the similarity at the output. It would thus be interesting to explore a multi-task learning approach that attempts at simultaneously registering and correcting for the artefact. Deep learning approaches are becoming more popular in medical image registration as well [53, 178], and the field of similarity learning [179, 180] could represent an interesting direction of investigation in a multi-task setting. Currently, the proposed algorithm is using NMI to drive the registration, being the standard similarity measure for

multimodal registration. A deep similarity learning approach could provide an alternative solution which is specific to the task and could be combined with the disentanglement of artefact and anatomy. Similarity learning is also advantageous as it can be integrated with traditional transformation models. Combined with the proposed rigidly-constrained algorithm, it would promote diffeomorphic and anatomically plausible transformations.

CT synthesis from MRI. Being associated to radiation exposure to the gonads, reducing the need for CT imaging while still providing the same amount of anatomical information is of paramount clinical relevance. The proposed framework offers a natural environment for the development of image synthesis methods, by which the CT image could be synthetically computed from the respective MRI without being acquired. Image synthesis is a flourishing field especially in radiotherapy treatment planning [181], and also in this case deep learning is establishing itself as the state-of-the-art approach [182, 183]. In hip arthroplasty, the main challenge to address is the presence of the implant and the respective artefacts in the image. The success of learning algorithms for this task would thus rely on the ability to achieve sufficiently accurate CT-MRI alignment in the training set and to train the network to compensate for the MRI missing signal using the CT information.

8.2.2 Clinical applications

The proposed automated segmentation framework has the unexplored potential of supporting surgical planning, as it provides patient's specific anatomical rendering. Although developed for post-surgical assessments, the pipeline can directly be applied on pre-operative data and thus provide relevant information in both primary and revision surgery. Additionally, it could be employed in longitudinal analyses, e.g. in pre-operative and post-operative comparisons for surgical outcome assessment, or for post-operative patient monitoring.

The imaging biomarkers discussed in this work represent only an initial step towards quantitative image analysis. Together with a more extended clinical validation of the proposed intramuscular fat fraction, other measures

could also be investigated. Particularly relevant could be the possibility of detecting muscle-to-bone attachments to verify their integrity. A probabilistic atlas-based estimation from CT only has been recently proposed by Fukuda *et al.* [184] and could represent a starting point for similar analysis from our framework. Another potential research direction is the use of these biomarkers to study the association between implant position and muscle atrophy. The work from Hart *et al.* [13] already showed how CT-derived measures of MoM prosthesis components are linked to an increase of blood metal ion levels. Whilst many factors contribute to the development of muscle atrophy and should therefore be carefully taken into account, analysing implant position and muscle damage could help characterise different surgical approaches and identify healthy-tissue sparing methods.

List of Publications

Ranzini MBM, Ebner M, Cardoso MJ, Fotiadou A, Vercauteren T, Henckel J, Hart A, Ourselin S, Modat M. *Joint multimodal segmentation of clinical CT and MR from hip arthroplasty patients.* In: Glocker B, Yao J, Vrtovec T, Frangi A, Zheng G. (eds) Computational Methods and Clinical Applications in Musculoskeletal Imaging. MSKI 2017. Lecture Notes in Computer Science, vol 10734. Springer, Cham. https://doi.org/10.1007/978-3-319-74113-0_7

Kläser K, Markiewicz P, **Ranzini MBM**, Li W, Modat M, Hutton BF, Atkinson D, Thielemans K, Cardoso MJ, Ourselin S. *Deep boosted regression for MR to CT synthesis.* In: Gooya A, Goksel O, Oguz I, Burgos N. (eds) Simulation and Synthesis in Medical Imaging. SASHIMI 2018. Lecture Notes in Computer Science, vol 11037. Springer, Cham. https://doi.org/10.1007/978-3-030-00536-8_7

Drobny D, **Ranzini MBM**, Ourselin S, Vercauteren T, Modat M. *Landmark-based evaluation of a block-matching registration framework on the RESECT pre- and intra-operative brain image data set.* In: Zhou L et al. (eds) Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention. LABELS 2019, HAL-MICCAI 2019, CuRIOUS 2019. Lecture Notes in Computer Science, vol 11851. Springer, Cham. https://doi.org/10.1007/978-3-030-33642-4_15

Molteni E, **Ranzini MBM**, Beretta E, Stefanoni G, Strazzer S, Modat M. *Individual growth curve study of functional recovery up to the 7th year after acquired brain injury in childhood.* Abstract accepted for oral presentation at Frontiers in Traumatic Brain Injury 2019.

Ranzini MBM, Henckel J, Ebner M, Cardoso MJ, Isaac A, Vercauteren T, Ourselin S, Hart A, Modat M. *Automated postoperative muscle assessment of hip arthroplasty patients using multimodal imaging joint segmentation.* Computer Methods and Programs in Biomedicine. 2020; 183:105062. https://doi.org/10.1016/j.cmpb.2019.105062.

Ranzini MBM, Groothuis I, Kläser K, Cardoso MJ, Henckel J, Ourselin S, Hart A, Modat M. *Combining multimodal information for Metal Artefact Reduction: An unsupervised deep learning framework.* In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI); 2020. p. 600-604. https://doi.org/10.1109/ISBI45749.2020.9098633.

Drobny D, **Ranzini MBM**, Isaac A, Vercauteren T, Ourselin S, Modat M. *Towards automated spine mobility quantification: A locally rigid CT to X-ray registration framework.* In: Špiclin Ž, McClelland J, Kybic J, Goksel O. (eds) Biomedical Image Registration. WBIR 2020. Lecture Notes in Computer Science, vol 12120. Springer, Cham. https://doi.org/10.1007/978-3-030-50120-4_7

Bibliography

[1] Woolf A, March L, Officer A, Cross MJ, Briggs AM, Hoy D, et al. WHO World Report on Ageing and Health Background Paper on Musculoskeletal Health and the Impact of Musculoskeletal Disorders in the Elderly. [online]. 2015;Available from: http://bjdonline.org/wp-content/uploads/2016/08/MSK-Health-and-Ageing_Report-prepared-for-the-WHO-World-Report-on-Ageing-and-Health-10-July-2015.pdf.

[2] Parsons S, Symmons DPM. The burden of musculoskeletal conditions. Medicine (United Kingdom). 2014;42(4):190–192. Available from: http://dx.doi.org/10.1016/j.mpmed.2014.01.009.

[3] Royal College of Surgeons. Surgery and the NHS in numbers. [online]. 2014;Available from: https://goo.gl/csnf94.

[4] 15th Annual Report 2018. National Joint Registry for England, Wales, Northern Ireland and the Isle of Man; 2018. December 2017. Available from: http://www.njrreports.org.uk/Portals/0/PDFdownloads/NJR15thAnnualReport2018.pdf.

[5] Su EP, Su SL. Adoption of Hip Resurfacing Arthroplasty at Hospital for Special Surgery: A Cohort Study. HSS Journal. 2012;8(3):283–286.

