Deep learning-based methods for 3D medical image registration

Théo Estienne

université
PARIS-SACLAY

Deep learning-based methods for 3D medical image registration

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 580, sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat: Mathématiques et Informatique
Unité de recherche: Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systèmes, 91190, Gif-sur-Yvette, France.
Référent: CentraleSupélec

Thèse présentée et soutenue à Paris Saclay, le 10 septembre 2021, par

Théo ESTIENNE

Composition du jury:

Marleen de Bruijne Professeur, Erasmus MC Rotterdam	Présidente et Examinatrice
Stéphanie Allassonnière Professeur, Université de Paris	Rapporteure et Examinatrice
Hervé Delingette Directeur de recherche, Inria Sophia-Antipolis	Rapporteur et Examinateur
Christos Davatzikos Professeur, University of Pennsylvania	Examinateur
Mattias Heinrich Professeur associé, University of Luebeck	Examinateur
Nikos Paragios Professeur, CentraleSupélec	Directeur
Eric Deutsch Professeur des universités - Praticien hospitalier, Université Paris-Saclay, Institut Gustave Roussy, Inserm	Co-directeur
Maria Vakalopoulou Maîtresse de conférence, CentraleSupélec	Co-encadrante

Thèse de doctorat

NNT: 2021UPASG055

Méthodes d'apprentissage profond pour le recalage 3D d'images médicales

Théo Estienne

Directeurs de thèse :

Nikos Paragios
Eric Deutsch
Maria Vakaloupoulou

Jury :

Présidente :	Marleen de Bruijne	Erasmus MC Rotterdam
Rapporteurs :	Hervé Delingette	Inria Sophia-Antipolis
	Stéphanie Allassonnière	Université de Paris
Examinateurs :	Christos Davatzikos	University of Pennsylvania
	Mattias Heinrich	University of Luebeck

Abstract

This thesis focuses on new deep learning approaches to find the best displacement between two different medical images, known as image registration. Its applications in the clinical pipeline include the fusion of different imaging types or the temporal follow-up of a patient, using methods such as diffeomorphic, graph-based or physical-based methods. Recently, deep learning-based methods were proposed using convolutional neural networks. These methods obtained similar results to non-deep learning methods while greatly reducing the computation time and enabling real-time prediction. This improvement comes from the use of graphics processing units (GPU) and a prediction phase where no optimisation is required. However, deep learning-based registration has several limitations, such as the need for large databases to train the network or the requirement of finding optimal hyperparameters to prevent noisy transformations. In this manuscript, we investigate various modifications to deep learning algorithms, for different imaging types and body parts.

We first study the combination of segmentation and registration tasks proposing a new joint architecture. We apply our joint network to brain MRI datasets, exploring different cases: brain without and with tumours. Our architecture comprises one encoder and two decoders, and the introduction of a supplementary loss reinforces the coupling. In the case of the brain without a tumour, we segment the brain structures while comparing our method to unsupervised and weakly-supervised registration approaches. In the presence of a tumour, the segmentation decoder predicts the binary tumour mask. We introduce it into the similarity loss; the registration focuses then only on healthy parts ignoring the tumour. We evaluate the deformation of the tumour by the deformation grid and compare to other deep learning-based registration methods.

Then, we shift to abdominal CT, a more challenging localisation, as there are natural organ's movements and deformations. We improve registration performances thanks to the introduction of various techniques. We use pretraining to benefit from the existence of many public datasets and we improve the pretraining by using pseudo segmentations generated by a neural network. We investigate the impact of new losses to provide better regularisation, penalising the Jacobian and the deformation's symmetry. Finally, we develop a multi-step strategy to refine the predicted deformation. We entered this strategy, submitting deformations for two among four proposed tasks, into the Learn2Reg 2020 Challenge organised in conjunction with MICCAI 2020. We were awarded 2nd position in the overall ranking.

We end this manuscript by analysing the explainability of registration networks using a linear decomposition and applying it to lung and hippocampus MR. Thanks to the late fusion strategy, images are projected to the latent space while calculating a new basis. This basis corresponds to elementary transformation which is studied qualitatively. Each elementary deformation focuses on special part of the body or on special movement. The connection between decomposition and clinical features, especially on the lung dataset, is also studied.

Résumé

Cette thèse se concentre sur des nouvelles approches d'apprentissage profond (aussi appelé deep learning) pour trouver le meilleur déplacement entre deux images médicales différentes, connu sous le nom de recalage d'images. Ses applications dans le traitement clinique incluent la fusion de différents types d'images médicales ou le suivi temporel d'un patient, à l'aide de méthodes telles que des méthodes basées sur des difféomorphismes, basées sur des graphes ou des équations physiques. Récemment, des méthodes basées sur l'apprentissage profond ont été proposées en utilisant des réseaux de neurones convolutifs. Ces méthodes ont obtenu des résultats similaires aux méthodes classiques tout en réduisant considérablement le temps de calcul et en permettant une prédiction en temps réel. Cette amélioration provient de l'utilisation de processeurs graphiques (GPU) et d'une phase de prédiction où aucune optimisation n'est requise. Cependant, le recalage à l'aide de deep learning a plusieurs limites, telles que le besoin de beaucoup de données pour entraîner le réseau ou le choix des bons hyperparamètres pour empêcher les déformations irrégulières. Dans ce manuscrit, nous étudions diverses modifications apportées aux algorithmes de recalage à l'aide d'apprentissage profond, pour divers types d'images et de parties du corps.

Nous étudions dans un premier temps la combinaison des tâches de segmentation et de recalage en proposant une nouvelle architecture conjointe. Nous appliquons notre réseau à des IRM cérébrales, en explorant deux cas : les cerveaux sans et avec tumeurs. Notre architecture comprend un encodeur et deux décodeurs, et l'introduction d'une fonction de coût supplémentaire renforce le couplage entre les deux tâches. Dans le cas des cerveaux sans tumeur, nous segmentons les structures cérébrales tout en comparant notre méthode à du recalage non supervisé et faiblement supervisé. En présence d'une tumeur, le décodeur de segmentation prédit le masque tumoral binaire. Nous introduisons ce masque dans la fonction de coût; le recalage se concentre alors uniquement sur les parties saines en ignorant la tumeur. Nous évaluons la déformation de la tumeur par la grille de déformation et la comparons à d'autres méthodes de recalage utilisant le deep learning.

Ensuite, nous nous concentrons sur le scanner abdominal, une localisation plus délicate, en raison des mouvements et des déformations naturelles des organes. Nous améliorons les performances de recalage grâce à l'introduction de plusieurs techniques. Nous utilisons le pré-entraînement pour profiter de l'existence de nombreuses données publiques et nous améliorons ce pré-entraînement en utilisant des pseudo segmentations générées par un réseau de neurones. Nous étudions aussi l'impact de nouvelles fonctions de coût pour fournir une meilleure régularisation, pénalisant le Jacobien et imposant une formulation symétrique. Enfin, nous développons une stratégie en plusieurs étapes pour affiner la déformation prédite. Nous avons évalué notre méthode en participant au challenge Learn2Reg 2020 organisé conjointement avec MICCAI 2020. Nous avons soumis nos solutions pour deux des quatre tâches proposées et obtenu la 2e position du classement général.

Nous terminons ce manuscrit en analysant l'explicabilité des réseaux de recalage en utilisant une décomposition linéaire et en l'appliquant à l'IRM pulmonaire et à l'hippocampe cérébrale. Grâce à notre stratégie de fusion tardive, les images sont projetées dans l'espace latent et une base de cette espace est calculée. Cette base correspond à des transformations élémentaires qui sont étudiées qualitativement. Chaque déformation élémentaire se concentre sur une partie particulière du corps ou sur un mouvement particulier. Le lien entre cette décomposition et certaines caractéristiques cliniques, en particulier sur la base de données de poumons, est également étudié.

Contents

List of Figures

List of Tables

Notations and conventions

M	Moving image (also called source)
F	Fixed image (also called reference or target)
M^{seg}, F^{seg}	Moving and fixed segmentations
$\widehat{M}$, $\widehat{M}^{seg}$	Warped (or deformed) moving image and segmentation
Ω	Spatial domain of images
$\mathbf{p}$, $\mathbf{q}$	voxel of Ω, or point, pixel, pixel location.
Φ	Deformation, function from Ω to itself
J_Φ	Jacobian matrix of Φ
$\lvert J_\Phi \rvert$	Determinant of the Jacobian matrix, called Jacobian;

Loss & Metrics

$\mathcal{L}_{sim}$	Similarity Loss (Mean Square Error or Local Cross Correlation)
$\mathcal{L}_{smooth}$	Smooth Loss
$\mathcal{L}_{seg}$	Segmentation Loss
$\mathcal{L}_{jac}$	Jacobian Loss
$\mathcal{L}_{inv}$	Inverse consistency Loss
MSE	Mean Square Error
LCC	Local Cross Correlation
SdLogJ	Standard Deviation of the Log Jacobian

Architecture

$\mathbf{E}$	Encoder
$\mathbf{D}$	Decoder
$\mathbf{D_{seg}}$	Segmentation Decoder
$\mathbf{D_{reg}}$	Registration Decoder
$\mathcal{W}(.)$	backward trilinear sampling
$\mathcal{M}$	Merging operation
c_M, c_F	Latent code of the moving and fixed images

Abbreviations

DL-based	Deep Learning based
MTL	Multi-task Learning
CNN	Convolutional Neural Networks
EM	Expectation Maximization algorithm
MRF	Markov Random Fields

Clinical

CT	Computed Tomorgraphy
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
HU	Hounsfield Unit

Chapter 1

Introduction

"Voilà ce qui va se passer ! "

Le visiteur du futur

Contents

1.1 Clinical Context

This thesis was a collaboration between two institutes: the engineering school CentraleSupelec and Radiotherapy Oncology within Gustave Roussy Cancer Campus. Gustave Roussy is a leading cancer research institute located in the Paris region. Our work was therefore significantly connected with cancer research and radiotherapy. In this chapter, we present cancer diagnostic and treatment pipelines and summarise clinical applications of registration algorithms focusing on medical imaging. The organisation and objectives of the thesis are also described, as well as the overall publications that have been produced during this thesis.

Medical Imaging Medical imaging is the technique and process of imaging the interior of a human body. It designates the set of techniques that noninvasively produce images, meaning no direct intervention on the patient's body. The field of medical imaging is directly connected with medical physics to develop devices and improve image acquisition; medicine to use and analyse images; mathematics and computer science to propose and develop algorithms for faster acquisition, reconstruction and processing of medical imaging. Medical images can be split into two types: functional imaging and structural imaging. The first type focuses on representing physiological processes such as brain or metabolic activities. The second type displays the anatomy of the body, such as the shape of organs. Among the different imaging techniques, the most common imaging types are radiography (or X-rays), Magnetic Resonance Imaging (MRI), nuclear medicine (such as Positron emission tomography or PET), ultrasound (US) and Computed Tomography (CT). These imaging types differ by physical processes, type of tissue they display, and medical application. For instance, bones have a high absorption of X-rays, and therefore radiography is an appropriate solution to detect bone fractures. On

the other hand, MRI is more suitable for neuro-imaging or joint disease as it produces high contrast for soft tissues such as muscle, fat or ligaments.

In this manuscript, we particularly focus on two types of medical imaging: Computed Tomography and Magnetic Resonance Imaging. Contrary to MRI, CT imaging can produce adverse effects to cumulative radiation exposure. However, the benefit, such as screening different diseases including cancer, makes their use in clinical practice essential. CT scans always have the same unit, the Hounsfield unit (HU), measuring the ability of a material to obstruct radiations or not. Air, water and bones have respective values around -1000 HU, 0 HU and superior to 300 HU. Radiologists observe CT scans using contrast windows, highlighting organs or body parts depending on the diagnosis they want to perform. On the contrary, MRIs have no fixed range of values, and normalisation is essential to correct intensity variations due to the imaging devices. The normalisation has little impact on the visual diagnostic produced by a radiologist but a major impact on automated image processing methods [Zwanenburg, 2020].

Cancer Cancer is not a unique disease but rather a group of diseases characterised by abnormal cell multiplication and propagation. It aggregates diseases with various localisation, properties and diagnosis. However, these diseases are characterised by six common landmarks, including the absence of cell death, a limitless cell division and the tissue invasion [Hanahan, 2000; Hanahan, 2011]. Cancerous cells first group together, forming the primary tumour, and then spread to others parts of the body, creating metastasis. Different causes explain the development of cancer, such as environmental causes (smoking, alcohol, obesity, exposure to pollutants), genetic and hereditary factors or infections. The gold standard for cancer diagnosis is based on a microscopic analysis of cells by an anatomical pathologist, even if CT or MRI can give significant indications and tools for the diagnosis. This microscopic analysis requires collecting cells using an invasive technique (biopsy or surgery). After the diagnosis, different cancer treatments exist, including surgery, chemotherapy, radiation therapy, or targeted therapy. Nowadays, many personalised treatments are in development based on genome sequencing, which helps determine the exact type of cancer. One current trend of cancer research is extracting imaging characteristics from different non-invasive modalities to allow personalised medicine without genetic sequencing.

Radiotherapy Radiotherapy is one method to treat cancer using radiation to kill cells or prevent their multiplication. The radiation damages cell DNA, resulting in cellular death. It is the most frequent cancer treatment with surgery, and a linear accelerator delivers the radiation. Radiotherapy has different uses: curing cancer totally (curative radiotherapy), combining with other treatment like chemotherapy or surgery (adjuvant or neo-adjuvant radiotherapy) or relieve symptoms (palliative radiotherapy).

One of the challenges of radiotherapy is to focus the radiation exposure only on the tumoral regions and not on the healthy tissue. Thus, during the treatment preparation, the radiotherapist segments the tumoral volume and the healthy organs near the tumour denoted in this context as Organs At Risk (OARs). The doctor prescribes a radiation dose in Grays (Gy), and a plan is decided

to deliver the dose to the tumour while minimising the one received by healthy tissue. This plan includes, for instance, the number of radiation beams or the shooting positions of the beams. A much larger dose is administered to the tumour by a precise optimisation of the radiation parameters than in the healthy tissue. However, it is always necessary to add margins to the irradiated volumes to correct position uncertainties. These uncertainties are caused by internal movements, such as breathing or digestion, organ's movements inside the body and the difference of the patient position during the treatment and at the pre-treatment scans, used to contour and plan the dose. The dose prescribed is not delivered once but is often split into several smaller doses. Dose fractioning reduce the toxic effects on healthy cells and increase the effect on tumoural cells.

Medical imaging is widely used during the radiotherapy workflow. It is needed to detect the tumoral regions, prescribe the dose and segment the tumour and OARs. Recent linear accelerators also include imaging devices to perform image-guided radiotherapy, aiming to correct the patient's motion during the treatment to decrease the margin and reduce the radiation received by OARs.

1.2 Registration

Clinical application of registration Image registration aims to find correspondences and map two or more images to a common space. Its development is strongly connected with the progress of medical imaging and particularly the emergence of various imaging modalities. The registration process can be applied to images from the same patient (intra-patient registration) or different patients (inter-patient registration). We should highlight three major application of image registration: multi-modality registration, longitudinal studies and anatomical variability studies [Hill, 2001; Maintz, 1998].

The goal of multi-modal registration is to better combine images by performing a spatial alignment between them. As the images have not been captured simultaneously, there will be some dissimilarity between them, for instance, due to the patient position. Before being computer-based, the fusion was performed manually by selecting and printing an image slice. However, a computer-based registration accelerates and improves the fusion and the diagnosis. Multi-modal registration is often performed to combine functional and structural imaging, such as CT and PET images, for cancer diagnosis [Oliveira, 2014]. The major difficulty of multi-modal registration is to find correspondences between anatomical points that have different intensities on each modality. Longitudinal studies intend to observe long-term changes, such as disease progression or tumour growth. The registration process deals with images from the same patient but at different time points. The modifications are not only due to the patient position but also from changes in the physiological state of the patients. Intra-patient registration also includes the fusion between pre and post-operative imaging to compare them and, for instance, measures the percentage of tumour removed by the surgery.

Finally, registration is also computed to obtain statistical information about a population of subjects. The goal is here to find common patterns and compare patients to each other or with a specific subject, also called the atlas. Atlas-based registration is another application that is often used as an unsupervised way of segmentation. It consists of solving the registration task with a unique

subject, the atlas, i.e. a medical volume resulted by averaging multiple medical volumes. Doctors have segmented the atlas, and the unknown segmentation is obtained by the backward deformation of the atlas's labels [Vakalopoulou, 2018; Cabezas, 2011; Rohlfing, 2005; Kalinic, 2008].

Registration & Radiotherapy Some applications of registration are specific to the radiotherapy pipeline. First studies, combining radiotherapy and registration, date back to the 1990s. In particular, Rosenman et al. [Rosenman, 1998] implemented a registration algorithm in the radiotherapy treatment planning system and then analysed the motivations for performing registration in the radiotherapy pipeline. The authors described four main reasons for using radiotherapy, mainly related to the contouring process. The first two reasons were that the tumour volumes had better definition on the MRI or the diagnostic CT than on the planning CT. The two other cases were if the tumour's delineation was not possible on the planning CT due to surgery or chemotherapy, but the tumour was well-defined on the pre-treatment scanner. For all these cases, the tumour was contoured on a different image (MRI, diagnostic CT or pre-treatment CT) than the planning CT, and then the registration was performed to warp the tumour volume to the planning CT. Recently, other registration applications have been developed and studied, such as dose accumulation or image-guided radiotherapy [Oh, 2017; Brock, 2017]. Dose accumulation gives a better estimation of the real dose delivered to the tumour than the planning dose. Registration is performed between the planning CT and the daily images acquired at each treatment step. The real dose is then evaluated by warping the planning dose, correcting motion and anatomical changes. Image-guided radiotherapy (IGRT) combines an imaging device with a linear accelerator to improve the accuracy of the radiation and reduce the dose delivered to OARs, taking the patient's motion into account. The correction can be online, adjusting the radiation's parameters during the treatment, or offline, determining the best position before starting the treatment. The registration algorithms are required to find the correspondence between the planning CT and images obtained during the treatment and shift the planning dose to the patient space. Online guidance involves a strong integration between image acquisition and the radiation process and almost real-time registration. Various imaging modalities are used for image-guided radiotherapy, including conventional CT, cone-beam CT (CBCT) or MRI.

1.3 Objectives and Contributions of the Thesis

This thesis focuses on the development and design of deep learning-based registration methods. More precisely, we investigate unsupervised and weakly-supervised approaches using modern 3D convolutional architectures. We apply them to several anatomical and medical imaging datasets. We explore different strategies aiming to improve and study the quality of registration. Here, we tackle the following questions :

- Can the registration benefit from its combination with segmentation task in a multi-task framework?

- How to design a registration algorithm that processes organs with abnormal regions such as tumour regions? More precisely, is it feasible to train a network to register only healthy structures?
- How evaluate registration's results properly? Particularly, how determine a trade-off between the deformation's regularity, the respect of anatomy, shape and organs and the registration's accuracy? Moreover, how to train properly deep learning-based registration methods on small datasets?
- Can we develop new methods for a better understanding of deep learning-based registration?

This manuscript is organised as follows :

In **Chapter 2**, we introduce the mathematical fundamentals of registration. It includes the formulation of the registration problem, the algorithms to solve it and the metrics to evaluate registration. We also detail different algorithms, first with non-deep learning methods (or iterative methods) and then with deep learning approaches. Finally, we summarise different deep learning-based frameworks, focusing on modern methods, such as unsupervised and weakly-supervised registration frameworks.

In **Chapter 3**, we study the combination of registration and segmentation and propose a deep learning joint formulation. These two tasks are among the most studied in the medical imaging field, and several researchers have already proposed approaches to benefit from their association. However, few deep learning methods have been developed to fuse them optimally. This work is inspired by classical registration algorithms and multi-task deep learning-based frameworks, where different problems are solved jointly. We propose UReSNet, a 3D convolutional neural network made of one encoder and two decoders, one for the segmentation and the other for the registration. The combination is reinforced by the introduction of a supplementary loss which influences both of the decoders. We present experiments on Oasis 3, a brain MRI dataset, and evaluate both the registration and segmentation part. This chapter is mainly based on our contribution to MICCAI 2019 in Shenzhen [Estienne, 2019].

In **Chapter 4**, we extend our previous formulation focusing on registration and segmentation of abnormal regions and, in particular, expanding to the more complex case of brain tumours. The presence of tumours in the brain introduces abnormal areas which do not have correspondences with healthy tissues. To deal with this issue, we propose a formulation that focuses only on the healthy regions to perform the registration while it does not deform tumour regions. The main contributions of this study focus on two points. First, our segmentation decoder predicts tumour masks while UReSNet outputs the segmentation of healthy brain structures. Secondly, we introduce a shared encoder and a merging operation to disjoin the latent representation of each image. Each MRI is passed independently through the encoder. Then the latent representations are passed separately to the segmentation decoder while they are merged together before passing through the registration

decoder. We experiment with two different brain MRI datasets, BraTS 2018 and Oasis 3, using the first for training and segmentation's evaluation and the second to evaluate registration performances. The contributions of this work are summarised to the journal Frontiers in Computational Neuroscience [Estienne, 2020].

In **Chapter 5**, we explore several improvements to our registration pipeline and apply our method to more challenging anatomy, in particular medical volumes covering the abdominal part of the body. First, we study the impact of pre-training, and the influence of pseudo-segmentations on registration's performances. Instead of using ground-truth segmentations produced by doctors, we train a segmentation network using publicly available datasets to obtain approximated labels and use them in the pre-training step. We also study the impact of various regularisation losses to respect topological properties better, constraining the transformation's symmetry and Jacobian determinant. Finally, we propose a multi-step approach to refine the predicted grid, taking advantage of the shared encoder proposed in chapter 4.

This chapter summarises our participation in the Learn2Reg Challenge, co-organised with the MICCAI 2020 conference. The organisers proposed four registrations tasks, including brain MR with intra-operative ultrasound, lung CT expiration-inspiration, abdominal CT and brain MR registration. We submitted our approach based on the symmetric formulation and the pseudo-segmentation and achieved second and third position for task 3 (abdominal CT) and 4 (hippocampus MRI) [Estienne, 2021a; Estienne, 2021b].

In **Chapter 6**, we investigate explainability that deep learning registration methods could offer. This is a major issue in medical imaging as we need to understand the reasons for a prediction. In particular, with the appropriate model architecture and by using a simple linear projection, we decompose the encoding space, generating a new basis that we empirically show that captures various decomposed anatomically aware geometrical transformations. We validate the performance of our method on two different datasets, lung and hippocampus MR, verifying that the elementary transformations focus on particular areas of the images. We also explore the connection between our decomposition and some clinical values. The results presented in this chapter have been submitted and accepted to the MICCAI 2021 DART Workshop (Domain Adaptation and Representation Transfer) [Estienne, 2021c].

1.4 Scientific Productions

Journal articles

- Nathalie Lassau, Théo Estienne, Philippe de Vomecourt, Mikael Azoulay, John Cagnol, et al. "Five Simultaneous Artificial Intelligence Data Challenges on Ultrasound, CT, and MRI". in: *Diagnostic and Interventional Imaging* 100.4 (Apr. 2019), pp. 199–209
- Théo Estienne, Marvin Lerousseau, Maria Vakalopoulou, Emilie Alvarez Andres, Enzo Battistella, et al. "Deep Learning-Based Concurrent Brain Registration and Tumor Segmentation".

In: *Frontiers in Computational Neuroscience* 14 (2020)

Conference Papers

- Théo Estienne, Maria Vakalopoulou, Stergios Christodoulidis, Enzo Battistela, Marvin Lerousseau, et al. "U-ReSNet: Ultimate Coupling of Registration and Segmentation with Deep Nets". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 310–319

- Théo Estienne, Maria Vakalopoulou, Enzo Battistella, Alexandre Carré, Théophraste Henry, et al. "Deep Learning Based Registration Using Spatial Gradients and Noisy Segmentation Labels". In: *Segmentation, Classification, and Registration of Multi-Modality Medical Imaging Data*. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 87–93

- Théo Estienne, Maria Vakalopoulou, Stergios Christodoulidis, Enzo Battistella, Théophraste Henry, et al. "Exploring Deep Registration Latent Spaces". In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 112–122

In submission

- Théo Estienne, Maria Vakalopoulou, Enzo Battistella, Theophraste Henry, Marvin Lerousseau, et al. "MICS : Multi-Steps, Inverse Consistency and Symmetric Deep Learning Registration Network". In: *arXiv:2111.12123 [cs]* (Nov. 2021). arXiv: 2111.12123 [cs]

Additional Contributions

- Siddhartha Chandra, Maria Vakalopoulou, Lucas Fidon, Enzo Battistella, Théo Estienne, et al. "Context Aware 3D CNNs for Brain Tumor Segmentation". In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 299–310

- Enzo Battistella, Maria Vakalopoulou, Théo Estienne, Marvin Lerousseau, Roger Sun, et al. "Gene Expression High-Dimensional Clustering Towards a Novel, Robust, Clinically Relevant and Highly Compact Cancer Signature". In: *Bioinformatics and Biomedical Engineering*. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 462–474

- Alexandre Carré, Guillaume Klausner, Myriam Edjlali, Marvin Lerousseau, Jade Briend-Diop, et al. "Standardization of Brain MR Images across Machines and Protocols: Bridging the Gap for MRI-Based Radiomics". In: *Scientific Reports* 10.1 (July 2020), p. 12340

- Marvin Lerousseau, Maria Vakalopoulou, Marion Classe, Julien Adam, Enzo Battistella, et al. "Weakly Supervised Multiple Instance Learning Histopathological Tumor Segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 470–479
- Théophraste Henry, Alexandre Carré, Marvin Lerousseau, Théo Estienne, Charlotte Robert, et al. "Brain Tumor Segmentation with Self-Ensembled, Deeply-Supervised 3D U-Net Neural Networks: A BraTS 2020 Challenge Solution". In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 327–339

Other activities

- Four months research visit at Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, under the supervision of de Christos Davatzikos and Anahita Fathi Kazerooni;
- Participation to the BraTS 2018 data challenge;
- Participation to the Learn2Reg 2020 data challenge, with an overall 2^{nd} position (tie) and an oral presentation at the MICCAI workshop;
- Co-organisation of a data challenge during the congress of the French Radiology Society (JFR 2018);
- Participation to the scientific vulgarization competition *Ma Thèse en 180 s*;
- Teaching assistant for Deep Learning and Python courses at CentraleSupelec.

Chapter 2

Deformable Registration and Deep Learning

*"**Bohort :** Il me semble que pour bien chercher il faut savoir ce qu'on cherche.*
[...]
***Perceval** Si Joseph d'Arimathie a pas été trop con, vous pouvez être sur que le Graal, c'est un bocal à anchois. "*

Kaamelott, En Forme de Graal

Contents

This chapter introduces the key concepts of medical image registration. We describe primary deep learning development in the medical imaging domain and its attendant challenges. Then, we explore the different components of the registration problem, such as the deformation model or the objective function, together with their mathematical formulations and properties. We also focus on the required properties of a deformation grid and the evaluation metrics of registration algorithms. Finally, we study registration algorithms, beginning with iterative algorithms and more traditional formulations and moving towards deep learning-based algorithms. Our focus is particularly unsupervised and weakly-supervised formulations of registration frameworks.

2.1 Introduction

Image registration is a challenging and widely researched task in medical imaging and computer vision. It aims to determine the best geometrical transformation to project or map two or more volumes to the same space. In chapter 1, we detailed how clinical applications need registration, especially radiotherapy treatment. In the current chapter, we focus on the mathematical formulations and the solving of the registration problem. A large variety of methods have been formulated to tackle registration problems. These methods included traditionally graph-based methods [Glocker, 2009; Parisot, 2012], diffusion-based approaches [Thirion, 1998; Vercauteren, 2009], symmetric formulation [Avants, 2008], flows of diffeomorphism [Beg, 2005; Yan Cao, 2005] or stationary velocity field based algorithm [Arsigny, 2006; Ashburner, 2007].

Recently, various deep learning (DL)-based methods have been proposed, reducing the computational time significantly and thus enabling real-time applications. These methods have currently gained much attention, providing very efficient and accurate performances while continuing to face some limitations. Among the first DL-based formulation, [Balakrishnan, 2018] proposed a network trained for atlas-based registration of brain MR images, [Stergios, 2018] presented a joint affine and deformable framework for lung MR images and [Krebs, 2018] developed a diffeomorphic formulation for cardiac MR registration. These methods take advantage of an unsupervised formulation, using only images and no ground truth deformations. [Hering, 2018] improved the registration process by the use of segmentation labels, thus developing weakly-supervised registration.

The presentation of this chapter is based on different review articles focusing on traditional and DL-based registration approaches. Two major analyses of classical registration algorithms are available in Sotiras et al. [Sotiras, 2013] and Oliveira et al. [Oliveira, 2014]. More precisely, Sotiras et al. [Sotiras, 2013] report an important number of algorithms, analysing the registration's three main components, namely the deformation model, the optimisation process and the objective function. Concerning DL-based registration, our analysis is based on three very recent review papers : [Fu, 2020; Haskins, 2020; Boveiri, 2020]. They report the classification of DL-based methods, the studied organs and modalities, the top databases used and the most active authors in the field. Finally, Krebs [Krebs, 2020] presents various deep learning-based registration approaches, focusing on reinforcement learning, diffeomorphic and variational methods.

This chapter is organised as follows : we first describe the recent development of deep learning in the medical imaging field (section 2.2), then the registration problem in general without implementation consideration (section 2.3). We also detail the popular and traditional approaches designed before deep learning (section 2.4) and to conclude, we focus on DL-based registration methods (section 2.5).

2.2 Deep Learning and Medical Imaging

The use of deep learning in the computer vision field has experienced exponential growth since the development of convolutional neural networks (CNNs), the availability of large datasets as well as the improvement of graphical processing units (GPU). The creation of public databases and computer vision competition (MNIST [Deng, 2012], ImageNet [Deng, 2009], CIFAR-100 [Krizhevsky,]) resulted in the conception of new neural network architectures (AlexNet [Krizhevsky, 2014], VGG [Simonyan, 2015] and ResNet [He, 2016] for instance). Before deep learning, researchers studied medical imaging with different tools. Machine learning algorithms, focusing mainly on random forest and SVM models, were used to perform computer-aided diagnostic, classification, or even segmentation [Wernick, 2010; Erickson, 2017; Giger, 2018]. The main difference between machine and deep learning is the way data are processed. In particular, deep learning methods generate hierarchical representations automatically, while classical machine learning methods are based on predefined handcraft features. These features have many forms such as statistical, geometrical, morphological or wavelet transform based [Kumar, 2014]. The high development of deep learning in the past few years reduced the work on feature extraction and machine learning. Yet, it is still used in cases where the number of images is too small, and thus the deep learning algorithms cannot be trained [Gillies, 2015; Lambin, 2017].

Authors in [Litjens, 2017; Shen, 2017] analysed the different tasks, modalities and anatomies where deep learning have been applied. Concerning the imaging types, among the most popular modalities, one could find: Magnetic Resonance Imaging (MRI), Computed Tomography scans (CT), Ultrasound (US), X-rays, Mammography (MG), nuclear imaging (PET scans), histopathology images, ophthalmic imaging with colour fundus imaging (CFI) and dermoscopic images. The main tasks focus on classification, detection, regression and segmentation. Two main subtasks of the classification task include: characterising a lesion/exam as malignant or not and recognising particular diseases. The goal of the detection task is to detect lesions, organs, anatomical structures or regions of interest. The detection task can be combined with classification or segmentation to develop computer-aided diagnosis software (CADX) that are applied in clinical practice. One of the main challenges for the detection problem includes identifying small areas of interest from huge 3D dimensional data such as CT and MRI modalities. Regression approaches aim to generate algorithms that predict real numbers, such as survival prediction or even predicting other clinical attributes like patients' age using brain MRI. This task is more challenging than the others due to the high number of outcomes ranges and the significant dataset needed. The segmentation task is among the most studied tasks for deep learning, resulting in many articles, challenges and medical databases. The

segmentation can include anatomical structures or abnormal regions. The segmentation task is often applied with a fully convolutional network and in a fully supervised way. Contrary to detection or classification tasks where labels are related to one patient or one image, for segmentation problems, the ground truth annotations include voxel-level annotations, which could be quite computationally expensive. Thanks to voxelwise labelling, deep learning approaches performed well for segmentation, even with a relatively small number of patients. However, recent researches explore unsupervised or weakly supervised segmentation using, for instance, bounding box or scribble as labels, making the annotation process less demanding [Lin, 2016; Rajchl, 2017; Kervadec, 2019].

An important point for the application of deep learning in medical imaging is the selection of the proper architecture (2D or 3D). Starting with the two dimensions deep learning architectures, medical imaging can benefit from all the progress made in other fields such as computer vision (i.e. non-medical). This includes the use of modern architectures such as ResNet [He, 2016], VGG [Simonyan, 2015] or EfficientNet [Tan, 2019] and networks pretrained on huge dataset like ImageNet. The 2D problem resolution also has fewer memory issues and an important number of already implemented functions in the main deep learning frameworks, such as data augmentation functions. However, one of the main disadvantages of 2D architectures is that they do not explore all the available information that exist in medical imaging.

On the other hand, working in three dimensions raise several difficulties. All the activation maps and the gradients are in 3D, and it constrains the number of layers and channels of the network used. While recent publications promote 3D pretrained network and architecture [Chen, 2019; Zhou, 2019b], a consensus remain to be achieved. Due to the memory problem, we are forced to use a small batch size, sometimes equal to 1, and a small patch size, which will not cover the whole image (for abdominal CT, for instance). Finally, in a 3D context, the input and output operations (I/O) are often the bottleneck of our training process, leading to an underuse of the GPU and a long training time. Despite these difficulties, 3D networks produce better representations, and results, as they combine multiples slices. Thus, three-dimension approaches became state-of-the-art for the majority of medical imaging problems.

The growth of deep learning applied to medical imaging has been followed by an increasing number of data challenges and public medical databases. This allows a fair comparison between different algorithms, as the choice of the training and testing set significantly impacts performances. One of the largest databases is the TCIA[1] [Clark, 2013], which contains a large type of imaging for many cancer pathologies. Different websites helps to organise data challenge with online submission systems[2][3]. However, these platforms are not specialised in biomedical and health. The website Grand Challenge gather the majority of the biomedical imaging challenge since 2010[4]. In Maier-Hein et al.

[1] https://www.cancerimagingarchivei.net/
[2] https://www.kaggle.com/
[3] https://codalab.org/
[4] https://grand-challenge.org/challenges/

[Maier-Hein, 2018], the authors studied the development of challenges and their reproducibility. The major task was segmentation, with 70% of the challenges, followed by classification and detection. The majority of the challenges took place during two conferences, the MICCAI conference and the IEEE International Symposium on Biomedical Imaging (ISBI). Following Maier-Hein et al. [Maier-Hein, 2020], MICCAI provides guidelines to standardise data challenges and improve their quality, reproducibility and interpretability. One important impact of data challenges after the fair comparison of algorithms is the release of public datasets for the community. Such datasets could be beneficial for training on the same problems or even be used to pre-train deep learning architectures for other tasks.

The difference between computer vision and medical imaging resulted in the creation of new architectures and networks designed especially for medical data. [Kamnitsas, 2017] proposed a 3D CNN called DeepMedic for semantic segmentation of brain tumours. This network comprises two multi-resolution branches, one processing the image at the normal resolution and the other at low resolution. The authors also used a conditional random field as post-processing and achieved top ranking performances on the BraTS 2015 and ISLES 2015 challenges [Menze, 2015; Maier, 2017]. The UNet [Ronneberger, 2015] was developed to segment neural structures and cells, obtaining high popularity. This auto-encoder architecture is composed of two symmetric paths joined with skip connections. In the encoder, max-pooling layers downsample the image, and in the decoder, upsampling convolutions restore the original resolution. The skip connections keep high dimensional information and help the network to generate a more precise segmentation. The UNet is fully convolutional (without dense layer), which allows the processing of any image size. Many changes have been proposed to this architecture, such as the 3D UNet [Çiçek, 2016] first developed for Xenopus kidney segmentation and the VNet [Milletari, 2016] with each convolutional blocks learning residual functions and convolutional layers used to perform downsampling. Researchers experimented with other modifications for segmentation challenges like the addition of regularisation layers, second decoder branch or attention blocks. In Isensee et al. [Isensee, 2019], the authors claim that the best performances are not achieved by architecture modifications but special care on the training process. They proposed a 3D UNet architecture called nnU-Net and achieved top performances to segmentation challenge such as brain tumor segmentation (BraTS Bakas et al. [Bakas, 2019]), multi organs segmentation (Medical Segmentation Decathlon Simpson et al. [Simpson, 2019]) or kidney tumor segmentation (KiTS, Heller et al. [Heller, 2020]).

2.3 Registration Algorithm

2.3.1 Mathematical formulation

In this section, we describe the registration problem and the general methodology of registration algorithms. The registration consists of finding the best correspondence between two images. These images are often referred to as moving or source for the first one and reference, target, or fixed for the second one. In this manuscript, we chose to define them as the moving image M and the fixed image F. Theses two volumes are defined over a spatial domain $\Omega = \mathcal{I}_1 \times \cdots \times \mathcal{I}_n$ with $\mathcal{I}_i$ being an interval of $\mathbb{N}$. The length of each interval $\mathcal{I}_i$ corresponds to the width, height or depth of M and F. For the case of medical imaging, n is mostly equal to 2 or 3. Moving and fixed images can be consider as a function of the domain Ω, which for each voxel $\mathbf{p} \in \Omega$ associate a value (or intensity) $M(p)$, respectively $F(p)$. In this manuscript, we refer interchangeably to the voxel $\mathbf{p}$ as voxels, points, pixels or pixel locations, and we use the term volume and image equivalently, even if we are always working in three dimensions.

The registration problem consists in finding the optimal transformation Φ to warp M to F, while Φ is included in a set of possible transformations $\mathcal{T}$. Mathematically, we want to solve the following optimisation problem :

$$\widehat{\Phi} = \arg\min_{\Phi \in \mathcal{T}} \mathcal{D}(F, M(\Phi)) + \lambda \mathcal{R}(\Phi) \tag{2.1}$$

where $M(\Phi)$ is the volume M warped by the transformation Φ. The previous optimisation problem consists in two part. The first term $\mathcal{D}$, called (dis)similarity or matching criterion, controls the similarity between F and the warped image $M(\Phi)$. The second term $\mathcal{R}$, designated as the regularisation term, aims to respect specific pre-defined properties such as the smoothness of the transformation, while λ is a real value controlling the influence of the regularisation term on the optimization process.

Following [Sotiras, 2013], we can divide an image registration problem into three parts: the deformation model, the objective function and the optimisation algorithm. The deformation model concerns the choice of the deformation space $\mathcal{T}$ where Φ is defined. Deformation models can be, for instance, graph-based models or physical-based models. The objective function refers to the formulation of $\mathcal{D}$ and $\mathcal{R}$ and the optimisation algorithm is the selected method to resolve the optimisation problem 2.1. The formulation of the dissimilarity criterion and the deformation model depends on the type or registration problem. Some matching criteria are not appropriated for multi-modality registration as moving and fixed images have different intensities. Depending on the objective function and the deformable model, registration is often an ill-posed problem. Indeed many solutions can exist to the same problem. This is particularly the case when we deal with deformable registration as there are a large number of parameters. The regularisation term $\mathcal{R}$ helps us to restraint the ill-posed nature of registration [Sotiras, 2013].

Warping and interpolation The transformation Φ can warp points p from the moving to fixed image (forward warping or mapping) or reversely (backward warping). While both of these strategies are theoretically possible, the backward mapping is more interesting due to a more efficient implementation [Sotiras, 2013]. The differences between the two warping strategies rely on the fact that our spatial domain Ω is defined over integers and not real values, and thus we need to perform interpolation in the moving or fixed image.

In the forward mapping, every voxel from the moving image shifts to a new position in the fixed image. As this new position, p' is a non-voxel location (being a non-integer), we need to split the intensity $M(p)$ between all the neighbour pixels of p'. Conversely, in the backward mapping, the deformed image's voxels q are linked to non-voxel locations p' of the moving image. Then, we need to find the intensity value $M(p')$, applying the interpolation in M. Figure 2.1 represents and compares the two mapping approaches. The interpolation aims to find the intensity values on the spatial domain Ω after warping all the points. Different interpolation algorithms can be applied, such as nearest-neighbour, bilinear, cubic or spline.

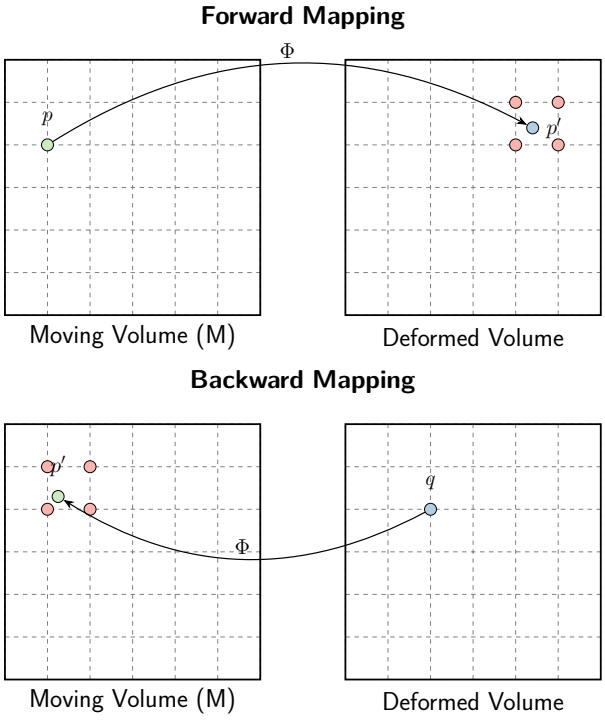


Figure 2.1: Comparison between the forward and backward warping presented in top and bottom respectively.

Deformation models The choice of the deformation models determine the number of parameters (or degrees of freedom), the model's complexity, and thus the optimisation's computational requirements. In the large variety of the deformation models, we choose to focus particularly on the two special cases: affine transformation and deformable transformations (also called elastic, non-rigid or dense). While affine transformations have only a few parameters, deformable transformations can have several million parameters depending on the image's size.

Mathematically speaking, an affine transformation is a transformation that preserves lines and parallelism. However, the distance and the angles are not always preserved. Among the different affine transformations, certain transformations exist, such as translation, rotation, reflection, scale, and shear. Rigid transformations are a subgroup of affine transformations which preserves the distance between every pair of points. This group comprises rotation, translations, reflections, and any composition of these. Reflection transformations are sometimes excluded from the rigid group to have only transformations which preserve the orientation. This smaller subgroup is denoted as rigid motions or proper rigid transformations. Affine transformations are formulated as a two-dimension matrix of shape 2×3 or 3×4, if the images are respectively two or three dimensions images. The transformations can be formulated as : $\forall p \in \Omega, \Phi(p) = A \begin{pmatrix} p_x \\ p_y \\ p_z \\ 1 \end{pmatrix}$ with A being the affine transformation matrix formulated in homogeneous coordinates. One method to enforce the transformation to be a rigid motion is to calculate it as a multiplication of translation matrix and rotation matrices. The registration algorithm directly predicts in this case the parameters of the translations and rotations. The expression of the transformation matrix is then $A = T \cdot R_x \cdot R_y \cdot R_z$ with T being the translation matrix, R_x, R_y and R_z being the three rotation matrices following x, y and z-axis. Their expressions in homogeneous coordinates are :

$$T = \begin{pmatrix} 0 & 0 & 0 & t_x \\ 0 & 0 & 0 & t_y \\ 0 & 0 & 0 & t_z \end{pmatrix} \qquad R_z = \begin{pmatrix} \cos\theta_z & -\sin\theta_z & 0 & 0 \\ \sin\theta_z & \cos\theta_z & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

with θ being the angle of the rotation and t_x, t_y and t_z the coordinates of the translation vector. Affine transformations can generate large displacements between two volumes and are often used as a preprocessing step before applying deformable models. The small number of parameters allows an easy understanding of affine transformations. However, they do not have the freedom of deformable displacement and may not be accurate for registering local regions or specific organs.

Affine registration is more effective when there is no change in the patient anatomy. However, in specific situations, high anatomy changes appear due to tumour growth or organs natural displacements. Deformable registration is then more suitable as it can perform local modifications. Deformable registration applies a different transformation for each images voxel p : $\Phi(p) = \begin{pmatrix} \Phi_x(p) \\ \Phi_y(p) \\ \Phi_z(p) \end{pmatrix}$. As the degree of freedom (DoF) is much more important, many more parameters require optimisation than they would in affine transformations. The number of parameters depends on the model's

choice. Interpolation-based models are a strategy to reduce the number of parameters and thus the computational complexity [Sotiras, 2013]. On the other hand, some deformation models consider one displacement vector for each voxel of Ω. Then, the displacement is modelled as a matrix of the shape $3 \times n \times m \times o$ (respectively $2 \times n \times m$ in 2D) with the moving and fixed images having the shape $n \times m \times o$ (resp. $n \times m$) and the first dimension corresponding to the displacement along the three axes x, y and z (resp. x and y).

2.3.2 Properties of the deformable registration methods

In the context of medical imaging and deformable models, the deformation field must respect several properties. Following [Sotiras, 2013], we study the next four properties that need to be validated: inverse consistency, symmetry, topology preservation and diffeomorphism. These properties can be enforced directly by the deformation model's choice or by the formulation of $\mathcal{R}$. Each additional term added to our regularisation formulation will have more or less impact depending on the weight we applied to it. The tuning of these weights requires a trade-off between the performance of the registration and the respect of the constraint.

Symmetry Most of the existing registration algorithms are asymmetric [Sotiras, 2013]. Consequently, if we reverse M and F in the formulation of the optimisation problem, we do not obtain the inverse transformation. Therefore, the evaluation of the performance could be different depending on the choice of the target domain. To overcome this limitation, different authors proposed new formulations of the objective function or the optimisation problem [Vercauteren, 2008; Lorenzi, 2013]. Two main strategies exist and include: i) building a matching criterion $\mathcal{D}$ involving simultaneously forward and backward warping and ii) applying the registration to a virtual mid-point between the moving and the fixed images [Sotiras, 2013].

Inverse consistency An important property for medical image registration consist of ensuring that the transformation applied on M can be reversed. The reverse grid Φ^{-1} can be estimated by resolving the reverse optimisation problem or by numerically inversing the grid Φ. Such strategies, even if they produce the reverse grid Φ^{-1}, augment the computational requirements of the proposed solutions. Designing a deformation model to output both the forward and the backward deformation may be a better solution. It requires modifying the objective function and the deformation model to deform both M to F and F to M. To guarantee that the two grids are effectively the inverse of each other, we need to add a new term to the equation 2.1. Formulations that are commonly used to ensure inverse consistency constraints include: i) penalising the difference between the composition of the forward and backward transformations and the identity transformation and ii) minimising the dissimilarity criterion $\mathcal{D}$ of the moving image and itself, which have been deformed by Φ and Φ^{-1} [He, 2003; Leow, 2005].

Topology preservation Without any constraint, the transformation can result in many non-relevant modifications. For instance, the deformation field can project two different voxels to the same output or encourage pixel's folding and crossings. A topological preserving algorithm should create a continuous, bijective transformation and with a continuous inverse. The Jacobian determinant is an important criterion to enforce these properties. The transformation Φ must be differentiable to calculate the Jacobian matrix and its determinant (also called Jacobian). The expression of the Jacobian matrix at a point $p \in \Omega$ will be :

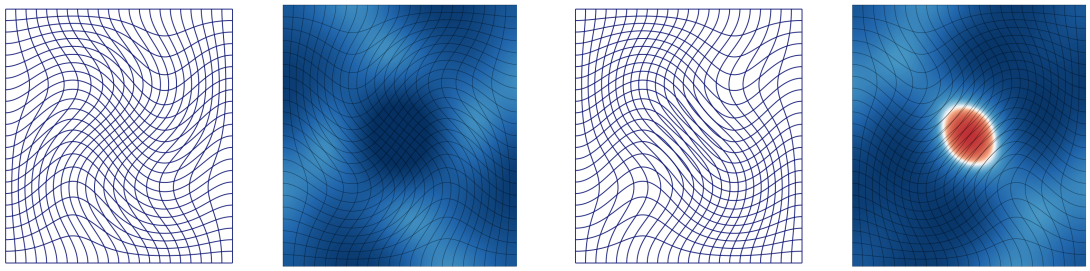
$$J_\Phi(\mathbf{p}) = \begin{pmatrix} \frac{\partial \Phi_x}{\partial x} & \frac{\partial \Phi_x}{\partial y} & \frac{\partial \Phi_x}{\partial z} \\ \frac{\partial \Phi_y}{\partial x} & \frac{\partial \Phi_y}{\partial y} & \frac{\partial \Phi_y}{\partial z} \\ \frac{\partial \Phi_z}{\partial x} & \frac{\partial \Phi_z}{\partial y} & \frac{\partial \Phi_z}{\partial z} \end{pmatrix}$$

If the Jacobian is non zero at the point p, then the transformation is locally invertible near p. Furthermore, the Jacobian informs us of the conservation of the orientation. If the Jacobian is negative, the grid Φ reverses the orientation, while it is preserved if the Jacobian is positive. Finally, the modification of the volume is given by the absolute value of the Jacobian.

To resume, a suitable transformation in terms of medical imaging registration should have a Jacobian strictly positive overall Ω. This can be enforced through an additional term in the equation 2.1.

Diffeomorphism One subset of topological preserving transformations are diffeomorphic transformations. Let define two manifolds U and V. A function f from U to V is a diffeomorphism if *(i)* f is a bijection from U to V, *(ii)* f is differentiable over U and *(iii)* the inverse f^{-1} is differentiable over V. The two manifolds U and V are then diffeomorphic. The Jacobian of a diffeomorphic transformation will be non zero everywhere on U as the function f is invertible. In the figure 2.2, we represented two different grids from $\Omega = [0, 1] \times [0, 1]$ to itself and their respective Jacobian. The first one is a diffeomorphic transformation, while the second one is not. The second grid's non-topological points correspond to the position where the Jacobian becomes equal to zero (white colour in the figure). The diffeomorphic character of a transformation can be enforced by the formulation of the regularisation term $\mathcal{R}$ or the deformation model's design. Among the diffeomorphic registration algorithms, we can quote the diffeomorphic Demons approach [Vercauteren, 2009], flows of diffeomorphisms like the large deformation diffeomorphic metric mapping (LDDMM) [Yan Cao, 2005; Beg, 2005] or B-splines approach with a penalisation of negative Jacobian values [Rueckert, 2006].

As we previously mentioned, the deformed image $M(\Phi)$ is obtained through interpolation, as the transformation generates non-integer positions. Because of interpolation and numerical approximation, we cannot always ensure topological properties are respected. Thus, a theoretically diffeomorphic formulation could still have zero values for its Jacobian and, therefore, non-topological points.



(a) Diffeomorphic transformation (b) Non diffeomorphic transformation

Figure 2.2: Comparison between two differents transformation, a diffeomorphic one and a non diffeomorphic one. For each transformation, we represent the deformation grid together with its Jacobian. Blue, white and red correspond to respectively positive, null and negative values of the Jacobian. Folding and crossings appears at the position where the Jacobian is equal to zero.

2.3.3 Evaluation Metrics

The evaluation of the registration performance is challenging due to the ill-posed nature of the problem. However, different metrics had been proposed in the literature. These metrics can be classified into different types: geometric metrics, intensity-based metrics and grid quality metrics.

Geometric metrics evaluate the registration performance through the displacements of organs or landmarks annotated by medical experts. More specifically, they measure the distance between the deformed structures and the target structures. Such metrics include the Dice Coefficient [Sorensen, 1948; Dice, 1945], which measures the overlap between two segmentations and the Hausdorff distance [Huttenlocher, 1993], which measures the greatest distance between a point in one set to another set and vice versa. Outliers, far away from the segmentation, penalise the Hausdorff distance, while these outliers do not significantly impact the Dice coefficient. Another geometric metric is the target registration error (TRE). To calculate it, we must first extract landmarks on both F and M. Then, we calculate the distance between the displaced landmarks and the target landmarks. Obtaining these landmarks in an automatic and reproducible way is a significant challenge of this metric. In general, geometric metrics need to have supplementary information to be calculated, such as segmentation masks for both fixed and moving images for Dice and Hausdorff and landmarks for the TRE.

The geometric metrics do not provide information about the grid's noise or if the transformation follows some topological properties. In recent papers [Kim, 2020; Mok, 2020a], the authors added grid quality metrics to better compare registration algorithms. The first of these metrics is the standard deviation of the Jacobian (or log Jacobian) written $\sigma(|J_\Phi|)$. As the Jacobian reflects the local behaviour of Φ, a high value of $\sigma(|J_\Phi|)$ reflect a noisy transformation. However, the standard deviation of the Jacobian does not characterise Φ as being a diffeomorphism or not. Indeed, we could have an important standard deviation but a Jacobian strictly positive everywhere on Ω. To evaluate the conservation of the orientation and the invertibility, we measure the percentage of voxels where the Jacobian is negative ($\%$ of $|J_\Phi|_{\leq 0}$). As this value gets smaller, the grid will preserve the topological properties better. A transformation having the Jacobian positive everywhere over Ω will

be a diffeomorphism.

Finally, intensity-based metrics perform a one to one comparison of each voxel and measure differences. Among these metrics, we can quote the mean absolute error or mean square error (MAE and MSE) [Sotiras, 2013; Brown, 1992], the local cross-correlation (LCC) [Jeongtae Kim, 2004; Avants, 2008] or mutual information (MI) [Wells, 1996; Collignon, 1995]. MAE and MSE require that similar structures have the same intensity range. Therefore they are not appropriate for multi modal registration or in the case the volumes have different noise level (coming from two different machines for instance). MI is inspired by the information theory and measures the mutual dependence between two random variables. As MI does not need similar intensities for the two images, it became very popular for multi-modal registration [Wells, 1996; Collignon, 1995]. Formulations of the MSE and the LCC are the folllowing :

$$MSE(D,F) = \frac{1}{|\Omega|}\sum_{p\in\Omega}|F_p - D_p|^2$$

$$LCC(D,F) = \frac{1}{|\Omega|}\sum_{p\in\Omega}\frac{\left(\sum_{p_i}(F_{p_i}-\overline{F}_p)\cdot(D_{p_i}-\overline{D}_p)\right)^2}{\sum_{p_i}(F_{p_i}-\overline{F}_p)^2\cdot\sum_{p_i}(D_{p_i}-\overline{D}_p)^2}$$

where D is the deformed image ($D = M(\Phi)$), $\overline{F}$ and $\overline{D}$ are the local mean value of F and D calculated over a small window of size $k \times k \times k$ centered on p, p_i are iterating over this volume. Recently, the use of the local cross-correlation as a similarity criterion reported better results combined with deep learning algorithms compared to mean square error, gaining more and more attention. Authors of [Balakrishnan, 2018] proposed an efficient numerical implementation of the LCC using convolutional kernel and thus ensuring differentiation.

Public evaluation The evaluation and comparison of registration algorithms is a challenging task. Some publications still set up a fair and reliable evaluation of registration on different body parts. In Klein et al. [Klein, 2009], authors evaluated 14 different deformable algorithms. The evaluation was performed on eighty brain MRIs, using eight different metrics and statistical tests. The best performing algorithms were ART [Ardekani, 2005], SyN [Avants, 2008], IRTK [Rueckert, 1999], and SPM's DARTELToolbox [Ashburner, 2007]. These algorithms were the most recent, with the highest number of parameters (at the time of the paper's publication). Following this framework, the EMPIRE10 challenge (Evaluation of Methods for Pulmonary Image REgistration 2010, [Murphy, 2011]) was organised in conjunction with the MICCAI 2010 conference. The twenty participating teams applied their algorithms to a common dataset of thirty pairs of thoracic CT. Organisers ranked the participants using several criteria: the alignment of lung boundaries, the alignment of major fissures, and the Jacobian determinant's negative values. More recently, six algorithms' performance has been assessed for registration of abdominal CT [Xu, 2016]. The dataset comprised thirteen segmented organs, hundreds of CT volumes while the selected algorithms were : FSL [Jenkinson,

2012], IRTK [Rueckert, 1999], NiftyReg [Modat, 2010], ANTs [Avants, 2008; Avants, 2009] and Deeds [Heinrich, 2013]. The metrics used in this challenge included the Dice similarity coefficient, the mean surface distance and the Hausdorff distance, calculated on each organ. Among the different benchmarked methods, Deeds achieved the best performance. However, this study showed the serious difficulty of abdominal registration compared to other anatomies. The dice coefficient was below acceptable values for most of the organs, and the algorithms produced huge local foldings. Even the best-performing algorithms provided insufficient results for clinical application.

More recently, two registration challenges have been proposed: the CuRIOUS Challenge [Xiao, 2020] and the Learn2Reg Challenge [Dalca, 2020; Hering, 2021]. The first one was a public competition on multimodal registration (MRI and US) and brain shift correction. The MRIs were preoperative images, and the ultrasounds were acquired during the surgery. The second one provided four different datasets to create a benchmark dataset to evaluate registration algorithms. The four tasks included brain MRI& US registration, CT lung registration between inspiration and expiration, CT abdominal inter-patient registration and hippocampus MRI registration. In the Learn2Reg, the organisers used six different metrics to rank the participants: the Dice coefficient, the 30% quantile of Dice, the Hausdorff distance, the TRE, the standard deviation of the Log Jacobian and the calculation time. Such efforts greatly facilitate the evaluation and the comparison of the registration methods, boosting the development on the topic.

2.4 Traditional Deformation Models

The registration task has been studied for many years by the computer vision and medical imaging communities. Therefore, there is a tremendous number of formulations and registration models that are based on non-deep learning algorithms. This section summarises some of these advances and is built on two significant analyses of the field [Oliveira, 2014; Sotiras, 2013].

In Sotiras et al. [Sotiras, 2013], the authors classify deformation models into three categories: physical inspired models, interpolation-based models and knowledge-based transformations. Other classifications can be produced based on the respect of desirable properties (such as symmetry or diffeomorphism, see 2.3.2) or on the choice of the optimisation methods (continuous optimisation with differentiable objective function versus discrete optimisation such as [Glocker, 2011]). Here, one should mention Demons, Large Deformation Diffeomorphic Metric Mapping (LDDMM), SyN, and stationary velocity field (SVF) among the most famous formulations.

Demons is a diffusion-based approach where the deformation is modeled by : $\Delta u + F = 0$. It was first introduced in Thirion [Thirion, 1998] and exploited by different researchers [Pennec, 1999; Vercauteren, 2007; Vercauteren, 2009]. In particular, Vercauteren et al. [Vercauteren, 2009] developed a diffeomorphic Demons approach. The LDDMM framework is a member of another group of deformation models, the flows of diffeomorphism. In this framework, we obtained the deformation by integrating its velocity over time [Joshi, 2000; Beg, 2005; Yan Cao, 2005; Ceritoglu, 2009; Hernandez, 2009]. It produces diffeomorphic deformation and provides a distance between

points following geodesics. Due to the integration over time, this framework requires a long calculation time and high memory usage. SyN is a similar approach to the LDDMM framework with the addition of preserving the properties of symmetry [Avants, 2008; Avants, 2009]. Researchers reduced the computational requirement of the LDDMM framework by considering a stationary velocity field. The integration is less demanding and often performed by the *squaling and squaring* approach ([Arsigny, 2006; Ashburner, 2007], see section 2.5.3).

Other algorithms include : free form deformations with b-splines [Rueckert, 1999; Schnabel, 2001]; elastic body models based on Navier-Cauchy equation [Davatzikos, 1997; Shen, 2002; Pennec, 2005]; viscous fluid models Navier-Stokes equation [Christensen, 1996]; statistical deformation models [Ashburner, 2000; Rueckert, 2003]. Recently many discrete registration algorithms have been proposed such as Deeds [Heinrich, 2013] or Drop [Glocker, 2008]. They are often based on graphical model and mainly on Markov Random Field (MRF). Similar discrete methods cand be found in Wyatt et al. [Wyatt, 2002], Glocker et al. [Glocker, 2009], and Parisot et al. [Parisot, 2012].

2.5 Deep Learning based Registration Models

The main difference between deep learning approaches and other approaches lies in the strategy used to register new pairs of images. While in non-deep learning (also referred to as *traditional*) approaches, the optimisation problem (equation 2.1) needs to be solved for each pair, for deep learning approaches, the optimisation problem is only resolved during the training phase. Similar to other learning schemes, the model construction is performed in two phases: the training and the prediction phase. In learning-based models, we model the registration by $g_\theta(F, M) = \Phi$ where θ are the set of learnable parameters and g is a function. During the training phase, we minimise an objective function similar to equation 2.1 to obtain the optimal parameters θ^* and during the prediction phase, we apply the function g_{θ^*} to new pairs of images M and F. Thus, DL-based registration has then two majors improvements. The first one concerns the calculation speed during the prediction phase, which is heavily decreased, allowing real-time registration. This acceleration comes not only from the training and prediction method but also from the GPU implementation of the deep learning-based frameworks. The second improvement is the benefit of learning from huge datasets, generating more robust and generalisable features instead of focusing on specific pairs of volumes, repeating the optimisation process per pair.

In the deep learning context, researchers mainly refer to the similarity and regularisation criteria of equation 2.1 as losses. Thus, we will refer to $\mathcal{D}$ as $\mathcal{L}_{sim}$ for similarity loss and to $\mathcal{R}$ as $\mathcal{L}_{reg}$ for regularisation loss in the rest of this thesis.

In [Boveiri, 2020], the authors analysed the different organs and modalities used in DL-based registration papers. The majority of the research has been performed on the brain by far, along with lung and cardiac anatomies. Several reasons include: the availability of datasets on brain anatomies, the wide range of clinical applications for brain and the registration's simplicity as the brain is surrounded by the skull, preventing large displacements. Most of the paper focuses on MRI and CT, with more than half on MRI. Following [Boveiri, 2020], we divide deep learning-based methods

into the following five categories: deep similarity metrics, reinforcement-based methods, supervised, unsupervised and semi-supervised methods. Other categories can be found in Fu et al. [Fu, 2020] and Haskins et al. [Haskins, 2020]. We detail the first three categories briefly and then focus particularly on unsupervised and semi-supervised approaches as they became the new trends and the state of the art [Boveiri, 2020].

2.5.1 Overview of Old Deep Learning Frameworks

Deep similarity metrics Deep similarity metrics methods were among the first methods, applying deep learning techniques for registration [Wu, 2013; Simonovsky, 2016; Cheng, 2018]. They focus on training a network to classify pairs of images/volumes as aligned or not. Once trained, this network could replace traditional metrics such as mutual information in traditional registration methods. Deep similarity metrics prove their strength for multimodal registration (MRI-CT or MRI-US) particularly [Fu, 2020; Boveiri, 2020]. A potential challenge is in defining a good metric to evaluate the registration of two images with different modalities, while it is less demanding to train a network to recognise the appropriate alignment o of these images. Various articles demonstrate that deep similarity methods outperform traditional metrics with several limitations [Fu, 2020; Boveiri, 2020]. This type of method relies on datasets with well-aligned pairs, and constructing such training datasets demands much effort. The challenge is then to ensure that these metrics respect the properties needed for optimisation. However, the largest limitation is that deep learning is only used to evaluate other algorithms' performances, and the registration is still performed by classical iterative registration algorithms, preventing real-time registration.

Reinforcement learning Reinforcement learning (RL) is an area of machine learning where an agent performs actions depending on a state to maximise a reward. Contrary to supervised learning, the agent does not need pairs of input and labels. In the registration context, the reward is the similarity measure between deformed and fixed volumes, and the actions are transformations (rotation, translation or pixel-wise deformations). The RL approach does not take the images as input but only the previous actions and the corresponding reward [Fu, 2020]. One limitation for the application of RL methods is the vast action space that needs to be explored for deformable approaches. Thus most of the RL approaches focus on affine transformations [Miao, 2018; Liao, 2017; Ma, 2017]. Still, in Krebs et al. [Krebs, 2017], the authors combined RL with deformable registration using a low dimension statistical deformation model. On top of the low dimension limitation, other ones are the iterative nature preventing fast registration and the optimisation which is performed for only one pair of images at a time [Fu, 2020]. With the development of supervised and unsupervised registration, deep similarity and reinforcement learning methods became less and less studied.

Supervised registration Supervised methods rely on ground-truths transformation to train the network [Cao, 2017; Rohé, 2017; Miao, 2016; Yang, 2017; Sokooti, 2017]. Thus they need a large number of image pairs with their deformation. The motivation behind this technique is the possibility of real-time prediction after the learning process. The ground truth transformation can be obtained from generated transformations or by using the ones obtained by traditional registration algorithms. Using traditional algorithms requires pre-calculating the transformation to construct the ground-truth database, thus taking a substantial preprocessing time. On the other hand, random transformations consist of applying a predefined transformation (affine or deformable) to one image and using the network to retrieve it. The assumption is that the network learns to generate real deformations after being trained with random ones. However, the performance of the generated displacement is highly correlated to the quality of the ground-truth grids. Furthermore, the network learns to imitate existing algorithms and thus does not learn the displacement by itself.

2.5.2 Unsupervised and Weakly-Supervised Registration

Unsupervised registration The idea of unsupervised registration lies in the optimisation of g_θ with a similarity loss calculated directly at the image level, between the fixed and the deformed image. Such modelling of the problem frees us from the requirement of building a dataset with ground truth deformations. To use deep learning and the backpropagation method, the warping operation, which produces the deformed image $\widehat{M}$ using Φ and M, must be differentiable. The development of spatial transformer [Jaderberg, 2016] made this operation applicable for deep learning architectures. The Spatial Transformer Network (depicted in figure 2.3) comprises a localisation network, a grid generator and a differentiable image sampler. The localisation network outputs θ the parameter of the transformation, the grid generator produces a grid with the new position of every pixel p, and the image sampler generates the deformed image using the grid and the initial image. In the case of DL-based registration, we keep only one part of the spatial transformer, the differentiable image sampler, while the registration network produces the grid, which has the same shape as the input image. Among the different ways to apply the sampling operator, the commonly used is the backward trilinear sampling operation where the expression of $\widehat{M}$ is :

$$\widehat{M}(\mathbf{p}) = \mathcal{W}(M, \Phi)(\mathbf{p}) = \sum_{\mathbf{q}} M(\mathbf{q}) \prod_{d \in \{x,y,z\}} \max(0, 1 - |\Phi(\mathbf{p})_d - \mathbf{q}_d|) \tag{2.2}$$

with $\mathbf{p}$ and $\mathbf{q}$ being the pixel location, $d \in \{x, y, z\}$ an axis on a 3D space, $\mathbf{q}_d$ the d-component of q and $\mathcal{W}$ the sampling operation. This formulation is differentiable, and the gradients' expression can be found in Jaderberg et al. [Jaderberg, 2016]. The maximum operation is used in the previous equation in order to restrain our interpolation to the neighbouring points of p. This equation could be rephrased by calculating the sum for the voxels q in the neighbourhood of p and removing the maximum. But the maximum is needed to have differentiable operations. Other interpolation differentiable formulations could be used, such as nearest neighbour warping.

Since the first publications combining spatial transformer and unsupervised registration [Balakrish-

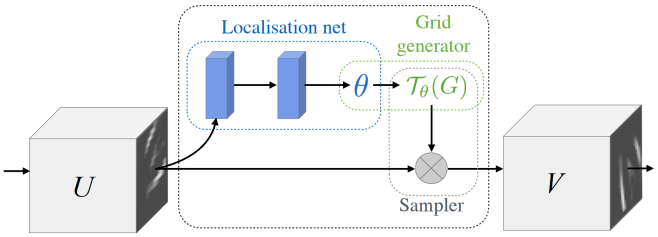
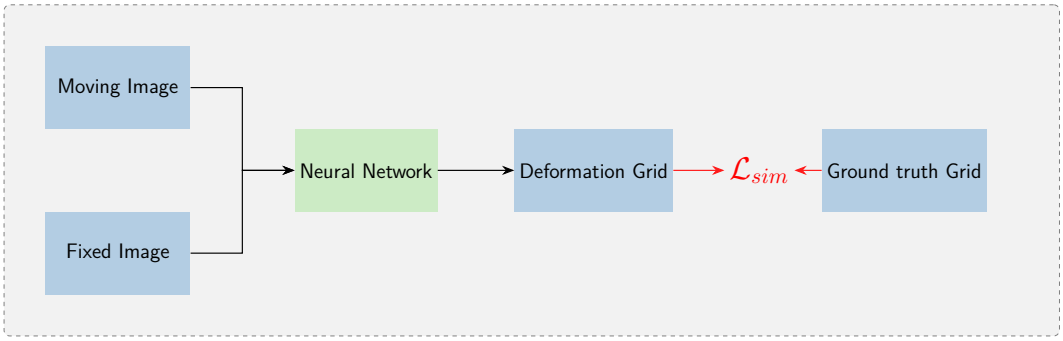


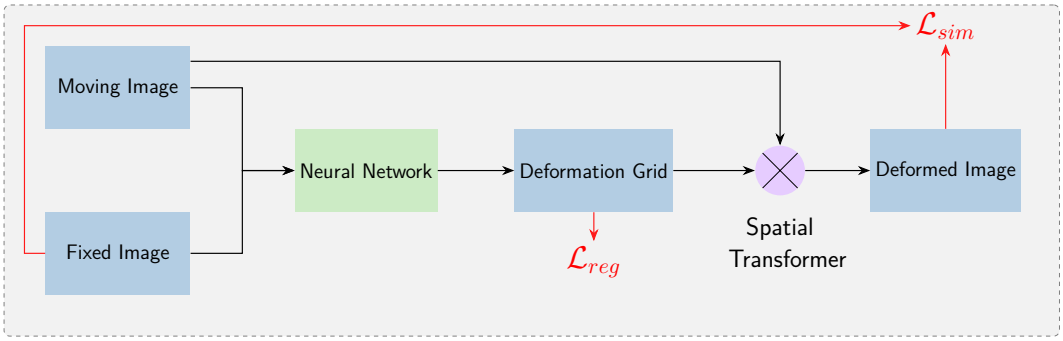
Figure 2.3: Spatial Transformer (Image from Jaderberg et al. [Jaderberg, 2016])

nan, 2018; de Vos, 2017; Yoo, 2017; Ghosal, 2017; Li, 2017; Stergios, 2018], the field has experienced a strong growth with many studies. Different extensions have been proposed to the general unsupervised pipeline such as : multi-modal registration [Xu, 2020; Wang, 2019; Qin, 2019]; multi-scale architecture [Hering, 2019; Fechter, 2020; Mok, 2020a]; cycle consistency loss with two networks predicting the forward and backward grid [Zhang, 2018; Kim, 2019; Kim, 2020; Mok, 2020b; Guo, 2020; Wang, 2020]; diffeomorphic approaches (more details in 2.5.3) and exploration on a better regularisation and deformation constraints [Kuang, 2019a; Mansilla, 2020; Hering, 2020]. Unsupervised approaches have been developed for different organs such as brain MRI [Balakrishnan, 2018; Dalca, 2018], cardiac MRI [Krebs, 2018; de Vos, 2019] or multimodal approach [Ferrante, 2018; Qin, 2019]. Two other main frameworks, inspired by unsupervised registration, are generative adversarial networks (GAN, [Goodfellow, 2014]) and weakly supervised frameworks.

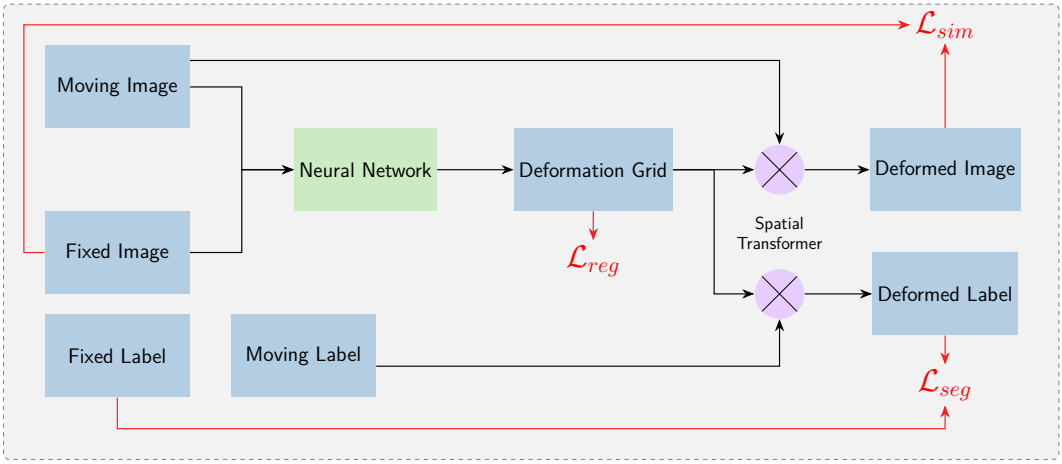
Adversarial frameworks In a classic GAN architecture, the generator is trained to create realistic images and fool the discriminator while the discriminator learns to distinguish real images from fake ones. This competitive strategy provides excellent performances in the computer vision field. In registration, we already have a network playing the role of the generator. Thus only the discriminator needs to be added to the registration network. Contrary to the traditional GAN framework, the deformations are not generated from random distributions but from the moving and fixed images [Hu, 2018a; Fan, 2018; Tanner, 2018; Mahapatra, 2018a]. The discriminator is trained for two different goals. On the one hand, it can increase the registration performance by separating well-aligned pairs from misaligned ones. On the other, it can provide a better regularisation by discriminating real images from warped ones, thus creating more realistic deformation. The GAN framework is particularly suited for multi-modality registration, where the different modalities have various ranges of intensities.



(a) General framework of supervised registration



(b) General framework of unsupervised registration



(c) General framework of weakly supervised registration

Figure 2.4: Comparison between different framework of DL-based registration

Weakly-supervised registration Weakly supervised approaches regroup methods using supplementary labels during the training process. These labels are mostly segmentation masks or anatomical landmarks. This additional information usually improves the registration performance as the networks identify correspondences not only based on the intensity but also based on specific organs and interesting locations, preserving better anatomical structures. One significant advantage of these approaches is that the supplementary labels are only used during the training process, not during the inference. Indeed, labels are not inputs of the network but are used to calculate the additional supervised loss (sometimes called segmentation loss). Thus, we only need imaging data to register new pairs of patients. Concretely, the network generates the deformation grid from the image pairs, and the labels are displaced following the grid. Let define M_{seg} and F_{seg} as the labels or segmentation masks of the moving and fixed volume. The moving labels are deformed using the predicted grid Φ and a similarity loss is calculated between the warped label $\widehat{M}_{seg}$ and the fixed label F_{seg}. The warped label formulation is : $\mathcal{W}(M^{seg}, \Phi)$. The similarity loss is different from the loss applied between images, and its formulation depends on the type of provided labels. In the case of segmentation labels, most weakly supervised methods used a Dice loss or a cross-entropy loss.

Some of the first weakly supervised approaches were published in [Hu, 2018b; Hering, 2018; Balakrishnan, 2019] for respectively prostate multimodal registration (MR-US), cardiac MR registration and brain MR registration. [Hering, 2018; Balakrishnan, 2019] add the loss of the deformed labels to the general loss formulation while [Hu, 2018b] optimised their method only with the loss of the label. Weakly supervised methods became incrising popular due to the impact on registration performance [Boveiri, 2020]. New registration strategies, combining images and labels, are currently being investigated. For instance, in [Mansilla, 2020], the authors proposed a novel approach with a supplementary encoder. They used segmentations to enforce an anatomical regularisation of the deformation and thus predict more realistic transformations.

2.5.3 Grid Formulation

The choice of the deformation model is one key element of the resolution of the registration problem. In this section, we focus on the formulation used in the context of DL models. In recent publications, we can find different grid formulations. These models differ in how the transformation Φ is produced from the output of the neural network. In order to use gradient backpropagation, all the calculation steps need to be differentiable, reducing the number of model's formulations compared to classic registration. The predicted transformation is more likely to respect suitable properties, or not, such as diffeomorphism, depending on the chosen formulation.

The first formulation is to output Φ directly by the neural network. This formulation does not impose any constraints on the grid except the one given by the regularisation loss $\mathcal{L}_{reg}$. Another very similar one is the displacement-based formulation. In this situation, the transformation at every position $\mathbf{x}$ is given by the addition of the identity transformation with a displacement field $\mathbf{u}$ and the neural network outputs the displacement field $\mathbf{u}$ [Balakrishnan, 2019]. For each $\mathbf{p} \in \Omega$, we define $\Phi(\mathbf{p}) = \mathbf{p} + u(p)$. In case of small deformation, we can approximate the reverse transformation

by $\Phi^{-1}(p) \approx \mathbf{p} - u(p)$. This formulation of the inverse grid is a major approximation and is even inaccurate in case of large deformation (see [Ashburner, 2007]). In this article, the author displayed the composition of the forward and the approximated backward and it resulted in a transformation far from the identity transform.

In order to have better topological preservation, a diffeomorphic approach are proposed in : [Dalca, 2018; Krebs, 2019; Mok, 2020b]. In the diffeomorphic approach, the neural network produce a stationary velocity field (SVF). This method assumes that the velocity field v is constant over time $t \in [0, 1]$. The velocity field and the transformations are linked through the following differential equation :

$$\frac{\partial \Phi_t}{\partial t} = v(\Phi_t) \tag{2.3}$$

To obtain the final deformation Φ at time $t = 1$, we integrate the stationary velocity field v over $[0, 1]$ using the initial condition that Φ is equal to the identity transformation at $t = 0$ ($\Phi_0(p) = p$). In terms of algebra, v is a member of one Lie algebra, and the deformation is the member of a Lie group with the relationship $\Phi_1 = \exp(v)$. The deformation field obtained through the exponentiation of a flow field has a Jacobian always positive. Φ is thus a diffeomorphism, and we can produce the inverse transformation.

One famous numerical efficient method to perform the integration is the *scaling and squaring* approach Arsigny et al. [Arsigny, 2006] and Ashburner [Ashburner, 2007]. This approach was first used to compute matrix exponentiation and is based on two main ideas. The initialisation is made using a first order Euler integration : $u(x + h) = u(x) + h \cdot u'(x)$, with u being a function, x a point and h the integration step. For the recurrence, we use a property of one-parameter subgroup : $\exp((t_1 + t_2)v) = \exp(t_1 v) \circ \exp(t_2 v)$, with $\circ$ being the composition operation. Finally, using a number of steps being a power of 2 and noticing that the deformation at $t = 0$ is the identity transformation, we obtained the following numerical scheme :

$$\begin{aligned}
\Phi_{1/2^N} &= p + \tfrac{v}{2^N} \\
\Phi_{1/2^{N-1}} &= \Phi_{1/2^N} \circ \Phi_{1/2^N} \\
&\vdots \\
\Phi_{1/2} &= \Phi_{1/4} \circ \Phi_{1/4} \\
\Phi_1 &= \Phi_{1/2} \circ \Phi_{1/2}
\end{aligned}$$

We should notice that we perform only N composition starting from a small initial deformation. The scaling term comes from the division of the stationary velocity field by 2^N to have a field close to zero, while the squaring term comes from the composition of the grid at time $t = 1/2^N$ by itself to obtain the grid at time $t = 1/2^{N+1}$. [Dalca, 2018; Krebs, 2018] developed simultaneously a *squaling and squaring layer* and made it possible to use this algorithm in a neural network framework. One supplementary benefit of this method, is that we can obtained the reverse transformation $\Phi^{-1} = \Phi_{t=-1}$ by backward integration, starting with $\Phi_{-1/2^N} = p - \frac{v}{2^N}$ and using the same numerical scheme.

A third formulation has been proposed by Stergios et al. [Stergios, 2018] and Shu et al. [Shu, 2018], using the gradient of the registration grid. Instead of predicting directly the deformation, the network predict its gradients $\nabla_x \Phi_x$, $\nabla_y \Phi_y$ and $\nabla_z \Phi_z$. The value predicted at one point p represented the displacement between this point and the previous pixel. If the predicted value is between 0 and θ, the distance between two pixels will decrease. On the other hand, if it is superior to θ, the two pixels will separate (θ being a positive real depending on hyper-parameters choice such as the last activation function). Enforcing the value of the spatial gradient to be positive will prevent non-topological deformation. Indeed, as the new position of pixel p is equal to the new position of $p-1$ plus the predicted value at p, the pixel's order is maintained on the deformed image. We reconstruct the deformation by an integration operation along each axis. A cumulative sum can approximate this operation. This formulation's advantage is the simplicity of the sum operation compared to the scaling and squaring operation of the diffeomorphic formulation. But having a constraint only on $\nabla_x \Phi_x$, respectively y and z, we ignore the other axis's influence on the displacement over one axis. Therefore, we put only constraints on the Jacobian matrix's diagonal term, and this formulation cannot enforce a positive Jacobian and thus a diffeomorphism.

2.6 Conclusion

In this chapter, we first described the development of deep learning in medical imaging, including the main tasks, architectures, databases and challenges. Then we presented the mathematical formulation of the registration problem, as well as the different components and properties of registration algorithms. To conclude, we gave an overview of the different registration algorithms, presenting both classical algorithms and deep learning-based algorithms. We detailed especially the progression of deep learning-based registration, with a special attention to unsupervised and weakly-supervised framework.

Chapter 3

Joint Segmentation-Registration into a Multi-Task framework

"*Souquez les artimuses !*
Du calme, ma fille, du calme ! En plus ça veut rien dire souquez les artimuses "

Astérix et Obelix : Mission Cléopatre

Contents

In chapter 2, we introduced the theoretical concepts of registration and several registration algorithms developed before and after the deep learning area. This chapter presents the concept of multi-task deep learning, and in particular, joint segmentation and registration. While the combination of these two tasks was widespread for classical registration algorithms, few DL-based formulations take advantage of their coupling. Our method consists of an architecture with one unique encoder and two independent decoders; one produces the segmentation mask, the other outputs the deformation grid. We reinforced the coupling by adding a supplementary loss. This chapter has been presented at the MICCAI 2019 conference in Estienne et al. [Estienne, 2019].