[6] Arthritis Research UK. State of Musculoskeletal Health 2017. [online]. 2017;p. 30. Available from: file:///Users/mariellabaxter/Downloads/State-of-musculoskeletal-health-2017.PDF.

[7] Toms AP, Marshall TJ, Cahir J, Darrah C, Nolan J, Donell ST, et al. MRI of early symptomatic metal-on-metal total hip arthroplasty: a retrospective review of radiological findings in 20 hips. Clinical Radiology. 2008;63(1):49–58.

[8] Berber R, Khoo M, Cook E, Guppy A, Hua J, Miles J, et al. Muscle atrophy and metal-on-metal hip implants: a serial MRI study of 74 hips. Acta orthopaedica. 2015;86(3):351–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4443469{&}tool=pmcentrez{&}rendertype=abstract.

[9] Pandit H, Whitwell D, Gibbons CLM, Athanasou N, Gill HS, Murray DW. Pseudotumours associated with metal-on- metal hip resurfacings. The Journal of Bone and Joint Surgery. 2008;90(7):847–851.

[10] Hart AJ, Sabah SA, Henckel J, Lloyd G, Skinner JA. Lessons learnt from metal-on-metal hip arthroplasties will lead to safer innovation for all medical devices. HIP International. 2015;25(4):347–354.

[11] MHRA of the Department of Health. Medical Device Alert. MHRA Database. 2017;p. 1–6. Available from: https://www.gov.uk/drug-device-alerts/all-metal-on-metal-mom-hip-replacements-updated-advice-for-follow-up-of-patients.

[12] Bogner EA, Sofka CM. CT evaluation of total hip arthroplasty complication: Dissociation of acetabular component. HSS Journal. 2007;3(1):112–114.

[13] Hart AJ, Skinner JA, Henckel J, Sampson B, Gordon F. Insufficient acetabular version increases blood metal ion levels after metal-on-metal hip resurfacing. Clinical Orthopaedics and Related Research. 2011;469(9):2590–2597.

[14] Robinson E, Henckel J, Sabah S, Satchithananda K, Skinner J, Hart A. Cross-sectional imaging of metal-on-metal hip arthroplasties.

Acta Orthopaedica. 2014;85(6):577–584. Available from: http://www.tandfonline.com/doi/full/10.3109/17453674.2014.964618.

[15] Smith-Bindman R, Lipson J, Marcus R, Kim K, Mahesh M, Gould R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Archives of internal medicine. 2009;169(22):2078–86. Available from: http://archinte.jamanetwork.com/article.aspx?articleid=415384.

[16] Siddiqui IA, Sabah SA, Satchithananda K, Lim AK, Cro S, Henckel J, et al. A comparison of the diagnostic accuracy of MARS MRI and ultrasound of the painful metal-on-metal hip arthroplasty. Acta orthopaedica. 2014;85(4):1–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24694273.

[17] Olsen RV, Munk PL, Lee MJ. Metal Artifact Reduction Sequence: Early Clinical Applications. Radiographics : a review publication of the Radiological Society of North America, Inc. 2000;20(3):699–712.

[18] Delso G, Wiesinger F, Sacolick LI, Kaushik SS, Shanbhag DD, Hullner M, et al. Clinical Evaluation of Zero-Echo-Time MR Imaging for the Segmentation of the Skull. Journal of Nuclear Medicine. 2015;56(3):417–422. Available from: http://jnm.snmjournals.org/cgi/doi/10.2967/jnumed.114.149997.

[19] Matharu GS, Mansour R, Dada O, Ostlere S, Pandit HG, Murray DW. Which imaging modality is most effective for identifying pseudotumours in metal-on-metal hip resurfacings requiring revision. Bone and Joint Journal. 2016;98B(1):40–48.

[20] Pedoia V, Majumdar S, Link TM. Segmentation of joint and musculoskeletal tissue in the study of arthritis. Magnetic Resonance Materials

in Physics, Biology and Medicine. 2016;29(2):207–221. Available from: http://dx.doi.org/10.1007/s10334-016-0532-9.

[21] Blemker SS, Asakawa DS, Gold GE, Delp SL. Image-based musculoskeletal modeling: Applications, advances, and future opportunities. Journal of Magnetic Resonance Imaging. 2007;25(2):441–451. Available from: http://dx.doi.org/10.1002/jmri.20805.

[22] Abdoli M, Mehranian A, Ailianou A, Becker M, Zaidi H. Assessment of metal artifact reduction methods in pelvic CT. Medical Physics. 2016;43(4):1588–1597. Available from: http://scitation.aip.org/content/aapm/journal/medphys/43/4/10.1118/1.4942810.

[23] Hargreaves BA, Worters PW, Pauly KB, Pauly JM, Koch KM, Gold GE. Metal-induced artifacts in MRI. American Journal of Roentgenology. 2011;197(3):547–555.

[24] Pfirrmann CWa, Notzli HP, Dora C, H J, Zanetti M. Abductor tendons and muscles assessed at MR imaging after total hip arthroplasty in asymptomatic and symptomatic patients. Radiology. 2005;235(3):969–976.

[25] Smith JJ, Sorensen AG, Thrall JH. Biomarkers in Imaging: Realizing Radiology's Future. Radiology. 2003;227(3):633–638. PMID: 12663828.

[26] Sabah SA, Mitchell AWM, Henckel J, Sandison A, Skinner JA, Hart AJ. Magnetic Resonance Imaging Findings in Painful Metal-On-Metal Hips. A Prospective Study. Journal of Arthroplasty. 2011;26(1):71–76.e2. Available from: http://dx.doi.org/10.1016/j.arth.2009.11.008.

[27] Bal BS, Lowe JA, Arthroplasty H, Evidence MRI, Hillard AE. Muscle damage in minimally invasive total hip arthroplasty: MRI evidence that it is not significant. Instructional course lectures. 2008;57:223–229.

[28] Vidt ME, Santago AC, Tuohy CJ, Poehling GG, Freehill MT, Kraft RA, et al. Assessments of fatty infiltration and muscle atrophy from a single magnetic resonance image slice are not predictive of 3-dimensional measurements. Arthroscopy - Journal of Arthroscopic and Related Surgery. 2016;32(1):128–139. Available from: http://dx.doi.org/10.1016/j.arthro.2015.06.035.

[29] Karlsson A, Rosander J, Romu T, Tallberg J, Grönqvist A, Borga M, et al. Automatic and quantitative assessment of regional muscle volume by multi-atlas segmentation using whole-body water-fat MRI. Journal of Magnetic Resonance Imaging. 2015;41(6):1558–1569.

[30] Skorupska E, Keczmer P, lochowski RM, Tomal P, Rychlik M, Samborski W. Reliability of MR-Based volumetric 3-D Analysis of pelvic muscles among subjects with low back with leg pain and healthy volunteers. PLoS ONE. 2016;11(7):1–15.

[31] Grimaldi A, Richardson C, Durbridge G, Donnelly W, Darnell R, Hides J. The association between degenerative hip joint pathology and size of the gluteus maximus and tensor fascia lata muscles. Manual Therapy. 2009;14(6):611–617. Available from: http://dx.doi.org/10.1016/j.math.2008.11.002.

[32] Grimaldi A, Richardson C, Stanton W, Durbridge G, Donnelly W, Hides J. The association between degenerative hip joint pathology and size of the gluteus medius, gluteus minimus and piriformis muscles. Manual Therapy. 2009;14(6):605–610. Available from: http://dx.doi.org/10.1016/j.math.2009.07.004.

[33] Le Troter A, Fouré A, Guye M, Confort-Gouny S, Mattei JP, Gondin J, et al. Volume measurements of individual muscles in human quadriceps femoris using atlas-based segmentation approaches. Magnetic Resonance Materials in Physics, Biology and Medicine. 2016;29(2):245–257.

[34] Nordez A, Jolivet E, Südhoff I, Bonneau D, De Guise JA, Skalli W. Comparison of methods to assess quadriceps muscle volume using magnetic resonance imaging. Journal of Magnetic Resonance Imaging. 2009;30(5):1116–1123.

[35] Engelken F, Wassilew GI, Köhlitz T, Brockhaus S, Hamm B, Perka C, et al. Assessment of fatty degeneration of the gluteal muscles in patients with THA using MRI: Reliability and accuracy of the goutallier and quartile classification systems. Journal of Arthroplasty. 2014;29(1):149–153. Available from: http://dx.doi.org/10.1016/j.arth.2013.04.045.

[36] Crawford RJ, Cornwall J, Abbott R, Elliott JM. Manually defining regions of interest when quantifying paravertebral muscles fatty infiltration from axial magnetic resonance imaging: a proposed method for the lumbar spine with anatomical cross-reference. BMC Musculoskeletal Disorders. 2017;18(1):1–11. Available from: http://dx.doi.org/10.1186/s12891-016-1378-z.

[37] Gadermayr M, Disch C, Müller M, Merhof D, Gess B. A Comprehensive Study on Automated Muscle Segmentation for Assessing Fat Infiltration in Neuromuscular Diseases. Magnetic resonance imaging. 2017;48(August 2017):20–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29269318.

[38] Karampinos DC, Baum T, Nardo L, Alizai H, Yu H, Carballido-Gamio J, et al. Characterization of the regional distribution of skeletal muscle adipose tissue in type 2 diabetes using chemical shift-based water/fat separation. Journal of Magnetic Resonance Imaging. 2012;35(4):899–907.

[39] Morrow JM, Sinclair CDJ, Fischmann A, Machado PM, Reilly MM, Yousry TA, et al. MRI biomarker assessment of neuromuscular dis-

ease progression: a prospective observational cohort study. The Lancet Neurology. 2016;15(1):65–77. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1474442215002422.

[40] Ugarte V, Sinha U, Malis V, Csapo R, Sinha S. 3D multimodal spatial fuzzy segmentation of intramuscular connective and adipose tissue from ultrashort TE MR images of calf muscle. Magnetic Resonance in Medicine. 2017;77(2):870–883.

[41] Lareau-Trudel E, Troter AL, Ghattas B, Pouget J, Attarian S, Bendahan D, et al. Muscle quantitative MR imaging and clustering analysis in patients with facioscapulohumeral muscular dystrophy type 1. PLoS ONE. 2015;10(7):1–16.

[42] Momose T, Inaba Y, Choe H, Kobayashi N, Tezuka T, Saito T. CT-based analysis of muscle volume and degeneration of gluteus medius in patients with unilateral hip osteoarthritis. BMC Musculoskeletal Disorders. 2017;18(1):1–7.

[43] Pham DL, Xu C, Prince JL. Current Methods in Medical Image Segmentation. Annual Review of Biomedical Engineering. 2000;2(1):315–337. Available from: https://doi.org/10.1146/annurev.bioeng.2.1.315.

[44] Van Leemput K, Maes F, Vandermeulen D, Suetens P. Automated model-based bias field correction of MR images of the brain. IEEE transactions on medical imaging. 1999;18(10):885–896.

[45] Cootes TF, Taylor CJ, Cooper DH, Graham J. Active Shape Models-Their Training and Application. Computer Vision and Image Understanding. 1995;61(1):38–59. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1077314285710041.

[46] Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2001;23(6):681–685.

[47] Cabezas M, Oliver A, LladÃş X, Freixenet J, Cuadra MB. A review of atlas-based segmentation for magnetic resonance brain images. Computer Methods and Programs in Biomedicine. 2011;104(3):e158 – e177. Available from: http://www.sciencedirect.com/science/article/pii/S0169260711002033.

[48] Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging. 2004;23(7):903–921.

[49] Cardoso JM, Leung K, Modat M, Keihaninejad S, Cash D, Barnes J, et al. STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation. Medical Image Analysis. 2013;17(6):671–684. Available from: http://dx.doi.org/10.1016/j.media.2013.02.006.

[50] Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models. International Journal of Computer Vision. 1988 Jan;1(4):321–331. Available from: https://doi.org/10.1007/BF00133570.

[51] Sethian JA. Fast Marching Methods and Level Set Methods for Propagating Interfaces; 1998. von Karman Institute Lecture Series.

[52] Boykov Y, Funka-Lea G. Graph cuts and efficient N-D image segmentation. International Journal of Computer Vision. 2006;70(2):109–131.

[53] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017;42(December 2012):60–88.

[54] Eaton-Rosen Z, Bragman FJS, Ourselin S, Cardoso M. Improving Data Augmentation for Medical Image Segmentation. In: Medical Image with Deep Learning MIDL 2018; 2018. Available from: https://openreview.net/forum?id=rkBBChjiG.

[55] Zeng G, Yang X, Li J, Yu L, Heng Pa. 3D U-net with Multi-level Deep Supervision: Fully Automatic Segmentation of Proximal Femur in 3D MR Images. In: MICCAI workshop on Machine Learning in Medical Imaging (MLMI 2017). vol. 10541; 2017. p. 274–282.

[56] Yang L, Zhang Y, Chen J, Zhang S, Chen DZ. Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, editors. Medical Image Computing and Computer Assisted Intervention ÃćÂĹÂŠ MICCAI 2017. Springer International Publishing; 2017. p. 399–407.

[57] Bai W, Oktay O, Sinclair M, Suzuki H, Rajchl M, Tarroni G, et al. Semi-supervised Learning for Network-Based Cardiac MR Image Segmentation. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, editors. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017. Springer International Publishing; 2017. p. 253–260.

[58] Arezoomand S, Lee WS, Rakhra KS, Beaulé PE. A 3D active model framework for segmentation of proximal femur in MR images. International Journal of Computer Assisted Radiology and Surgery. 2015;10(1):55–66.