3.1 Introduction

Among the different medical imaging tasks, registration and segmentation are two well-studied problems. Image segmentation consists in label each pixel as belonging to an organ, a lesion or a type of tissue. Image registration search for the best transformation that map two different images to the same shared space. The combination of these two tasks was popular before the development of deep learning algorithms. For instance, segmentation could be performed by warping an image to an atlas using registration and then warping the atlas labels using the reverse transformation. Several researchers focus on developing a stronger relationship between these two tasks hypothesising that this will improve each task's results [Yezzi, 2001; Wyatt, 2003]. Joint formulation approaches were particularly investigated concerning brain MRI. It could concern brains without any lesions, segmenting brain structures or images in the presence of lesions in this case. In this situation, the registration task becomes challenging as the lesions create an absence of correspondences [Parisot, 2012; Gooya, 2011b]. This chapter focuses on images without tumours while we address the opposite situation in chapter 4. Nowadays, deep learning-based algorithms have become state of the art for both tasks. The development of 3D fully convolutional architecture set up a new baseline for the segmentation task [Çiçek, 2016; Milletari, 2016], while the introduction of the spatial transformer allowed differentiable interpolation and thus unsupervised registration, removing the need for ground-truths transformation [Jaderberg, 2016; Balakrishnan, 2018]. DL based methods achieve similar performance, but they also reduce prediction time, taking advantage of GPU implementation. However, registration and segmentation coupling have been less studied in the deep learning framework than with classical approaches.

In this chapter, we present a new strategy based on deep learning for simultaneous optimisation of registration and segmentation. Our work is inspired by the multi-task formulation, where one network solves different problems jointly. We expect that learning both tasks help generate more informative features and obtain more robust performances when applying to new unseen datasets. Our deep learning framework takes as input the two volumes and predicts, at the same time, segmentation masks and deformation grids. Contrary to other methods, the two tasks are not processed by two different networks, but they share a common encoder and are trained together. A specific decoder produces the moving segmentation, while the fixed segmentation is generated by the combination of the predicted segmentation and the deformation grid. In addition to the classical segmentation and registration losses, we reinforce the joint formulation with a new loss function, which updates both parts of the network. Our experiments are performed on brain MRI, with the segmentation network outputting three major brain structures and the registration being evaluating on 15 annotated anatomical structures. Finally, we compare our joint formulation with unsupervised and weakly supervised registration as well as with segmentation networks alone.

Our main contributions are :

- A multi-task framework which predicts at the same time segmentation masks and deformation grid;

- A simultaneous training, with the implementation of a new loss to reinforce the joint formulation;
- Combining both rigid and elastic transformations.

This chapter is organised as follows: In section 3.2, we present similar work focusing on multi-task frameworks and joint segmentation-registration for deep and non-deep learning algorithms. Our methodology is introduced in section 3.3 including the joint formulation, the network architecture and the training process. Finally, section 3.4 presents our experiences and results, while we discuss and conclude this chapter in section 3.5.

3.2 Related Work

3.2.1 Multi-task Learning

Multi-task Learning (MTL) is a machine learning approach that combines different tasks, expecting to improve the model performance on all the tasks and obtain a better generalisation on unseen datasets [Caruana, 1997]. The intuition behind this lies in the fact that the different tasks should share a common representation, and by combining them, the obtained trained representations could be more informative. Multi-task learning is often decomposed in two different approaches: the hard parameter sharing and the soft parameter sharing [Ruder, 2017; Zhang, 2021]. The hard parameter sharing consists of sharing the first layers of the network for all the tasks while having task dependant layers at the end of the network. Conversely, in the soft parameter sharing, each task has its own independent network, but a loss is introduced to enforce the network's parameters to have similar values. The MTL framework was first applied in the computer vision field. For instance, Zhang et al. [Zhang, 2014] combined landmarks detection's task with some auxiliary classification tasks, including additional attributes such as appearance characteristics (wearing or not glasses), gender, expression or head pose. Similarly, Luvizon et al. [Luvizon, 2018] proposed a method to predict action recognition and pose estimation simultaneously, and Kendall et al. [Kendall, 2018] built an architecture, returning semantic segmentation, instance segmentation and depth estimation.

Different algorithms using multi-task learning in the medical image domain also exist. Moeskops et al. [Moeskops, 2016] proposed an architecture combining different segmentation tasks for different available modalities. A unique CNN was used to produced anatomical structures, including brain tissues in brain MRI, pectoral muscle in breast MRI and coronary artery segmentation in cardiac CT angiography. Yan et al. [Yan, 2019] developed a multi-task network for lesion analysis inspired by the Mask R-CNN framework [He, 2018] and applied it on the DeepLesion dataset [Yan, 2018]. Their network has a backbone network to extract features and three head branches: one detection branch to perform bounding box regression and lesion classification, one tagging branch to predict different clinical attributes such as body part or lesion type and a segmentation branch to provide fine masks. Due to the Covid-19 pandemic, many researchers focused on the segmentation of Covid-19 lesions. In [Amyar, 2020], the authors combined three Covid related tasks in a multi-task framework. Their architecture consisted of a shared encoder and one decoder to perform CT reconstruction, a

second decoder to segment the lesions and fully connected layers to classify Covid positive patients from Covid negative. Similarly, Myronenko [Myronenko, 2019] add a reconstruction branch to a segmentation network to regularise the shared encoder. The same two-branch network was used for brain tumour segmentation and won the BraTS 2018 challenge [Bakas, 2019].

Most of the cited multi-task approaches applied the hard parameter sharing, using one encoder or features extractor part and multiples decoder or task-specific layers.

3.2.2 Approaches for joint Segmentation and Registration

The fusion of segmentation and registration tasks has been investigated for many years. Significant researches have been published in the early 2000s. They mainly focus on brain MRI and the segmentation of healthy brain structures such as white matter or ventricles. Yezzi et al. [Yezzi, 2001] proposed an approach based on active contours and energy minimization, Wyatt et al. [Wyatt, 2003] used a Markov random field framework and Pohl et al. [Pohl, 2006] an Expectation Maximization-based algorithm (EM).

Combining the two tasks become more challenging in the context of brain gliomas. Indeed, the tumour creates a lack of correspondence between the two different images. Different approaches can be found in the literature. Stefanescu et al. [Stefanescu, 2004] first segmented pathological regions and then performed the registration while increasing the regularity in these regions. Bach Cuadra et al. [Bach Cuadra, 2006] combined a variational flow for the registration and a tumour growth model to perform the registration. In Gooya et al. [Gooya, 2011b] and Gooya et al. [Gooya, 2012], the authors proposed a method to register an atlas with a cancerous brain MRI. The original atlas was modified using an algorithm that simulates the tumour growth. The modified atlas is then registered to the patient space. Both registration and growth model are obtained using the EM algorithm. Parisot et al. [Parisot, 2012; Parisot, 2014] proposed a different approach based on a discrete graphical model. The segmentation and registration are performed onto a sparse grid. Depending on their classification, the healthy and non-healthy nodes will be processed differently by the registration algorithm.

The coupling of segmentation and registration can be found in more recent deep learning approaches. Methodologies differ by being fully DL-based or not and by the proposed architectures. In Vakalopoulou et al. [Vakalopoulou, 2018], authors registered lung CT images to N different atlas and then trained N independent segmentation networks. They improved segmentation accuracy by warping them backwards and combining. However, the registration task is performed through a graphical approach and does not benefit from the segmentation task. Mahapatra et al. [Mahapatra, 2018b] and Elmahdy et al. [Elmahdy, 2019] proposed adversarial approach for respectively chest Xray and prostate CT joint registration-segmentation. These two studies have some limitations: In Elmahdy et al. [Elmahdy, 2019], segmentations are not predicted, but ground truths masks are used to introduce a supplementary loss, which corresponds to weakly supervised registration and not joint registration-segmentation (see section 2.5.2). In Mahapatra et al. [Mahapatra, 2018b], the segmentation masks are obtained using Otsu's thresholding on the registration network's activation map. This

article's major limitation lies in the use of ground truths grids generated by artificial deformation. [Li, 2019] and [Xu, 2019] are the closet methods to our. Both of them use two different networks, one for registration and one for segmentation, and implement an additional loss that optimises both networks. There is still an important difference with the work presented in this chapter: our framework consists of a unique network, while they employ two. Finally, [Sinclair, 2020] proposed a deep learning framework to construct atlas by learning both registration and segmentation.

3.2.3 Segmentation

As pointed in section 2.2, deep learning methods have become state-of-the-art for medical imaging segmentation. Different challenges have been organised among which the multimodal brain tumor segmentation challenge (BraTS) [Menze, 2015; Bakas, 2019], the medical segmentation decathlon (MSD) [Simpson, 2019] or the kidney tumor segmentation (KiTS19) [Heller, 2020]. Among the different deep learning architectures, fully convolutional autoencoders are widely used for segmentation, such as the UNet [Ronneberger, 2015; Çiçek, 2016] or the VNet [Milletari, 2016]. The organisation of segmentation challenges brought many variations to the original UNet architecture. Among these modifications, we can quote: the use of a second decoder to perform reconstruction regularisation [Myronenko, 2019], an ensemble of different models [Kamnitsas, 2018; Zhou, 2019a], multiple stages and multi-view cascaded networks [Wang, 2018; Jiang, 2020], or simply a list of good practices for training a segmentation network [Zhao, 2020b]. Other top-performing methods at the BraTS competitions used generic UNet architecture with data augmentation and post-processing [Isensee, 2019], dilated convolutions and label uncertainty loss [McKinley, 2019], context aggregation and localisation pathways [Isensee, 2018] or a combination of deep learning architectures together with algorithms such as conditional random fields (CRFs) [Chandra, 2019]. A more detailed comparison and presentation of the last years' challenges on BraTS is presented and summarised in [Bakas, 2019].

Recently, the nnUNet architecture was proposed in [Isensee, 2019; Isensee, 2021]. The authors claimed that except for the different variations on the architecture, a very important part of the performance lies in the training process, data handling (pre and post-processing) and data augmentation processes. nnUNet reports very high performances for various medical challenges and is now considered as state of the art for segmentation processes. Finally, the majority of the segmentation networks are trained with objective functions specific to medical image segmentation, mainly the Dice loss [Sudre, 2017]. Other classification losses could be used, such as cross-entropy, but they often achieve worse results.

Nowadays, research on segmentation focuses more on detecting small lesions like metastasis or weakly and unsupervised training. Indeed, the major drawback of current segmentation pipelines is the need for ground truths masks. Obtaining these masks requires much time from clinicians and radiologists. New algorithms explore the use of inequality constraints or weak annotations such as bounding boxes, scribbles or estimated percentage of labels [Kervadec, 2019; Lerousseau, 2020].

3.3 Methodology

This study proposes a joint architecture to predict both segmentation and registration based on a convolutional neural network (CNN). Our network takes two brains MRI, depicted as moving image M and fixed image F, and predicts the grid to warp M to F and the segmentation of the moving image M_{seg}. Our architecture comprises one shared encoder and two independent decoders, represented in Figure 3.1. Contrary to other methods, the two tasks share parameters, and the optimisation is performed simultaneously.

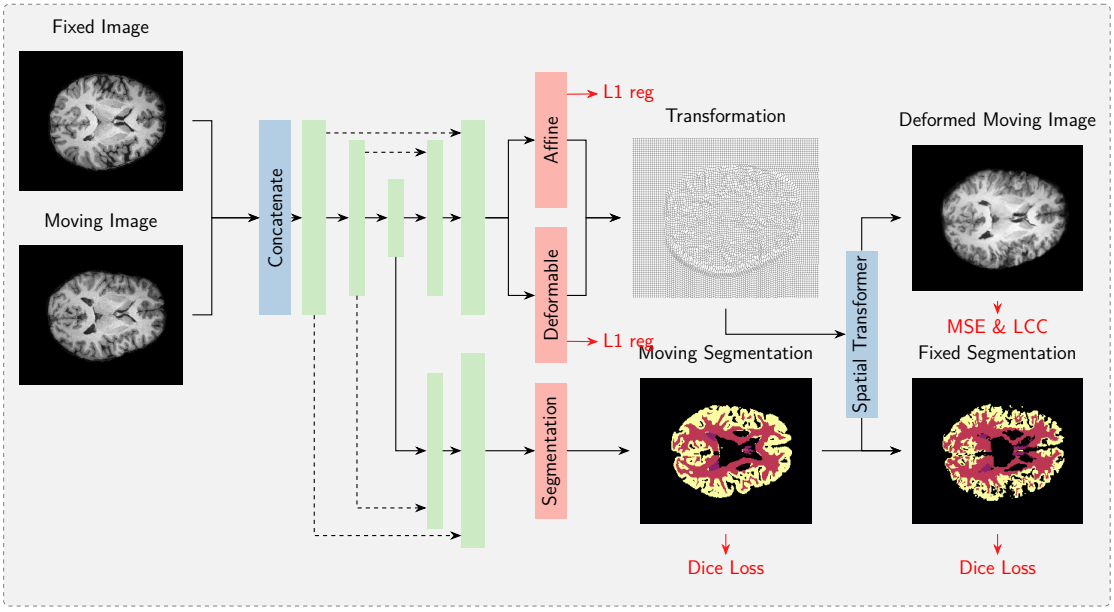


Figure 3.1: The employed architecture with the registration and segmentation parts. Green layers: convolutional blocks with successive Convolution, LeakyReLu and down/upsampling operations.

3.3.1 Joint Formulation

Our network can be defined into three parts: the encoder $\mathbf{E}$, the registration decoder $\mathbf{D_{reg}}$ and the segmentation decoder $\mathbf{D_{seg}}$. Sharing a unique encoder allows us to take advantage of the coupling better. The two images M and F are concatenated before being passed through $\mathbf{E}$.

We formulated our registration network following the architecture proposed in [Stergios, 2018]. Our predicted deformation is decomposed into a rigid transformation and an elastic transformation. The output of the registration decoder $\mathbf{D_{reg}}$ goes to two different sub-networks. The affine block generated a 3×4 matrix A, while the deformable block produced a matrix with the same shape as the input images.

In this manuscript, we follow the gradient formulation from [Stergios, 2018; Shu, 2018]. Our deformable block outputs the deformation gradient along the three-axis x, y and z. Then, we obtain the actual grid Φ by applying an integration operation of $\nabla_x \Phi_x$, $\nabla_y \Phi_y$ and $\nabla_z \Phi_z$ along their respective axis. We approximate the integration by a cumulative sum operation. Predicting the gradient, instead of estimating the sampling coordinates directly, reduce non-topological deformations. Indeed, we constrain the gradient to be positive by applying a sigmoid function, and thus we prevent crossings of the deformed voxels. Two pixels p and $p+1$ move closer, maintain distance, or move apart in the warped image, if $\nabla_d \Phi_d(p)$ is respectively less than 0.5, equal to 0.5, or greater than 0.5 (with d being an axis in x, y, z).

After obtaining A and Φ, we apply them sequentially to the moving image : first the linear transformation, then the deformable one. A 3D spatial transformer deforms (or warps), the moving image M with the affine grid and the deformation grid Φ. In details, the warped moving image $\widehat{M}$ is equal to :

$$\widehat{M} = \mathcal{W}\left(\mathcal{W}(M, A), \Phi\right) \tag{3.1}$$

where $\mathcal{W}(., \Phi)$ indicates the sampling operation under the deformation Φ, which is done using a backward trilinear interpolation sampling (see [Jaderberg, 2016] and equation 2.2.

Our segmentation is produced by using the same encoder $\mathbf{E}$ and a different decoder. It worth noting that we produce segmentation of brain structures and not lesions or tumours. As depicted in figure 3.1, the two images M and F are treated differently from the segmentation point of view. The segmentation decoder predicts the moving image's mask, $\widehat{M}_{seg}$, while the fixed image's mask will be generated by applying the predicted deformation as in equation 3.1. The explicit formulation of our predicted segmentation masks is :

$$\begin{aligned} \widehat{M}_{seg} &= \mathbf{D_{seg}}\left(\mathbf{E}(M, F)\right) \\ \widehat{F}_{seg} &= \mathcal{W}\left(\mathcal{W}(\widehat{M}_{seg}, A), \Phi\right) \end{aligned} \tag{3.2}$$

with $\mathbf{E}$ and $\mathbf{D_{seg}}$ being the encoder and segmentation decoder and $\widehat{M}_{seg}$ and $\widehat{F}_{seg}$ respectively the predicted moving and fixed segmentation. The formulation of the predicted fixed segmentation comes together from the segmentation decoder and the registration decoder. Indeed, the output of the segmentation decoder $\widehat{M}_{seg}$ is deformed by the affine and deformable transformations, coming from the registration decoder. We train our segmentation network together with the registration one in an end to end process.

3.3.2 Optimisation Strategy

To train the network, we use a combination of multiple loss functions, each of them related to a specific task. Following the studies of [Stergios, 2018; Dalca, 2018], we use two distinct losses to ensure and validate the deformation's performance, namely the mean square error and the local cross-correlation (LCC). The two losses corresponding to the difference between the fixed and deformed image and the grid's regularity have different designation depending on articles. In this manuscript, we denominate them as $\mathcal{L}_{sim}$ and $\mathcal{L}_{smooth}$. The expression of the similarity loss $\mathcal{L}_{sim}$ is :

$$\mathcal{L}_{sim} = ||F - \widehat{M}||^2 + \text{LCC}(F, \widehat{M}). \tag{3.3}$$

Moreover, in order to ensure that the predicted deformations are smooth for both parts we use two different regularisation terms. In particular, we regularise the predicted deformations to be close to the identity A_I and Φ_I, calculating the L_1 norm for the two obtained displacements

$$\mathcal{L}_{smooth} = \alpha||A - A_I||_1 + \beta||\nabla\Phi - \nabla\Phi_I||_1. \tag{3.4}$$

where the regularisation parameters α and β are essential to the joint optimisation. Too low values will make the network diverge, generating irregular deformations, while too high values prevent the network from creating significant deformations by keeping them very close to the identity.

For the segmentation, we calculate two different dice coefficient losses using the predicted moving and fixed segmentation, $\widehat{M}_{seg}$ and $\widehat{F}_{seg}$.

$$\mathcal{L}_{seg} = \gamma\text{Dice}(\widehat{M}_{seg}, M_{seg}) + \delta\text{Dice}(\widehat{F}_{seg}, F_{seg}) \tag{3.5}$$

The first term of $\mathcal{L}_{seg}$ influence only the encoder and segmentation decoder, while the second term impact also the registration decoder $\mathbf{D_{reg}}$, due to the expression of $\widehat{F}_{seg}$ in equation 3.2. Finally, as the network is trained end to end, its final optimisation is performed by the summation of the Equations 3.3,3.4 and 3.5 :

$$\mathcal{L} = \mathcal{L}_{sim} + \mathcal{L}_{smooth} + \mathcal{L}_{seg} \tag{3.6}$$

3.3.3 Network Architecture

Our convolutional architecture is based on the 3D VNet [Milletari, 2016] for the autoencoder part and on [Stergios, 2018] for the registration part. The original VNet architecture is composed of 4 blocks of convolution for both the encoder and the decoder with $[32, 64, 128, 256]$ channels in each block. In our original paper, [Estienne, 2019], we implemented a smaller version with 40 thousand parameters due to our input images' large size. The encoder was composed of only two blocks (with 8 and 16 channels). In this thesis manuscript, we perform supplementary experiments using a bigger network. Each block comprises a convolution with kernel size $2 \times 2 \times 2$, downsampling the image size by 2, and convolution with kernel size $3 \times 3 \times 3$. A LeakyReLu activation layer follows each convolution layer. The segmentation and registration decoder $\mathbf{D_{seg}}$ and $\mathbf{D_{reg}}$ have the same structure. They are

symmetric to the encoder with transposed convolutions to perform upsampling and skip connections. As illustrated in Figure 3.1, the skip connections are linked with the two decoders.

Our architecture has three supplementary blocks compared to a classic VNet, called the affine, deformable and segmentation blocks. The affine and deformable blocks are connected to the registration decoder, while the segmentation block is linked to the segmentation decoder. Both registration and segmentation blocks are composed of 3 convolutions with kernels size $3 \times 3 \times 3$ followed by LeakyReLu activation. The registration block's last layer has three channels and a sigmoid activation. These three channels correspond to the displacements over the three-axis x, y and z. The segmentation block ends with a softmax activation and four output channels. We should mention that these four channels correspond to three different types of anatomies (and the background), but this can be modified and changed depending on the application. The affine block is composed of one global average pooling and one dense layer. The dense layer outputs a matrix with 12 parameters corresponding to a 3D affine transformation matrix's parameters.

3.4 Experimental Results

We evaluated our method's performance in both registration and segmentation tasks by calculating the Dice coefficient metric for 15 different brain structures for the registration and 3 different brain structures for the segmentation. For the registration task, the 15 brain structures were: the brain stem (BS), CSF (CSF), fourth ventricle (4V), amygdala (Am), caudate (Ca), cerebellum cortex (CblmC), cerebellum white matter (CblmWM), cerebral cortex (CeblC), cerebral white matter (CeblWM), hippocampus (Hi), lateral ventricle (LV), pallidum (Pa), putamen (Pu), ventral DC (VDC) and third ventricle (3V), while for the segmentation task, we only predicted and evaluated on the CeblWM (or white matter), the CeblC (or grey matter) and the LV.

Our original implementation was based on Keras. We used 4 Nvidia GeForce GTX 1080 GPUs for each experiment with a batch size of 4. We updated our implementation, rewriting the code in Pytorch and using one Nvidia Tesla V100 GPU card. The experimental results presented here are not the original one from [Estienne, 2019], but the ones obtained by the new implementation. For training, we used Adam optimiser and a learning rate of 10^{-3}. In order to improve the convergence of the network, we initialised them with zeros weights and a bias corresponding to the identity transformation. Moreover, to prevent overfitting, we randomly shuffled the training set to generate different pairs of moving and reference images per epoch, fixing the ones on validation and testing. We performed a grid search for all experiments to find the optimal values of the losses weights β, γ and δ on the validation set, and then we calculated the different metrics on the testing set. The affine regularisation weight α was kept equal to 0.1 for all experiences. We trained our network for approximately 80 epochs, needing approximately one day, and selected the epoch having the smallest overall loss on the validation set.

3.4.1 Data and Preprocessing

For our experiments, we used the T1 brain MR images from the publicly available OASIS 3 [Marcus, 2009] dataset. Other brain datasets are available online such as ANDI, LONI, ABIDE or Harvard GSP (see [Boveiri, 2020; Balakrishnan, 2019] for instance). The majority of these datasets deal with neural pathologies like Parkison, Alzheimer or autism. All the modalities have already been resampled to a 1mm voxel grid, and the skull has been removed, resulting in a volume of size $256 \times 256 \times 256$. Moreover, for these modalities, 47 different structures for the brain's left and right sides are provided. These annotations have been automatically produced by Freesurfer [Fischl, 2012]. We used 520, 67 and 156 images for respectively the training, validation, and test set of our experiments. Moreover, we performed some additional pre-processing of the images, including $\mathcal{N}(0,1)$ normalisation, cropping of the images to a $160 \times 176 \times 208$ and translation of the volumes such that the centre of mass of the brain is moved to the centre of the volume.

3.4.2 Evaluation of the Registration

We compared the proposed approach with two different methods without the segmentation decoder. The **Unsupervised** formulation considers only the similarity and regularisation losses $\mathcal{L}_{sim}$ and $\mathcal{L}_{smooth}$, setting γ and δ to 0, thus using no segmentation labels to optimise the networks. The **Weakly-supervised** strategy exploits segmentation masks for the three selected structures but without having a segmentation branch. The new expression of the deformed reference mask is $\widehat{F}_{seg} = \mathcal{W}\left(\mathcal{W}(M_{seg}, A), \Phi\right)$ with M_{seg} being the ground truth segmentation of the moving image and not the one predicted by the segmentation decoder. The equation 3.5 is modified as a consequent : $\mathcal{L}_{seg} = \delta Dice(\widehat{F}_{seg}, F_{seg})$ removing the part corresponding to the segmentation decoder. The weights were fixed to $\beta = 0.1$ for the unsupervised alternative, $\beta = 1$, $\gamma = 0$ and $\delta = 10$ for the weakly-supervised approach and $\beta = 0.1$, $\gamma = \delta = 1$ for the proposed joint segmentation-registration network. The unsupervised approach is similar to the framework presented in [Stergios, 2018] and [Dalca, 2018], while the use of segmentation was introduced in different papers, including [Hu, 2018b; Hering, 2018; Balakrishnan, 2019]. One major difference between our approach and the one presented in [Dalca, 2018] is the choice of the fixed image. The authors provided an atlas-based approach, having the fixed images always equal to the same MRI (the atlas), while we trained our network for a more complex task, predicting the registration between random pairs of patients.

In Table 3.1, the evaluation in terms of the mean and the standard deviation of the dice coefficient is presented for the testing set. With rigid, we indicated the dice coefficient after the translation of the volumes such that the centre of the brain mass is placed in the centre of the volume. The results in Table 3.1 shows that the weakly-supervised approach overperformed the unsupervised and our proposed method. The weakly supervised improved the results not only for the structures that are used in the segmentation loss (CeblWM, CeblC and LV) but also for other structures such as the amygdala, hippocampus or caudate. The proposed approach has similar results with the unsupervised for most of the structures with improvements for the structures used in $\mathcal{L}_{seg}$, indicating that the joint

formulation concentrates more on the segmented areas than others.

Figure 3.2 shows the results of UReSNet on the registration for different patients of the testing set. We depict the moving, reference and deformed images on the axial plane as well as its deformation in the axial and coronal plane. The transformation is represented as a warped grid, and we provided a two-dimension representation. Therefore we ignore the displacement following the cutting plane. Other representations include displaying the norm of the deformation field or an RGB representation where each colour corresponds to the displacements along one axis X, Y or Z.

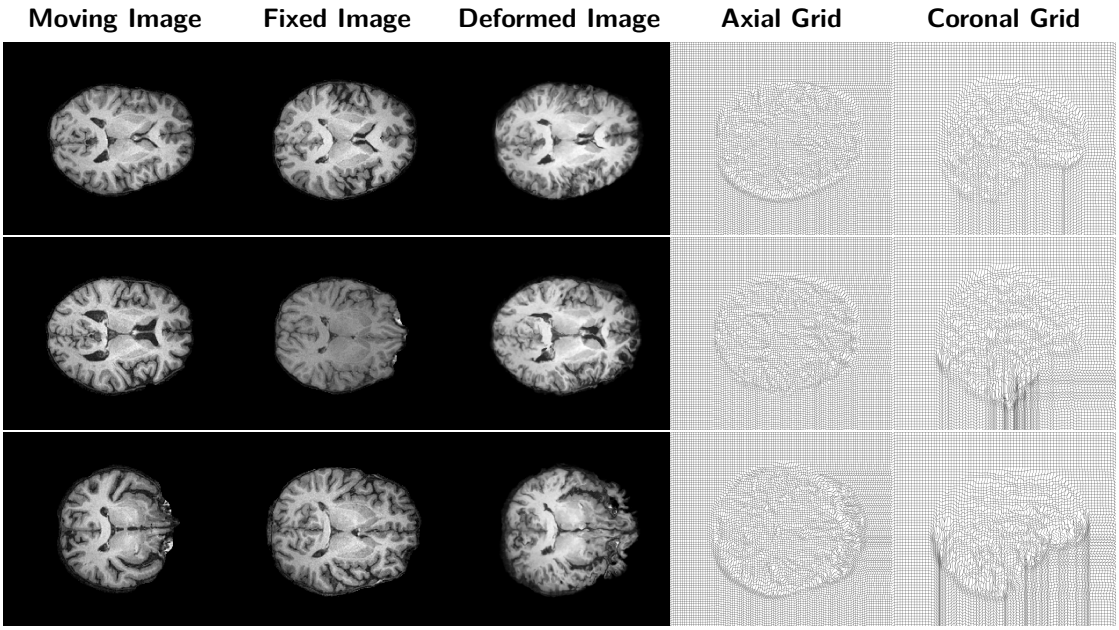


Figure 3.2: A MR slice example from the evaluation of the registration task on the test set for the proposed method. From left to right: the moving, reference, deformed image and the computed displacements in axial and sagittal planes.

3.4.3 Evaluation of the Segmentation

We trained the network to provide segmentation maps for 3 different brain structures: the cerebral white matter (CeblWM), the lateral ventricle (LV) and the cerebral cortex or grey matter (CeblC), merging the annotations of both left and right regions. After evaluating the reported predictions of our proposed method, we compared with a different version removing the registration decoder. In the **only segmentation** version, the loss $\mathcal{L}_{sim}$ and $\mathcal{L}_{reg}$ were discarded, the weights α, β and δ were set equal to 0 and the expression of $\mathcal{L}$ in the equation 3.6 became : $\mathcal{L} = Dice(\widehat{M}_{seg}, M_{seg})$. We reported dice scores on the testing set equal to 0.84 (± 0.02), 0.75 (± 0.04) and 0.83 (± 0.1) for each class respectively for UReSNet and 0.91 (± 0.02), 0.85 (± 0.03) and 0.92 (± 0.05) for the only segmentation variant. Our proposed approach did not reach similar scores than the only segmentation network, reporting still better results than other studies [Chen, 2018]. It should be noticed here that

	Rigid	Unsupervised	Weakly-supervised	UReSNet
BS	0.57 ± 0.17	0.65 ± 0.14	**0.66** ± 0.14	0.63 ± 0.15
CSF	0.39 ± 0.11	0.4 ± 0.13	0.41 ± 0.12	**0.41** ± 0.11
CblmC	0.46 ± 0.15	0.55 ± 0.13	**0.57** ± 0.12	0.56 ± 0.12
CblmWM	0.50 ± 0.06	0.5 ± 0.14	**0.51** ± 0.15	0.48 ± 0.15
CeblWM	0.40 ± 0.18	0.61 ± 0.056	**0.65** ± 0.054	0.63 ± 0.057
Pu	0.43 ± 0.16	0.47 ± 0.17	**0.51** ± 0.16	0.47 ± 0.17
VDC	0.47 ± 0.14	0.5 ± 0.13	**0.55** ± 0.11	0.5 ± 0.12
Pa	0.33 ± 0.18	0.39 ± 0.19	**0.42** ± 0.19	0.38 ± 0.2
Ca	0.26 ± 0.21	0.38 ± 0.21	**0.42** ± 0.2	0.38 ± 0.21
LV	0.39 ± 0.15	0.58 ± 0.15	**0.66** ± 0.14	0.63 ± 0.15
Hi	0.35 ± 0.18	0.4 ± 0.16	**0.44** ± 0.16	0.38 ± 0.17
3V	0.35 ± 0.18	0.45 ± 0.18	0.49 ± 0.18	**0.52** ± 0.18
4V	0.16 ± 0.17	0.22 ± 0.19	**0.23** ± 0.2	0.22 ± 0.2
Am	0.24 ± 0.22	0.26 ± 0.22	**0.32** ± 0.22	0.26 ± 0.21
CeblC	0.36 ± 0.04	0.43 ± 0.054	**0.48** ± 0.055	0.46 ± 0.058

Table 3.1: Impact of the degree of supervision on the network performance. The mean and standard deviation of the dice coefficient for the 15 different categories for the different evaluated methods. The methods vary by the formulation of the optimisation penalty.

we selected the best hyperparameters on the registration performance and not on the segmentation performance.

Figure 3.3 display different MRI, their ground truth segmentations, and the predicted masks for UReSNet and the version without the registration decoder. The hardest brain part for the proposed network is the white matter/grey matter border, with the predicted segmentation of the grey matter exceeds the white matter part and is less sharp. On the opposite, the only segmentation network has better results on the edges.

3.4.4 Impact of the network size

Our implementation on the original UReSNet article was based on Keras [Estienne, 2019]. We trained our network using 4 Nvidia GeForce GTX 1080 GPUs for each experiment with a batch size of 4. Due to the size of the images processed $(160 \times 176 \times 208)$, we had to limit our architecture to a small number of parameters (≈ 40.000). To quantify the impact of the number of parameters, we achieved new experiments using novel GPUs. We used one GPU Nvidia Tesla V100 GPU with 32GB. We designed three different experiments using for both of them batch size 4. For the first one, we kept the patch size $(160 \times 176 \times 208)$ and increased the network size with four blocks having 16, 32, 64 and 128 channels, leading to 2.1 million parameters. We denote this experiment as **big-UReSNet**. For the second experiment, called **Big-UReSNet-128**, we modified the cropping strategy. In most brain registration articles, authors pass the full brain to the network, constraining the number of parameters. However, in deep learning-based segmentation, small patches are often used during the training. We experimented with the use of random patches of size $128 \times 128 \times 128$ for the registration. This is only possible as the images have already been registered through an affine transformation.

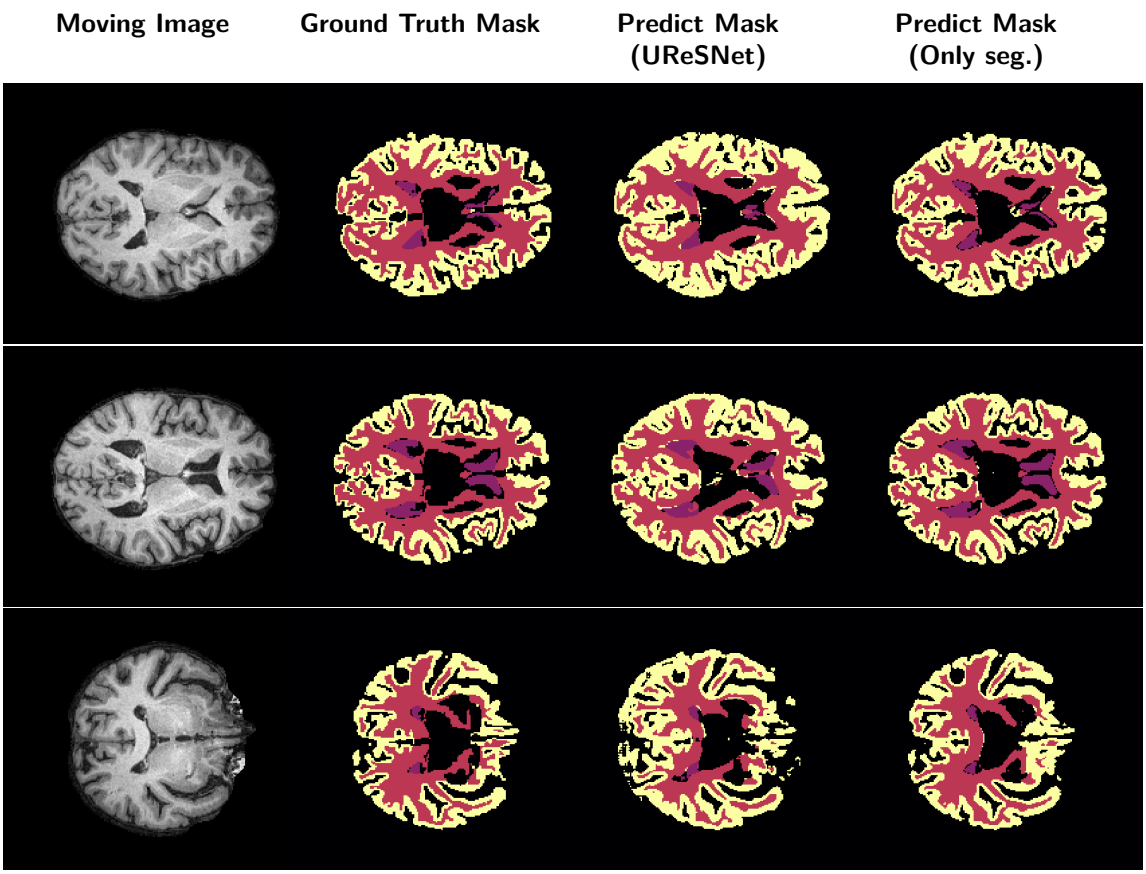


Figure 3.3: A MR slice from the test set for evaluating the segmentation task for the proposed method. From left to right: the moving image, the moving segmentation, the predicted segmentation using UReSNet and the predicted segmentation removing the registration decoder. With yellow color the Cerebral Cortex, pink color the Cerebral White Matter and purple the Lateral Ventricles are indicated.

Indeed, we needed to crop the same part of the brain in both the moving and the fixed image. During the inference phase, the entire volume is processed and warped by the network with a bigger patch. Indeed, saving the gradient is not required during the inference, and thus more memory is available. Using random crop and smaller patches, we could use a network with four blocks and 32, 64, 128, 256 channels having 8.3 million parameters. Finally, to better measure the influence of the cropping strategy and the network's size, we designed an experiment with the same numbers of parameters as **big-UReSNet** but cropping random patch of size 128. This experiment is a fusion of the previous two. All these experiments were performed with the same parameters as the proposed approach : batch size 4, 80 training epochs, learning rate equal to $1e^{-3}$ and $\beta = 0.1$, $\gamma = \delta = 1$.

In the Table 3.2, we compared the original results, the **big-UReSNet**, **big-UReSNet-128** and the **Big-UReSNet-128**. Comparing first the original version and the one with 2.1 million parameters (first two columns), we have a better dice for all the structures with an increase between 0.07 (fourth

	original UReSNet	big-UReSNet (2.1M #)	big-UReSNet-128 (2.1M #)	Big-UReSNet-128 (8.3M #)
BS	0.63 ± 0.15	0.7 ± 0.15	0.7 ± 0.15	**0.77** ± 0.12
CSF	0.41 ± 0.11	0.5 ± 0.12	0.53 ± 0.11	**0.59** ± 0.096
CblmC	0.56 ± 0.12	0.63 ± 0.12	0.63 ± 0.12	**0.68** ± 0.12
CblmWM	0.48 ± 0.15	0.57 ± 0.13	0.58 ± 0.13	**0.66** ± 0.12
CeblWM	0.63 ± 0.057	0.71 ± 0.048	0.73 ± 0.05	**0.78** ± 0.046
Pu	0.47 ± 0.17	0.63 ± 0.12	0.63 ± 0.14	**0.69** ± 0.12
VDC	0.5 ± 0.12	0.59 ± 0.11	0.59 ± 0.12	**0.66** ± 0.098
Pa	0.38 ± 0.2	0.52 ± 0.19	0.51 ± 0.19	**0.55** ± 0.2
Ca	0.38 ± 0.21	0.6 ± 0.13	0.59 ± 0.16	**0.66** ± 0.13
LV	0.63 ± 0.15	0.8 ± 0.1	0.78 ± 0.11	**0.85** ± 0.086
Hi	0.38 ± 0.17	0.52 ± 0.14	0.53 ± 0.14	**0.62** ± 0.13
3V	0.52 ± 0.18	0.59 ± 0.15	0.64 ± 0.14	**0.72** ± 0.12
4V	0.22 ± 0.2	0.29 ± 0.24	0.3 ± 0.25	**0.47** ± 0.26
Am	0.26 ± 0.21	0.38 ± 0.24	0.38 ± 0.24	**0.47** ± 0.24
CeblC	0.46 ± 0.058	0.54 ± 0.056	0.56 ± 0.058	**0.62** ± 0.059

Table 3.2: Impact of the network size on the performance. The mean and standard deviation of the dice coefficient for the 15 different categories for the different evaluated methods. The methods differ by the number of parameters and the cropping strategy. # indicates the number of parameters

ventricle and brain stem) and 0.22 (caudate). This demonstrates the importance of the number of parameters and the network's size in the registration performance. We compared then the two experiments with the same number of parameters (2^{nd} and 3^{rd} columns of table 3.2). We noticed only a few results modifications due to the new cropping strategy, proving that registration networks do not require processing the full brain during the training. However, an affine registration step is necessary. Finally, taking full advantage of the random crop approach, we increased the number of parameters to 8.3 million. This network outperformed all other experiments highly, confirming the importance of the network's size.

3.5 Discussion and Conclusion

In this chapter, we develop a novel deep learning framework to produce segmentation and deformation grids jointly. Our architecture, called UReSNet, outputs three brain structures' contours and the transformations between two random brains. Our method was one of the first attempts to apply multi-task learning for image registration, reporting very promising results. We provide experiments using our architecture and a public brain MRI dataset on registration and segmentation tasks. We also explore the network's size's impact comparing our original architecture that had only 400000 parameters with network architectures with more parameters (up to 8M). However, the performance of our joint method is inferior to weakly supervised registration. Different hypothesises can explain it. One possible explanation could be that the encoder pays more attention to the three selected structures used in the segmentation branch, thus ignoring other validated structures. Moreover,

another possible explanation is that the modelling of the two tasks reduces the expression power of the architecture (especially in the case of a small number of parameters), degrading its performance. Other experiments are needed concerning the joint formulation and the network architectures. An interesting idea to explore would be to use a non-symmetric network with bigger decoders than the joint encoder. It would allow sharing the network's first layers while having more task-specific layers for registration and segmentation.

Our proposed methods have several limitations. Concerning the registration, the main limitation is the lack of respect for relevant registration properties (see section 2.3.2). Despite the gradient formulation, we cannot claim that our transformation is diffeomorphic. Moreover, our approach does not enforce symmetry or inverse consistency. Regarding segmentation, one of our formulation's drawbacks is the absence of symmetry. Indeed our encoder inputs both the moving and the fixed images, but the decoder predicts only the segmentation for the moving image. Therefore, the network must understand that it needs to produce only the moving image's contours. Developing a symmetrical segmentation formulation could help produce better results and introduce a segmentation consistency loss. This loss would be calculated between the deformed moving segmentation and the predicted reference segmentation and also boost both segmentation and registration performances. Our approach is also trained to exploit only three brain areas, white and grey matter and the ventricles. As 15 structures are available on the exploited dataset thanks to Freesurfer, we could upgrade our experiments using all the structures. We choose to work on these three as they are the most representative brain anatomy structures.