[59] Buzug T. Computed tomography: From photon statistics to modern cone-beam CT. Springer Science & Business Media; 2008.

[60] Schreiber JJ, Anderson PA, Rosas HG, Buchholz AL, Au AG. Hounsfield units for assessing bone mineral density and strength: a tool for osteoporosis management. The Journal of bone and joint surgery American volume. 2011;93(11):1057–63.

[61] Whitmarsh T, Treece GM, Poole KES. Automatic Segmentation and Discrimination of Connected Joint Bones from CT by Multi-atlas Regis-

tration. In: Yao J, Klinder T, Li S, editors. Computational Methods and Clinical Applications for Spine Imaging: Proceedings of the Workshop held at the 16th International Conference on Medical Image Computing and Computer Assisted Intervention, September 22-26, 2013, Nagoya, Japan. Cham: Springer International Publishing; 2014. p. 199–207. Available from: https://doi.org/10.1007/978-3-319-07269-2_17.

[62] Chu C, Chen C, Liu L, Zheng G. FACTS: Fully Automatic CT Segmentation of a Hip Joint. Annals of Biomedical Engineering. 2015;43(5):1247–1259.

[63] Pauchard Y, Fitze T, Browarnik D, Eskandari A, Pauchard I, Enns-Bray W, et al. Interactive graph-cut segmentation for fast creation of finite element models from clinical ct data for hip fracture prediction. Computer Methods in Biomechanics and Biomedical Engineering. 2016;19(16):1693–1703. Available from: http://dx.doi.org/10.1080/10255842.2016.1181173.

[64] Cheng Y, Zhou S, Wang Y, Guo C, Bai J, Tamura S. Automatic segmentation technique for acetabulum and femoral head in CT images. Pattern Recognition. 2013;46(11):2969–2984. Available from: http://dx.doi.org/10.1016/j.patcog.2013.04.006.

[65] Guo H, Song S, Wang J, Guo M, Cheng Y, Wang Y, et al. 3D surface voxel tracing corrector for accurate bone segmentation. International Journal of Computer Assisted Radiology and Surgery. 2018;13(10):1549–1563. Available from: https://doi.org/10.1007/s11548-018-1804-9.

[66] Seim H, Kainmüeller D, Heller M. Automatic Segmentation of the Pelvic Bones from CT Data Based on a Statistical Shape Model. Eurographics Workshop on Visual Computing for Biomedicine. 2008;p. 93–100. Available from: https://diglib.eg.org/EG/DL/WS/VCBM/

VCBM08/093-100.pdf.abstract.pdf;internal{&}action=action.digitallibrary.ShowPaperAbstract$\delimiter"026E30F$nhttp://www.zib.de/zachow/publications/vcbm2008.pdf.

[67] Kainmueller D, Lamecker H, Zachow S, Hege HC. An articulated statistical shape model for accurate hip joint segmentation. Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009. 2009;p. 6345–6351.

[68] Yokota F, Okada T, Takao M, Sugano N, Tada Y, Tomiyama N, et al. Automated CT segmentation of diseased hip using hierarchical and conditional statistical shape models. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2013;8150 LNCS(PART 2):190–197.

[69] Pettersson J, Knutsson H, Borga M. Automatic hip bone segmentation using non-rigid registration. Proceedings - International Conference on Pattern Recognition. 2006;3:946–949.

[70] Besler BA, Michalski AS, Forkert ND, Boyd SK. Automatic Full Femur Segmentation from Computed Tomography Datasets Using an Atlas-Based Approach. In: Glocker B, Yao J, Vrtovec T, Frangi AF, Zheng G, editors. MICCAI workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging (MSKI 2017); 2017. p. 116–27.

[71] Krčah M, Székely G, Blanc R. Fully automatic and fast segmentation of the femur bone from 3D-CT images with no shape prior. Proceedings - International Symposium on Biomedical Imaging. 2011;p. 2087–2090.

[72] Chu C, Bai J, Wu X, Zheng G. MASCG: Multi-Atlas Segmentation Constrained Graph method for accurate segmentation of hip CT images. Medical Image Analysis. 2015;26(1):173–184. Available from: http://dx.doi.org/10.1016/j.media.2015.08.011.

[73] Huang J, Griffith JF, Wang D, Shi L. Graph-cut-based segmentation of proximal femur from computed tomography images with shape prior. Journal of Medical and Biological Engineering. 2015;35(5):594–607.

[74] Kim JJ, Nam J, Jang IG. Fully automated segmentation of a hip joint using the patient-specific optimal thresholding and watershed algorithm. Computer Methods and Programs in Biomedicine. 2018;154:161–171. Available from: https://doi.org/10.1016/j.cmpb.2017.11.007.

[75] Gangwar T, Calder J, Takahashi T, Bechtold JE, Schillinger D. Robust variational segmentation of 3D bone CT data with thin cartilage interfaces. Medical Image Analysis. 2018;47:95–110. Available from: https://doi.org/10.1016/j.media.2018.04.003.

[76] Noguchi S, Nishio M, Yakami M, Nakagomi K, Togashi K. Bone segmentation on whole-body CT using convolutional neural network with novel data augmentation techniques. Computers in Biology and Medicine. 2020;121:103767. Available from: https://doi.org/10.1016/j.compbiomed.2020.103767.

[77] Leydon P, O'Connell M, Greene D, Curran KM. Bone Segmentation in Contrast Enhanced Whole-Body Computed Tomography. arXiv e-prints. 2020;Available from: https://arxiv.org/abs/2008.05223.

[78] Lindgren Belal S, Sadik M, Kaboteh R, Enqvist O, Ulén J, Poulsen MH, et al. Deep learning for segmentation of 49 selected bones in CT scans: First step in automated PET/CT-based 3D quantification of skeletal metastases. European Journal of Radiology. 2019;113:89 – 95. Available from: https://doi.org/10.1016/j.ejrad.2019.01.028.

[79] Schnider E, Horváth A, Rauter G, Zam A, Müller-Gerbl M, Cattin PC. 3D Segmentation Networks for Excessive Numbers of Classes: Distinct Bone Segmentation in Upper Bodies. In: Liu M, Yan P, Lian C, Cao X, editors. Machine Learning in Medical Imaging. Springer International

Publishing; 2020. p. 40–49. Available from: https://doi.org/10.1007/978-3-030-59861-7_5.

[80] Zhao C, Keyak JH, Tang J, Kaneko TS, Khosla S, Amin S, et al. A Deep Learning-Based Method for Automatic Segmentation of Proximal Femur from Quantitative Computed Tomography Images. arXiv e-prints. 2020;Available from: https://arxiv.org/abs/2006.05513.

[81] van Eijnatten M, van Dijk R, Dobbe J, Streekstra G, Koivisto J, Wolff J. CT image segmentation methods for bone used in medical additive manufacturing. Medical Engineering & Physics. 2017;0:1–11. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1350453317302631.

[82] Schmid J, Magnenat-Thalmann N. MRI bone segmentation using deformable models and shape priors. Medical image computing and computer-assisted intervention : MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention. 2008;11(Pt 1):119–126.

[83] Chandra SS, Xia Y, Engstrom C, Crozier S, Schwarz R, Fripp J. Focused shape models for hip joint segmentation in 3D magnetic resonance images. Medical Image Analysis. 2014;18(3):567–578. Available from: http://dx.doi.org/10.1016/j.media.2014.02.002.

[84] Damopoulos D, Lerch TD, Schmaranzer F, Tannast M, Chênes C, Zheng G, et al. Segmentation of the proximal femur in radial MR scans using a random forest classifier and deformable model registration. International Journal of Computer Assisted Radiology and Surgery. 2019;14(3):545–561. Available from: https://doi.org/10.1007/s11548-018-1899-z.

[85] Xia Y, Fripp J, Chandra SS, Schwarz R, Engstrom C, Crozier S. Automated bone segmentation from large field of view 3D MR images of the hip joint. Physics in Medicine and Biology. 2013;58(20):7375–7390.

[86] Liu F, Zhou Z, Jang H, Samsonov A, Zhao G, Kijowski R. Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging. Magnetic Resonance in Medicine. 2017;00. Available from: http://doi.wiley.com/10.1002/mrm.26841.

[87] Deniz CM, Xiang S, Hallyburton RS, Welbeck A, Babb JS, Honig S, et al. Segmentation of the Proximal Femur from MR Images using Deep Convolutional Neural Networks. Scientific Reports. 2018;8(16485). Available from: https://doi.org/10.1038/s41598-018-34817-6.

[88] Zhang W, Liu J, Yao J, Summers RM. Segmenting the thoracic, abdominal and pelvic musculature on CT scans combining atlas-based model and active contour model. SPIE Medical Imaging International Society for Optics and Photonics. 2013;8670(March 2013):867006–867008. Available from: http://dx.doi.org/10.1117/12.2007970.

[89] Lee H, Troschel FM, Tajmir S, Fuchs G, Mario J, Fintelmann FJ, et al. Pixel-Level Deep Segmentation: Artificial Intelligence Quantifies Muscle on Computed Tomography for Body Morphometric Analysis. Journal of Digital Imaging. 2017;30(4):487–498.

[90] Chung H, Cobzas D, Birdsell L, Lieffers J, Baracos V. Automated segmentation of muscle and adipose tissue on CT images for human body composition analysis. Progress in Biomedical Optics and Imaging - Proceedings of SPIE. 2009;7261(March 2009):72610K–72610K–8. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-67149108708{&}partnerID=40{&}md5=708b4cc73f7ab4ed595f737b9f302a4a{%}5Cnhttp://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1335704.

[91] Kamiya N, Zhou X, Chen H, Muramatsu C, Hara T, Yokoyama R, et al. Automated segmentation of psoas major muscle in X-ray CT images

by use of a shape model: Preliminary study. Radiological Physics and Technology. 2012;5(1):5–14.

[92] Wei Y, Xu B, Tao X, Junfeng Qu. Paraspinal muscle segmentation in CT images using a single atlas. In: 2015 IEEE International Conference on Progress in Informatics and Computing (PIC); 2015. p. 211–215. Available from: http://ieeexplore.ieee.org/document/7489839/.

[93] Otake Y, Yokota F, Fukuda N, Takao M, Takagi S, Yamamura N, et al. Patient-Specific Skeletal Muscle Fiber Modeling from Structure Tensor Field of Clinical CT Images. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, editors. Medical Image Computing and Computer Assisted Intervention (MICCAI) 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I. Cham: Springer International Publishing; 2017. p. 656–663. Available from: https://doi.org/10.1007/978-3-319-66182-7_75.

[94] Yokota F, Otake Y, Takao M, Ogawa T, Okada T, Sugano N, et al. Automated muscle segmentation from CT images of the hip and thigh using a hierarchical multi-atlas method. International Journal of Computer Assisted Radiology and Surgery. 2018;(1). Available from: http://link.springer.com/10.1007/s11548-018-1758-y.

[95] Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945;26(3):297–302.

[96] Hiasa Y, Otake Y, Takao M, Ogawa T, Sugano N, Sato Y. Automated Muscle Segmentation from Clinical CT using Bayesian U-Net for Personalized Musculoskeletal Modeling. IEEE Transactions on Medical Imaging. 2019;p. 1–1.

[97] Blemker SS, Delp SL. Three-Dimensional Representation of Complex Muscle Architectures and Geometries. Annals of Biomedical Engineer-

ing. 2005;33(5):661–673. Available from: http://dx.doi.org/10.1007/s10439-005-1433-7.

[98] Chèze, Laurence, Moissenet, Florent, Dumas, Raphaël. State of the art and current limits of musculo-skeletal models for clinical applications. Mov Sport Sci/Sci Mot. 2015;(90):7–17. Available from: https://doi.org/10.1051/sm/2012026.

[99] Wang C, Teboul O, Michel F, Essafi S, Paragios N. 3D knowledge-based segmentation using pose-invariant higher-order graphs. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2010;6363 LNCS(PART 3):189–196.

[100] Baudin P, Azzabou N, Carlier PG, Paragios N, Bm I. Automatic skeletal muscle segmentation through randomwalks and graph-based seed placement. Optimization. 2012;p. 1036–1039.

[101] Jurcak V, Fripp J, Engstrom C, Walker D, Salvado O, Ourselin S, et al. Automated segmentation of the quadratus lumborum muscle from magnetic resonance images using a hybrid atlas based - geodesic active contour scheme. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference. 2008;2008:867–870.

[102] Prescott JW, Best TM, Swanson MS, Haq F, Jackson RD, Gurcan MN. Anatomically anchored template-based level set segmentation: Application to quadriceps muscles in MR images from the osteoarthritis initiative. Journal of Digital Imaging. 2011;24(1):28–43.

[103] Ahmad E, Yap MH, Degens H, McPhee JS. Atlas-registration based image segmentation of MRI human thigh muscles in 3D space. Progress in Biomedical Optics and Imaging - Proceedings of SPIE.

2014;9037:90371L. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84901812643{&}partnerID=tZOtx3y1.

[104] Ogier A, Sdika M, Foure A, Le Troter A, Bendahan D. Individual muscle segmentation in MR images: A 3D propagation through 2D non-linear registration approaches. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. 2017;p. 317–320.

[105] Yao J, Kovacs W, Hsieh N, Liu CY, Summers RM. Holistic Segmentation of Intermuscular Adipose Tissues on Thigh MRI. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, editors. Medical Image Computing and Computer Assisted Intervention, MICCAI 2017: 20th International Conference, Proceedings, Part I. Cham: Springer International Publishing; 2017. p. 737–745. Available from: https://doi.org/10.1007/978-3-319-66182-7_84.