In the next chapters, we extend our proposed method to integrate diseased regions' segmentation, such as tumour areas. These regions confuse the registration process, as they create a mismatch between the two volumes. We also explore modifications to make our method symmetric and introduce new losses during the training to output a more regular grid.

Chapter 4

Joint Segmentation-Registration for patients with abnormalities

"Bonsoèr monsieur Leblanc, George Van Brugel à l'appareil, je vous appelle parce que je suis producteur n'est ce pas, j'arrive de Belgique une foè et je suis très intéressé par votre roman, ..."

Le Dîner de cons

Contents

In chapter 3, we studied the combination of registration and segmentation tasks, and we proposed a new neural network formulation to perform them simultaneously. However, our experiments were focused on brain MRI without abnormal regions, and we predicted only brain structures contours. In this chapter, we deal with the problem of registration in the presence of tumours. As we register two different patients, the tumour will create a lack of correspondence between healthy and tumorous tissues.

Following the previous chapter, we proposed a joint segmentation-registration network that will process registration while ignoring the tumour. To do so, we modified the similarity loss to take into account only healthy tissue. Our approach does not need ground-truth tumour masks for the inference, due to the joint implementation that we proposed. The work presented in this chapter has been published in the journal Frontiers in Computational Neuroscience in Estienne et al. [Estienne, 2020].

4.1 Introduction

Brain tumours and, more specifically, gliomas as one of the most frequent types, are across the most dangerous and rapidly growing types of cancer [Holland, 2001]. Multimodal magnetic resonance imaging (MRI) is the primary screening method and glioma diagnosis in clinical practice. While gliomas are commonly stratified into Low grade and High grade due to different histology and imaging aspects, prognosis and treatment strategy, radiotherapy is one of the mainstays of treatment [Sepúlveda-Sánchez, 2018; Stupp, 2014]. However, radiotherapy treatment planning relies on the tumour's manual segmentation by physicians, making the process tedious, time-consuming, and sensitive to bias due to low inter-observer agreement [Wee, 2015].

In order to overcome these limitations, numerous methods have been proposed recently that try to provide tools and algorithms that will make the process of gliomas segmentation automatic and accurate [Parisot, 2016; Zhao, 2018]. Towards this direction, the multimodal brain tumour segmentation challenge (BraTS) [Bakas, 2017a; Bakas, 2017b; Bakas, 2017c; Menze, 2015] is annually organised in order to highlight efficient approaches and indicate the way towards this challenging problem. In recent years, most of the approaches that exploit BraTS have been based on deep learning architectures using 3D convolutional neural networks (CNNs) similar to UNet or VNet [Çiçek, 2016; Milletari, 2016]. We give a detailed overview of segmentations network in general and brain tumor segmentation in the chapter 3 sub-section Segmentation.

As we described previously, deep learning-based registration became the state for the art, thanks to the breakthrough made by unsupervised registration using spatial transformer [Krebs, 2018; Dalca, 2018; Hering, 2018]. However, when it comes to anatomies that contain abnormalities such as tumoral areas, these methods fail to register the volumes at certain locations due to a lack of similarity between the volumes. This, most of the time ends to complete distortion of the tumour area of the deformed image. Classical registration algorithms proposed formulation to couple registration and segmentation and tackled the issue of registration in the presence of tumours. The authors in [Parisot, 2012; Parisot, 2014] and [Gooya, 2011a; Gooya, 2012] focus on the special challenge of brain MRI and gliomas. Their methods are founded upon graphical models and tumour classification for the first and tumour growth model applied to an atlas with EM algorithm for the second. However, to the best of our knowledge, there is no deep learning architecture focusing on the challenge of brain registration in the presence of gliomas.

In this chapter, we propose a dual deep learning-based architecture that addresses registration and tumour segmentation simultaneously, reducing the registration criterion inside the predicted tumour areas and thus ignoring them. This new architecture shares key concepts with the previous chapter [Estienne, 2019], having, however, a major distinction. The segmentation decoder provides the segmentation of tumour lesions instead of normal brain structures, and we decouple the segmentation of the moving and the fixed images by introducing a shared encoder formulation. We also introduce a new loss such as tumour areas do not influence the registration decoder optimisation. Our framework has similarities with the work presented in [Parisot, 2012] where a Markov Random Field (MRF)

framework has been proposed to address both tumour segmentation and image registration jointly. Their method requires approximately 6 minutes to register one pair while providing a computationally efficient and accurate method, taking advantage of deep learning and GPU calculation. Our experiments are performed on two publicly available brain MRI databases, OASIS 3 and BraTS 2018. We evaluate the segmentation performance by calculating the Dice over three tumour classes and comparing them with public results. We benchmark our registration performance by deforming 14 structures in a similar way to the preceding chapter. Our goal is not only to benefit from the coupling formulation but also to deform only healthy parts of the brain. Thus, we calculate the tumour volume modifications between moving and deformed images, showing that our proposed loss outperforms other approaches.

Our main contributions are :

- Proposing a multi-task architecture to obtain on the same time deformation grids for two patients together with their tumour maps;
- Performing the registration focusing only on normal brain structures and ignoring the tumoral part of the brain;
- Introducing a shared encoder formulation and a merge operator to disjoin the moving and fixed images' segmentation.

This chapter is split into four sections. We first present our methodology in section 4.2, specially the shared encoder formulation and the merging operation and the new registration loss. We describe our experimental results on segmentation with the BraTS cohort and on registration with the OASIS 3 cohort in section 4.3 and conlude with a discussion in section 4.4. Concerning the related works on registration, segmentation and joint formulation, the reader should consult chapters 2 and 3, particularly sections 2.4, 2.5 and 3.2.

4.2 Materials and Methods

We consider a pair of medical volumes from two different patients, a moving M, and a fixed F together with their annotations for the tumour areas (M_{seg} and F_{seg}). The framework consists of a bi-cephalic structure with shared parameters, depicted in Figure 4.1. During training, the network uses as input a moving M and a fixed F volumes and outputs their brain tumour segmentation masks $\widehat{M}_{seg}$ and $\widehat{F}_{seg}$ and the optimal elastic transformation Φ which will project or map the moving volume to the fixed volume. In this section, we present the details for each of the blocks as well as our final formulation for the optimisation.

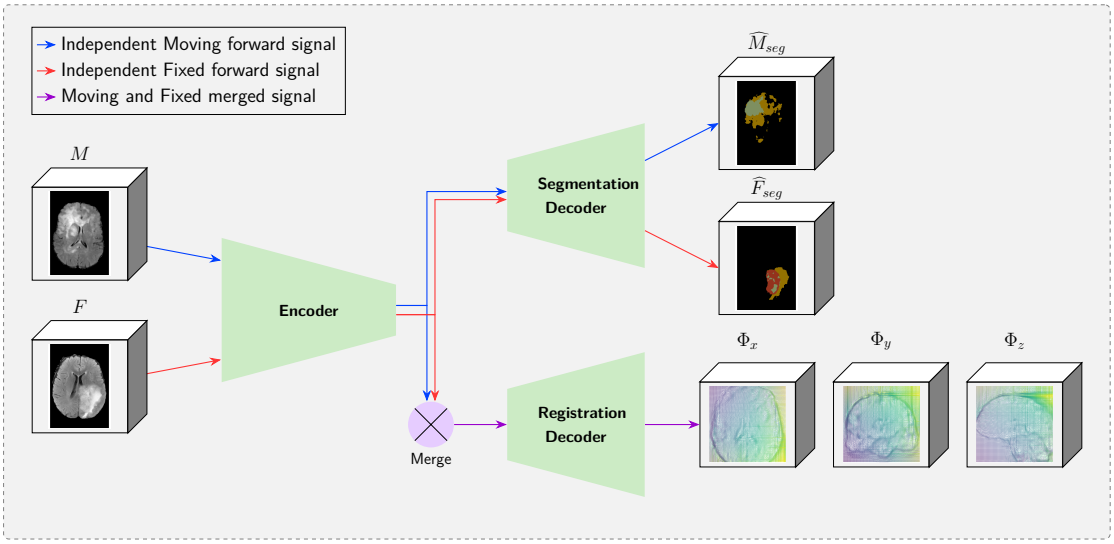


Figure 4.1: A schematic representation of the proposed framework. The framework comprises two decoders, one which provides tumour segmentation masks for both M and F images, and one provides the optimal displacement grid Φ that will accurately map the M to the F image. The merge bloc combines the forward signal of the moving input and the fixed input (forwarded independently in the encoder).

4.2.1 Shared encoder

One of the main differences of the proposed formulation with other registration approaches in the literature is how the moving and fixed volumes are combined. In particular, instead of concatenating the two initial volumes, these volumes are independently forwarded in a unique encoder, yielding two sets of features maps (called *latent codes*) c_M and c_F for the moving and the fixed volumes, respectively. These two codes are then independently forwarded into the segmentation decoder, providing the predicted segmentation maps M_{seg} and F_{seg}. Simultaneously, the two codes are merged before being forwarded to the registration decoder. This operation is depicted in the *"Merge"* block in Figure 4.1. The motivation behind adopting this strategy is to force the encoder to extract meaningful representations from individual volumes instead of a pair of volumes. This is equivalent to asking the encoder to discover a template, "deformation-free" space for all volumes, and encoding each volume against this space [Shu, 2018], instead of decoding the deformation grid between every possible pair of volumes. Besides, from the segmentation point of view, there is no relationship between the tumour maps of the moving volume and the fixed volume, so the codes to be forwarded into the segmentation decoder should not depend on each other. It was indeed one limitation of the previous chapter formulation.

We tested two merging operators, namely concatenation and subtraction. Both moving and fixed images are $4D$ volumes whose first dimension corresponds to the 4 different MRI modalities used per subject. After the forward to the encoder, the codes c_M and c_F are also 4D volumes with the first

dimension corresponding to n_f, which is the number of convolutional filters of the last block of the encoder. Before c_M and c_F are inserted into the registration decoder, they are merged, outputting one 4D volume of size $2 \times n_f$ in the case of the concatenation, and of size n_f for the elementwise subtraction operator, both leaving the rest of the dimensions unchanged. We must also notice that we merge the encodings from all the block of the encoder in order to keep the same skip connections as in the classical VNet architecture : $c_{i,i=1..4} = \mathcal{M}(M_i, F_i)$ with $\mathcal{M}$ being the merge operation, M_i and F_i being the encoding coming from the ith block of the VNet. In particular, the subtraction presents the following natural properties for every coding image c_I:

- $\forall c_I \in \mathbb{R}^n : \mathcal{M}(c_I, c_I) = 0$
- $\forall c_I, c_J \in \mathbb{R}^n \times \mathbb{R}^n : \mathcal{M}(c_I, c_J) = -\mathcal{M}(c_J, c_I)$

The subtraction operation respect suitable mathematical properties. Indeed, if we calculated the registration between two identical images, the subtraction will give a zero tensor which will then generate identity transformation. Furthermore, the backward transformation Φ^{-1} is obtained by inverting c_I and c_J in the subtraction. In this framework, the encoder can be seen as playing the role of a feature extractor, as the two images are processed independently. Another late fusion strategy can be found in [Heinrich, 2019].

4.2.2 Registration and Segmentation Decoder

Inspired by the latest advances reported on the BraTS 2018 dataset, we adopt a powerful autoencoder architecture. The segmentation and registration decoders share the same encoder (Section 4.2.1) for feature extraction, and they provide brain tumour segmentation masks ($\widehat{M}_{seg}$ and $\widehat{F}_{seg}$) for the moving and the fixed images and the deformation Φ. These masks refer to valuable information about the regions that cannot be registered properly as there is no corresponding anatomical information on the pair. This information is integrated into the optimisation of the registration component, relaxing the similarity constraints and preserving to a certain extent the geometric properties of the tumour.

In this chapter, we keep the same registration strategy as in the previous, with the main components being the 3D spatial transformer and the prediction of spatial gradients from [Stergios, 2018; Shu, 2018]. The registration decoder outputs a 4D matrix with three channels corresponding to $\nabla_x \Phi_x$, $\nabla_y \Phi_y$, $\nabla_z \Phi_z$. Then we apply a cumulative sum operation approximating the integration operation, and a spatial transformer warps the moving image M and its segmentation M_{seg} using Φ. More details on the gradient formulation are given in section 2.5.3 and section 3.3.1. Contrary to the method described in chapter 3, we consider here only a deformable transformation, removing the affine block. The mathematical expression of our predicted segmentation and deformation is

$$\begin{aligned} \widehat{M}_{seg} &= \mathbf{D_{seg}}(\mathbf{E}(M)) \\ \widehat{F}_{seg} &= \mathbf{D_{seg}}(\mathbf{E}(F)) \\ \nabla\Phi &= \mathbf{D_{reg}}(\mathcal{M}(\mathbf{E}(M), \mathbf{E}(F))) \\ D = \widehat{M} &= \mathcal{W}(M, \Phi) \end{aligned} \tag{4.1}$$

with $\mathbf{E}$ being the shared encoder, $\mathbf{D_{reg}}$ being the registration decoder, $\mathbf{D_{seg}}$ the segmentation decoder, $\mathcal{W}(\cdot)$ the warping operation, D or $\widehat{M}$ the deformed image.

4.2.3 Network Architecture

Our network architecture is a modified version of the fully convolutional VNet [Milletari, 2016] for the underlying encoder and decoders parts, maintaining the depth of the model and the rest of the filter's configuration unchanged. The model, whose computational graph is displayed in Table 4.4, comprises several sequential residual convolutional blocks made of one to three convolutional layers, followed by downsampling convolutions for the encoder part and upsampling convolutions for the decoder part. We replaced the initial $5 \times 5 \times 5$ convolutions filter-size by $3 \times 3 \times 3$ in order to reduce the number of parameters without changing the depth of the model. We also replace PReLu activations with ReLU ones. To speed up its convergence, the model uses residual connections between each encoding and corresponding decoding stage for both the segmentation and the registration decoder (see Figure 4.1). This allows every layer of the network, particularly the first ones, to be trained more efficiently since the gradient can flow easier from the last layers to the first ones with less vanishing or exploding gradient issues. The encoder part deals with 4-inputs per volume, representing the 4 different MRI modalities available on the BraTS dataset. Thus, an extra $1 \times 1 \times 1$ convolution is added to fuse the initial modalities. Moreover, the architecture contains two decoders of identical blocks, one dedicated to the segmentation of tumours for the moving and fixed image and one dedicated to the optimal displacement that will map the moving to the fixed image, defined respectively as $\mathbf{D_{seg}}$ and $\mathbf{D_{reg}}$. The difference between this architecture and the preceding chapter lies in the number of convolution and deconvolution blocks, the number of channels for each block, and the network numbers of parameters.

4.2.4 Optimization

The network is trained to minimise the segmentation and registration loss functions jointly. Various loss functions have been proposed in the literature for the semantic segmentation of 3D medical volumes. In this chapter, we performed all our experiments using weighted categorical cross-entropy loss and optimising 3 different segmentation classes for the tumour area provided by the BraTS dataset. In particular,

$$\mathcal{L}_{seg} = CE(M_{seg}, \widehat{M}_{seg}) + CE(F_{seg}, \widehat{F}_{seg}) \tag{4.2}$$

where CE denotes the weighted cross-entropy loss. The cross-entropy is calculated for both the moving and fixed images, and the overall segmentation loss is the sum of the two. Here we should note that other segmentation losses can be applied as, for example, the dice coefficient [Sudre, 2017], focal loss [Lin, 2018], e.t.c.

For registration, the classical optimization scheme is to combine a similarity loss $\mathcal{L}_{sim}$ and a smooth loss $\mathcal{L}_{smooth}$. The similarity loss minimize the Frobenius norm between the fixed F and deformed D image intensities:

$$\mathcal{L}_{sim} = ||(F - D)||_2 \tag{4.3}$$

Here, to better achieve overall registration, the Frobenius norm within the regions predicted to be tumours is excluded from the loss function. We argue that by doing this, the model does not focus on tumour regions, which might produce very high norm due to their texture, but rather focuses on the overall registration task by looking at regions outside the tumour which contain information more pertinent to the alignment of the volumes. Here we should mention that on $\widehat{M}_{seg}$, we apply the same displacement grid as on M, resulting in $D_{seg} = \mathcal{W}(\widehat{M}_{seg}, \Phi)$. Further, let $\widehat{F}^0_{seg}$ and D^0_{seg} be binary volumes indicating the voxels which are predicted to be outside any segmented regions ($\widehat{F}^0_{seg}(p) = 0$ if p is a pixel corresponding to tumor tissue, and 1 otherwise). Then, the registration loss can be written as

$$\mathcal{L}^{\star}_{sim} = ||(F - D) \cdot D^0_{seg} \cdot \widehat{F}^0_{seg}||_2 \tag{4.4}$$

where $\cdot$ is the element-wise multiplication, $||\cdot||_2$ indicates the Frobenius norm. The multiplication by D^0_{seg} and $\widehat{F}^0_{seg}$ cancel any gradients in the tumour areas, and thus, they do not influence the registration. A key point of our proposed loss $\mathcal{L}^{\star}_{sim}$ lies in the use of the predicted segmentation $\widehat{F}_{seg}$ and D_{seg} to calculate the binary masks and ignore the tumour areas in the registration loss and not the ground truth segmentation. Thus, we do not need to provide ground truth segmentation during the inference phase, and the network is learning which area matters for the registration.

The use of regularisation on the displacements Φ is essential in order to constrain the network to predict smooth deformation grids that are anatomically more meaningful while at the same time regularise the objective function towards avoiding local minimum. The smooth loss is formulated as :

$$\mathcal{L}_{smooth} = \|\nabla\Phi - \nabla\Phi_I\|_1 \tag{4.5}$$

where $\nabla\Phi_I$ is the spatial gradient of the identity deformation.

Finally, the final optimisation of the framework is performed by minimising the segmentation and registration loss functions jointly :

$$\mathcal{L} = \mathcal{L}_{sim} + \alpha\mathcal{L}_{smooth} + \beta\mathcal{L}_{seg}$$

where α is the regularisation hyperparameter, and β is a weight that indicates the influence of the segmentation on the joint network optimisation. Both α and β were defined after grid search.

For the training process, the initial learning rate was $2 \cdot 10^{-3}$ and subdued by a factor of 5 if the performance on the validation set did not improve for 30 epochs. The training procedure stops when there is no improvement for 50 epochs. The regularization weights α and β were set to 10^{-10} and 1 after grid search. As training samples, random pairs among all cases were selected with a batch size limited to 2 due to the limited memory resources on the GPU. The network's performance was evaluated every 100 batch, and both proposed models converged after nearly 200 epochs. The overall training time was calculated to ~ 20 hours, while the time for inference of one pair, using 4 different modalities was ~ 3 sec, using an NVIDIA GeForce GTX 1080 Ti GPU.

4.3 Experiences and Results

Our contributions in the study are threefold: multi-task segmentation and registration, registration with a shared encoder and latent space merge operator, as well as the loss $\mathcal{L}^{\star}_{reg}$ (Equation 4.4) that alleviates the registration modifications of tumour tissues in both moving and fixed patients. Our experiments were intended to weigh these novelties' impact on tumour segmentation and registration of MRIs with tumour areas.

4.3.1 Datasets

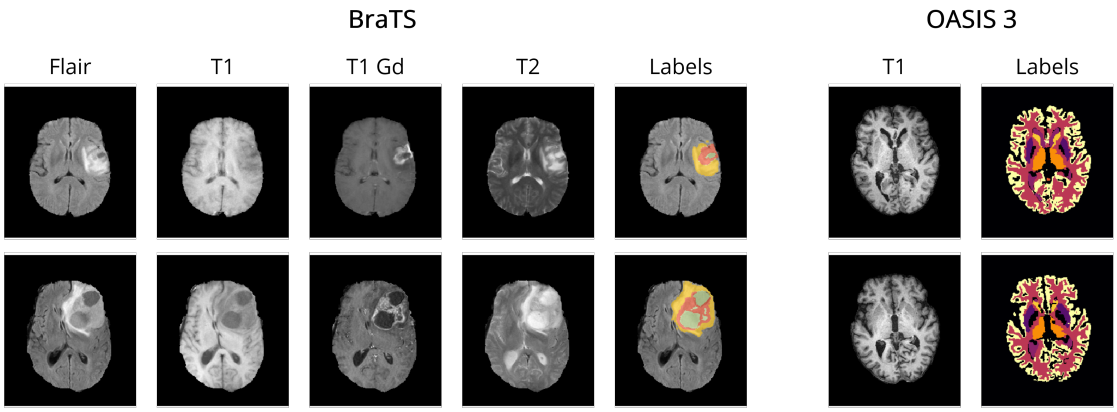


Figure 4.2: Illustration of one slice from two examples from both BraTS and OASIS 3 datasets. The data from BraTS are 3D spatial volumes with four modalities (T1, T1 gadolinium, T2, T2 FLAIR), along with voxelwise annotations for the three tumour tissue subclasses depicting the overall extent of tumours. OASIS 3 contains 3D volume only for the T1 modality, and images are provided with voxelwise annotations of 13 normal brain structures for patients without brain tumours.

We evaluated the performance of our method using two publicly available datasets, namely the Brain tumour Segmentation (BraTS) [Bakas, 2019] and Open Access Series of Imaging Studies

(OASIS 3) [Marcus, 2009] datasets. BraTS contains multi-institutional pre-operative MRI scans of whole brains with visible gliomas, intrinsically heterogeneous in their imaging phenotype (shape and appearance) and histology. The MRIs are all pre-operative and consist of 4 modalities, i.e. four 3D volumes, namely a) a native T1-weighted scan (T1), b) a post-contrast Gadolinium T1-weighted scan (T1Gd), c) a native T2-weighted scan (T2), and d) a native T2 Fluid Attenuated Inversion Recovery scan (T2-FLAIR). The BraTS MRIs are provided with voxelwise ground-truth annotations for five disjoint classes denoting a) the background, b) the necrotic and non-enhancing tumour core (NCR/NET), c) the GD-enhancing tumour (ET), d) the peritumoral oedema (ED) as well as invaded tissue, and finally e) the rest of the brain, i.e. brain with no abnormality nor invaded tissue. Each hospital was responsible for annotating their MRIs, with a central validation by domain experts. We use the original dataset split of BraTS 2018, which contains 285 training samples and 66 for validation. In order to perform our experiments, we split this training set into 3 parts, i.e. train, validation and test sets (199, 26 and 60 patients, respectively), while we used the 66 unseen cases on the platform to report the performance of the proposed and the benchmarked methods. Moreover, and especially for the registration task, we evaluated the performance of the models trained on BraTS on the OASIS 3 dataset to test the method's generalisation. This dataset consists of a longitudinal collection of 150 subjects who were characterised as either nondemented or with mild cases of Alzheimer's disease (AD) using the Clinical Dementia Rating (CDR). Each scan is made of 3 to 4 individual T1-weighted MRIs, intended to reduce the signal-to-noise ratio visible with single images. The scans are also provided with annotations for 47 different structures for the brain's left and right side generated with FreeSurfer. Some datasets samples can be seen in Figure 4.2.

The same pre-processing steps have been applied for both datasets. MRIs were resampled to voxels of volume $1mm^3$ using trilinear interpolation. Each scan is then centred by automatically translating their barycenter to the centre of the volume. Ground-truth masks of training and validation steps were accordingly translated. Each modality of each scan has been standardised, i.e. the values of the voxels of the 3D sub scans were of zero mean and unit variance. This normalisation step is done independently, patient-wise and channel-wise, considering each channel equally since modalities have voxels values in completely different ranges. Finally, these consequent scans are cropped into $(144, 208, 144)$ sized volumes.

4.3.2 Statistical evaluations

Methods benchmarked

We, therefore, benchmark multiple versions of our proposed approach with a subset of these novelties to assess their impact on both registration and segmentation. We notably derive 2 variants for both merging operators subtraction and concatenation. The first variant is our fully proposed architecture with a shared encoder for registration and one decoder for segmentation whose tumour predictions are used to implement the proposed loss $\mathcal{L}^{\star}_{reg}$. These models are named "Proposed concatenation with $\mathcal{L}^{\star}_{sim}$" and "Proposed subtraction with $\mathcal{L}^{\star}_{sim}$". The second variant of models does not use the proposed

loss and are identified with "w/o $\mathcal{L}_{sim}^{\star}$". Finally, we also derive a third variant of our approach, yielding one method per merging operator, by discarding the segmentation decoder. Because the proposed loss uses the predicted tumour maps from a segmentation decoder, this variant does not rely on it. These latter methods are named "Proposed concatenation only reg." and "Proposed subtraction only reg." and are primarily benchmarked to assess the performance of the segmentation decoder and the loss $\mathcal{L}_{sim}^{\star}$ respecting our fully proposed architecture.

We also benchmark baseline methods without any of the proposed contributions. Since our deep learning architecture is derived from the Vnet [Milletari, 2016], this model is used as a baseline for segmentation. This comparison seems fair since the fully proposed approach can be seen as a Vnet for the segmentation part: the shared encoder and the proposed loss are primarily designed for registration and have no direct impact on the segmentation apart from the features learnt in the encoder. For completeness, the top-performing results on the BraTS [Bakas, 2019] challenge are reported, although we argue that the comparison is unfair since our deep learning architecture is entirely built using the Vnet [Milletari, 2016], which is not specifically designed to perform well on the BraTS segmentation task. Finally, we also report the performance of Voxelmorph [Dalca, 2018], a well-performing brain MRI registration neural network-based approach, although their entire deep learning structure, as well as their grid formulation, is different.

Performance assessment

For performance assessment of the segmentation task, we reported the Dice coefficient metric and Hausdorff distance to measure the performance for the tumour classes tumour Core (TC), Enhancing tumour (ET) and Whole tumour (WT) as computed and provided from the BraTS submission website. These classes are the ones used in the BraTS challenge [Bakas, 2019], but differ from the original ones provided in the BraTS dataset: TC is the same as the one labelled in the BraTS dataset for necrotic core (NCR/NET), ET is the disjoint union of the original classes NCR/NET and ET, while WT refers to the union of all tumoral and invaded tissues.

For the registration, we evaluated the change on the tumour area together with the Dice coefficient metric for the following categories of the OASIS 3 dataset: brain stem (BS), cerebrospinal fluid (CSF), fourth ventricle (4V), amygdala (Am), caudate (Ca), cerebellum cortex (CblmC), cerebellum white matter (CblmWM), cerebral cortex (CeblC), cerebral white matter (CeblWM), hippocampus (Hi), lateral ventricle (LV), pallidum (Pa), putamen (Pu), ventral DC (VDC) and third ventricle (3V) categories. Here we should mention that for the experiments with the OASIS 3 dataset, we trained a model only with the T1-weighted MRIs of the BraTS dataset in order to match the available modalities of the OASIS 3 dataset. This evaluation is important as *i)* BraTS does not provide anatomical annotations to evaluate the registration performance quantitatively, and *ii)* the generalisation of the proposed method on an unseen dataset is evaluated. For the registration of tumour tissues, which might not exist in the moving or fixed MRIs, we expect the model to maintain the tumours geometric properties. In particular, we do not expect the tumour areas to stay completely unchanged. However, we expect that the volume of the different tumour types would change with a ratio similar to the

one that the entire moving to the fixed volume changes. We calculate this ratio by computing $\frac{D^j_{seg}}{S^j_{seg}}$ where $j = \{0, 1, 2, 3\}$ corresponds to the entire brain and the different tumour classes (NCR/NET, ET and ED). We then assess the tumour's change by calculating the absolute value of the difference between $j = 1$ and every other tumour class. Ideally, we expect a model which preserves the tumour geometry and shape during inference to present a zero difference between the entire brain and tumour class ratio. We independently calculate this difference for each tumour class to monitor each class's behaviour and after merging the entire tumour area.

For statistical significance evaluations between any two methods, we compute independent t-tests as presented in [Rouder, 2009], defining as the null hypothesis the evaluation metrics of the two populations to be equal. We then report the associated p-value and Cohen's d [Rice, 2005], which we use to measure the effect size. Such statistical significance evaluation is reported in the form ($t(n)$; p; d) where n is the number of samples for each population, $t(n)$ is the t-value, p is the p-value and d is Cohen's d. We defined the difference of the two population means as statistically significant if the associated p-value is lower than 0.005, and consider, as a rule of thumb, that a value of d of 0.20 indicates a small effect size, 0.50 for medium effect size and 0.80 for large effect size. All of the results in this chapter have been computed on unseen testing sets, and the performance of all benchmarked models has been assessed once.

For rigour and each t-test conducted, we ensure that the underlying distributions meet the following assumptions: observations are independent and identically distributed, the outcome variable follows a normal distribution in the population (with [Jarque, 1980]), and the outcome variable has equal standard deviations in two considered (sub)populations (using Levene's test [Schultz, 1985]). Finally, when comparing two populations, each made of several subpopulations, we merge such subpopulations into a single set, then compute t-tests on the obtained two gathered populations.

4.3.3 Evaluation of the Segmentation

	Average		Dice			Hausdorff95		
Method	Dice	Hausdorff95	ET	WT	TC	ET	WT	TC
Baseline segmentation	0.79 ±0.29	7.0 ± 9.6	0.73 ±0.29	0.87 ±0.13	0.75 ±0.24	4.7 ±8.2	7.2 ±9.4	9.2 ±8.9
Proposed								
concatenation w/o $\mathcal{L}^{\star}_{sim}$	0.74 ±0.29	8.3 ± 10.4	0.70 ±0.29	0.87 ±0.11	0.65 ±0.29	6.2 ±9.8	7.8 ±11.1	11.3 ±7.1
concatenation with $\mathcal{L}^{\star}_{sim}$	0.73 ±0.29	7.6 ± 9.9	0.68 ±0.30	0.87 ±0.12	0.66 ±0.28	6.3 ±9.9	5.6 ±4.2	10.8 ±6.6
subtraction w/o $\mathcal{L}^{\star}_{sim}$	0.76 ± 0.27	7.8 ± 10.3	0.71 ± 0.28	0.88 ±0.10	0.70 ±0.24	6.5 ±10.8	7.4 ±11.0	10.0 ±7.4
subtraction with $\mathcal{L}^{\star}_{sim}$	0.76 ±0.27	7.9 ± 10.1	0.71 ±0.29	0.88 ±0.10	0.69 ±0.25	5.8 ±9.6	7.7 ±11.5	11.1 ±8.3

Table 4.1: Evaluation of the Segmentation. Quantitative results of the different methods on the segmentation task on the BraTS 2018 validation dataset. Dice and Hausdorff95 are reported for the three classes Whole tumour (WT), Enhancing tumour (ET) and tumour Core (TC) together with their average values. Results are reported with mean across patients (MRIs) along with the associated standard deviation. We upload our predictions on the official leaderboard of the validation set (66 patients).

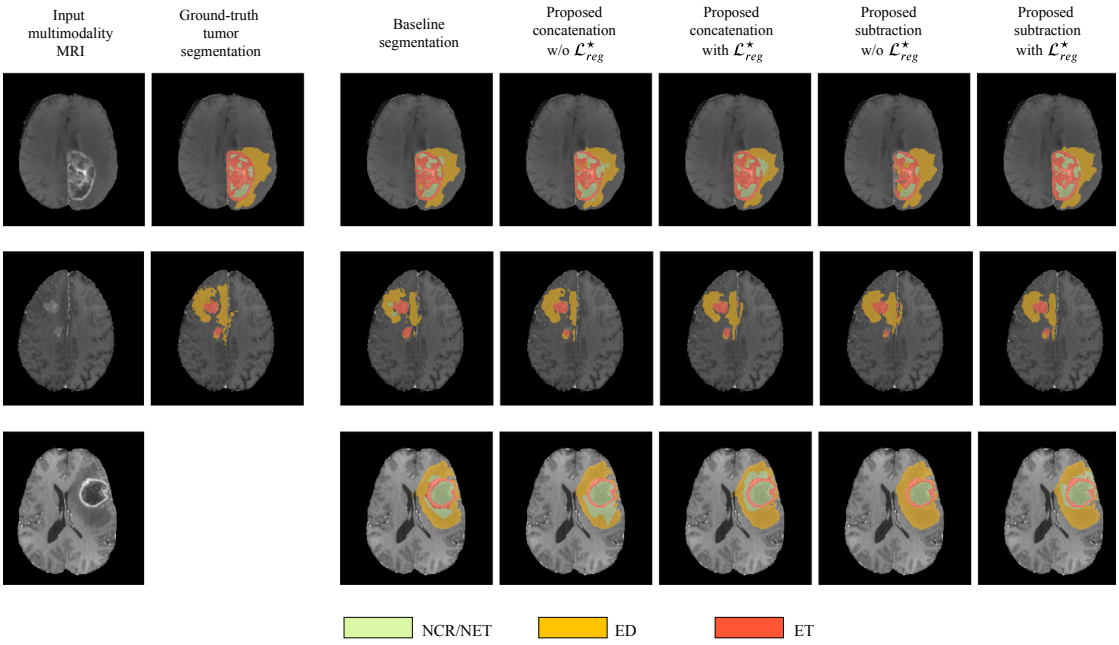


Figure 4.3: The segmentation maps produced by the different evaluated methods displayed on post-contrast Gadolinium T1-weighted modalities. We present the provided segmentation maps both on the our test dataset and on the BraTS 2018 validation dataset. NCR/NET: necrotic core, ET: GD-enhancing tumor, ED: peritumoral edema.

Segmentation results for the tumour regions are displayed in Table 4.1 for the case of the same autoencoder architecture trained only with a segmentation decoder (Baseline *segmentation*) and the proposed method using different merging operations and with or without $\mathcal{L}^{\star}_{sim}$. One can observe that all evaluated methods perform quite similarly with Dice higher than 0.66 for all the classes and models. The *baseline segmentation* model reports slightly better average Dice coefficient and average Haussdorf distance measurements, with an average Dice 0.03 higher and an average Hausdorff95 distance 0.6 higher than the proposed with concatenation merging operator. However, none of these differences is statistically significant, as indicated in Table 4.5. As an illustration, for Dice, the minimum received p-value was $p = 0.24$, reported between *baseline segmentation* and *proposed concatenation with* $\mathcal{L}^{\star}_{sim}$ together with an associated Cohen's $d = 0.21$ indicating a small size effect. Similarly, for Hausdorff95, the minimum received p-value was $p = 0.46$, reported this time between *baseline segmentation* and *proposed concatenation w/o* $\mathcal{L}^{\star}_{sim}$ with $d = 0.13$ also indicating a small size effect, which indicated that the means differences between those two models and any other two models are not statistically significant.

This is very promising if we consider that our proposed model is learning a far more complex architecture addressing both registration and segmentation, with the same volume of training data without a significant drop in the segmentation performance.

The superiority of the *baseline segmentation* seems to be presented mainly due to higher perfor-

mance for the TC class (*baseline segmentation* and *proposed subtraction with* $\mathcal{L}^{\star}_{sim}$: $t(66) = 1.41$; $p = 0.16$; $d = 0.24$). Moreover, the concatenation operation seems to perform slightly better for the tumour segmentation than the subtraction, with at least 0.02 improvement for average Dice coefficient, although this improvement is not statistically significant (*proposed concatenation with* $\mathcal{L}^{\star}_{sim}$ and *proposed subtraction with* $\mathcal{L}^{\star}_{sim}$: $t(66) = 0.62$; $p = 0.53$; $d = 0.11$).

Moreover, even if one of the main goals of our method is the proper registration of the tumoral regions, we perform a comparison with the two best-performing methods presented in BraTS 2018 [Myronenko, 2019; Isensee, 2019] evaluated on the validation dataset of BraTS 2018. In particular, the [Myronenko, 2019] reports an average dice of 0.82, 0.91 and 0.87 for ET, WT and TC respectively, while [Isensee, 2019] reports 0.81, 0.91 and 0.87. Both methods outperform our proposed approach on the validation set of BraTS 2018 by integrating novelties specifically designed to the tumour segmentation task of BraTS 2018. In this study, we based our architecture on a relatively simple and widely used 3D fully convolutional network [Milletari, 2016]. However, different architectures with tumour specific components (trained on the evaluated tumour classes), trained on more data (similar to the ones that are used from [Isensee, 2019]), or even integrating post-processing steps can be easily integrated, boosting the performance of our method considerably.

Finally, in Figure 4.3 we represent the ground truth and predicted tumour segmentation maps comparing the *baseline segmentation* and our proposed method using the different components and merging operators. We present three different cases, two from our custom test set, on which we have the ground truth information and one from the validation set of the BraTS submission page. One can observe that all the methods provide quite accurate segmentation maps for all three tumour classes.

4.3.4 Evaluation of the Registration

Evaluation on anatomical structures

The registration performance has been evaluated on an unseen dataset with anatomical information, namely OASIS 3. Table 4.2 presents the mean and standard deviation of the Dice coefficient for the different evaluated methods. With rigid, we indicate the Dice coefficient after the translation of the volumes such that the centre of the brain mass is placed in the centre of the volume. It can be observed that the performance of the evaluated methods are quite similar, indicating that the additional tumour segmentation decoder does not decrease the performance of the registration. On the other hand, it provides additional information about the areas of the tumour in the image. From our experiments, we show that the proposed formulation can provide registration accuracy similar to the recent state-of-the-art deep learning-based methods [Dalca, 2018] with approximate the same average Dice values, that is 0.50 for [Dalca, 2018] and 0.49 for all but one of the proposed variants. Moreover, again this difference in the performance between [Dalca, 2018] and the proposed method is not statistically significant with $t(150) = 0.64$; $p = 0.52$; $d = 0.07$. From our comparisons, the only significant difference on the evaluation of the registration task was reported between the proposed

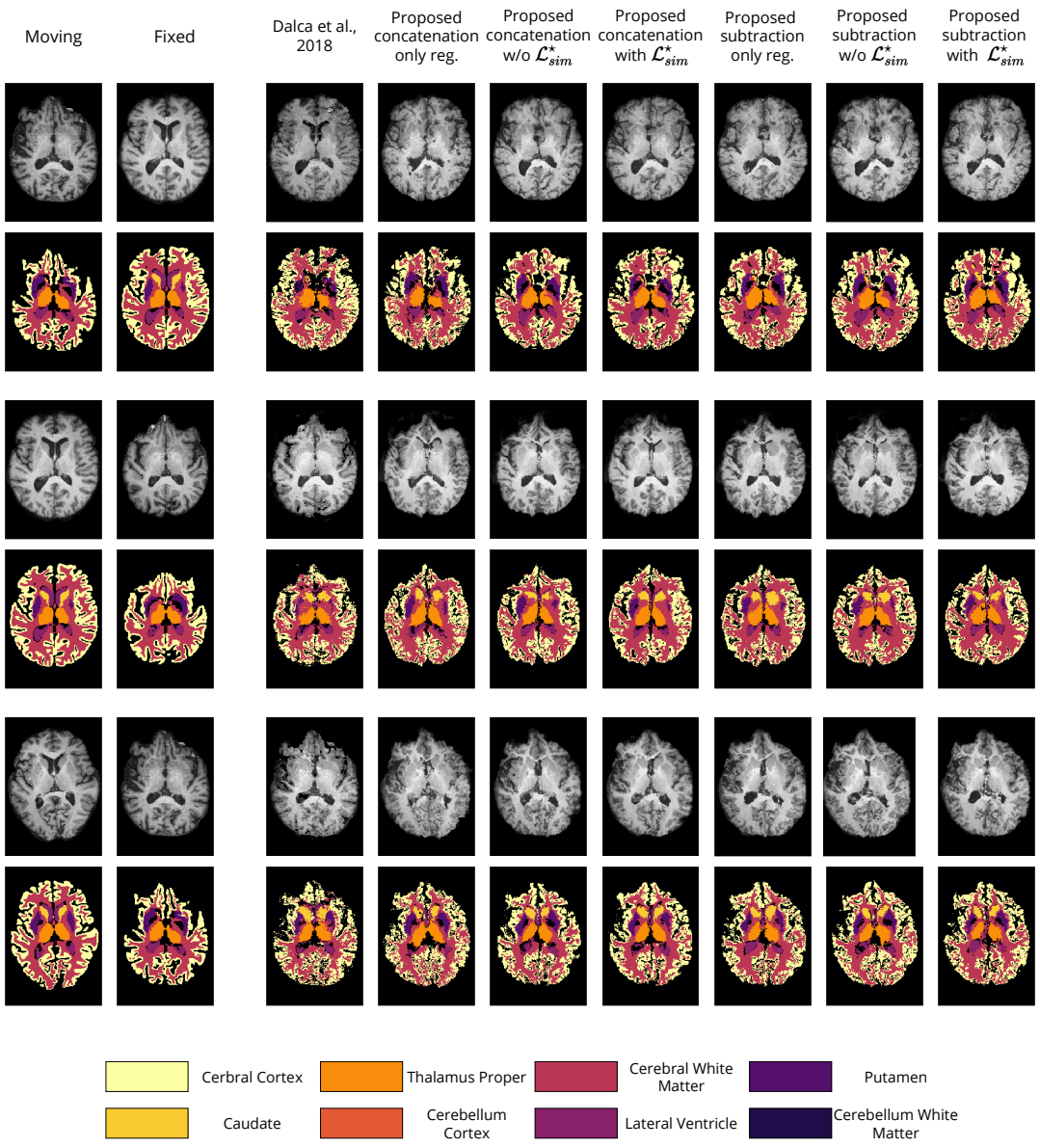


Figure 4.4: Qualitative evaluation of the registration performance for the different evaluated methods, displayed on T1 modalities. For an easier visualisation, we group left and right categories and only display the following 9 classes: caudate (Ca), cerebellum cortex (CblmC), cerebellum white matter (CblmWM), cerebral cortex (CeblC), cerebral white matter (CeblWM), lateral ventricle (LV), pallidum (Pa), putamen (Pu), ventral DC (VDC).

Method	Rigid	Voxelmorph	Proposed					
			concatenation			**subtraction**		
			only reg.	w/o $\mathcal{L}^{\star}_{sim}$	with $\mathcal{L}^{\star}_{sim}$	only reg.	w/o $\mathcal{L}^{\star}_{sim}$	with $\mathcal{L}^{\star}_{sim}$
BS	0.58 ± 0.15	0.69 ± 0.12	0.65 ± 0.15	0.72 ± 0.13	0.7 ± 0.15	0.71 ± 0.13	0.7 ± 0.13	**0.72± 0.12**
CSF	0.39 ± 0.11	**0.46± 0.13**	0.34 ± 0.1	0.42 ± 0.1	0.44 ± 0.12	0.41 ± 0.1	0.41 ± 0.1	0.4 ± 0.11
CblmC	0.46 ± 0.13	**0.63± 0.11**	0.58 ± 0.11	0.61 ± 0.11	0.6 ± 0.13	0.61 ± 0.12	0.6 ± 0.11	0.61 ± 0.11
CblmWM	0.40 ± 0.14	**0.57± 0.13**	0.48 ± 0.14	0.51 ± 0.12	0.52 ± 0.14	0.53 ± 0.13	0.52 ± 0.12	0.53 ± 0.12
CeblWM	0.49 ± 0.05	**0.73± 0.083**	0.6 ± 0.056	0.63 ± 0.056	0.66 ± 0.06	0.66 ± 0.058	0.65 ± 0.057	0.64 ± 0.058
Pu	0.44 ± 0.13	0.42 ± 0.14	0.46 ± 0.12	0.47 ± 0.14	0.47 ± 0.14	0.47 ± 0.12	**0.48± 0.13**	0.47 ± 0.12
VDC	0.47 ± 0.13	0.5 ± 0.11	0.47 ± 0.12	0.51 ± 0.12	0.52 ± 0.13	0.5 ± 0.11	**0.53± 0.11**	0.51 ± 0.11
Pa	0.35 ± 0.17	0.33 ± 0.14	0.38 ± 0.14	0.37 ± 0.16	0.38 ± 0.16	0.37 ± 0.15	**0.39± 0.15**	0.38 ± 0.15
Ca	0.27 ± 0.15	0.42 ± 0.17	0.35 ± 0.15	**0.44± 0.15**	0.42 ± 0.16	0.43 ± 0.14	0.43 ± 0.14	0.41 ± 0.15
LV	0.40 ± 0.13	0.62 ± 0.14	0.54 ± 0.14	**0.65± 0.13**	0.65 ± 0.14	0.63 ± 0.12	0.64 ± 0.13	0.63 ± 0.13
Hi	0.34 ± 0.13	0.38 ± 0.13	0.35 ± 0.13	**0.42± 0.14**	0.4 ± 0.15	0.4 ± 0.13	0.41 ± 0.13	0.43 ± 0.13
3V	0.39 ± 0.17	**0.53± 0.18**	0.4 ± 0.16	0.46 ± 0.17	0.51 ± 0.19	0.47 ± 0.16	0.49 ± 0.17	0.44 ± 0.17
4V	0.15 ± 0.15	**0.32± 0.23**	0.21 ± 0.17	0.31 ± 0.22	0.3 ± 0.22	0.34 ± 0.22	0.3 ± 0.22	0.3 ± 0.22
Am	0.24 ± 0.18	0.25 ± 0.17	0.27 ± 0.18	0.31 ± 0.19	0.28 ± 0.2	0.29 ± 0.19	0.29 ± 0.18	**0.33± 0.18**
CeblC	0.36 ± 0.04	**0.6± 0.084**	0.46 ± 0.051	0.48 ± 0.052	0.49 ± 0.058	0.49 ± 0.054	0.48 ± 0.053	0.48 ± 0.054
Average	0.38 ± 0.13	**0.5± 0.14**	0.44 ± 0.13	0.49 ± 0.13	0.49 ± 0.14	0.49 ± 0.13	0.49 ± 0.13	0.49 ± 0.13

Table 4.2: Evaluation of the Registration. The mean and standard deviation of the dice coefficient for the 15 different classes of OASIS 3 dataset for the different evaluated methods. The first two rows are baseline methods. The rest of the rows present the results of our proposed method evaluating the different variants and merging operators. The names of the columns represent various brain structures, namely: brain stem (BS), cerebrospinal fluid (CSF), 4th ventricle (4V), amygdala (Am), caudate (Ca), cerebellum cortex (CblmC), cerebellum white matter (CblmWM), cerebral cortex (CeblC), cerebral white matter (CeblWM), hippocampus (Hi), lateral ventricle (LV), pallidum (Pa), putamen (Pu), ventral DC (VDC) and 3rd ventricle (3V).

method *concatenation only reg.* with an average difference of dice reaching 0.05% and with maximum p-values calculated with *concatenation with* $\mathcal{L}^{\star}$ ($t(200) = 3,33$; $p < 10^{-3}$; $d = 0,38$). From our experiments, we saw that the merging operation affects a lot the performance of the *only reg.* model, with the concatenation reporting the worst average Dice than the rest of the methods.

We present in figure 4.4 some qualitative evaluation of the registration component by plotting three different pairs and their registration from all the evaluated models. The first two columns of the figure depict the moving and fixed volumes together with their tissue annotations. The rest of the columns display the deformed moving volume and the deformed tissue annotations for each evaluated method. Visually, all methods perform well on the brain's overall shape, with the higher errors in the deformed annotations being presented at the cerebral white matter and cerebral cortex classes.

Finally, we should also mention that the subjects of the OASIS 3 dataset do not contain regions with tumours. However, our proposed formulation provides tumour masks to evaluate the robustness of the segmentation part. Indeed, our model for all the different combinations of merging operations and loss functions reported a precision score of more than 0.999, indicating its robustness for the tumour segmentation task.

Evaluation on the tumour areas

Method	NCR/NET	ET	ED	Combined
[Dalca, 2018]	2.27 ± 2.68	0.67 ± 0.55	1.96 ± 3.03	0.62 ± 0.51
Proposed				
concatenation only reg.	0.51 ± 0.61	0.26 ± 0.19	0.71 ± 0.94	0.22 ± 0.15
concatenation w/o $\mathcal{L}^{\star}_{sim}$	1.35 ± 1.14	0.64 ± 0.41	1.80 ± 1.82	0.64 ± 0.42
concatenation with $\mathcal{L}^{\star}_{sim}$	0.26 ± 0.20	0.26 ± 0.13	0.30 ± 0.28	0.21 ± 0.12
subtraction only reg.	1.34 ± 0.77	0.77 ± 0.59	2.02 ± 1.65	0.68 ± 0.52
subtraction w/o $\mathcal{L}^{\star}_{sim}$	1.74 ± 1.35	0.72 ± 0.72	2.38 ± 1.74	0.74 ± 0.76
subtraction with $\mathcal{L}^{\star}_{sim}$	**0.24 ± 0.17**	**0.25 ± 0.13**	**0.23 ± 0.22**	**0.20 ± 0.11**

Table 4.3: Quantitative estimates on tumour shrinking. The measure used is the average over 200 testing pairs of patients of the distance between the ratio of the volumes of the deformed moving ground-truth mask and the original ground-truth mask for each original class of the BraTS 2018 dataset (NCR/NET, ET and ED), and the ratio of the fixed brain volume over the moving brain volume. In this context, the best performance reachable is 0 for each class. Additionally, ground-truth masks are binarised into Whole tumour masks, with a value of 1 if and only if a voxel is annotated as one of the 3 tumour classes, and the same measure is computed in the last column ("Combined"), which should indicate the overall impact of tumour shrinking of the whole tumour without considering swapping of intra-tumoral classes.

Even if the proposed method reports very similar performance with models that perform only registration, we argue that it addresses better the registration of the tumour areas, maintaining their geometric properties, as can be inferred in Table 4.3. This statement is also supported by the statistical tests we performed to evaluate the difference in performance between the methods, while registering tumour areas (Table 4.6). In particular, for each of the tumour classes NCR/NET, ET and ED the difference between the [Dalca, 2018] and the proposed method *subtraction with* $\mathcal{L}^{\star}_{sim}$ was significant with NCR/NET: $t(200) = 10.69$; $p < 10^{-3}$; $d = 1.07$ | ET: $t(200) = 10.51$; $p < 10^{-3}$; $d = 1.05$ | ED: $t(200) = 8.05$; $p < 10^{-3}$; $d = 0.81$. The similar behavior was obtained when the evaluation was performed by merging all 3 tumour classes into one (denoted *Combined*). Again, we reported significant differences between [Dalca, 2018] and the proposed method: $t(200) = 11.38$; $p < 10^{-3}$; $d = 1.14$.

To evaluate the performance of the different variants of our proposed method, we compared the performance of the proposed *subtraction with* $\mathcal{L}^{\star}_{sim}$ and *concatenation with* $\mathcal{L}^{\star}_{sim}$ that reported the best performances. Indeed, we did not find significant changes between the two different components except the edema class ($t(200) = 2.78$; $p < 10^{-3}$; $d = 0.28$). Moreover, the proposed *concatenation only reg.* reports also competitive results without using the segmentation masks. In particular, even if the specific method does not report very good performance on the registration evaluated on anatomical structures (Section 4.3.4), it reports very competitive performance on the *Combined* and the smallest in size tumour class (*ET*). However, for the other two classes the difference on the

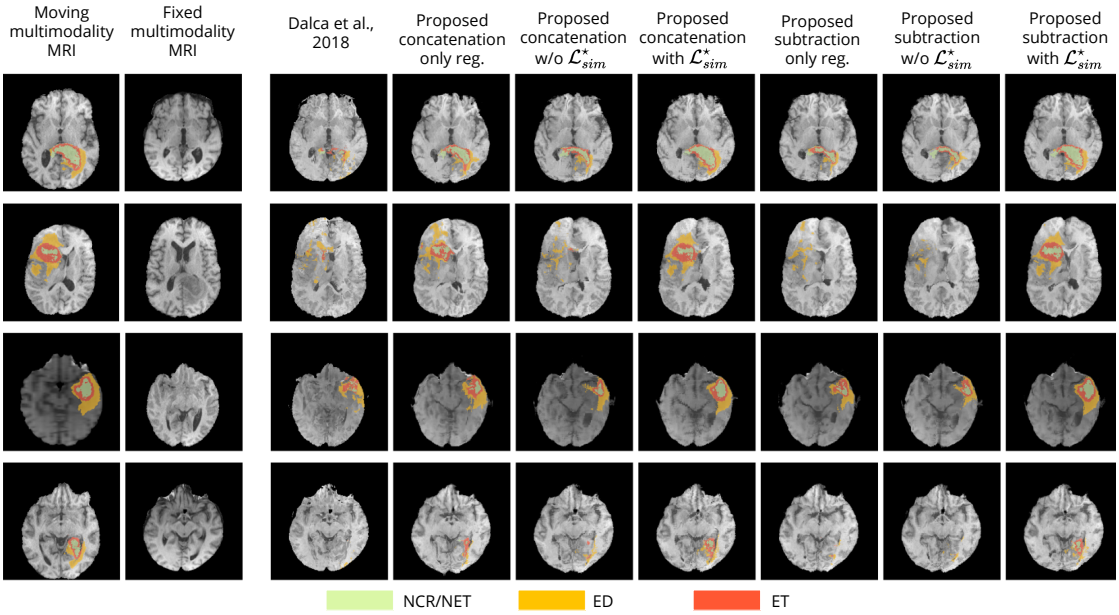


Figure 4.5: Qualitative evaluation of the tumour deformation of different evaluated methods, displayed on T1 modalities. Each line is a sample, with moving MRI in the first column registered on fixed MRI in the second column. BraTS ground-truth annotations are plotted onto the moving MRI. 7 models are benchmarked, one for each of the remaining columns, which display the result of applying the predicted grid onto the moving MRI. For each model and each line, the moving ground-truth annotation masks of the moving MRI were also registered with the predicted deformation grid, and the consequently obtained deformed ground-truth were plotted onto each deformed moving MRI to illustrate the impact of all methods regarding the preservation of tumour extent.

performance that it reports in comparison to the proposed variant *subtraction with* $\mathcal{L}^{\star}_{sim}$ is significant different: NCR/NET: $t(200) = 6,03$; $p < 10^{-3}$; $d = 0,60$ | ED: $t(200) = 7,03$; $p < 10^{-3}$; $d = 0,70$). Here we should mention that even though *subtraction only reg.* works very well for the registration of the anatomical regions (Section 4.3.4), it reports one of the worst results about tumour preservation, with values close to the ones reported by [Dalca, 2018]. This indicates again that the *only reg.* model is highly sensitive to the merging operation, and it cannot simultaneously provide good performance on tumour areas and registration of the entire volume, proving its inferiority to the proposed method using the *with* $\mathcal{L}^{\star}_{sim}$.

Independently of the merging operation with both registration and segmentation tasks, ie with or without $\mathcal{L}^{\star}_{sim}$, we find that the proposed approach works significantly better in preserving tumour areas when optimized with $\mathcal{L}^{\star}_{sim}$ than without (NCR/NET: $t(200) = -14.33$; $p < 0.005$; $d = 1.43$ | ET: $t(200) = -9.99$; $p < 0.005$; $d = 1.00$ | ED: $t(200) = -14.17$; $p < 0.005$; $d = 1.42$ | Combined: $t(200) = -10.94$; $p < 0.005$; $d = 1.09$).

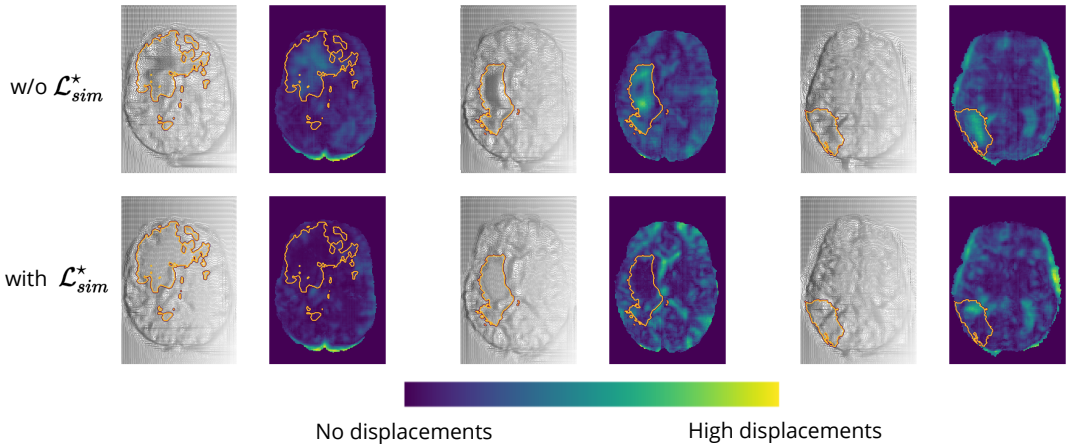


Figure 4.6: Comparison of the registration grid of the proposed model using the subtraction operation with and w/o $\mathcal{L}^{\star}_{sim}$. This figure is obtained by sampling three random pairs of test patients and computing the predicted registration fields, which are displayed byline for the two models, and in consecutive columns, one for each of the three dimensions, showing the registration field as a warped grid (grayscale) and as a coloured map obtained by computing its norm pixels (blue-green map). Furthermore, the whole tumour's contour is plotted on top of each image, obtained from the ground truth segmentation.

Figure 4.5 presents some qualitative examples from the BraTS 2018 to evaluate the performance of the different methods. The first two columns present the pair of images to be registered and segmented, and the rest of the columns the deformed moving image with the segmented tumour region superimposed. One can observe that most of the methods that are based only on registration ([Dalca, 2018], proposed concatenation and subtraction *only reg.*) together with the proposed concatenation and subtraction *w/o* $\mathcal{L}^{\star}_{sim}$ do not preserve the geometry of the tumour, tending to significantly reduce the area of tumour after registration, or intermix the different types of tumour. On the other hand, the proposed *with* $\mathcal{L}^{\star}_{sim}$ behaviour seems to be much better, with the tumour area properly maintained in the deformed volume.

Moreover, in Figure 4.6 we provide a better visualisation for the displacement grid inside the tumour area, highlighting the importance of Eq. 4.4. Indeed, one can observe that the displacements inside the tumour area are much smoother and relaxed when we use the information about tumour segmentation.

4.4 Discussion and Conclusion

In this chapter, we proposed a novel deep learning-based framework to address segmentation and registration simultaneously. The framework combines and generates features, integrating valuable information from both tasks within a bidirectional manner while it takes advantage of all the available modalities, making it quite robust and generic. Our model's performance indicates highly promising results comparable to recent state-of-the-art models that address each of the tasks separately [Dalca, 2018]. However, we reported a better behaviour of the model in the proximity of tumour regions. This behaviour has been achieved by training a shared encoder that generates meaningful features for both registration and segmentation problems. Simultaneously, these two problems have been coupled in a joint loss function, forcing the network to focus on regions that exist in both volumes.

Even if we could not do a proper comparison with [Parisot, 2012] which shares similar concepts, our method provides very good improvements. In particular, we train both problems at the same time, without using pre-calculated classification probabilities. The method proposed in [Parisot, 2012]) is based on a pre-calculated classifier indicating the tumoral regions. The authors provided their segmentation results by adapting the Gentle Adaboost algorithm and using different features, including intensity values, texture such as Gabor filters and symmetry. After training the classifier, they defined an MRF model to optimise their predictions by considering pairwise relations. By adopting this strategy, the used probabilities for the tumour regions are not optimised simultaneously with the registration, something that is not the case in our methodology. In particular, by sharing representation between the registration and segmentation tasks, we argue that we can create more complex and useful features that come from both problems. Using a deep learning architecture that is end-to-end trainable, we can automatically extract features that are suitable to deal with both problems. Moreover, our implementation is modular and scalable, permitting easy integration of multiple modalities, which is not so straightforward with [Parisot, 2012] as it is more complicated to adapt and calculate the different similarity measures and classifiers taking into account all these modalities. Finally, we should mention that our method takes advantage of GPU implementation, needing only a few seconds to provide segmentation and displacement maps, while the method in [Parisot, 2012] needs approximately 6 minutes.

Both qualitative and quantitative evaluations of the proposed architecture highlight the great potentials of the proposed method reporting more than 0.66 Dice coefficient for the segmentation of the different tumour areas, evaluated on the publicly available BraTS 2018 validation set. Our formulation reported similar behaviour to the model with only the segmentation block, which indicates that the joint formulation did not affect the tumour segmentation's performance. Nevertheless, it provides more complex models providing tumour segmentation masks for two images simultaneously, predicting simultaneously optimal displacements between them. Moreover, both concatenation and subtraction operators report similar performances, an expected result for the specific segmentation

task, since the merging operation is mainly used on the registration decoder, even if it affects the learned parameters of the encoder and thus indirectly the segmentation decoder.

Concerning the comparison between top-performing tumour segmentation methods, although our formulation underperforms the winning methods of BraTS 2018, we want to highlight two major points. First of all, our formulation is modular because different network architectures with optimised components for tumour segmentation can be evaluated depending on the application and the problem's goals. For our experiments, we chose a simple VNet architecture [Milletari, 2016] proving that the registration components do not significantly hinder the segmentation performance and indicating the soundness of our method. However, any other encoder-decoder architecture can be used and evaluated. Secondly, our method's main goal was the proper registration and segmentation of the tumoral regions and the rest of the anatomical structures, and that was the main reason we did not optimise our network architecture according to the winning methods of BraTS 2018. However, we demonstrated that we could register properly tumoral and anatomical structures with a very simple architecture while segmenting with more than 76% of Dice the tumoral regions.

Continuing with the evaluation of the registration performance, once more, the joint multi-task framework reports similar and without statistical difference performance with formulations that address only the registration task evaluated on anatomical regions that exist on both volumes. However, we argue that abnormal region registration is better addressed in terms of qualitative and quantitative metrics. Moreover, from our experiments, we observed that subtraction of the tumours' coding features reports higher performances for registering the tumour areas. This indicates that subtraction can capture and code more informative features for the registration task. What is more, we achieved very good generalisation for all the deep learning-based registration methods, as they reported very stable performance in a completely unseen dataset (part of the OASIS3).

Even if, from our experiments, the competence of our proposed method for both registration and segmentation tasks is indicated, we report a much better performance for the registration of the tumoral regions. In particular, in one joint framework, we could efficiently and accurately produce tumour segmentation maps for both moving and fixed images and their displacement maps that register the moving volume to the fixed volume space. Our experiments indicated that the proposed method with the $\mathcal{L}^{\star}_{sim}$ variant register the anatomical properly together with the tumoral regions with statistical significance compared to the rest of the methods. Both qualitative and quantitative evaluations of the different components indicate the superiority of the with $\mathcal{L}^{\star}_{sim}$ variant of the proposed method for brain MRI registration with tumour extent preservation. Using such a formulation, the network focuses on improving local displacements on tissues anywhere in the common brain space instead of minimising the loss within the tumour regions, which are empirically the regions with the highest registration errors. Consequently, the network improves its registration performance on non-tumour regions (as discussed in Section Evaluation on anatomical structures) while also relaxing the obtained displacements inside those predicted tumour regions.

Some limitations of our method include the number of parameters that have to be tuned during the training due to the multi-task nature of our formulation, namely α and β that affect the performance of the network. Moreover, due to the multimodal input and the two decoders, the network cannot be very deep due to GPU memory limitations. Our joint formulation could be improved, especially concerning the registration part. We resolved the previous chapter issue concerning the non-symmetric formulation of the segmentation decoder by introducing the shared encoder. Although, we did not change the registration formulation. In particular, the proposed registration do not respect symmetric, inverse consistency and diffeomorphism properties. In the next chapter, we investigate these points more precisely. Adding such properties to the deformation should result in more realistic warped images.

Future work involves extending our registration pipeline to new organs and imaging modalities and the relationship between registration and clinical applications. We plan to explore longitudinal registration and if the deformation grid can capture information concerning tumour growth. Finally, automatically obtaining the training parameters α and β should be investigated together with adversarial losses to improve the optimisation.

Although the pipeline was built using different patients for the registration task as a proof of concept, such a tool could have numerous clinical practice applications, especially when applied in different images acquired from the same patient. Regarding the radiotherapy treatment planning, several studies have shown that significant changes of the targeted volumes in the brain occurred during radiotherapy, raising the question of replanning treatment to reduce the amount of healthy brain irradiated in case of tumour reduction or readapting the treatment for brain tumours that grow during radiation [Champ, 2012; Yang, 2016; Mehta, 2018]. Since MR-guided linear accelerator will offer the opportunity to acquire daily images during RT treatment, the proposed tool could help with automatic segmentation and image registration for replanning purposes, and it could also allow accurate evaluation of the dose delivered in targeted volumes and healthy tissues by taking into account the different volume changes. Moreover, while imaging features under treatment are known to be associated with treatment outcomes in several cancer diseases [Fave, 2017; Vera, 2014], the registration grid computed from two same-patient acquisitions realised at different times allows an objective and precise evaluation of the tumour changes.

4.5 Appendix

Name	Input	Res. input	Operations	Output shape
			Encoder	
Enc^1	4D MRI		$\text{Conv}_{1,8}$, ReLU, ($\text{Conv}_{3,8}$, ReLU), AddId,	(144, 208, 144, 8)
Enc^2	Enc^1		$\text{Conv}_{2,16}$, ReLU, ($\text{Conv}_{3,16}$, ReLU)$*2$, AddId	(72, 104, 72, 16)
Enc^3	Enc^2		$\text{Conv}_{2,32}$, ReLU, ($\text{Conv}_{3,32}$, ReLU)$*3$, AddId	(36, 52, 36, 32)
Enc^4	Enc^3		$\text{Conv}_{2,64}$, ReLU, ($\text{Conv}_{3,64}$, ReLU)$*3$, AddId	(18, 26, 18, 64)
Enc^5	Enc^4		$\text{Conv}_{2,128}$, ReLU, ($\text{Conv}_{3,128}$, ReLU)$*3$, AddId	(9, 13, 9, 128)
			Segmentation decoder	
Dec^4_{seg}	Enc^5	Enc^4	$\text{DeConv}_{2,64}$, ReLU, ResConc, ($\text{Conv}_{3,64}$, ReLU)$*3$, AddId	(18, 26, 18, 64)
Dec^3_{seg}	Dec^4_{seg}	Enc^3	$\text{DeConv}_{2,32}$, ReLU, ResConc, ($\text{Conv}_{3,32}$, ReLU)$*3$, AddId	(36, 52, 36, 32)
Dec^2_{seg}	Dec^3_{seg}	Enc^2	$\text{DeConv}_{2,16}$, ReLU, ResConc, ($\text{Conv}_{3,16}$, ReLU)$*2$, AddId	(72, 104, 72, 16)
Dec^1_{seg}	Dec^2_{seg}	Enc^1	$\text{DeConv}_{2,8}$, ReLU, ResConc, ($\text{Conv}_{3,8}$, ReLU), AddId	(144, 208, 144, 8)
Dec^0_{seg}	Dec^1_{seg}		$\text{Conv}_{1,4}$, Softmax	(144, 208, 144, 4)
			Registration decoder	
Merge	Enc^i_M, Enc^i_F		For all $1 \leq i \leq 5, \text{MEnc}^i = \text{Enc}^i_M \oplus \text{Enc}^i_F$	
Dec^4_{reg}	MEnc^5	MEnc^4	$\text{DeConv}_{2,64}$, ReLU, ResConc, ($\text{Conv}_{3,64}$, ReLU)$*3$, AddId	(18, 26, 18, 64)
Dec^3_{reg}	Dec^4_{reg}	MEnc^3	$\text{DeConv}_{2,32}$, ReLU, ResConc, ($\text{Conv}_{3,32}$, ReLU)$*3$, AddId	(36, 52, 36, 32)
Dec^2_{reg}	Dec^3_{reg}	MEnc^2	$\text{DeConv}_{2,16}$, ReLU, ResConc, ($\text{Conv}_{3,16}$, ReLU)$*2$, AddId	(72, 104, 72, 16)
Dec^1_{reg}	Dec^2_{reg}	MEnc^1	$\text{DeConv}_{2,8}$, ReLU, ResConc, ($\text{Conv}_{3,8}$, ReLU), AddId	(144, 208, 144, 8)
Dec^0_{reg}	Dec^1_{reg}		$\text{Conv}_{1,3}$, Sigmoid	(144, 208, 144, 3)

Table 4.4: Layer architecture of the encoder, the segmentation and the registration decoders. The sub-architectures are grouped into blocks, one per table line, whose names are indicated in the first column. Each block processed a forward signal as input identified by the second column. Additionally, both decoders have residual connections from different encoder stages, identified by the third column. The blocks are made of a set of successive operations where $\text{Conv}_{w,f}$ (resp. $\text{DeConv}_{w,f}$) stands for a convolutional (resp. deconvolutional) layer with weight size $w \times w \times w$ and f filters, ReLU - Rectified Linear Unit, AddId - intra-block residual connection with the output of the first activated convolution of the corresponding block, ResConc - encoder to decoder - residual connection from the output of the third column block to the current signal, Softmax and Sigmoid - finale output activation. $*$ indicates successive repetition of the previous operations in parenthesis. For convolutions and deconvolutions layers, strides is $1 \times 1 \times 1$ except for the $\text{Conv}_{2,\cdot}$ which is $2 \times 2 \times 2$. The first layer of the registration decoder indicates the merging operation of the moving and the fixed signals, which are obtained by inferring them successively in the encoder; $\oplus$ indicates elementwise subtraction or channel-wise concatenation of the moving and fixed list of tensors (forward network signal and four residual connection signals). The last column indicates each block output shape (channels last).

Method	Average		Dice			Hausdorff95		
	Dice	Hausdorff95	ET	WT	TC	ET	WT	TC
Baseline segmentation	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Proposed								
concatenation w/o $\mathcal{L}^{\star}_{sim}$	0.32	0.46	0.55	1.00	0.03	0.34	0.74	0.14
concatenation with $\mathcal{L}^{\star}_{sim}$	0.24	0.72	0.33	1.00	0.05	0.31	0.21	0.24
subtraction w/o $\mathcal{L}^{\star}_{sim}$	0.55	0.65	0.69	0.62	0.24	0.28	0.91	0.58
subtraction with $\mathcal{L}^{\star}_{sim}$	0.55	0.60	0.69	0.62	0.16	0.48	0.79	0.21

Table 4.5: Statistical significance of the proposed methods with [Milletari, 2016] on the BraTS segmentation task. For each model (line) and each performance measure (column), the displayed value is the p-value, up to 2 significant figures, of the statistical significance between the model and [Milletari, 2016] for the corresponding measure (Dice or Hausdorff95) on the corresponding tumour class (ET, WT, TC, or the union of the 3 latter in the two columns *Average*) on the 66 testing samples of BraTS. No p-values are statistically significant between all of the proposed variants and [Milletari, 2016]. Blue line represents the fixed model, red cells indicate no statistical significant p-values (cutoff 0.005) while green color represent statistical significant p-values.

Method	NCR/NET	ET	ED	Combined
[Dalca, 2018]	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
Proposed				
concatenation only reg.	$< 10^{-3}$	0.540	$< 10^{-3}$	0.130
concatenation w/o $\mathcal{L}^{\star}_{sim}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
concatenation with $\mathcal{L}^{\star}_{sim}$	0.282	0.442	0.006	0.386
subtraction only reg.	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
subtraction w/o $\mathcal{L}^{\star}_{sim}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
subtraction with $\mathcal{L}^{\star}_{sim}$	1.000	1.000	1.000	1.000

Table 4.6: Statistical significance of the proposed methods and [Dalca, 2018], with the best proposed variant *subtraction with* $\mathcal{L}^{\star}_{sim}$ regarding the tumour shrinking preservation on the OASIS 3 registration task. For each model (line) and each performance measure (column), the displayed value is the p-value, up to 3 significant figures, of the statistical significance between the model and *subtraction with* $\mathcal{L}^{\star}_{sim}$ for the tumour preservation measure on the corresponding tumour class (NCR/NET, ET, ED, and the union of the 3 latter in the column *Combined*) on the 200 testing pairs of OASIS 3. Blue line represents the fixed model, red cells indicate no statistical significant p-values while green color represent statistical significant p-values.

Chapter 5

Multi-steps, Symmetric and Inverse-Consistency Registration Network

"Changer le monde, changer le monde
vous êtes bien sympathiques mais faudrait
déjà vous levez le matin. [...]
C'est le vrai monde dehors et le vrai monde
il va chez le coiffeur."

OSS 117 : Rio Ne Répond Plus

Contents

In chapter 3 and 4, we propose a joint segmentation and registration network, working with brain MRI. In this chapter, we shift to a new anatomical: abdominal CT. This anatomical part is very challenging as there are many natural deformations and non-rigid-organs. As registration is an ill-posed problem, we focus the design of our algorithm on introducing different registration properties: inverse consistency, symmetry and conservation of the orientation. We investigate the impact of diverse losses while we propose a multi-step strategy and a pretraining step using pseudo-segmentations to identify the best deformations. We evaluate our method on a dataset used during the Learn2Reg challenge, allowing a fair comparison with published methods. This chapter resulted in participating in the Learn2Reg Challenge, organised in parallel with the MICCAI 2020 conference, with a 2^{nd} position and an oral presentation. We also published in the workshop proceedings [Estienne, 2021a] and submitted an extension to a journal [Estienne, 2021b].

5.1 Introduction

One significant difficulty of registration is the evaluation of its performance. Registration is an ill-posed problem as many different transformations could warp one image to another, even using affine transformation. Evaluating the performance in terms of organs or landmarks correspondences is one of the standard ways to benchmark the performance of the registration problem. Such an evaluation does not reflect the quality of the deformation grid. To overcome this problem and to measure the grid's plausibility, different metrics have been proposed, such as the standard deviation of the Jacobian [Dalca, 2020; Hering, 2021].

In machine learning and deep learning, the benchmark of algorithms' performances has often been performed through the organisation of challenges and evaluation on publicly available datasets. Indeed, the problem of overfitting makes it essential to compare algorithms with identical training and testing datasets. Concerning registration, different challenges have been performed on brain MRI [Klein, 2009], on thoracic CT during the EMPIRE10 challenge [Murphy, 2011] and for abdominal CT [Xu, 2016]. Recently Learn2Reg, a registration challenge, was organised during the Miccai conference [Dalca, 2020; Hering, 2021]. It comprises four tasks: multimodal interoperative brain registration (CT-US), CT lung inspiration-expiration registration, abdominal CT registration and MRI hippocampus registration. Among the best solutions were a multi-scale deep learning approach [Mok, 2020a], a probabilistic dense displacements network [Heinrich, 2019], a classical approach using Markov Random Field [Heinrich, 2013] and our solution, a deep learning approach based on pretraining and spatial gradients [Estienne, 2021a]. In this chapter, we present our participation in the third and fourth tasks of the Learn2Reg Challenge. Then, we upgrade our original method and compare it with the best-performing methods. Evaluating our approach on a public dataset allows a fair comparison with our proposed method and other algorithms.

Recently, different reviews detail deep learning-based registration, its development, the different approaches and the future trends [Haskins, 2020; Boveiri, 2020; Fu, 2020]. More precisely, Boveiri et al. [Boveiri, 2020] analysed the most studied organs that appeared on papers for registration. Brain MRI is far ahead of other organs, with more than 70 papers addressing this problem by publication date. The brain has several advantages: the absence of large deformation due to the skull's surface or large public datasets easing network training. We explore here the application of deep learning registration to abdominal CT. This task is more difficult than brain registration because of the large deformation in the abdomen and the lack of correspondence.

In the preceding chapters, we introduced a joint architecture for both segmentation and registration, and we applied it to brain MRI with and without abnormalities. In this chapter, we concentrate on increasing the registration's accuracy while preventing the emergence of many irregularities in the deformation field. We propose a deep learning algorithm that learns to register abdominal CTs while respecting suitable topological properties. Our convolutional network project both images to a common latent space and merge them to output the deformation gradient. To raise the registration performance, we investigate a multi-step strategy and the use of pretraining using many public

abdominal datasets. The pretraining is also upgraded by the introduction of pseudo-segmentations produced by a segmentation network. The pseudo-segmentations allow weakly supervised pretraining while the public datasets collected do not have segmentation masks. Concerning the smoothness of the deformation grid, we study the impact of two new regularisation losses, and we introduce symmetry constraints as well as a multi-step formulation to refine and identify small deformations. The training is split into two parts, the pretraining with supplementary data and pseudo segmentations and the fine-tuning using only the challenge's data. Our main contributions are the following :

1. Optimising our network with three different regularisation losses to respect symmetry, inverse consistency and local orientation conservation;

2. Developing a multi-step framework to refine the predicted deformation with respect to our symmetric formulation.

3. Taking advantage of publicly available data to pretrain and fine-tune our network.

In the section 5.2, we present existing research on deep learning registration and regularisation strategies. Section 5.3 describes our methodology, including our symmetric approach and the multi-steps formulation. Section 5.4 presents our experiments performed on abdominal CT and the results, and in the section 5.5, we discuss and analyse the presented results.

5.2 Related work

Previously, we described the different categories of deep learning-based registration (supervised, unsupervised and weakly supervised) and the various architecture published following unsupervised and weakly-supervised framework. Among the existing modifications, we focus in this chapter on those who bring more regularisation on the predicted transformations. Different regularisation techniques have been explored in classical registration algorithms, such as flows of diffeomorphisms with the LDDMM framework [Beg, 2005; Yan Cao, 2005] or symmetric normalisation formulation [Avants, 2008]. Another regularisation strategy is to respect several properties, including symmetry, inverse consistency, or diffeomorphism [Sotiras, 2013]. In deep learning approaches, constraints are often created by the addition of supplementary losses or by modifications to the network architecture. In this chapter, we evaluate the influence of various losses on the regularisation of the grid.

Among the constraints, different researchers explored methods to regularise the determinant of the Jacobian matrix. Negative values of the Jacobian reveal local non-topological behaviour of the deformation, while it also provides us with valuable information about the local conservation of the orientation. Thus, enforcing the Jacobian to obtain positive values help to generate more realistic deformations. Different formulation of Jacobian constraints can be found in Mok et al. [Mok, 2020b], Kuang et al. [Kuang, 2019b], and Zhang et al. [Zhang, 2020]. Hering et al. [Hering, 2020] proposed a different approach by controlling the changes of volume. Another formulation was developed by Mansilla et al. [Mansilla, 2020] to take into account anatomical constraints. They added a

supplementary denoising autoencoder to the registration pipeline, and they constrained the encoding of the warped segmentation to be as close as possible to the encoding of the target segmentation. Applying the loss at the encoding scale will create a global matching witch is not produced by pixel level loss.

Recent formulations introduced symmetry properties through cycle-gan approaches. In this framework, two networks are needed. The first one maps the volumes from the X space to the Y, while the other does the inverse. Two cycle-consistency losses are then introduced, compelling the composition of the two networks to return to the original image. Cycle-gans have produced excellent results for image-to-image translation [Zhu, 2017]. Kim et al. [Kim, 2019] introduced the cycle formulation to ensure topological preservation and tested it for registration of the liver anatomy. Cycle formulations are well suited for multi-modality registration, where the moving and target images are in different spaces. This was proposed by Wang et al. [Wang, 2019] to perform T1-weighted to T2 Flair brain MRI registration. Finally, [Wang, 2020] combined cycle formulation and weakly supervised registration, using two cycle-losses, one calculated on images and the other one on segmentation masks.

Multi-scale or multilevel formulations are another strategy to smooth the transformation and reduce local perturbations. UNet architecture is widely spread among unsupervised registration methods. It could be considered as a multilevel formulation as down-sampling, and up-sampling operations are applied during the forward pass. However, in the registration context, we consider a formulation as multi-scale if deformations at different resolutions are explicitly calculated and applied to the source image. Such approaches are proposed in [Mok, 2020a; Hering, 2019; Fechter, 2020]. They differ by the way the different level transformations are combined and the training strategy.

Finally, multi-step formulations refine predicted deformations by applying the registration network between the deformed image and the fixed image. These approaches are also known as cascade networks, recursive registration networks or multi-stage networks. Such formulations were applied to the liver and the heart and proposed in [Zhao, 2019; Zhao, 2020a; de Vos, 2019]. Differences between these approaches lie in the use of one or several independent networks or in the training procedure, which can be performed in an end-to-end way or one network at a time and then freezing it before passing to the next one.

[Mok, 2020b] is the closest method to our study. They proposed a deep learning diffeomorphic and symmetric approach. Their diffeomorphic formulation is based on stationary velocity field and scaling and squaring procedures. Thus, our method has some significant differences, including weakly supervision, the symmetric formulation, the inverse consistency loss and the multi-step formulation.

5.3 Method

5.3.1 Registration

In the rest of this manuscript, the images are denominated as F and M for respectively fixed and moving images. However, in this chapter, we choose to designate them as A and B, as they have a symmetrical role in our approach. The registration goal is to find the best deformation Φ to warp one image into another. Our network g_θ takes as input A and B and return the two transformations $\Phi_{A\to B}$ and $\Phi_{B\to A}$. As in the previous chapters of this thesis, we use a gradient-based approach to predict our deformation [Stergios, 2018; Shu, 2018; Estienne, 2020]. Others formulation are based on outputting the displacement field u or use the *diffeomorphic* formulation [Dalca, 2018; Krebs, 2018; Mok, 2020b], outputting a stationary velocity field v and obtaining Φ by scaling and squaring algorithm [Ashburner, 2007; Arsigny, 2006]. More details on the different registration formulation are given in section 2.5.3 and section 3.3.1.

One major difference between other deep learning methods and ours is how images are passed through the network. Most of the methods in the literature concatenate the two volumes before passing them through the network [Balakrishnan, 2019; de Vos, 2019; Krebs, 2019]. The network process then a four dimension volume including x, y and z axis and two channels representing each volume. In the previous chapter, we introduced a late fusion strategy [Estienne, 2020; Estienne, 2021a]. Lets consider that our network g_θ is divided into one encoder $\mathbf{E}$ and one decoder $\mathbf{D}$. We pass the images A and B independently through the encoder $\mathbf{E}$. Then we merge the encoding of A and B through the subtraction operation, and we obtain the gradients by applying the decoder to it. Finally, the expression of our spatial gradient is :

$$\begin{aligned}\nabla\Phi_{A\to B} &= \mathbf{D}\left(\mathbf{E}(A) - \mathbf{E}(B)\right)\\ \nabla\Phi_{B\to A} &= \mathbf{D}\left(\mathbf{E}(B) - \mathbf{E}(A)\right)\end{aligned} \tag{5.1}$$

This formulation has the advantage to require only one network to generate the forward and backward transformation, while cycle gan methods use one network for the forward transformation and one network for the backward [Kim, 2019; Wang, 2020]. We depict a schematic representation of our formulation on Figure 5.1.