[106] Mesbah S, Shalaby AM, Stills S, Soliman AM, Willhite A, Harkema SJ, et al. Novel stochastic framework for automatic segmentation of human thigh MRI volumes and its applications in spinal cord injured individuals. PLoS ONE. 2019;14(5):1–20.

[107] Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis. 2017;36:61 – 78. Available from: http://www.sciencedirect.com/science/article/pii/S1361841516301839.

[108] Klemt C, Modat M, Pichat J, Cardoso MJ, Henckel J, Hart A, et al. Automatic assessment of volume asymmetries applied to hip abductor muscles in patients with hip arthroplasty. SPIE Medical Imaging. 2015;9413:94131M. Available

from: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2082341.

[109] Gilles B, Magnenat-Thalmann N. Musculoskeletal MRI segmentation using multi-resolution simplex meshes with medial representations. Medical Image Analysis. 2010;14(3):291–302. Available from: http://dx.doi.org/10.1016/j.media.2010.01.006.

[110] Kohout J, Clapworthy GJ, Zhao Y, Tao Y, Gonzalez-Garcia G, Dong F, et al. Patient-specific fibre-based models of muscle wrapping. Interface Focus. 2013;3(2):20120062–20120062. Available from: http://rsfs.royalsocietypublishing.org/cgi/doi/10.1098/rsfs.2012.0062.

[111] Hemke R, Buckless CG, Tsao A, Wang B, Torriani M. Deep learning for automated segmentation of pelvic muscles, fat, and bone from CT studies for body composition assessment. Skeletal Radiology. 2019;.

[112] Liang TT, Tsutsui S, Gao L, Lu JJ, Sun M. Combining Pyramid Pooling and Attention Mechanism for Pelvic MR Image Semantic Segmentaion. arXiv e-prints. 2018 Jun;p. arXiv:1806.00264.

[113] Sakamoto M, Hiasa Y, Otake Y, Takao M, Suzuki Y, Sugano N, et al. Automated Segmentation of Hip and Thigh Muscles in Metal Artifact-Contaminated CT using Convolutional Neural Network-Enhanced Normalized Metal Artifact Reduction. arXiv e-prints. 2019 Jun;p. arXiv:1906.11484.

[114] Maurer CR, Jr, Fitzpatrick JM. A Review of Medical Image Registration. In: Interactive imageguided neurosurgery; 1993. p. 17–44.

[115] Rueckert D, Schnabel JA. Medical Image Registration. In: Deserno T, editor. Biomedical Image Processing. Biological and Medical Physics, Biomedical Engineering. Springer, Berlin, Heidelberg; 2010. p. 131–154.

[116] Modat M. Efficient dense non-rigid registration using the Free-Form Deformation framework [PhD dissertation]. University College London; 2012.

[117] Charnley N, Morgan A, Thomas E, Wilson S, Bacon S, Wilson D, et al. The use of CT-MR image registration to define target volumes in pelvic radiotherapy in the presence of bilateral hip replacements. British Journal of Radiology. 2005;78(931):634–636.

[118] Kerkhof EM, Raaymakers BW, van der Heide UA, van de Bunt L, Jürgenliemk-Schulz IM, Lagendijk JJW. Online MRI guidance for healthy tissue sparing in patients with cervical cancer: An IMRT planning study. Radiotherapy and Oncology. 2008;88(2):241–249.

[119] Wells WM, Viola P, Atsumi H, Nakajima S, Kikinis R. Multi-modal volume registration by maximization of mutual information. Medical image analysis. 1996;1(1):35–51.

[120] Rivest-Hénault D, Greer P, Fripp J, Dowling J. In: Erdt M, Linguraru MG, Oyarzun Laura C, Shekhar R, Wesarg S, González Ballester MA, et al., editors. Structure-Guided Nonrigid Registration of CT–MR Pelvis Scans with Large Deformations in MR-Based Image Guided Radiation Therapy. Cham: Springer International Publishing; 2014. p. 65–73. Available from: https://doi.org/10.1007/978-3-319-05666-1_9.

[121] Noorda YH, Bartels LW, Huisman M, Nijenhuis RJ, van den Bosch MA, Pluim JP. Registration of CT to pre-treatment MRI for planning of MR-HIFU ablation treatment of painful bone metastases. Physics in medicine and biology. 2014;59(15):4167–79. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25004143.

[122] Cao X, Gao Y, Yang J, Wu G, Shen D. Learning-Based Multimodal Image Registration for Prostate Cancer Radiation Therapy. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. Medical Image

Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III. Cham: Springer International Publishing; 2016. p. 1–9. Available from: https://doi.org/10.1007/978-3-319-46726-9_1.

[123] Tanner C, Schnabel JA, Chung D, Clarkson MJ, Rueckert D, Hill DLG, et al. Volume and Shape Preservation of Enhancing Lesions when Applying Non-rigid Registration to a Time Series of Contrast Enhancing MR Breast Images. Lecture Notes in Computer Science. 2000;1935:327–337. Available from: http://www.springerlink.com/index/10.1007/978-3-540-40899-4{_}33{%}5Cnpapers2://publication/doi/10.1007/978-3-540-40899-4{_}33.

[124] Arsigny V, Pennec X, Ayache N. Polyrigid and polyaffine transformations: A novel geometrical tool to deal with non-rigid deformations - Application to the registration of histological slices. Medical Image Analysis. 2005;9(6):507–523.

[125] Seiler C, Pennec X, Ritacco L, Reyes M. Femur specific polyaffine model to regularize the log-domain demons registration. Proceedings of SPIE. 2011;p. 79620G–79620G–6. Available from: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=724526.

[126] Commowick O, Wiest-Daesslé N, Prima S. Automated diffeomorphic registration of anatomical structures with rigid parts: application to dynamic cervical MRI. In: 15th International Conference on Medical Image Computing and Computer Assisted Intervention. vol. 15 of LNCS. Nice, France: Springer; 2012. p. 163–70. Available from: http://www.hal.inserm.fr/inserm-00716106.

[127] Staring M, Klein S, Pluim JPW. A rigidity penalty term for nonrigid registration. Medical Physics. 2007;34(11):4098–4108.

[128] Reaungamornrat S, Wang aS, Uneri a, Otake Y, Khanna aJ, Siewerdsen JH. Deformable image registration with local rigidity constraints for cone-beam CT-guided spine surgery. Physics in medicine and biology. 2014;59(14):3761–87.

[129] Dowling JA, Sun J, Pichler P, Rivest-Hénault D, Ghose S, Richardson H, et al. Automatic substitute computed tomography generation and contouring for magnetic resonance imaging (MRI)-alone external beam radiation therapy from standard MRI sequences. International Journal of Radiation Oncology Biology Physics. 2015;93(5):1144–1153.