5.3.2 Network architecture

Our network is based on the 3D UNet architecture [Çiçek, 2016]. It consists of a symmetrical encoder-decoder architecture with four blocks with 64, 128, 256 and 512 channels, resulting in 20 million parameters. Each block includes 3D convolution layers with a kernel of size 3, instance normalisation layer, leaky ReLU activation function. The up and down-sampling operations are performed by 3D convolution with kernel size 2 and stride 2, and the encoder and decoder are connected through skip connection. The final layer predicts a three channels tensor between 0 and 1 thanks to the

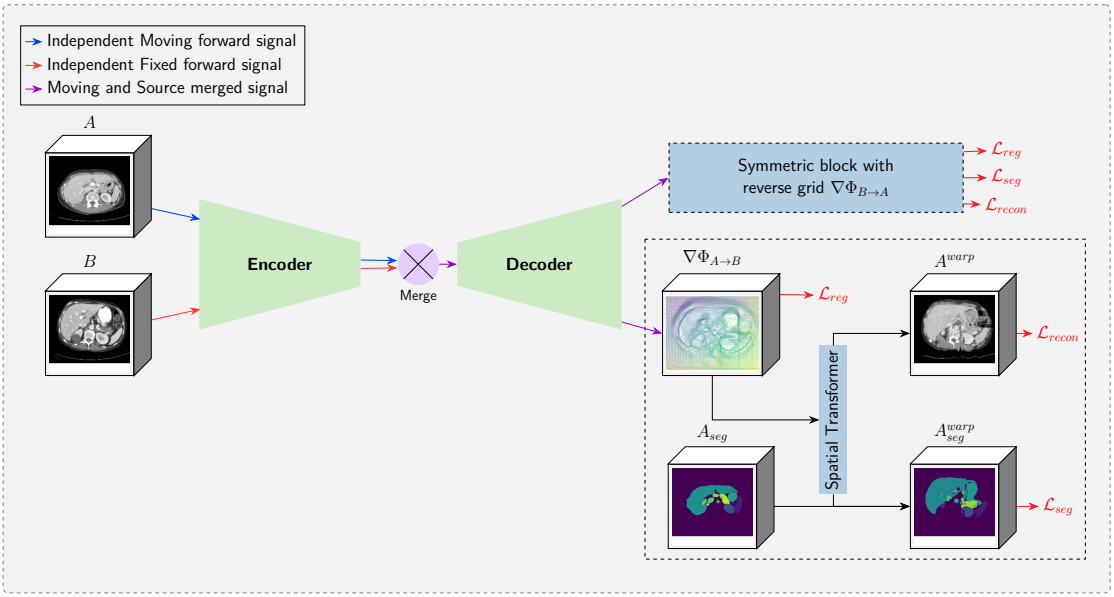


Figure 5.1: Representation of our symmetric formulation. The framework comprises one encoder and one decoder. The two volumes A and B are passed independently through the encoder E and the merge bloc combines the two forward signals. We used a symmetrical formulation, predicting both the displacements $\Phi_{A\to B}$ and $\Phi_{B\to A}$.

sigmoid operation, which is integrated through cumulative sum to obtain the transformation. The warped image and segmentation are then produced using the spatial transformer. The two main differences between this network and the one from the preceding chapter are the addition of instance normalisation layers and the multiplication by 4 of the channels' numbers.

5.3.3 Multi-steps formulation

To improve the registration produced by our network, we implement a multi-steps strategy. The idea is to refine the deformation by predicting the transformation between the deformed image and the target. This strategy can be applied either during the training of the network or during the inference. Lets define the initial grid as the grid at step one : $\nabla\Phi^1_{A\to B} = \mathbf{D}\left(\mathbf{E}(A) - \mathbf{E}(B)\right)$. The warped image A at step one is then : $\widehat{A}^1 = \mathcal{W}\left(A, \Phi^1_{A\to B}\right)$. We can now define recursively the deformation grid and the deformed image at step i :

$$\begin{aligned} \nabla\Phi^i_{A\to B} &= \mathbf{D}\left(\mathbf{E}(\widehat{A}^{i-1}) - \mathbf{E}(B)\right) \\ \widehat{A}^i &= \mathcal{W}\left(\widehat{A}^{i-1}, \Phi^i_{A\to B}\right) \end{aligned} \tag{5.2}$$

The multi-step deformation of the image B is obtained by inverting A and B in the equation 5.2. Our multi-step formulation takes advantage of our strategy to combine the two images (eq 5.1).

Indeed, we only need to pass the deformed image at step i through the encoder and then merge it with the target's encoding. Thus, we gained calculation time and memory usage.

5.3.4 Pretraining and Pseudo segmentations

Transfer learning is a common strategy to improve the performance of deep learning algorithms. In two dimensions, people often initialise networks with weights coming from the ImageNet dataset. However, in three dimensions, there is no consensus about pretraining strategies. Thus recent publications proposed three dimensions pretrained networks with tasks such as segmentation [Chen, 2019] or unsupervised pretraining [Zhou, 2019b].

To improve transfer learning performance, we proposed a pretraining strategy using directly the registration task instead of using a different task like segmentation or reconstruction. Our proposed pipeline is: First, train the registration task with a large dataset built from public medical databases and then fine-tune the network with only the official dataset. One advantage of registration compared to other tasks is that the network can be trained in an unsupervised way. Thus we can only collect data and do not need to have corresponding labels. However, weakly supervised registration has proved to outperform unsupervised training [Hu, 2018b; Hering, 2018; Balakrishnan, 2019]. For this reason, we experiment with the influence of weakly supervised pretraining.

As the supplementary data do not have the same segmentation mask as our original dataset, we have to design a strategy to obtain these labels. One possibility would have been to generate the segmentation masks by clinicians, who would have also been required to devote an excessive amount of time to perform annotations. Instead, we decided to train a segmentation network to generate the segmentation labels automatically without human interaction. As these labels have been acquired by an automatic and trained algorithm, we designate them as *pseudo labels*. Our segmentation network is a 3D UNet [Çiçek, 2016] trained with the following parameters : batch size equal to 6, learning rate equal to $1e^{-4}$, leaky ReLU activation functions, instance normalisation layers and random crop of patch of size $144 \times 144 \times 144$. Depending on the dataset, each image has different organs segmented by doctors. Thus we used a modified Dice loss to back-propagate only the available labels. We apply various post-processing steps to improve the pseudo labels' quality: keep the ground-truths labels for the organs available, keep only the biggest connected component of the predicted label to remove small segmentation and manual inspection of the predicted segmentation to remove outlier results.

Our pipeline is defined as follows: *i)* Train a segmentation network using available masks *ii)* Predict the pseudo labels using the trained network *iii)* Train a new network on the registration task, using supplementary data and pseudo segmentations *iv)* Fine-tune the registration network using only the challenge data for the task in question and ground-truths segmentations.

5.3.5 Loss Function

We train our network by minimising a combination of five different losses. These losses have two different goals: obtaining the best transformation to warp A to B and ensuring that the grid respect desirable topological properties. As we developed a symmetrical approach to the registration problem, each loss will also have a symmetrical formulation.

The first two losses, the similarity loss $\mathcal{L}_{sim}$ and the segmentation loss $\mathcal{L}_{seg}$, constrain the network to produce the best deformation. The similarity loss is the mean square error between the deformed image and the target image. Other approaches substitute the mean square error by the local cross-correlation for $\mathcal{L}_{sim}$ [Balakrishnan, 2019; Mok, 2020b; Mansilla, 2020]. However, for our problem, local cross-correlation did not help during the training. The segmentation loss consists of a Dice loss [Milletari, 2016] between the ground-truth segmentation and the deformed segmentation and thus is a supervised loss contrary to $\mathcal{L}_{sim}$. Many recent articles show that weakly-segmentation registration outperforms unsupervised registration [Hu, 2018b; Hering, 2018; Balakrishnan, 2019]. The network can then learn to produce deformation not only in function of image intensity but also from organs. Their expression is the following :

$$\mathcal{L}_{sim} = ||\widehat{A} - B||^2 + ||\widehat{B} - A||^2 \tag{5.3}$$

$$\mathcal{L}_{seg} = \text{Dice}\left(\widehat{A}_{seg}, B_{seg}\right) + \text{Dice}\left(\widehat{B}_{seg}, A_{seg}\right) \tag{5.4}$$

We add three supplementary regularisation losses to force our network to produce realistic deformations. The smooth loss $\mathcal{L}_{smooth}$ control the smoothness of the predicted deformation. The Jacobian loss $\mathcal{L}_{jac}$ impose the positivity of the Jacobian and the inverse consistency loss $\mathcal{L}_{inv}$ force the predicted grids to be the inverse of each other.

The regularisation loss (or smooth loss) is one of the most common losses in deep learning-based registration. Indeed, minimising only $\mathcal{L}_{sim}$ and $\mathcal{L}_{seg}$ could create non-realistic deformations. This loss penalises high values of the gradient to enforce the smoothness of the grid. Contrary to most of the approaches, our network predicts the gradient directly (see section 5.3.1).

However, the smooth loss is not enough to have a grid that respects topological properties. It does not prevent folding and wrong orientation. A new loss was proposed in recent papers [Mok, 2020b; Kuang, 2019b; Zhang, 2020] to impose positive values of the determinant of the Jacobian matrix (also called the Jacobian). The expression of the Jacobian matrix of the deformation Φ at one voxel p is the following :

$$J_\Phi(\mathbf{p}) = \begin{pmatrix} \frac{\partial \Phi_x}{\partial x} & \frac{\partial \Phi_x}{\partial y} & \frac{\partial \Phi_x}{\partial z} \\ \frac{\partial \Phi_y}{\partial x} & \frac{\partial \Phi_y}{\partial y} & \frac{\partial \Phi_y}{\partial z} \\ \frac{\partial \Phi_z}{\partial x} & \frac{\partial \Phi_z}{\partial y} & \frac{\partial \Phi_z}{\partial z} \end{pmatrix} (\mathbf{p}) \tag{5.5}$$

We can then calculate the determinant for each voxel p and obtain the Jacobian in the same form

as the moving and fixed volumes. The Jacobian characterise two properties of the local behaviour of the deformation. First, the sign of the Jacobian informs us of the local orientation of the deformation field. If its value at voxel p is negative, the registration reverses the orientation locally around p. On the opposite, it conserves the orientation if the Jacobian is positive. Secondly, the transformation is locally invertible around p if the Jacobian is non-zero at p. Thus, we want to compel the Jacobian to be strictly positive. The Jacobian loss $\mathcal{L}_{jac}$ is the sum of all negative values of the Jacobian.

Our last loss constrains the network to generate symmetric transformations which are inverse to each other. Many recent deep learning formulations do not respect these properties. Our formulation is symmetric by construction, as we predict both $\Phi_{A\to B}$ and $\Phi_{B\to A}$. Though, we do not have guarantees that the two transformations are effectively the inverse of each other. Therefore, we implemented a new loss to respect the inverse consistency properties. This loss consists in penalise the difference between the composition of the two transformations and the identity transformation. The composition of the transformation is performed using the spatial transformer. Another formulation of the inverse consistency loss was proposed in Zhang [Zhang, 2018].

The mathematical formulation of our three regularisation loss is the following :

$$\mathcal{L}_{smooth} = ||\nabla\Phi_{A\to B}|| + ||\nabla\Phi_{B\to A}|| \tag{5.6}$$

$$\mathcal{L}_{jac} = \frac{1}{N}\sum_{p\in\Omega} \max\left(0, -|J_{\Phi_{A\to B}}|\right) + \max\left(0, -|J_{\Phi_{B\to A}}|\right) \tag{5.7}$$

$$\mathcal{L}_{inv} = ||\Phi_{A\to B}\circ\Phi_{B\to A} - \Phi_{Id}|| + ||\Phi_{B\to A}\circ\Phi_{A\to B} - \Phi_{Id}|| \tag{5.8}$$

Finally, our total loss will be the combination of these five different loss with their respective weight α, β, γ, δ and ϵ :

$$\mathcal{L} = \alpha\mathcal{L}_{sim} + \beta\mathcal{L}_{seg} + \gamma\mathcal{L}_{reg} + \delta\mathcal{L}_{jac} + \epsilon\mathcal{L}_{inv} \tag{5.9}$$

In the multi-steps formulation, we applied the total loss $\mathcal{L}$ to each of the warped images, deformation and grid : $\mathcal{L}_{multi} = \sum_i \mathcal{L}(\nabla\Phi^i_{A\to B}, \nabla\Phi^i_{B\to A}, \widehat{A}^i, \widehat{B}^i)$. As the registration is an ill-posed problem, the tuning of the weights will have a high impact on our network's performance. A high value for the regularisation weights will produce a grid close to the identity and poor performance in terms of organs registration. In contrast, a too-small value will produce unrealistic deformed images. We explore the contribution of these weights in the following sections.

5.4 Experiments

5.4.1 Experimental parameters

Training parameters We performed experiments to study the impact of the pseudo segmentations, the different regularisation losses and the multi-step strategy. For each experiment, we first pretrained the network with a big dataset and pseudo segmentations and then fine-tuned with a smaller dataset and the ground-truths segmentations. For all our experiments, we choose the following hyper-parameters: The learning rate was set equal to $1e^{-4}$, the network processed randomly cropped patches of size $128 \times 128 \times 128$ during the training and $256 \times 160 \times 192$ during the prediction. The batch size was equal to 4 and 2 for the training with respectively one step and two steps. We applied three HU windows to the CT scans, the abdominal, lung and bones windows. The corresponding width and level values are $W = 400$, $L = 40$ (abdominal), $W = 1400$, $L = -500$ (lung), $W = 1000$, $L = 400$ (bones). Therefore, the network's input is a three channels volume, and the similarity loss is calculated with these three windows. We did not apply any data augmentation procedures, as it seems to lower the measured performance. Other hyper-parameters such as the values of the weights α, β, γ, δ and ϵ are given in each subsection.

Dataset To generate the pseudo labels for our registration problem, we built a dataset combining many publicly available datasets. The first dataset comes from the Learn2Reg Challenge [Dalca, 2020; Hering, 2021]. This dataset was first used in a comparison of registration algorithms [Xu, 2016]. It comprises 50 abdominal CT and thirteen abdominal organs segmented: spleen (Spl), right kidney (RKid), left kidney (LKid), gall bladder (GBla), oesophagus (Oes), liver (Liv), stomach (Sto), aorta (Aor), inferior vena cava (InfVe), portal and splenic vein (P&SVein), pancreas (Pan), left adrenal gland (LAd), and right adrenal gland (RAd). All images have different characteristics (pixel spacing, image size, for instance). Thus, organisers applied the following processing steps: affine registration, resampling to 2 mm voxel size, cropping to a common shape of $256 \times 192 \times 160$. We use this dataset for the test, as public comparisons are available.

We collected three supplementary abdominal datasets : the Medical Segmentation Decathlon (MSD Simpson et al. [Simpson, 2019]), the Kits 2019 dataset [Heller, 2020] and the TCIA Pancreas dataset [Roth, 2016; Clark, 2013]. The MSD challenge was a segmentation challenge where the participants had to segment different structures (organs, tumours, vessels) with different modalities. We selected five sub-datasets of the MSD corresponding to abdominal CT: Liver (Task 3), Pancreas (Task 7), and Spleen (Task 9) with respectively 200, 420, 190, 443 and 61 volumes. The Kits 19 dataset contains 300 abdominal CTs with the segmentation of kidneys and kidney's tumours. The TCIA Pancreas comprises 82 CT scans with the segmentation of 8 different organs: spleen, left kidney, gallbladder, oesophagus, liver, stomach, pancreas and duodenum. The segmentations have been published in parallel with two articles, first focusing only on the pancreas [Roth, 2015], then on seven supplementary organs [Gibson, 2018]. We kept only the segmentations already available in the L2R dataset. In the case of tumour segmentation, we chose to merge the tumour's labels with

the corresponding organs. All these datasets have different imaging characteristics. We standardised them by resampling them to 2 mm voxels size (similar to the L2R dataset) and affine registration with the software Ants [Avants, 2009]. The affine registration was performed with the same image of the training set of the L2R dataset. A summary of the different images used is depicted in Table 5.1.

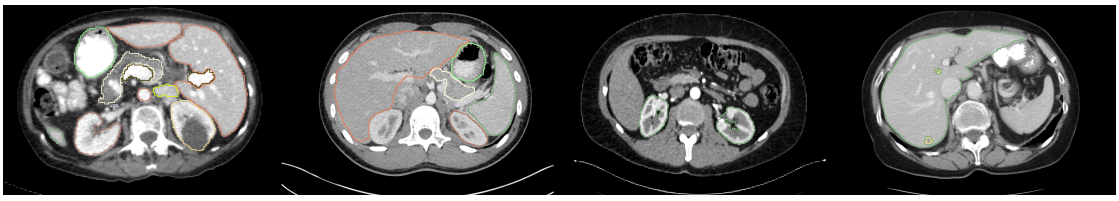


Figure 5.2: Presentation of the different datasets used with their labels. Col 1-4 : Learn2Reg Task 3, TCIA Pancreas, MSD Liver, Kits 19. Slices in an axial view are displayed.

Dataset	Segmentations	Number of Volumes
Learn2Reg Task 3	13 organs	50
TCIA Pancreas	8 organs	82
Kits 19	Left and Right Kidneys + Tumours	300
MSD Liver	Liver + Tumours	201
MSD Pancreas	Pancreas + Tumours	420
MSD Spleen	Spleen	61
Learn2Reg Task 4	Hippocampus Head& Body	394
Oasis 3	Not used	788

Table 5.1: An overview of the different dataset used for this study

This chapter focuses mainly on abdominal registration, but we also perform experiments on another localisation and modality: hippocampus MR. This dataset was used for Task 4 of the Learn2Reg challenge and comes from the publicly available dataset Medical Segmentation Decathlon [Simpson, 2019]. This cohort comprises brain T1-MRIs showing only the hippocampus part. All the images were given with segmentations of two small structures: the head and the tail of the hippocampus. All the images have been resampled to a voxel size of 1 mm and have a volume shape of $64 \times 64 \times 64$. We apply an extra preprocessing step to this dataset: we deleted the padding and replaced it with a constant value equal to 0. The dataset has been split into 200, 60 and 131 for training, validation and test. Additionally, to the data given for the challenge, we used the publicly available dataset OASIS 3 used in the previous chapters [Marcus, 2009]. We performed the same normalisation strategy for both the hippocampus dataset and OASIS 3 dataset: $\mathcal{N}(0, 1)$ normalisation, clipping values outside of the range $[-5, 5]$ and finally min-max normalisation to stay to the range $(0, 1)$. The significant difference between these two datasets is the cropping strategy: For the hippocampus dataset, we pass the image in full size to the neural network, while we performed random cropping of size $128 \times 128 \times 128$ for the OASIS 3 dataset to fit into memory. This is possible because an affine pre-registration step has been performed on the brain.

Metrics To evaluate our registration performance, we use different metrics. We need to measure both the accuracy and robustness of the method as well as the smoothness of the deformation. We follow the Learn2Reg organisers using the Dice score, 30% of the lowest Dice score, Hausdorff distance and standard deviation of the log Jacobian (noted SdLogJ in this chapter). Using the same metrics allow a fair comparison with others results published on this challenge. In order to better assess the regularity of the grid, we add a supplementary metric, the percentage of voxels where the Jacobian is negative ($\%$ of $|J_\Phi|_{\leq 0}$).

5.4.2 Participation to the Learn2Reg Challenge

In Table 5.2, we report our results to the Learn2Reg Challenge and the best performing methods [Dalca, 2020; Hering, 2021]. We participated in Task 3 and Task 4, respectively, CT abdominal registration and MR hippocampus registration. Our proposed solution was based on the symmetric formulation (eq. 5.1) and the pretraining with pseudo segmentations for task 3 [Estienne, 2021a]. We did not use the multi-steps formulation as well as the jacobian and the inverse-consistency losses $\mathcal{L}_{jac}$ and $\mathcal{L}_{inv}$ (eq. 5.6). We also provide an ablation study of our method on the validation set for both the task 3&4.

Our proposed method achieved the 3^{rd} position on tasks 3 and 4 and a 2^{nd} position to the overall challenge. The winner of the challenge proposed a deep learning method based on a Laplacian pyramid and multi-scale optimisation [Mok, 2020a; Mok, 2021]. His network is decomposed into three parts, each predicting the deformation grid at a different resolution level. The three levels are not trained simultaneously but independently, starting from the lowest resolution and freezing it when training the next level. The predicted transformation at one resolution level is passed as input of the following level. The top-performing methods were deep learning-based for task 4, while Deeds, a classical method based on MRF, obtained the 2^{nd} position on task 3 [Heinrich, 2013]. Two reasons can justify this: task 4 had a bigger dataset and focus on small deformation, while task 3 had only 30 patients for the training set and many complex deformations. Our network achieved similar results to the Laplacian pyramid approach (even if lower) for the Dice score, the 30% of lowest Dice score and the Hausdorff distance. The main difference was on the regularity of our grid and the calculation time where we achieved on the abdominal task 1.53 for the SdLogJ and 6.21 seconds for the prediction time compared to 0.12 and 1.83 seconds for the winner. The slower prediction is probably due to the size of our network, which has around 20 million parameters. The low regularity of our transformation is explained by the weight of our smooth loss $\mathcal{L}_{smooth}$, which was set $\gamma = 0.01$ and the non-use of inverse-consistency and Jacobian losses.

Dataset	Methods	Dice	Dice30	Hd95	SdLogJ
Task 3					
Val	Unregistered	0.23	0.01	46.1	
	Baseline	0.38	0.35	45.2	1.70
	Baseline + sym.	0.40	0.36	45.7	1.80
	Baseline + sym. + pretrain	0.52	0.50	42.3	0.32*
	Baseline + sym. + pretrain + pseudo labels	0.62*	0.58*	39.3*	1.77
Test	Proposed	0.64	0.40	37.1	1.53
	Laplacian Pyramid [Mok, 2020a; Mok, 2021]	**0.67**	**0.48**	**0.5**	0.12
	Deeds [Heinrich, 2013]	0.51	0.29	39.8	**0.11**
	PDD-Net [Heinrich, 2020; Hansen, 2021]	0.46	0.22	42.1	.43
Task 4					
Val	Unregistered	0.55	0.36	3.91	
	Baseline	0.80	0.78	2.12	0.067*
	Baseline + sym.	0.83	0.82	1.68	0.071
	Baseline + sym. + pretrain	0.84*	0.83*	1.63*	0.093
Test	Proposed	0.85	0.84	1.51	0.09
	Laplacian Pyramid [Mok, 2020a; Mok, 2021]	**0.88**	**0.86**	**0.3**	**0.05**
	[Wodzinski, 2021]	0.79	0.76	2.2	0.08
	PDD-Net [Heinrich, 2020; Hansen, 2021]	0.78	0.76	2.23	0.07

Table 5.2: Evaluation of our method for the Tasks 3&4 of Learn2Reg Challenge on the validation set (Val) and the test set (Test). We present our ablation study on the validation set and our result [Estienne, 2021a] and other top-performing methods to the challenge. We indicate the best metrics on our ablation study with a star and the best results among the challenge participant in bold.

5.4.3 Impact of the pseudo-segmentations

We performed experiments to measure the impact of the pseudo-segmentations during the pretraining and present the results in Figure 5.3. For all experiments, we kept the same training parameters. The weights of the different losses were set to $\alpha = \beta = \gamma = 1$ and $\delta = \epsilon = 0$, meaning that only the smooth loss $\mathcal{L}_{smooth}$ is used to regularise the transformations. The pretraining was performed during 24 hours (approximatively 100 epochs) with 760 different images, and the network was fine-tuned with 300 epochs using only the 20 patients of the Learn2Reg dataset and the ground-truth segmentations. The training lasted in total 38 hours, and we selected the epoch with the minimal loss on the validation set.

We compared five different experiments: the registration network without any pretraining, an unsupervised pretraining without using the pseudo-segmentations, and three supervised pretraining using respectively 1 (Liv), 4 (Liv, Spl, RKid, LKid) and 11 organs as pseudo-segmentations. The experiment with 11 organs corresponds to using all the organs as pseudo-segmentations except the right and left adrenal glands as these organs are small, and the segmentation network could not predict them with very good accuracy. After the training, our segmentation network achieve the following performances in terms of dice : 0.92 (Spl), 0.90 (RKid), 0.91 (LKid), 0.94 (Liv) 0.83 (Sto),

0.74 (Pan), 0.72 (GBla), 0.89 (Aor), 0.76 (InfV), 0.62 (PorV) and 0.61 (Oes). The segmentation's validation set comprises 21 patients coming from the Learn2Reg and the TCIA Pancreas dataset. The best Dice performances were achieved by the biggest organs such as the liver or spleen, and the organs present in large quantity in the dataset we built.

In Figure 5.3, we represent a box plot of the registration performances for the five experiments in terms of the Dice coefficient. We calculate the Dice on the ten patients of the Learn2Reg Validation (45 image pairs) with the script given by the challenge's organisers. The Dice coefficient is evaluated for the 13 organs, and we also represent its average value and the smoothness of the transformations using the SdLogJ. We can draw several conclusions from this figure. The unsupervised pretraining improves the performance on the majority of the organs but has a higher impact on voluminous organs such as the liver and spleen. When we use some organs' pseudo segmentation during the pretraining, the performance is increased mainly on these organs but also on others. For instance, the experiment which used only the liver during the pretraining resulted in Dice's improvement not only for the liver but also for the spleen, kidneys or stomach. Adding only a few organs produce a sort of overfitting on them. Indeed, if we compare the experiments with 1, 4 and 11 organs, we find a Dice's decrease for the liver and spleen and kidneys. The network concentrates more on the supervised organs than on other body parts, and thus the performance decrease when we add new organs. However, the overall Dice continues to improve. The experiment using only the liver and the one with 11 organs produce transformations with low regularity (high SdLogJ). In one case, the network focuses too much on the liver, producing a noisy grid. In the other case, the grid becomes more complex and, thus, more irregular. For all the experiments, it is important to recall that all the organs were used during the fine-tuning step. From these experiments, we can conclude that adding organs helps register, even with approximated segmentations. The most voluminous organs have the biggest impact.

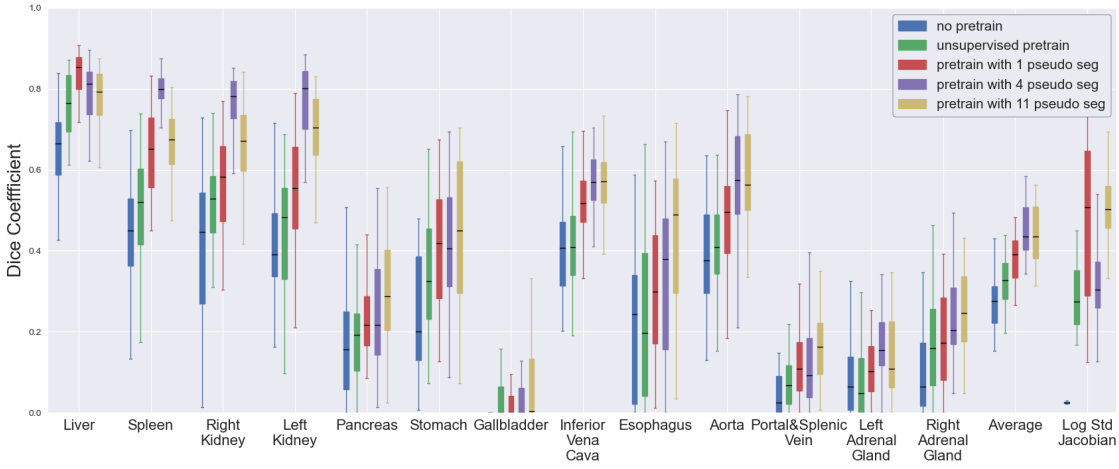


Figure 5.3: Comparison of the impact of the number of pseudo-segmentations during the pretraining. The Dice coefficient is represented in a box plot for the 13 evaluated organs as the average Dice and the SdLogJ. The results are displayed for the ten patients included in the Learn2Reg validation set (45 pairs), and five different pretraining strategies are compared.

5.4.4 Impact of the regularisation losses

We explored the impact of the different regularisation losses $\mathcal{L}_{smooth}$, $\mathcal{L}_{jac}$ and $\mathcal{L}_{inv}$. We kept the same training parameters as for the previous section, changing only the values of γ, δ and ϵ : pretraining with pseudo-segmentations of 11 organs for 24 hours and then fine-tuning for 300 epochs. The value of the weights was the same during the pretraining and the fine-tuning. We present the results in the Figure 5.4. The metrics are calculated with the script given by the challenge organisers and on the ten patients of the validation set (45 image pairs). This figure aims to study the correlation between registration's performance in terms of Dice and the grid's smoothness with respect to SdLogJ. The best transformations correspond to a high dice and a low standard deviation, which is the bottom right corner of the graph. The bottom left corner corresponds to the identity transformation (very regular grid and low Dice), and the top right corner to a good registration in terms of Dice but with a poor regularity.

The blue curve of Figure 5.4 represents the performance with δ and ϵ equal to 0 and γ varying between $1e^{-3}$ and $1e^{1}$. We show that there is a strong correlation between the smoothness and performance using only $\mathcal{L}_{smooth}$. Especially, low values of the weight γ result in noisy transformations while they do not improve much the registration performance in terms of Dice. Thus the smooth loss $\mathcal{L}_{smooth}$ is insufficient to obtain a high Dice together with regular deformations.

For the red and green curve of Figure 5.4, we set γ equal to $1e^{-1}$, as this value reaches high performance in terms of dice and smoothness., and we varied the δ and ϵ values. For the green one, δ iterates between $1e^{-4}$ and $1e^{0}$ while ϵ was constantly equal to 0 and inversely for the red curve δ is equal to 0 and ϵ goes from $1e^{-3}$ to $1e^{2}$. The two regularisations losses have a strong impact on the smoothness of the deformations. They allow decreasing the SdLogJ while keeping the Dice around 0.55. However, when δ and ϵ reach high values, the deformation is close to the identity transformations. Moreover, the Jacobian loss has a stronger impact than the inverse consistency loss. Indeed the predicted transformations get closer to identity transformations with δ around $1e^{-1}$, while we observed the same results for values of ϵ around $1e^{2}$. We obtained the best compromise between the registration performances and the smoothness of the grid for values of δ between $1e^{-2}$ and $1e^{-3}$ and values of ϵ between 1 and 10.

On the Figures 5.7, 5.8 and 5.9 in the supplementary material, we present some visual representation of one pair of volumes produced for respectively the blue, green and red curve, displaying the deformed image, predicted transformations, Jacobian and the composed grid $\Phi \circ \Phi^{-1}$ for one pair of the validation set. These figures illustrate the complexity of tuning the regularisation weights for abdominal registration: low values of these weights create negative Jacobian and non-topological points and high values result in inaccurate registration and transformations close to identity transformations. The grids depicted on these figures are in agreement with the conclusions drawn previously from Figure 5.4. Indeed, we see that the smooth loss $\mathcal{L}_{smooth}$ is not able to produce a grid regular and smooth at the same time (blue curve on Figure 5.4 and Figure 5.7) and outputs the identity grid if γ is too strong. Adding $\mathcal{L}_{inv}$ and $Ljac$, we obtained smoother grid (red and green curve on Figure 5.4 and Figures 5.8 and 5.9). However, we are still facing negative Jacobian and irregular points.

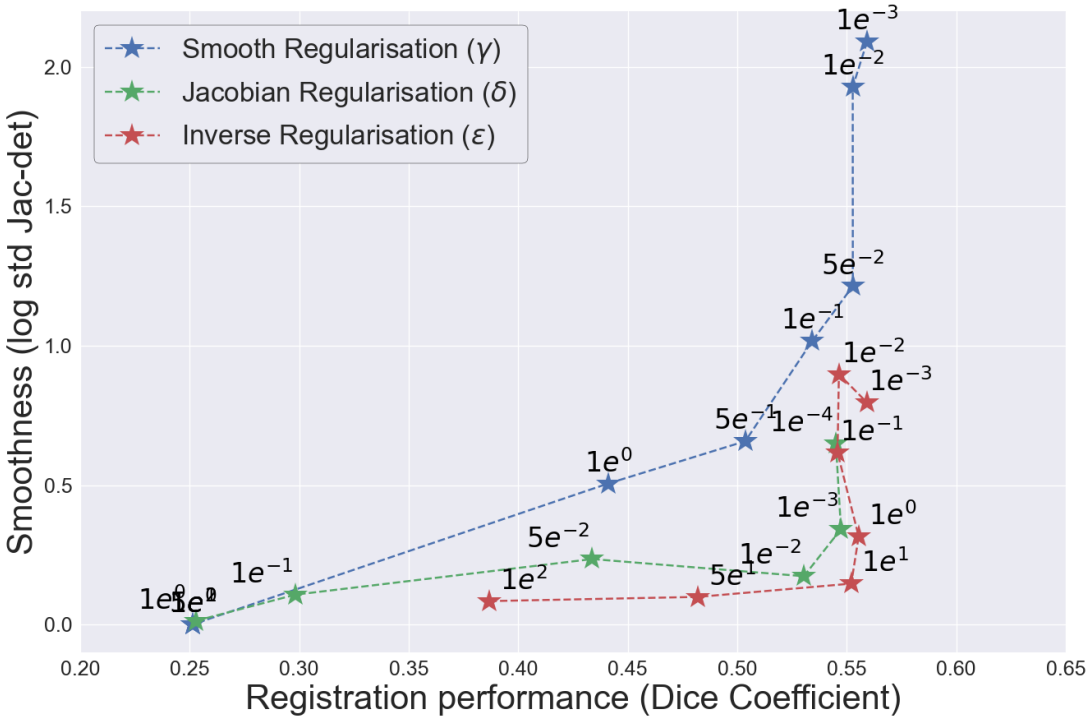


Figure 5.4: Representation of the smoothness of the grid in function of the registration performances for different regularisations weights. The metrics were calculated for the 10 patients from the Learn2Reg validation set (45 pairs). The value of the weights γ, δ and ϵ are indicated on the graph and correspond to the blue, green and red curve.

5.4.5 Impact of the multi-steps strategy

We also explored the influence of the multi-step strategy on the grid's smoothness and the registration's performance. The multi-step formulation is given in equation 5.2, and it consists of finding the transformation between the fixed image and the already deformed image.

We investigate first the consequence of applying the multi-step formulation during the training phase or during the inference phase. For these experiments, we set the weights to $\alpha = \beta = 1$, $\gamma = 1e^{-1}$, $\delta = 1e^{-2}$ and $\gamma = 1e^{1}$, we pretrain the network during 24 hours using 11 organs as pseudo-segmentations and 760 patients and fine-tune the network with the Learn2Reg training set and the ground-truth segmentations. In Table 5.3, we present the results both in Dice and SdLogJ for two networks trained with 1 and 2 steps, while the inference's steps are between 1 and 4. The results for the network trained with 1 step shows that using the multi-step strategy during the inference has a negative impact. Indeed, the Dice coefficient decreases from 0.35 to 0.30 while the grid regularity is reduced significantly, with the SdLogJ going from 0.13 to 0.76. Concerning the two-steps approach, we obtained different results. First, looking at the first line (1 step inference), we obtain a higher Dice and a similar SdLogJ when the network is trained with two steps than one, with a Dice equal to

0.54 compared to 0.35. This demonstrates the impact of the dual-step training, even if the inference is performed with one step. For all these experiments, we kept the same hyper-parameters to have fair comparisons between the different setups (1 or 2 steps). Studying the inference with two steps, it increases the Dice coefficient (0.54 to 0.60) but reduces the regularity (0.12 to 0.26 for the SdLogJ). Finally, performing the inference with 3 and 4 steps do not improve the Dice but increase the noise. This experiment demonstrates that training the network with a 2 step formulation improve the results, even if the inference is performed with 1 step. We also show that applying the inference with more steps than the training does not ameliorate the grid but instead makes it noisier.

Training / Inference	1 step		2 steps	
	Dice	SdLogJ	Dice	SdLogJ
1 step	0.35 ± 0.078	0.13 ± 0.019	0.54 ± 0.070	0.12 ± 0.009
2 steps	0.33 ± 0.067	0.52 ± 0.116	0.60 ± 0.059	0.26 ± 0.050
3 steps	0.30 ± 0.067	0.76 ± 0.146	0.61 ± 0.058	0.34 ± 0.062
4 steps	\	\	0.61 ± 0.057	0.41 ± 0.069

Table 5.3: Study of the impact of the multi-steps formulation during the training or the inference phase. Two models are compared trained with 1 or 2 steps and four inference formulation. The metrics are calculated on the Learn2Reg validation set using 45 pairs. The average Dice over 13 organs and the Standard deviation of the Log Jacobian are presented.

In a second time, we examine the impact of the 2-step strategy for different regularisations. For these experiments, we set the loss weights to $\alpha = \beta = 1$, $\gamma = 0.1$ and different choices for δ and ϵ. These two parameters correspond to the Jacobian and inverse consistency loss $\mathcal{L}_{jac}$ and $\mathcal{L}_{inv}$. We choose three different combinations corresponding to a strong regularisation ($\delta = 5e^{-2}$, $\epsilon = 5e^{1}$), a medium regularisation ($\delta = 1e^{-2}$, $\epsilon = 1e^{1}$) and a weak regularisation ($\delta = 1e^{-3}$, $\epsilon = 1e^{0}$). These weights were selected using the Figure 5.4. Following the results presented in Table 5.3, we select the same number of steps for the training and inference, comparing the 1-step and 2-step formulations. We did not experiment on a bigger number of steps, mainly because of memory limitations. We represent the results in terms of registration's performance with the Dice and grid's smoothness with the SdLogJ in Figure 5.5, and we also give the corresponding numerical values in Table 5.4. In Figure 5.5, we draw an arrow going from the one step's results to the two steps' results. From this figure, we see that the 2-step approach boost the results for all the choice of regularisation parameters. Moreover, it raises the performance on the Dice coefficient while keeping the smoothness relatively low. It is also interesting to notice that the 2-step formulation improved more the Dice coefficient when the regularisation is stronger, and the 1-step method produces deformation very close to the identity transformation. Finally, we demonstrate that the dual-step strategy is a positive improvement of registration formulation, as it increases the accuracy while preserving the smoothness.

Weights		1 Step		2 Steps	
δ	ϵ	Dice	SdLogJ	Dice	SdLogJ
$1e^{-3}$	$1e^{0}$	0.57 ± 0.066	0.25 ± 0.098	0.62 ± 0.055	0.38 ± 0.091
$1e^{-2}$	$1e^{1}$	0.34 ± 0.078	0.13 ± 0.019	0.60 ± 0.059	0.26 ± 0.050
$5e^{-2}$	$5e^{1}$	0.29 ± 0.070	0.066 ± 0.002	0.44 ± 0.076	0.080 ± 0.004

Table 5.4: Comparison of the 1 step and 2 steps strategy for different values of the regularisation weights. The metrics are calculated on the Learn2Reg validation set using 45 pairs. The Dice coefficient and the Standard deviation of the Log Jacobian are given. For each regularisation weights, the 2 steps strategy improve the Dice coefficient. The values correspond to the points depicted in Figure 5.5.

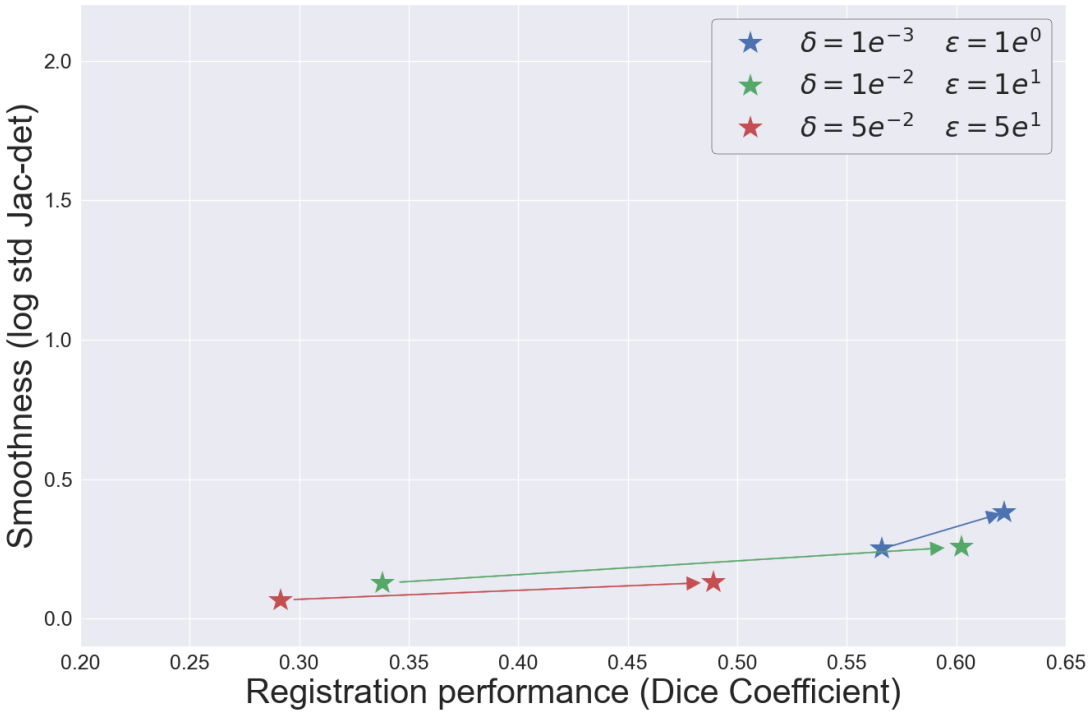


Figure 5.5: Impact of the multi-step on the smoothness and the Dice coefficient. The arrows go from the results of the 1 step to the results of the 2 steps. Three different training have been made with different values of the regularisation weights δ and ϵ. The metrics were calculated for the 10 patients from the Learn2Reg validation set (45 pairs). The X and Y axis have the same limits as in Figure 5.4 and the numerical values can be found in Table 5.4.

5.4.6 New evaluation on Learn2Reg

In this section, we compare our new approach using the dual-steps formulation and the new regularisation losses with the results of the Learn2Reg Challenge 2020 and our original participation [Estienne, 2021a]. The results are presented in Table 5.5. For our approach, we set the regularisation weights to: $\alpha = \beta = 1$, $\gamma = 1e^{-1}$, $\delta = 1e^{-2}$ and $\epsilon = 1e^{1}$ and use our proposed dual-steps strategy.

We choose these regularisations weights as they provide a good trade-off between the smoothness and the registration. We pretrain the network during 48 hours using 11 organs as pseudo-segmentations and 760 patients and fine-tune the network with the Learn2Reg training set and the ground-truth segmentations. Other training parameters are given in Section 5.4.1. The difference with our original participation lies in the dual-step formulation and the introduction of $\mathcal{L}_{jac}$ and $\mathcal{L}_{inv}$.

Methods	Dice	Dice30	Hd95	SdLogJ
Baseline + sym. + pretrain + pseudo labels [Estienne, 2021a]	0.64	0.40	37.1	1.53
Laplacian Pyramid [Mok, 2020a]	**0.67**	0.48	**36.5**	0.12
Deeds [Heinrich, 2013]	0.51	0.29	39.8	**0.11**
PDD-Net [Heinrich, 2020]	0.46	0.22	42.1	0.43
Ours **(val)**	0.66	**0.61**	38.8	0.21
Ours **(test)**	0.64	0.55	41.6	0.20

Table 5.5: Comparison of our method (MICS) with the top-performing methods of Task 3 of Learn2Reg Challenge. The metrics are calculated on the test set of the Learn2Reg 2020 Challenge. The first three rows are results obtained by different teams during the challenge. Bold indicates best values.

Our method outperforms three teams of the challenge, obtaining better Dice and Dice30 [Heinrich, 2013; Heinrich, 2020; Estienne, 2021a]. Concerning the smoothness of the grid, measured by the SdLogJ, we obtain satisfactory results, overpassing two competitors. Especially if we compare with the results of our first participation [Estienne, 2021a], which have a close result for the Dice, our SdLogJ is seven times smaller (1.53 vs 0.21), showing the impact of our supplementary regularisation losses. Two methods obtained more plausible deformations, one deep learning-based [Mok, 2020a] and one iterative method [Heinrich, 2013].

Our method reports very promising results, particularly for the Dice30, with a score of 0.61, while the best method had 0.48, attesting to the robustness of our method, which performed well even for the most challenging moving-fixed images pairs. The method proposed by Mok et al. [Mok, 2020a], based on deep learning and a multi-scale pyramid approach, surpass our results with similar results for Dice and Hausdorff distance but mostly a SdLogJ twice lower. Multi-scales formulations seem to be very efficient to generate accurate and smooth deformation, especially for localisation with high displacements such as abdominal CT. These methods will become state-of-the-art for deep learning-based registration.

In Figure 5.6, we display our results for the three different pairs, keeping the same fixed image and changing the moving image. We represent the moving CT, the deformed image and segmentation, the deformation grid and the Jacobian. Even if we trained our network using a specific loss to reduce negative Jacobian, we have negative values (in red in the figure) as well as crossing and folding in the grid.

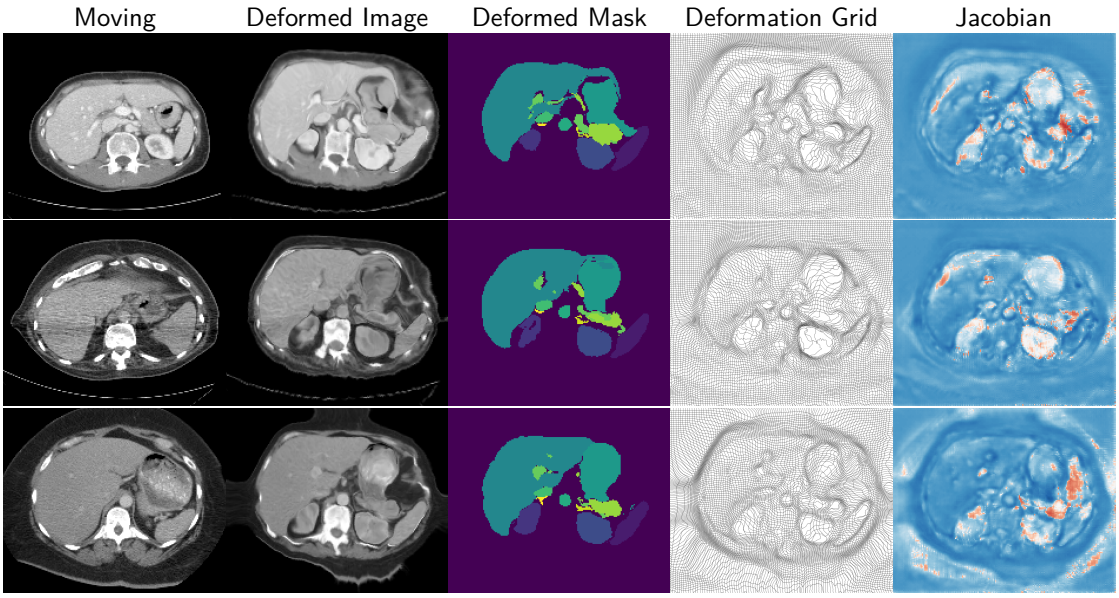


Figure 5.6: Representation of the registration results for three different moving images and the same fixed image. From left to right: the moving CT, the deformed CT, the deformed segmentation, the transformation Φ and the Jacobian $|J_\Phi|$. The Jacobian is depicted in blue, white and red for respective positive, zero and negative values.

5.5 Discussion and conclusion

In this chapter, we investigated the challenging problems of the hippocampus and abdominal registration. First, we developed a method using spatial gradients, a symmetrical formulation and pseudo segmentations and participated in the Learn2Reg Challenge. We obtained a 3^{rd} position on tasks 3 and 4 and a 2^{nd} position to the overall challenge. The best participant developed a multi-scale registration algorithm using a Laplacian pyramid [Mok, 2020a]. Secondly, we introduce more registration properties, including inverse consistency and local non-topological deformations. We studied, in particular, the impact on the accuracy and the plausibility of the deformations of three different regularisation losses, the smooth loss, the Jacobian loss and the inverse consistency loss. In addition, we provide an extensive study of the effect and impact of the pseudo labels. Finally, we benchmarked our new formulation with the Learn2Reg 2020 Challenge dataset, comparing it with the top-performing results of the challenge and our initial submission. We demonstrate that both the use of regularisation losses and the multi-step formulation improve the performances of the registration approach.

This chapter also highlighted the difficulty of evaluating registration algorithms. How should we rank two algorithms, one being irregular but accurate, the other being smooth but less accurate? This issue is of particular importance in the case of deep learning-based registration, where networks manage to elaborate complex but noisy deformations. During the Learn2Reg Challenge [Dalca, 2020; Hering, 2021], the organisers set various weights for each metric (calculation time, accuracy, smoothness,

robustness), but altering these weights could lead to a new winner solution. The representation of both the smoothness and the accuracy in one graph allows a new visual evaluation and a better comparison of different methods. We provided such representations in Figure 5.4 and 5.5.

One important limitation of our work is the presence of negative Jacobian, despite the Jacobian loss. A stronger regularisation could remove them. However, it would produce a deformation close to the identity transformations. In the same way, the composition of the forward and backward transformations lead to a grid unequal to the identity transformations. The deformed images are finally irregular and not applicable in a clinical application, demonstrating the requirement of great improvement for abdominal registration. In addition to the multi-step formulation described in this chapter, multi-scale approaches obtained promising results for different localisation [Hering, 2019; Fechter, 2020; Mok, 2020a]. These approaches produce accurate and smooth deformations thanks to the combination of transformations at different resolutions levels.

In the future, we want to explore a clinical application of our method. More specifically, we would like to apply our network to multi-temporal follow-up of patients, monitoring disease progression. Temporal registration of the same patient should produce less noisy deformations than inter-patient registration. One application of abdominal registration could be the case of liver tumour, to monitor the tumour growth inside the liver. The major difficulty would be to disjoin the deformations due to the tumour growth from the one related to natural deformations (digestion or breathing, for instance). Finally, we want to investigate the predictive power of registration networks for other clinical tasks such as survival prediction in cancerology. Three methodologies could be examined: i) using the registration network as a pretraining step for a network that predicts clinical parameters, ii) exploring the correlation between the deformation grid predicted by registration networks and clinical features and iii) using the latent space which encode the deformation grid and connecting a simple classifier to it.

5.6 Appendix

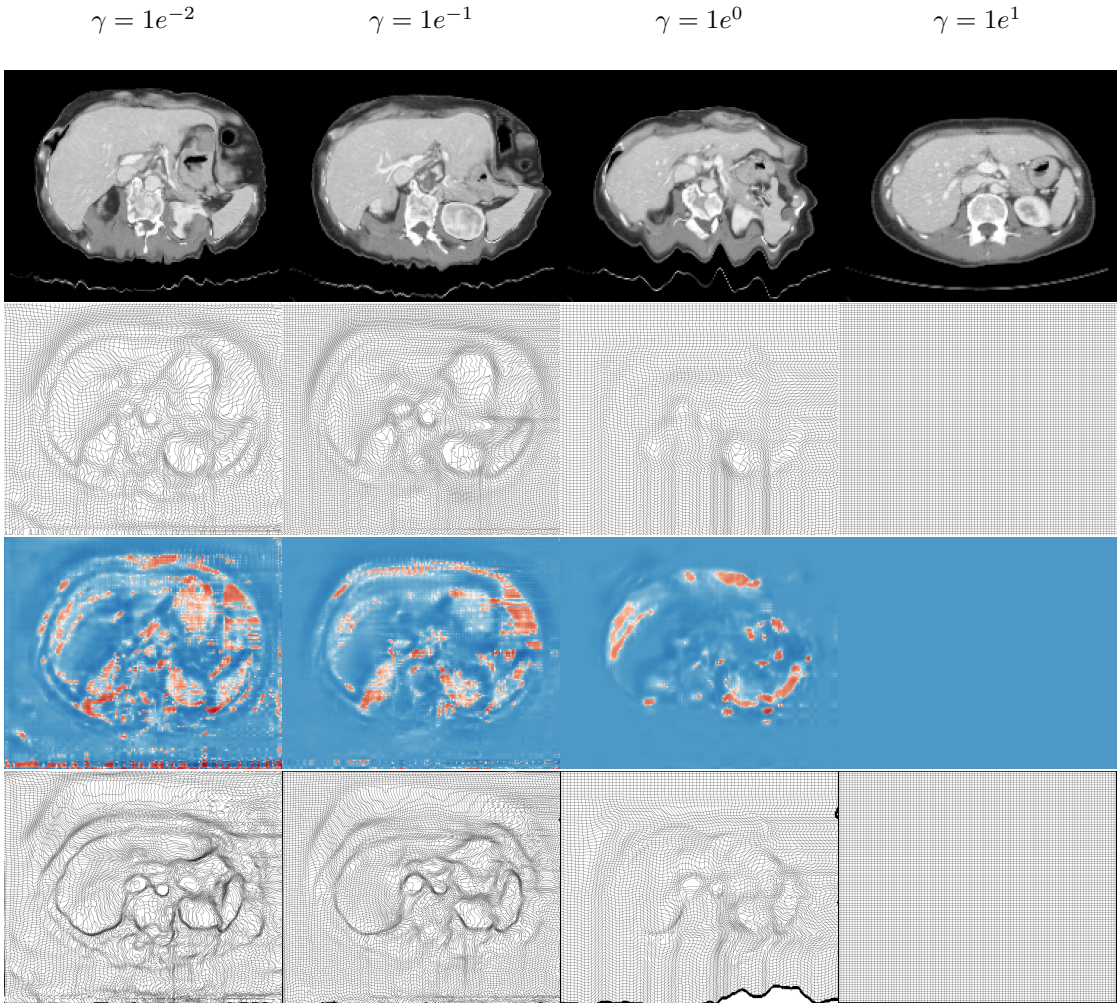
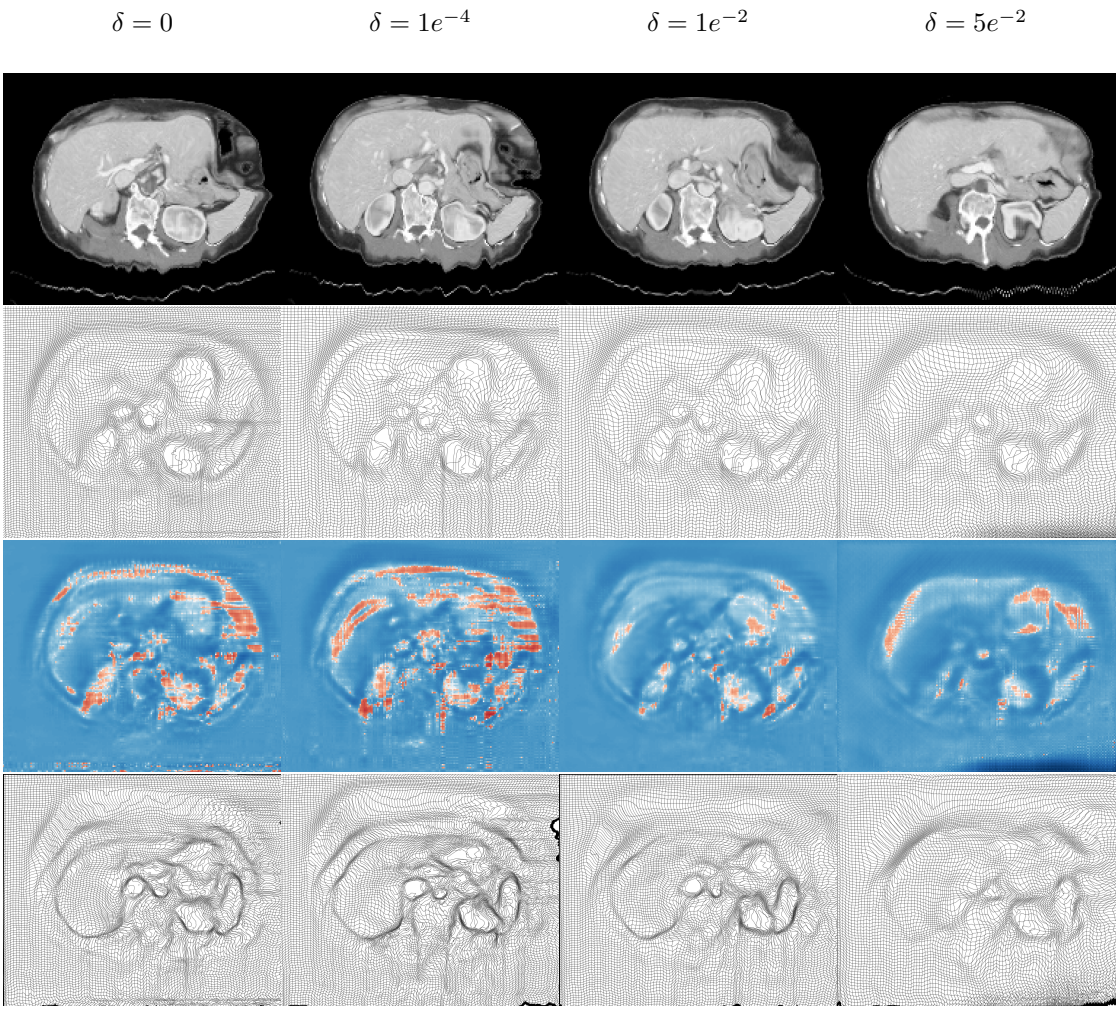
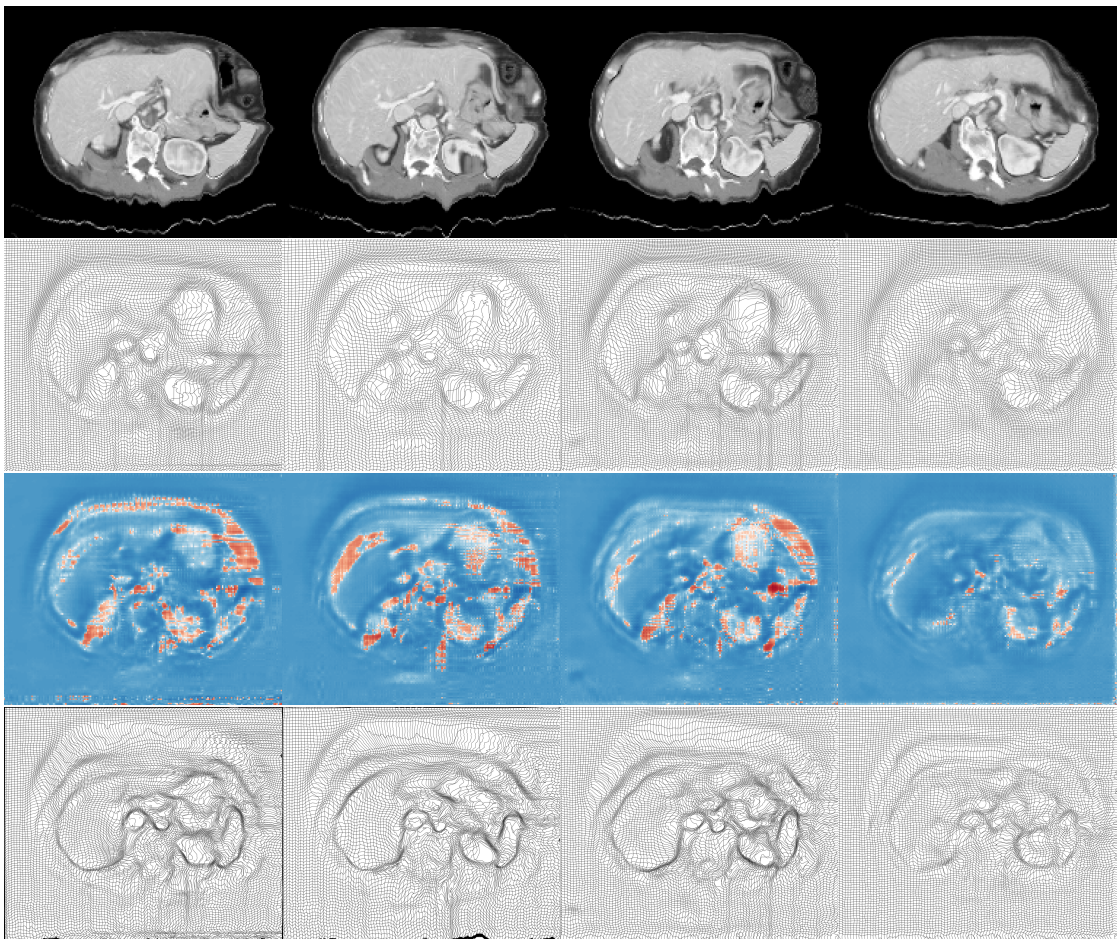


Figure 5.7: Representation of the impact of the smooth loss $\mathcal{L}_{smooth}$ on the predicted transformation. From top to bottom: the deformed CT, the transformation Φ, the Jacobian $|J_\Phi|$ and the composition $\Phi \circ \Phi^{-1}$ for one pair of the validation set. The Jacobian is depicted in blue, white and red for respective positive, zero and negative values. The average Dice and Standard Deviation of the Log Jacobian on the validation pairs are shown. δ and ϵ were set to 0 while γ values are indicated in the figure.



Dice	0.53	0.55	0.55	0.43
SdLogJ	1.01	0.65	0.34	0.23

Figure 5.8: Representation of the impact of the jacobian loss $\mathcal{L}_{jac}$ on the predicted transformation. From top to bottom: the deformed CT, the transformation Φ, the Jacobian $|J_\Phi|$ and the composition $\Phi \circ \Phi^{-1}$ for one pair of the validation set. The Jacobian is depicted in blue, white and red for respective positive, zero and negative values. The average Dice and Standard Deviation of the Log Jacobian on the validation pairs are shown. γ and ϵ were set $1e^{-1}$ and 0 while δ values are indicated in the figure.



Dice	0.53	0.55	0.56	0.48
SdLogJ	1.01	0.62	0.32	0.10

Figure 5.9: Representation of the impact of the inverse consistency loss $\mathcal{L}_{inv}$ on the predicted transformation. From top to bottom: the deformed CT, the transformation Φ, the Jacobian $|J_\Phi|$ and the composition $\Phi \circ \Phi^{-1}$ for one pair of the validation set. The Jacobian is depicted in blue, white and red for respective positive, zero and negative values. The average Dice and Standard Deviation of the Log Jacobian on the validation pairs are shown. γ and δ were set $1e^{-1}$ and 0 while ϵ values are indicated on the figure.

Chapter 6

Explainability of Registration Networks

"Martoni veut un hélico dans les 10 minutes sinon il la bute. Il dit que vous bluffez.
Dites-lui que j'ai plus de genoux
Il dit qu'il a plus de genoux. Il dit qu'il voit pas le rapport.
Bon ça suffit je compte 5, 4, 3, 2, 1 et à 0, paf ! Je lui explose la tête comme une pastèque."

La Cité de la Peur

Contents

In this chapter, we explore the problem of explainability. In the medical context, deep learning methods cannot be used as *blackboxes*, but we need to understand how they work. The explainability of deep neural networks is one of the most challenging and interesting problems in the field. Using the independent encoder approach, proposed in chapter 4 and used in chapter 5, we project every image to the encoding space. Then, with a simple linear projection, we generate a new basis that captures various anatomically aware geometrical transformations. We perform experiments using two different datasets focusing on MRI with hippocampus and lungs anatomies. We prove that such an approach can decompose the highly convoluted latent spaces of registration pipelines in an orthogonal space with several interesting properties. We hope that this work could shed some light on a better understanding of deep learning-based registration methods. This chapter's content have been accepted to the MICCAI 2021 DART Workshop (Domain Adaptation and Representation Transfer) [Estienne, 2021c].

6.1 Introduction

Deep learning methods currently provide the state of the art performance in variety of fields, as we previously detailed. This is due to their inherent property to generate highly abstract representations hierarchically. These representations are building on top of each other, making it possible to encode highly non-linear manifolds. Even though such hierarchies can outperform traditional methods, they lack explainability, making their translation difficult to solve real-life problems. The field of interpretable deep learning was first explored for computer vision problem (i.e. non-medical), where approaches such as Cam or GradCam were proposed [Selvaraju, 2017]. These methods produce heatmaps, showing areas of the input image which activate the most a specific label. Understanding deep learning algorithms is of great significance in the medical field and especially for the algorithms that are intended to be adapted to clinical practice, addressing problems of precision medicine [Holzinger, 2019; Castro, 2020]. For these reasons, it is essential to identify ways to understand better the high throughput operations that are applied.

In previous chapters, we studied the development of DL-based registration algorithms. These methods became more and more popular since the introduction of the differentiable spatial transformer [Jaderberg, 2016]. They reduce the computing time, allowing real-time applications while reporting better performance than traditional methods [Balakrishnan, 2019; Stergios, 2018; Mok, 2020b]. Current researches focus on applying to new and more complex body parts and on the regularity of the transformations, thanks to the introduction of new losses. Meanwhile, the deformation field, which is predicted by deformable registration networks, have been shown to encode not only the spatial correspondences but also clinical relevant information. The deformation field study could develop new approaches for different clinical applications, such as survival assessment or anomaly detection [Ou, 2015]. For instance, in the chapter 3 and 4 of this manuscript, we studied the relationship between registration and medical image segmentation [Estienne, 2019; Estienne, 2020]. However, according to our knowledge, there are not many efforts focusing on understanding and analysing DL-based registration. The study of the encoding information could be a first step for the explainability of this field.

In this chapter, we propose a framework for interpreting the encoded representation of deep learning-based registration methods. In particular, after training a network for medical image registration, we project every image to the latent space. This projection is made possible by the use of our proposed independent encoder formulation. Then, using a principal components analysis (PCA), we decompose the encoding space, generating a new basis and we empirically show that it captures various geometrical operations. This decomposed encoding space is then driving the generation of the deformation field. Our main contributions are threefold :

- Exploring the explainability of deep learning-based registration through the study of the latent space;
- Empirical analysis of the projection, using two datasets, and showing that these projections decompose the original transformations into *elementary transformations*;

- Studying the relationship between our decomposition and specific clinical parameters.

This chapter is organised as follows: in the section 6.2, we present an analysis of the deep learning explainability field. Section 6.3 describes our methodology, including the training of our registration network and the latent space's decomposition algorithm. Section 6.4 presents our experiments performed on two datasets, lung MR and hippocampus MR. The results are presented through qualitative analysis. Finally, in section 6.5, we discuss and analyse the presented results.

6.2 Related Work

Explaining how deep neural networks function is a matter of extensive research in the recent years. The question is not only to have a well performing model but to know if the predictions are based on valid reasons. The deep learning interpretability domain is often based on qualitative evaluations and gradient-based methods. However, other approaches exist such as, the generation of textual explanations [Hendricks, 2016]. Detailed classifications of explainability approaches are proposed in different reviews such as [Arya, 2019; Singh, 2020]. GradCam [Selvaraju, 2017] is one of the most popular methods widely used and can provide some insights on deep neural networks for many applications, including medical imaging. GradCam is able to highlight the regions of the input image that contribute the most to the final prediction, producing coarse heatmaps based on the gradients. Similar to GradCam, there are several additional methods based on the gradient [Zhou, 2015; Chattopadhay, 2018; Springenberg, 2015] that are commonly used for the explainability of the models. Moreover, in [Fong, 2017] the authors proposed a general framework of explanations as meta-predictors while they also reinterpret the network's saliency providing a natural generalisation of the gradient-based saliency techniques. Even though such approaches can provide information on where the models attend, they can be mostly utilised in classification or detection schemes. There are at the present days few works focusing on interpretability in the medical domain. However, Singh et al. [Singh, 2020] presented a survey of explainable the existing deep learning methods applied to medical imaging.

Representation disentangling methodologies is a concurrent field of research also investigating explainability topics. Instead of producing visualisation, such approaches mainly focus on generating interpretable latent representation by enforcing several constraints. This can be achieved either using architecture tricks [Shu, 2018; Sahasrabudhe, 2019] or with appropriate loss functions [Chen, 2016; Kingma, 2014]. One interesting property of disentangling algorithms is the image-to-image translation. By modifying only a few latent space values, we can generate synthetic images close to the original one, adding, for instance, glasses or a new hair cut. In medical image computing, several studies focus on approaches for generating disentangled or decomposed representations. The decomposition is often performed from a modality point of view. Indeed, separating features related to the modality from features related to anatomy could allow training model with multimodal databases. Given the complexity of building consequent medical image databases, such approaches help to benefit from more data. In [Yang, 2019], the authors developed a liver segmentation network using both CT

and MRI images. The latent space is decomposed into a modality invariant space and a modality-specific space. A similar decomposition is proposed in [Qin, 2019] to perform multimodal image registration. The authors performed experiences on two different multimodal datasets, lung with CT and MR and brain with T1 and T2 weighted MRI. Also, in [Chartsias, 2018], the authors proposed a different decomposition of the latent space, working on cardiac MR. They split into a spatial and discrete representation, encoding shape and segmentation, and a continuous representation, encoding textural information. Their decomposition is obtained thanks to the cycle-consistency principle and adversarial loss, and they demonstrated the decomposition's interest in applying it to a semi-supervised segmentation task.

Our method shares many common points with the aforementioned approaches, such as the particular interest in studying the latent space. However, we focus more on understanding its properties while disentangled representation algorithms use it for another task. This chapter has similar aspects with Schutte et al. [Schutte, 2021], such as generating new images to interpret neural network. The authors trained a StyleGan [Karras, 2019] on medical images and changed the latent space following a particular direction given by a linear model. Then, they generated artificial images and verified that these synthetic images correspond to an increase or decrease of the disease classified by the linear model.

6.3 Methodology

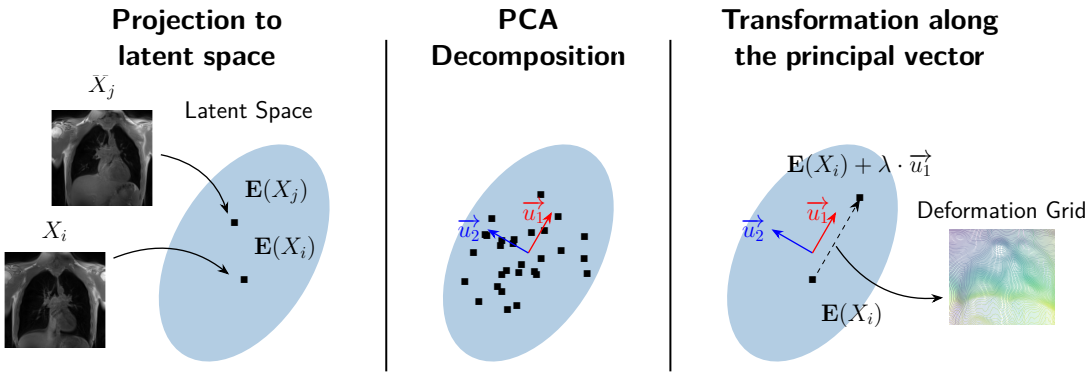


Figure 6.1: Overall overview of our proposed framework. The different subjects (X_i,X_j) are projected on the latent representation by the encoder $\mathbf{E}$ and then a linear decomposition of this latent space is calculated to identify a new vector space ($\vec{u}_i$).

Deep learning-based registration methods have received much attention in the last few years [Balakrishnan, 2019; Stergios, 2018]. As in the rest of this manuscript, we denoted our two volumes as moving M and fixed F, our neural network as g_θ and the optimal parameters for the network θ^*. Moreover, we considered only a non-rigid deformation field Φ, and we decomposed g_θ into an encoding $\mathbf{E}$ and a decoding $\mathbf{D}$ part.

Most of the DL-based registration approaches use an early fusion strategy on which the two volumes are concatenated before the pass-through g_θ. However, in chapter 4, we proposed a late fusion strategy. In this approach, the two volumes pass independently through the encoder, and a merging operation is performed at the encoding level using concatenation or subtraction operation [Estienne, 2020]. Thanks to this formulation, each volume has a unique encoding representation. In this context, the encoder played the role of a feature extractor. This chapter adopts our late fusion strategy using the subtraction operation since it carries several attractive properties for linear latent space decomposition. In figure 6.1 the overall scheme is presented.

6.3.1 Deep learning-based registration scheme

To perform our experiments and obtain our embeddings, we employ a 3D UNet based network [Çiçek, 2016] similar to the network used in chapters 4 and 5. The encoder and the decoder are composed of a fixed number of blocks with 3D convolution layers (stride 3, padding 1), instance normalisation layer and leaky ReLU activation function. The down and up-sampling operations are performed with a 3D convolution layer with stride and padding of 2. The main difference in our architecture with previous chapters and the original 3D UNet is the skip connections' removal. Indeed, we want to enforce that all information pass through the last encoding layer without any leak due to the skip connections. This modification led us to reduce the number of blocks from four to three for the lung dataset. Indeed removing the skip connections and using four blocks create a bottleneck with activation map of size $(256, {}^{H}/_{16}, {}^{W}/_{16}, {}^{D}/_{16})$ with 256 corresponding to the number of channels and H, W and D to the original size of the input image. The registration training process could not be completed properly with such a small bottleneck.

In section 2.5.3, we enumerated the different formulations proposed to generate the deformation from deep learning schemes, such as displacement field formulation [Balakrishnan, 2019], diffeomorphic formulations [Dalca, 2018; Krebs, 2019] and formulations based on the spatial gradients [Stergios, 2018]. In this chapter, we kept the spatial gradients formulation, with our network regressing the spatial gradients $\nabla_x\Phi_x$, $\nabla_y\Phi_y$ and $\nabla_z\Phi_z$, while the final deformation field is obtained through a cumulative sum operation. We also followed the symmetric formulation proposed in chapter 5 predicting both the moved to fixed and fixed to moved deformations: $\nabla\Phi_{M\rightarrow F} = \mathbf{D}(\mathbf{E}(M) - \mathbf{E}(F))$ and $\nabla\Phi_{F\rightarrow M} = \mathbf{D}(\mathbf{E}(M) - \mathbf{E}(F))$ [Estienne, 2021a].

The network was trained with a combination of four losses, one focusing on the intensity similarity using normalised cross-correlation ($\mathcal{L}_{sim}$), one focusing on anatomical structures using dice loss ($\mathcal{L}_{seg}$) and two losses for regularisation of the displacements. The first one was the Jacobian loss which is exploited on different works such as [Mok, 2020b; Kuang, 2019b; Zhang, 2020] ($\mathcal{L}_{jac}$) and the second one enforcing smooth gradients similar to [Estienne, 2020] ($\mathcal{L}_{smooth}$). More details on the impact of the different losses, particularly the jacobian one, were given in chapter 5. The final optimisation is the weighted sum of these components. As such our final loss is:

$$\begin{aligned}\mathcal{L} = (\mathcal{L}_{reg} + \mathcal{L}_{seg} + \alpha\mathcal{L}_{smooth} + \beta\mathcal{L}_{jac})_{M\to F} \\ +(\mathcal{L}_{reg} + \mathcal{L}_{seg} + \alpha\mathcal{L}_{smooth} + \beta\mathcal{L}_{jac})_{F\to M}\end{aligned} \tag{6.1}$$

with α and β being the weights of the regularisation losses.

6.3.2 Decomposition of latent space

Let $A_{train} = \{ X_i \mid i \in [0, n] \}$ be the set of our n training samples. The proposed formulation apply the encoder independently to each volume, and thus we obtain the set of latent vectors: $\mathbf{E}(A_{train}) = \{ \mathbf{E}(X_i) \mid i \in [0, n] \}$. Then, we decompose this space using principal components analysis (PCA). That way, we obtain a set of principal vectors $\mathcal{U}_K = (\overrightarrow{u_1}, \cdots, \overrightarrow{u_K})$ with K being a hyperparameter fixing the number of principal components. It is worth noting that each vector $\vec{u}_i$ has the same size as the encoder's last layer's activation map. This size depends on the number of channels, the input images' size, and the downsampling operations. We flatten each encoding representation from its four dimensions representation (channel dimension and the three spatial dimensions) to a one-dimensional array to perform the PCA. Thus, the PCA is not calculated channel-wise, but all the channels are considered together. Each principal vector $\vec{u}_i$ can be converted to a deformation grid ϕ_i using the corresponding decoder $\mathbf{D}$: $\phi_i = \mathbf{D}(\vec{u}_i)$. Therefore, we obtained a set of elementary transformations $\{\phi_i\}_{i=1\cdots K}$. These elementary transformations generate a basis that can be used to approximate and decompose every new deformation.

Using such a PCA decomposition, we obtain a representation in small dimensions of every training volume X_i. These representations are obtained by the projection of $\mathbf{E}(X_i)$ to each principal vector: $a_i^j = \mathbf{E}(X_i) \cdot \overrightarrow{u_j}$. For every volume of our training set we have the approximation: $\mathbf{E}(X_i) \approx \sum_{j=1}^{K} a_i^j \overrightarrow{u_j}$. After calculating the vector of the principal components $\mathcal{U}_K$ with the training set, we projected each image of the validation set to obtain its PCA representation. As the PCA decomposition is performed on the set of latent vectors, the set $\mathcal{U}_K$ depends on the optimisation process, the training parameters (batch size, number of epochs, for example) and the loss functions.

6.3.3 Implementation and Training Details

The Adam optimiser was used for our training, with a constant learning rate set to $1e^{-4}$, a batch size equal to 4 and 8 for lung and hippocampus, respectively. Our models were trained for 600 epochs, and it lasted approximately 4 and 9 hours for the lung and hippocampus dataset. The higher training time is due to a more significant number of samples for the hippocampus dataset. Concerning data augmentation, we applied random flip, rotation, translation and zoom. We did not perform data augmentation for the lung dataset as the input volumes are bigger, which would slow down the training too much. Moreover, the weights of the different loss components were set to 1 except the loss for smoothness set to $\alpha = 0.1$ for both datasets and the weight for the jacobian loss β that was discarded for the hippocampus dataset. Indeed, hippocampus registration was not

created non-topological points as the task is easier. During the training process, we registered random pairs of different patients as in previous chapters. However, the validation was executed with fixed pairs, particularly inspiration-expiration pairs for the lung dataset. Our training has been performed using the framework PyTorch and one GPU card Nvidia Tesla V100 with 32G memory. The PCA decomposition was calculated using the library scikit-learn, and the number of principal components K was set to 32. Using 32 components, our decomposition covered 95% and 93% of the variance ratio for the lung and hippocampus dataset, respectively, while 42% and 62% are covered by the first four components for each dataset, respectively.

6.4 Experiments and Results

6.4.1 Data and Preprocessing

We perform our experiments on two different datasets, one public and one private. Starting with the public dataset, we conduct experiments with the hippocampus[1] [Simpson, 2019]. This dataset has been published for the Medical Segmentation Decathlon (MSD) and was recently used as a subtask of a registration challenge, Learn2Reg [Dalca, 2020; Hering, 2021]. It is composed of 394 MRI with the segmentations of two small structures, the head and the body hippocampus. The images have been cropped around the hippocampus into small patches of $64 \times 64 \times 64$ voxels. The second dataset is a private dataset acquired during a study on pulmonary fibrosis detection. It comprises 41 lung MRI patients (12 healthy and 29 diseased with pulmonary fibrosis) together with their lung segmentations. Each patient had been acquired in two states, the inspiration and the expiration, resulting in 82 MRI. Each volume has been resampled to 1.39mm on the x and z-axis and 1.69 on the y-axis and cropped to $128 \times 64 \times 128$ volumes. More information about this dataset can be found in Stergios et al. [Stergios, 2018]. The same normalisation strategy has been applied for the two datasets: $\mathcal{N}(0, 1)$ standardisation, clip to $[-5, 5]$ to remove outliers values and min-max normalisation to $(0, 1)$. Both datasets were split into training and validation, resulting in 200 and 60 patients for the hippocampus and 28 and 13 patients for the lung dataset.

6.4.2 Evaluation of the Registration Performance

As the first step of our evaluation, we benchmarked the performance of the registration network g_θ, on which our decomposition is based on. More specifically, we obtained a Dice coefficient of 0.90 ± 0.04 for the lungs and 0.76 ± 0.05 for the hippocampus, while the initial unregistered cases reported a Dice of 0.74 ± 0.14 and 0.59 ± 0.15 respectively. Moreover, we calculated the registration for g_θ with the skip connections to measure their impact on the registration. The Dice is then equal to 0.92 ± 0.02 and 0.85 ± 0.03 respectively. Thus, by removing the skip-connections, we decrease the performance of the registration, slightly on the lungs, more importantly, on the hippocampus. However, both strategies register the pair of volumes properly.

[1] http://medicaldecathlon.com/

6.4.3 Qualitative Evaluation

To understand and evaluate the calculated components of $\mathcal{U}_K$ per dataset, we perform a qualitative analysis. In particular, for each principal vector $\overrightarrow{u_i}$, we calculated the corresponding deformation $\{\phi_i\}_{i=1\cdots K}$ and we applied to the moving image M together with its corresponding segmentation map M_{seg}. More formally, the deformed contour corresponds to $\mathcal{W}(\mathbf{D}(\lambda\overrightarrow{u_i}), M_{seg})$ with $\mathcal{W}$ being the warping operation and λ the parameter to control the strength of the displacements for better visualisation.

In figure 6.2, we show the principal components obtained for one validation subject for the lung dataset. Interestingly, one can observe that each ϕ_i corresponds to a different elementary deformable transformation. More precisely, the 1^{st} component is associated with translation from top to bottom, the 2^{nd} with a deformation focusing on the bottom of the lungs, the 3^{rd} with a deformation on the right lung focusing also on the heart region and lastly the 4^{th} with a deformation focusing on the top region of the lung and shoulders.