[130] Haber E, Heldmann S, Modersitzki J. A computational framework for image-based constrained registration. Linear Algebra and its Applications. 2009;431(3):459 – 470.

[131] Fitzpatrick JM, West JB, Maurer CR. Predicting error in rigid-body point-based registration. IEEE Transactions on Medical Imaging. 1998 Oct;17(5):694–702.

[132] Gjesteby L, De Man B, Jin Y, Paganetti H, Verburg J, Giantsoudi D, et al. Metal Artifact Reduction in CT: Where Are We After Four Decades? IEEE Access. 2016;4:5826–5849.

[133] Lewis M, Reid K, Toms AP. Reducing the effects of metal artefact using high keV monoenergetic reconstruction of dual energy CT (DECT) in hip replacements. Skeletal Radiology. 2013 Feb;42(2):275–282.

[134] Treece G. Refinement of clinical X-ray computed tomography (CT) scans containing metal implants. Computerized Medical Imaging and Graphics. 2017;56:11 – 23. Available from: http://www.sciencedirect.com/science/article/pii/S0895611117300058.

[135] Meyer E, Raupach R, Lell M, Schmidt B, Kachelriess M. Normalized metal artifact reduction (NMAR) in computed tomography. Med-

ical Physics. 2010;37(10):5482–5493. Available from: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.3484090.

[136] Naranjo V, Lloréns R, Alcañiz M, López-Mir F. Metal artifact reduction in dental CT images using polar mathematical morphology. Computer Methods and Programs in Biomedicine. 2011;102(1):64–74.

[137] Wang J, Zhao Y, Noble JH, Dawant BM. Conditional Generative Adversarial Networks for Metal Artifact Reduction in CT Images of the Ear. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Cham: Springer International Publishing; 2018. p. 3–11.

[138] Gjesteby L, Shan H, Yang Q, Xi Y, Claus B, Jin Y, et al. Deep Neural Network for CT Metal Artifact Reduction with a Perceptual Loss Function. In: The fifth international conference on image formation in X-ray computed tomography; 2018. p. 439–443.

[139] Zhu L, Han Y, Li L, Xu Y, Xi X, Yan B, et al. Metal artifact reduction based on fully convolutional networks in CT image domain. In: Tian Y, Xiao T, Liu P, editors. Second Symposium on Novel Technology of X-Ray Imaging. vol. 11068. International Society for Optics and Photonics. SPIE; 2019. p. 448 – 453. Available from: https://doi.org/10.1117/12.2524635.

[140] Huang X, Wang J, Tang F, Zhong T, Zhang Y. Metal artifact reduction on cervical CT images by deep residual learning. BioMedical Engineering Online. 2018;17(1):1–15. Available from: https://doi.org/10.1186/s12938-018-0609-y.

[141] Park HS, Lee SM, Kim HP, Seo JK, Chung YE. CT sinogram-consistency learning for metal-induced beam hardening correction. Medical Physics. 2018;45(12):5376–5384.

[142] Gjesteby L, Yang Q, Xi Y, Zhou Y, Zhang J, Wang G. Deep learning methods to guide CT image reconstruction and reduce metal artifacts. In: Flohr TG, Lo JY, Schmidt TG, editors. Medical Imaging 2017: Physics of Medical Imaging. vol. 10132. International Society for Optics and Photonics. SPIE; 2017. p. 752 – 758. Available from: https://doi.org/10.1117/12.2254091.

[143] Zhang Y, Yu H. Convolutional Neural Network Based Metal Artifact Reduction in X-Ray Computed Tomography. IEEE Transactions on Medical Imaging. 2018;37(6):1370–1381.

[144] Lin WA, Liao H, Peng C, Sun X, Zhang J, Luo J, et al. DuDoNet: Dual Domain Network for CT Metal Artifact Reduction. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2019. p. 10504–10513.

[145] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks. arXiv e-prints. 2014 Jun;p. arXiv:1406.2661.

[146] Wang K, Gou C, Duan Y, Lin Y, Zheng X, Wang F. Generative adversarial networks: introduction and outlook. IEEE/CAA Journal of Automatica Sinica. 2017;4(4):588–598.

[147] Cohen JP, Luck M, Honari S. Distribution Matching Losses Can Hallucinate Features in Medical Image Translation. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Cham: Springer International Publishing; 2018. p. 529–536.

[148] Hu Z, Jiang C, Sun F, Zhang Q, Ge Y, Yang Y, et al. Artifact correction in low-dose dental CT imaging using Wasserstein generative adversarial networks. Medical Physics. 2019;46(4):1686–1696.

[149] Liao H, Lin WA, Yuan J, Zhou SK, Luo J. Artifact Disentanglement Network for Unsupervised Metal Artifact Reduction. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, et al., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Cham: Springer International Publishing; 2019. p. 203–211.

[150] Anderla A, Culibrk D, Delso G, Mirkovic M. MR Image Based Approach for Metal Artifact Reduction in X-Ray CT. The Scientific World Journal. 2013;2013:1–8.

[151] Delso G, Wollenweber S, Lonn A, Wiesinger F, Veit-Haibach P. MR-driven metal artifact reduction in PET/CT. Physics in medicine and biology. 2013;58(7):2267–80. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84875293082{&}partnerID=tZOtx3y1{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/23478566.

[152] Park PC, Schreibmann E, Roper J, Elder E, Crocker I, Fox T, et al. MRI-based computed tomography metal artifact correction method for improving proton range calculation accuracy. International Journal of Radiation Oncology Biology Physics. 2015;91(4):849–856. Available from: http://dx.doi.org/10.1016/j.ijrobp.2014.12.027.

[153] Nielsen JS, Edmund JM, Leemput KV. CT metal artifact reduction using MR image patches. In: Lo JY, Schmidt TG, Chen GH, editors. Medical Imaging 2018: Physics of Medical Imaging. vol. 10573. International Society for Optics and Photonics. SPIE; 2018. p. 190 – 199. Available from: https://doi.org/10.1117/12.2293815.

[154] Koch KM, Lorbiecki JE, Hinks RS, King KF. A multispectral three-dimensional acquisition technique for imaging near metal implants. Magnetic Resonance in Medicine. 2009;61(2):381–390. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.21856.

[155] Lu W, Pauly KB, Gold GE, Pauly JM, Hargreaves BA. SEMAC: Slice encoding for metal artifact correction in MRI. Magnetic Resonance in Medicine. 2009;62(1):66–76.

[156] Lee MY, Song KH, Lee JW, Choe BY, Suh TS. Metal artifacts with dental implants: Evaluation using a dedicated CT/MR oral phantom with registration of the CT and MR images. Scientific Reports. 2019;9(1):1–10. Available from: http://dx.doi.org/10.1038/s41598-018-36227-0.

[157] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: Improved N3 bias correction. IEEE Transactions on Medical Imaging. 2010;29(6):1310–1320.

[158] Ebner M, Chouhan M, Patel PA, Atkinson D, Amin Z, Read S, et al. Point-Spread-Function-Aware Slice-to-Volume Registration: Application to Upper Abdominal MRI Super-Resolution. In: Zuluaga MA, Bhatia K, Kainz B, Moghari MH, Pace DF, editors. Reconstruction, Segmentation, and Analysis of Medical Images. Cham: Springer International Publishing; 2017. p. 3–13.

[159] Ranzini MBM, Ebner M, Cardoso MJ, Fotiadou A, Vercauteren T, Henckel J, et al. Joint Multimodal Segmentation of Clinical CT and MR from Hip Arthroplasty Patients. In: Glocker B, Yao J, Vrtovec T, Frangi A, Zheng G, editors. Computational Methods and Clinical Applications in Musculoskeletal Imaging. Cham: Springer International Publishing; 2018. p. 72–84. Available from: https://doi.org/10.1007/978-3-319-74113-0_7.

[160] Ranzini MBM, Henckel J, Ebner M, Cardoso MJ, Isaac A, Vercauteren T, et al. Automated postoperative muscle assessment of hip arthroplasty patients using multimodal imaging joint segmentation. Computer

Methods and Programs in Biomedicine. 2020;183:105062. Available from: https://doi.org/10.1016/j.cmpb.2019.105062.

[161] Arsigny V, Commowick O, Pennec X, Ayache N. A Log-Euclidean Framework for Statistics on Diffeomorphisms. In: Larsen R, Nielsen M, Sporring J, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 924–931.

[162] Ashburner J. A fast diffeomorphic image registration algorithm. NeuroImage. 2007;38(1):95–113.

[163] Ourselin S, Roche A, Subsol G, Pennec X, Ayache N. Reconstructing a 3D structure from serial histological sections. Image and Vision Computing. 2001;19(1):25 – 31.

[164] Modat M, Cash DM, Daga P, Winston GP, Duncan JS, Ourselin S. Global image registration using a symmetric block-matching approach. Journal of medical imaging. 2014;1(2):024003. Available from: http://medicalimaging.spiedigitallibrary.org/article.aspx?articleid=1909636.

[165] Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ. Non-rigid registration using free-form deformations: application to breast MR images. IEEE Transactions on Medical Imaging. 1999;18(8):712–21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10534053.

[166] Modat M, Ridgway GR, Taylor ZA, Lehmann M, Barnes J, Hawkes DJ, et al. Fast free-form deformation using graphics processing units. Computer Methods and Programs in Biomedicine. 2010;98(3):278–284.

[167] Modat M, Daga P, Cardoso MJ, Ourselin S, Ridgway GR, Ashburner J. Parametric non-rigid registration using a stationary velocity field. In: Proceedings of the Workshop on Mathematical Methods in Biomedical Image Analysis; 2012. p. 145–150.

[168] Gorgolewski K, Burns C, Madison C, Clark D, Halchenko Y, Waskom M, et al. Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. Frontiers in Neuroinformatics. 2011;5:13.

[169] Studholme C, Hill DLG, Hawkes DJ. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognition. 1999;.

[170] Ranzini MBM. The use of 3D imaging for musculoskeletal disease computational anatomy [Master's thesis]. University College London; 2016.

[171] Hiasa Y, Otake Y, Takao M, Matsuoka T, Takashima K, Prince JL, et al. Cross-modality image synthesis from unpaired data using CycleGAN: Effects of gradient consistency loss and training data size. arXiv e-prints. 2018 Mar;p. arXiv:1803.06629.

[172] Takao M, Ogawa T, Yokota F, Otake Y, Hamada H, T S, et al. Preoperative fatty degeneration of gluteus minimus predicts falls after tha. Bone Joint J. 2017;99(SUPP 6):39–39.

[173] Reito A, Elo P, Nieminen J, Puolakka T, Eskelinen A. Gluteal muscle fatty atrophy is not associated with elevated blood metal ions or pseudotumors in patients with a unilateral metal-on-metal hip replacement. Acta Orthopaedica. 2016;87(1):29–35.

[174] Ranzini MBM, Groothuis I, Kläser K, Cardoso MJ, Henckel J, Ourselin S, et al. Combining Multimodal Information for Metal Artefact Reduction: An Unsupervised Deep Learning Framework. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI); 2020. p. 600–604. Available from: https://doi.org/10.1109/ISBI45749.2020.9098633.

[175] Cachier P, Bardinet E, Dormont D, Pennec X, Ayache N. Iconic feature based nonrigid registration: the PASHA algorithm. Computer Vision

and Image Understanding. 2003;89(2):272 – 298. Nonrigid Image Registration. Available from: http://www.sciencedirect.com/science/article/pii/S107731420300002X.

[176] Odena A, Dumoulin V, Olah C. Deconvolution and Checkerboard Artifacts. Distill. 2016;Available from: http://distill.pub/2016/deconv-checkerboard.

[177] Dorent R, Joutard S, Modat M, Ourselin S, Vercauteren T. Hetero-Modal Variational Encoder-Decoder for Joint Modality Completion and Segmentation. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, et al., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Cham: Springer International Publishing; 2019. p. 74–82.

[178] Haskins G, Kruger U, Yan P. Deep learning in medical image registration: a survey. Machine Vision and Applications. 2020;31(1):8. Available from: https://doi.org/10.1007/s00138-020-01060-x.

[179] Simonovsky M, Gutiérrez-Becker B, Mateus D, Navab N, Komodakis N. A Deep Metric for Multimodal Registration. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Cham: Springer International Publishing; 2016. p. 10–18.

[180] Cheng X, Zhang L, Zheng Y. Deep similarity learning for multimodal medical images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 2018;6(3):248–252.

[181] Burgos N, Guerreiro F, McClelland J, Nill S, Dearnaley D, deSouza N, et al. Joint Segmentation and CT Synthesis for MRI-only Radiotherapy Treatment Planning. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. Medical Image Computing and Computer-Assisted

Intervention – MICCAI 2016. Cham: Springer International Publishing; 2016. p. 547–555.

[182] Nie D, Cao X, Gao Y, Wang L, Shen D. Estimating CT image from MRI data using 3D fully convolutional networks. In: International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis; 2016. p. 170–178.

[183] Kläser K, Markiewicz P, Ranzini M, Li W, Modat M, Hutton BF, et al. Deep Boosted Regression for MR to CT Synthesis. In: Gooya A, Goksel O, Oguz I, Burgos N, editors. Simulation and Synthesis in Medical Imaging. Cham: Springer International Publishing; 2018. p. 61–70.

[184] Fukuda N, Otake Y, Takao M, Yokota F, Ogawa T, Uemura K, et al. Estimation of attachment regions of hip muscles in CT image using muscle attachment probabilistic atlas constructed from measurements in eight cadavers. International Journal of Computer Assisted Radiology and Surgery. 2017;12(5):733–742.