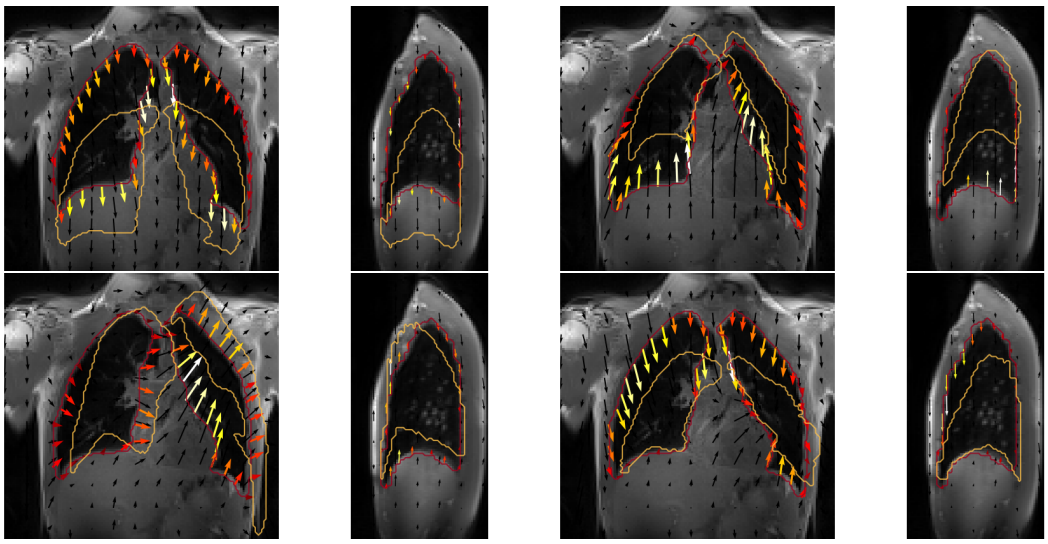


Figure 6.2: Visualisation of the displacements following the first four principal components. For each component, we depicted coronal and sagittal views. In red, the contours of the lungs of the M image, and in gold, the deformed lung's contours, $\mathcal{W}(\mathbf{D}(\lambda\overrightarrow{u_i}), M_{seg})$. The deformation field is represented with arrows. The arrows' norm is represented with a colour map, red being the smallest and white the largest. Other patients and components are displayed on supplementary materials. The components 1 and 2 are represented on the first line, the components 3 and 4 on the second line.

In Figure 6.3, we show the effect of the values of λ. In the figure, we present the original moving image's lung contours (in red) and the corresponding warped image and segmentation for the first and fourth components. As we have indicated, the $1st$ component is associated with translation, which can also be observed in this visualisation. In particular, for this experiment we sample λ from the values $\{-200, -100, 0, 100, 200\}$. One can observe that for a value of 0 we retrieve a near identity deformation, while for negative and positive values, the lung moves up and down, respectively. On

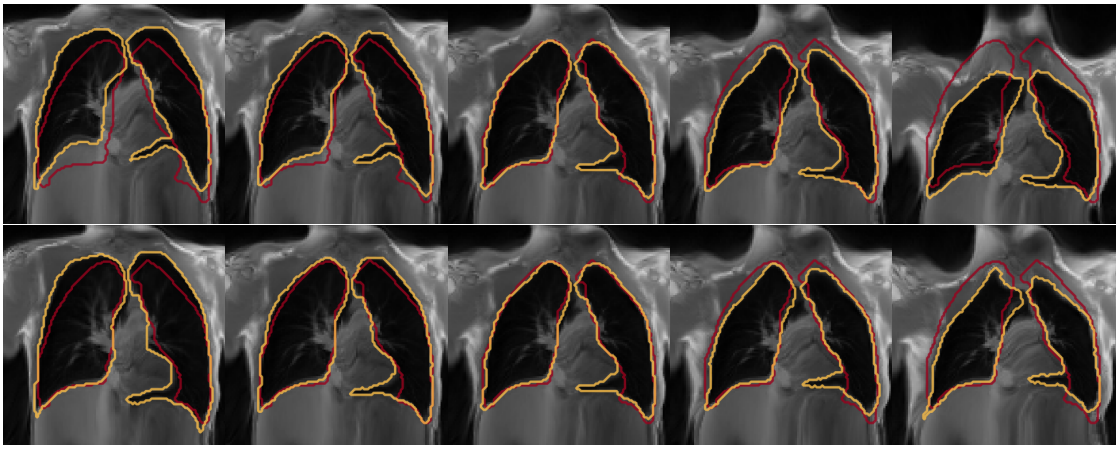


Figure 6.3: Representation of the deformed MR together with its lung contours following the first and fourth principal components u_1, u_4. The red contour represents the position of the lung segmentation of the input image while the gold contour the position of the deformed lung. The values of lambda range from: -200, -100, 0, 100 and 200 (left to right). Negative values of lambda correspond to an upward translation, while positive values to a downward translation. The first row represented the first component u_1, the second row the fourth component u_4.

the other hand, the fourth component is responsible for the displacement of the upper part of the lungs. Indeed, we can observe that through the different λ values, the top right lung region is the one that reports the most changes.

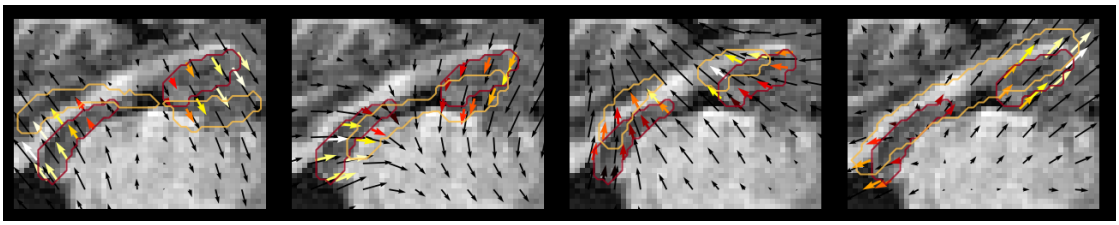


Figure 6.4: Visualisation of the displacements following the first four principal components. We depicted a sagittal view of one patient of the validation set of the hippocampus dataset. We represented the ground truth hippocampus contours (red) and the deformed one (gold), following the principal components.

In Figure 6.4, similarly, the 4 deformations produced by the first 4 principal components of the hippocampus dataset are presented. In this case, the 1^{st} component seems to capture rotation on the sagittal plane, the 2^{nd} translation and shrinking towards the bottom right, while the 3^{rd} seems to be the same operation towards the top left corner. Finally, the 4^{th} seems to be related to scaling, inflating the hippocampus's head and tail. We observe that the decomposition of the two datasets created different elementary transformations ϕ_i, with transformations closer to affine transformation for the hippocampus and more complex for the lung.

Finally, to verify the obtained decomposition, we performed a case study for all the hippocampus

dataset validation subjects. More specifically, we applied some predefined rigid transformations: translation using 10 pixels on the z axis, rotation using 20 degrees on the z axis and scaling using a factor of 0.2, transforming each subject X to X'. Then we calculated the difference between the projection of $\mathbf{E}(X)$ and $\mathbf{E}(X')$ on the PCA decomposition. In Figure 6.5, a box plot for all the validation subject of the absolute difference is presented. Specifically, the amount $||a^j_{\mathbf{E}(X)} - a^j_{\mathbf{E}(X')}||$ is shown for each principal component j, with $a^j_{\mathbf{E}(X)}$ being the projection of $\mathbf{E}(X)$ on the principal vectors $\mathcal{U}_K$, for the three different applied deformations. One can observe that for rotation and translation, only one component is significantly different from the rest. In the case of scaling, however, two components seem to be more activated. Moreover, these findings are in accordance with Figure 6.4, especially for the $1st$ component, which visually corresponds to a rotation. However, we should remind that the predefined rigid transformations are distinct from the elementary deformable transformation ϕ_i. This approach is more complex to produce for the lung dataset, as the elementary transformations are further from rigid transformations than in the hippocampus case.

In supplementary materials, we upgraded the Figure 6.5 by comparing the network with and without skip-connections (Figure 6.9). For each of the three selected transformations, we displayed the amount $||a^j_{\mathbf{E}(X)} - a^j_{\mathbf{E}(X')}||$ for two different decompositions. One decomposition obtained for a network trained without skip connections, an other with a network without skip connections. Contrary to our proposed formulation (without the skip), where only one or two components are activated, many components are activated with the skip-connections. This demonstrates the necessity of removing the skip connections to have all informations inside the bottleneck and thus a good decomposition.

6.4.4 Evaluation of the clinical pertinence

To measure our latent representation's predictive power, we performed different experiments on the lung MR dataset. On this dataset, we have different clinical information: the status of the patient, healthy or not, the volumes of the lung and the status of the images, inspiration and expiration. We conduct classification experiments on the status of the patient using the latent representation projection on the principal components vector (a^j_i). As our number of principal components was set to $K = 32$, we have a small input vector, giving the possibility to use a simple machine learning algorithm and removing the need for features selection algorithms. We use the same training set for the registration network training, the PCA calculation, and machine learning algorithms training. We trained only a logistic regression algorithm to demonstrate that even a linear algorithm could benefit from our PCA decomposition. The prediction of an image being an inspiration or expiration achieves an accuracy of 0.96 and 1.0 on the training and validation set with as many inspiration images as expiration in both sets. The disease status prediction obtained a balanced accuracy of 1.0 and 0.66 and an F1 score of 1.0 and 0.65 on the training and validation set. The validation set was composed of 13 patients with six healthy and seven non-healthy. 4 healthy patients have been misclassified while all the non-healthy patients have been correctly classified.

We compared our classification of the disease status with two different approaches using the latent space. These two approaches used the same logistic regression algorithm but differed by the features

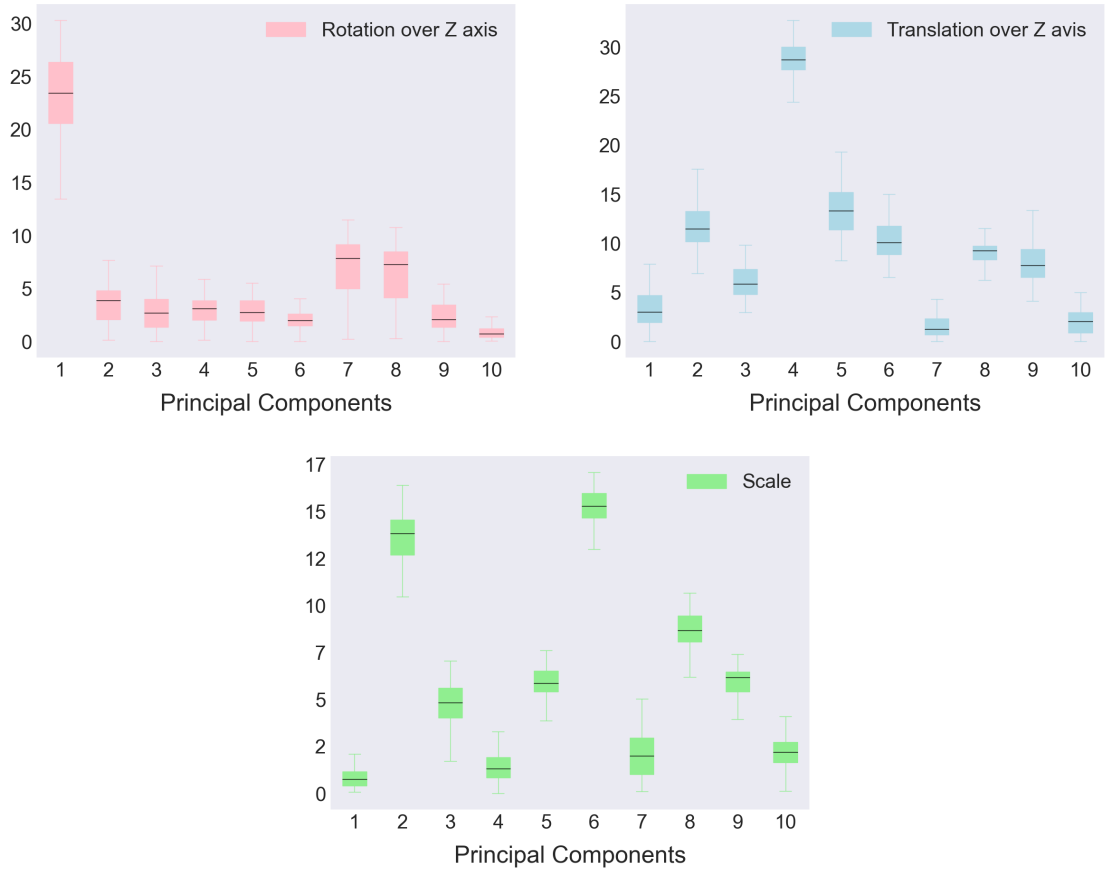


Figure 6.5: Visualisation of the differences between the components of a reference image and the same image to which we applied a predefined transformation. The first ten components have been displayed. From right to left: rotation, translation along Z-axis and scaling

selection algorithm. We explored simple features selection methods to be a fair comparison with the PCA decomposition. We applied a global max-pooling operation on the latent space representation for the first one, obtaining 256 features for each input volume. The second one was a univariate features selection, selecting the 32 best features. The global pooling and the univariate feature selection approaches obtained a balanced accuracy on the validation set of respectively 0.75 and 0.83, surpassing our proposed approach's performance. However, due to the small size of the validation set, it does not seem easy to conclude. Indeed, the difference between these three selection features is only one patient misclassified. Moreover, our methods allow a representation and analysis of the features shown in the previous section, while the two others can not be interpreted. Therefore, the evaluation of the clinical pertinence of this chapter needs to be pursued with more extensive datasets.

Finally, we displayed on Figure 6.6 the first and second principal values a_i^1 and a_i^2 for every patient X_i and we observed a singular organisation of the inspiration-expiration pairs. More particularly, the distance between the two-time points seemed to be related to the inspired volume. Thus, we

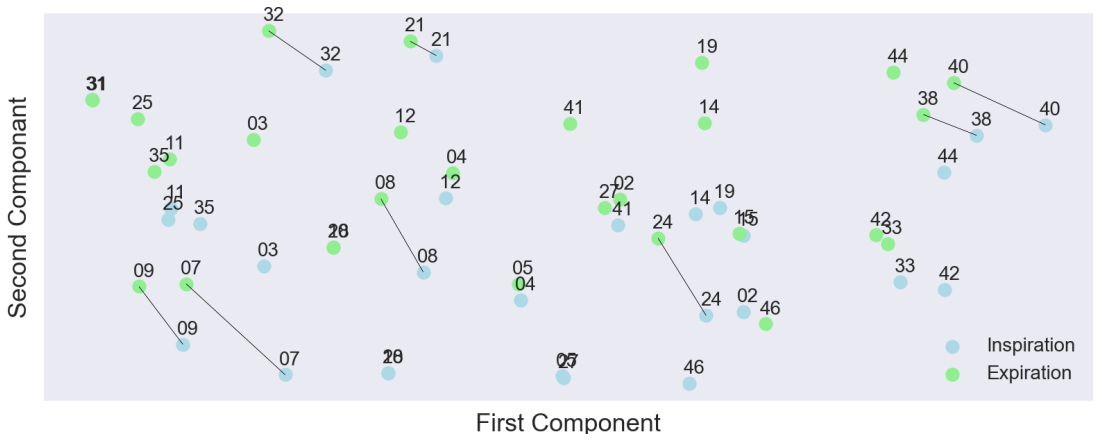


Figure 6.6: Representation of the 1^{st} and 2^{nd} components of each volumes of the lung dataset. Each patient is indicated with its number and is displayed twice, with the inspiration (in green) and expiration (in blue) MRI. We draw arrows connecting the inspiration and expiration points, showing only a few for readability's reasons.

calculated the correlation between the inspiration and expiration point on the latent space and the volume difference. For each patient, we calculated $\Delta V = V_{inspi} - V_{expi}$ and $||a_{\mathbf{E}(X_i)} - a_{\mathbf{E}(X_e)}||$, where X_i and X_e are the inspiration and expiration MRI. This norm corresponding to the length of the arrows depicted on Figure 6.6, however, taking into account all components. We obtained a spearman correlation of 0.91 and 0.76 for the training and validation set. These experiments show a relationship between our linear decomposition and some clinical features, while the decomposition is obtained in an unsupervised way.

6.5 Discussion and Conclusion

In this chapter, we proposed an approach to decompose and explain the representations of deep learning-based registration methods. The proposed method utilises a linear decomposition on the latent space projecting it to principal components closely associated with anatomically aware deformations. Our method's dynamics are demonstrated in two different MRI datasets, private and public, focusing on lung and hippocampus anatomies. We hope that these results will be able to take some steps towards a better understanding of latent representations learned by the deep learning registration architectures. Moreover, such projections can be used to drive the decoder to produce anatomically aware augmentations of the moving images.

We also explored a direct application of the PCA on the deformation's grid instead of the latent representation. However, we did not observe any qualitative correlations with types of deformations, which is the case for our proposed formulation. Indeed the different components of the PCA do not correspond to deformations.

We must recognise several limitations to the proposed method. The main one is the difficulty of the

quantitative evaluation of the decomposition and the analysis of the elementary transformation. We only evaluated our approach qualitatively, and transformations ϕ_i are often challenging to interpret, especially for the lung case. Another limit is the use of the lung and hippocampus segmentation mask. A fair comparison with fully unsupervised training is needed to understand if the decomposition is driven by the use of contours or not. However, as we applied a weakly-supervision registration training, the validation can be performed without masks, and even the training could be done with masks for a small percentage of the volumes. Finally, the analysis of the clinical pertinence of the PCA decomposition showed that other methods outperformed ours. However, a more robust clinical evaluation must be performed.

Our future steps include the more extensive evaluation of our method, both quantitatively and clinically, and the extension to new anatomies such as abdominal volumes. More specifically, we want to apply our approach to multi-temporal follow-up of patients, monitoring diseases' progression. We hope that our decomposition will be able to separate organs' displacement from tumour growth. Currently, the longitudinal registration during treatment is challenging to interpret for some localisation with high displacements such as the liver.

6.6 Appendix

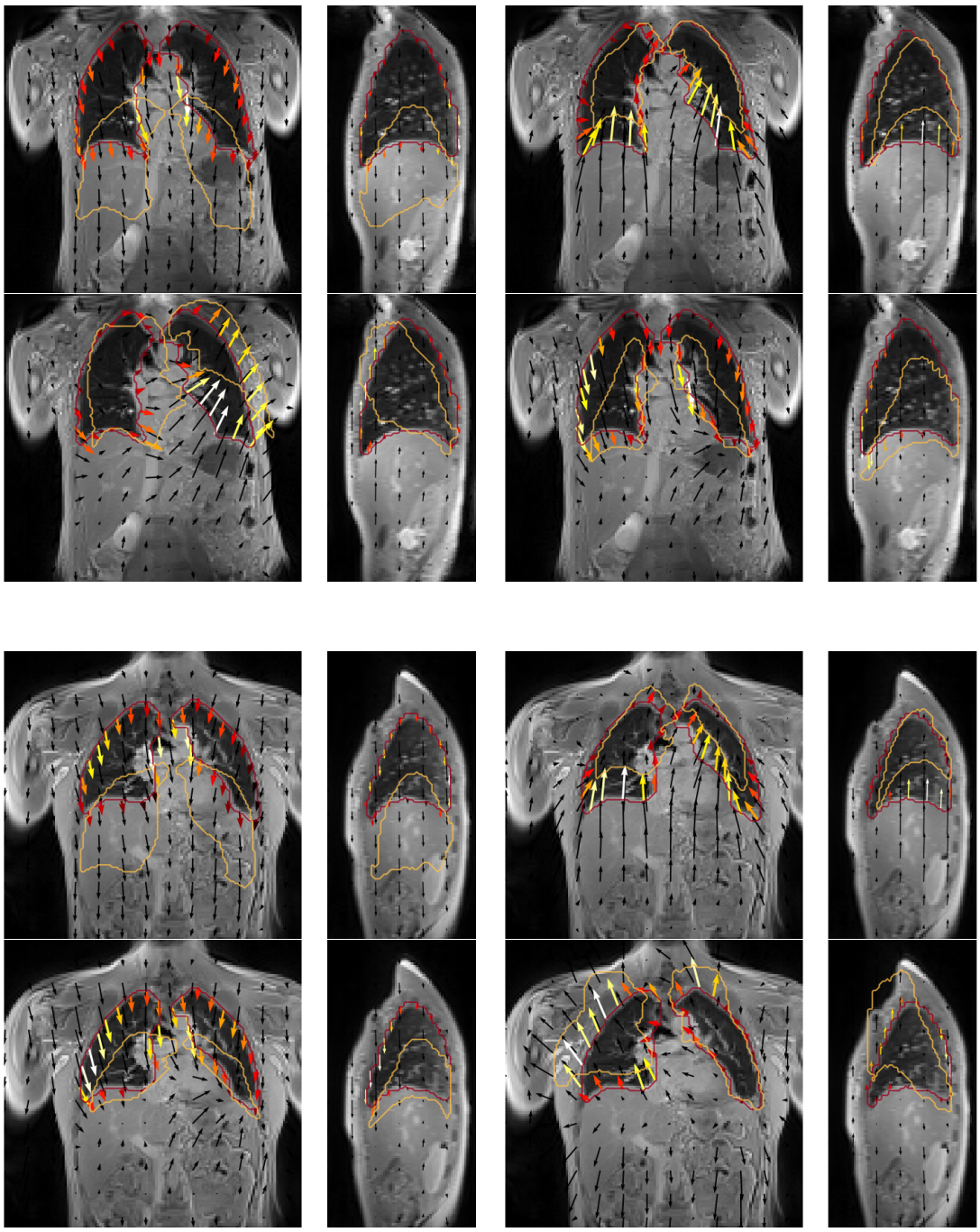


Figure 6.7: Extension of the figure 6.2 with two others patients and the component 1 to 4.

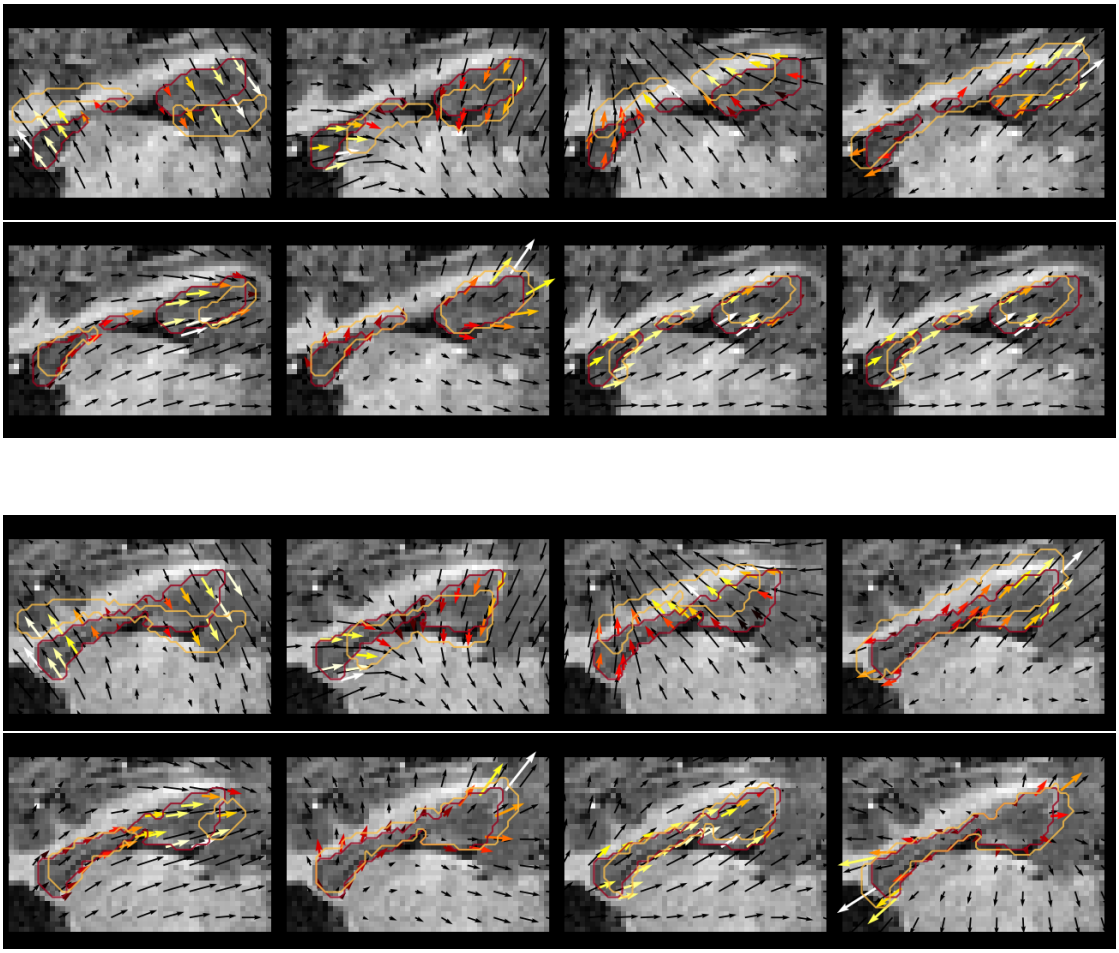


Figure 6.8: Extension of the Figure 6.4 with two other patients and the components 1 to 8 (first row 1-4, second row 5-8).

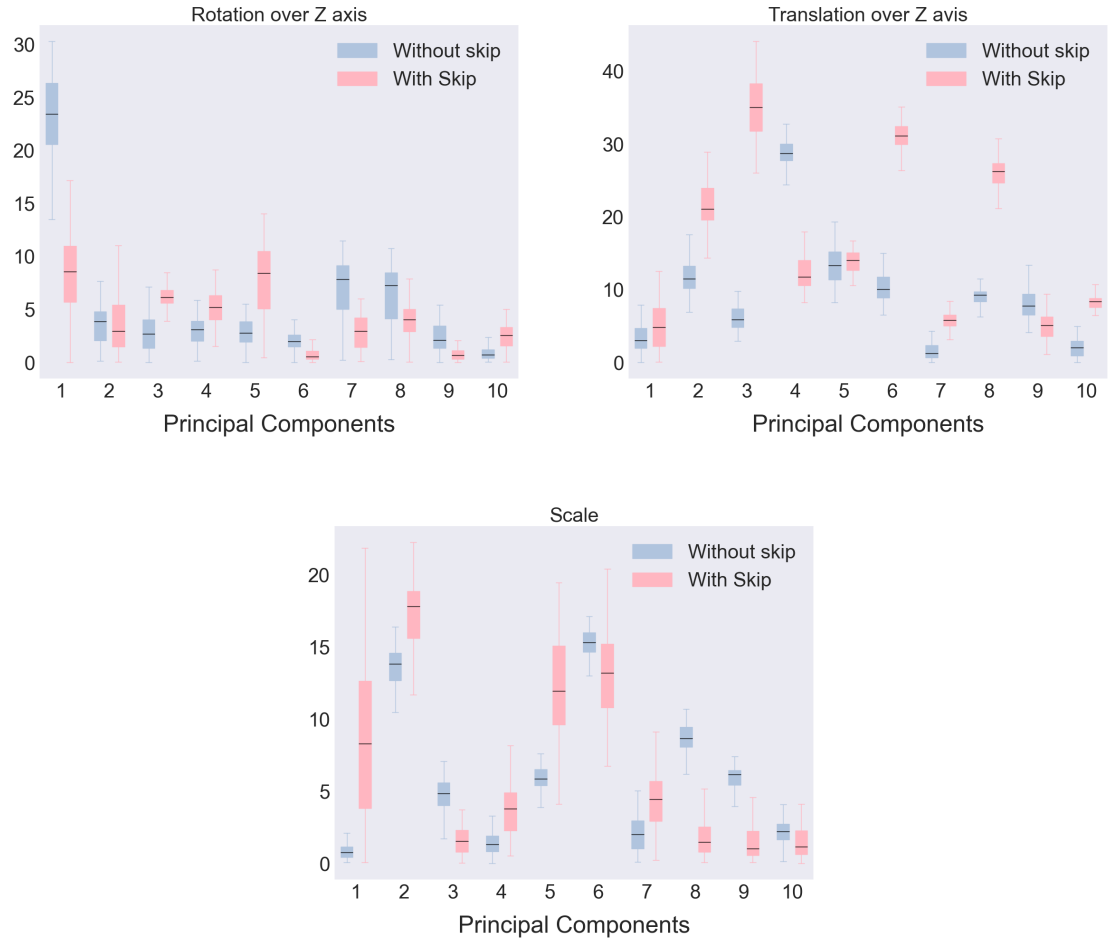


Figure 6.9: Comparison between the networks with and without skip connections. We displayed the difference between the components of an image and the same image to which we applied a predefined transformation. Only one or two components are activated without the skip-connections, while many of them are with the skip. The first ten components have been displayed. The results are in blue and pink for respectively without and with the skip-connections. From right to left: rotation, translation along Z-axis and scaling.

Chapter 7

Conclusions

"Hello, I'm the Doctor.
Basically... run!"

Doctor Who, The Eleventh Hour

7.1 Main Contributions

In this thesis, we investigated different methods to ameliorate deep learning-based registration. We built our research on the recent development of registration, including unsupervised and semi-supervised frameworks, but also classical registration and other deep learning frameworks such as multi-task learning or pretraining. We worked with different anatomies and modalities, such as brain MRI, abdominal CT and lung MRI.

First, we presented in **Chapter 2** the registration key concepts, such as the deformation model, the optimisation strategy and the objective functions. We also introduced the use of deep learning in medical imaging and the development of deep learning-based registration. In this chapter, we described particularly unsupervised and weakly supervised registration as we based our work on these frameworks.

In **Chapter 3** and **4**, we investigated the multi-task learning framework (MTL), combining segmentation and registration tasks. We proposed a joint architecture composed of one encoder and two task-specific decoders and trained our network in an end-to-end way. We aim to improve the robustness and the quality of the encoder features, thanks to this joint formulation. We also introduced a new loss function that modifies both the segmentation and the registration decoders.

In Chapter 3, we applied our architecture to the case of the brain without abnormalities and our segmentation network output three main brain structures (white matter, grey matter and ventricles). In Chapter 4, we worked with a more challenging dataset: brain with tumours. In this situation, we aimed to perform the registration only on healthy parts of the brain. To do so, our segmentation decoder predicts the tumour segmentation and the binary masks are integrated into the similarity loss. We compared our method to other DL-based registration algorithms and demonstrated that we managed to register only normal areas while keeping the tumours intact.

In **Chapter 5**, we investigated new localisations working with abdominal CTs and hippocampus MRIs. First, we modify our formulation to register both the moving and the fixed image, and we developed a new pretraining strategy using pseudo-segmentations generated by a segmentation network. Using this methodology, we participated to the task 3 and 4 of the Learn2Reg 2020 Challenge and obtained the 2^{nd} position on the overall competition. In a second time, we investigated more precisely the impact of the pseudo-segmentations as well as the introduction of two regularisation losses. We aimed to keep the same accuracy while increasing the smoothness and plausibility of the deformations. We also designed and studied a multi-step strategy, which improved the results despite strong regularisations weights. Finally, we obtained better results for the Dice coefficient as well as for the smoothness of the deformation.

In **Chapter 6**, we explored the topic of explainability of registration networks. Recent approaches on interpretable deep learning include qualitative evaluation and gradient-based methods such as GradCam. It produces heatmaps showing which areas of the input image have the most impact on the final prediction. Here, we explored the use of our network latent space and its connection with interpretability. More precisely, we projected every image to the latent space using the trained encoder and produced a linear decomposition of this space (PCA). Each vector of the basis corresponds to an elementary deformation, focusing on specific areas of the image. We produced a qualitative evaluation of our decomposition and also studied its predictive power, performing classification experiments on a lung MRI dataset.

7.2 Future Applications

We developed several registration methods in this thesis and demonstrated the high potential of deep learning in this field. Depending on the task, there are still many challenges to be addressed in order to make the registration techniques ready for clinical practice. Future work is therefore required to increase the accuracy and mostly the smoothness of registration methods. We investigated some ideas to improve deep learning-based registration, but other possibilities exist, including multi-scale networks, hyperparameter learning or cycle-gans approaches. Despite the recent progress of DL-based registration, classic iterative methods should not be forgotten as they provide good results. Recent publications focus on accelerating iterative methods using graphics processing units, obtaining similar calculation time than DL-based methods while keeping good results. A positive step should also be to extend DL-based registration to other anatomies and imaging modalities, for instance, head and neck or pelvis.

Regarding deep learning methodologies, joint formulation and multi-task learning should be studied more heavily. Our experiments in the chapter 3 did not demonstrate a clear advantage of the joint architecture compared to weakly supervised approaches. However, we did not explore the robustness of the learned representation completely. We evaluated our joint formulation on one of the tasks for which we had trained our network. Maybe the joint segmentation-registration network demonstrates the quality of the learned features if we evaluate on a third unknown and more complex task, freezing the encoder and training few task-specific layers. Another interesting usage of DL-based registration could be to reduce overfitting and increase generalisation. Two main fields where registration could be useful are imaginable: pretraining or data augmentation. For pretraining, one major advantage of unsupervised registration is that we can perform it without labels and with a huge number of pairs. Then, it could be used for pretraining without augmenting the complexity of the data collection or data preparation. Other pretraining tasks without supplementary labels include in/out-painting, reconstruction or local shuffling. We assume that registration could learn more robust features as it learns structures and organs, but an exhaustive comparison is demanded. Concerning data augmentation, the goal is to generate more various data than with current data augmentation methods. Indeed, they include mostly rotation, translation or zoom, which only modify the images globally. Using an already trained registration network, we could generate fast elastic deformation and thus expand the dataset. However, some researchers have already implemented elastic deformations in their training loop without deep learning and GPU acceleration.

Concerning clinical applications, one could name two very important applications of the registration: multi-modal registration and temporal registration. These two topics concerned most of the time intra-patient registration, demanding much effort to construct a database, as we need two images for the same patient. Multi-modal formulations often comprise two different encoders or projectors and one decoder. The encoders produce a common representation for the different modalities, and the decoder merges them and produces the deformation grid. During this thesis, we start exploring temporal follow-up concerning, particularly cancer patients. We registered pre and post-treatment CTs from the same patient and studied the deformation field generated using a trained network. We assumed that the transformation encodes the clinical response of the patient. However, the deformation field incorporates both the tumour growth and natural abdominal deformation, and the decomposition remains challenging. The linear decomposition described in Chapter 6 could help to obtain only the tumoral deformation and carry on the research on temporal registration. Finally, a last clinical application should incorporate DL-based registration in radiotherapy treatment, for instance, for image-guided radiotherapy and dose adaptation.

Résumé français

Dans cette thèse, nous nous intéressons à des nouvelles approches pour trouver le meilleur déplacement entre deux images médicales différentes. Ce domaine de recherche, dénommé recalage d'images, a de nombreuses applications cliniques, comme par exemple la fusion de plusieurs images de différentes modalités ou encore le suivi temporel d'un patient au cours de son traitement. Le recalage est étudié depuis de nombreuses années avec différentes méthodes, comme les méthodes basées sur des difféomorphismes, utilisant la théorie des graphes ou les équations physiques. Récemment, des méthodes basées sur l'apprentissage profond (ou deep learning) et utilisant des réseaux de neurones convolutionnels ont été proposé. Ces méthodes ont obtenu des résultats similaires aux méthodes classiques (c'est-à-dire sans deep learning) tout en réduisant fortement le temps de calcul et rendant possible une utilisation clinique en temps réel. Cette accélération provient de l'utilisation de processeurs graphiques (GPU) et du design de la phase de prédiction qui ne nécessite pas d'optimisation. Cependant, le recalage basé sur le deep learning a plusieurs limitations, comme le besoin de très grandes bases de données pour entraîner le réseau ou le réglage des hyper-paramètres de régularisation pour empêcher des transformations trop irrégulières. Dans ce manuscrit de thèse, nous étudions différentes modifications aux algorithmes de deep learning, comme par exemple l'association du recalage avec des réseaux de segmentation, l'utilisation de la technique de pré-entraînement, l'introduction de nouvelles fonctions de coût et l'étude de l'explicabilité des réseaux de neurones.

Dans un premier temps, nous étudions la relation entre deux domaines majeures de l'imagerie médicale : la segmentation et le recalage. En s'inspirant de méthodes non-deep learning qui résolvent ces deux problèmes simultanément, nous avons développé une nouvelle architecture composée d'un encodeur et de deux décodeurs. Cette architecture génère une transformation non-rigide et une matrice de segmentation. Contrairement à d'autres méthodes jointes, nous utilisons un seul réseau et nous renforçons le couplage entre les deux tâches en appliquant la grille de déformation au masque de segmentation prédit. Ceci est similaire au recalage faiblement supervisé (weakly-supervised) où la transformation est améliorée en déformant la segmentation. Nous avons aussi introduit une nouvelle fonction de coût qui modifie les poids des deux décodeurs. Nous évaluons les performances de notre méthode sur un jeu de données public composé d'IRM cérébrale et comparons avec des architectures qui n'utilisent pas le couplage. La performance du recalage est mesurée en calculant le coefficient Dice sur différentes structures cérébrales, et les résultats sont confrontés aux approches non supervisées et faiblement supervisées.

Ensuite, nous nous intéressons au cas des cerveaux avec une tumeur, alors que le précédent réseau était entraîné uniquement avec des cerveaux sans lésion. La présence d'une tumeur perturbe l'optimisation de l'algorithme de recalage, en créant une non-concordance entre les deux images. Ici, notre but est de recaler des IRMs cérébrales en se basant uniquement sur les parties saines du cerveau et en ignorant les zones tumorales. Des travaux similaires ont été publiés avec des algorithmes classiques de recalage, en utilisant l'association du recalage et de la segmentation. Nous avons modifié notre architecture multitâche pour prédire la segmentation de la tumeur et rajouter celle-ci dans la fonction de loss de similarité. Ainsi, le réseau est entraîné uniquement sur les régions en bonne santé. Nous avons aussi changé notre formulation conjointe de recalage et segmentation pour respecter de meilleures propriétés. Nous adoptons une nouvelle stratégie pour fusionner les deux images, appelée fusion tardive. Chaque image est traitée indépendamment par l'encodeur et elles sont combinées avant d'être passées dans le décodeur spécialisé dans le recalage. Grâce à cette nouvelle stratégie, la prédiction de chaque masque de segmentation dépend uniquement de l'IRM correspondante et non plus de la concaténation des deux. Avec nos expériences, nous évaluons à la fois le recalage, la segmentation des tumeurs et l'absence de déformation de la tumeur. Pour les deux tâches, nous comparons à des architectures références. Notre formulation obtient des performances similaires sur le recalage et la segmentation tout en gardant la tumeur intacte.

Après avoir travaillé sur les IRM cérébrales, nous étudions une nouvelle localisation et méthode d'imagerie : des scanners abdominaux (CT). Cette partie du corps est plus compliquée pour le recalage, car il les organes ont une tendance naturelle à se déplacer et sont facilement déformables. Ces mouvements sont dus par exemple à la digestion ou à la respiration. En conséquence, les déformations prédites sont souvent soit fortement bruitées ou bien très proche de la transformation identité. Dans cette partie, nous développons plusieurs techniques pour améliorer les performances du recalage. Parmi ces approches, nous utilisons la technique de pré-entraînement pour profiter de l'existence de nombreux jeux de données publiques ainsi que des pseudo-segmentations générées par un réseau de neurones. Nous analysons aussi l'impact de nouvelles fonctions de coût pour améliorer la régularité des déformations. Ces fonctions pénalisent les valeurs négatives du Jacobien ainsi que la symétrie des transformations. Enfin, nous développons une stratégie multi étapes pour raffiner la déformation. Ce travail a donné lieu à une participation au challenge Learn2Reg organisé à l'occasion de la conférence MICCAI. Nous avons obtenu une 3e position à deux tâches parmi les 4 proposées, la tâche 3 et 4 concernant le recalage de CT abdominaux et d'IRM cérébrales (hippocampe).

Finalement, dans le dernier chapitre, nous tentons de comprendre les liens entre le recalage à l'aide de deep learning et l'explicabilité des réseaux de neurones. La compréhension du fonctionnement de ces algorithmes est capitale pour l'imagerie médicale, car les médecins ne peuvent pas se baser sur une « boîte noire ». Notre approche se fonde sur une décomposition linéaire de l'espace latent en utilisant une analyse aux composantes principales (PCA). Chaque image est projetée dans l'espace latent grâce à notre stratégie de fusion tardive et nous obtenons une base de cet espace latent. Cette base permet de générer des transformations élémentaires en utilisant le décodeur et nous évaluons qualitativement ces transformations. Nous montrons sur deux différents jeux de données, IRM pulmonaire et cérébrale (hippocampe), que ces transformations élémentaires se spécialisent sur certaines parties du corps ou certains déplacements. Nous explorons aussi les liens entre notre la décomposition obtenue et des variables cliniques, étudiant par exemple la corrélation entre volume de respiration et la position dans l'espace latent.

Remerciements

4 mois après ma soutenance et la fin officielle de ma thèse, je me décide enfin à écrire mes remerciements de thèse. Partie difficile, car il faut remercier toutes les personnes qui m'ont accompagné depuis le 11 décembre 2017 et même avant. Spoiler alert : je vais oublier des gens, alors si vous n'êtes pas dans les paragraphes qui suivent, d'abord toutes mes excuses, et ensuite merci à vous.

Beaucoup de choses se sont passées pendant cette thèse (en plus de la lecture d'article scientifique et de l'écriture de code Python). Évidemment de l'escalade, des spectacles d'improvisation et des soirées avec les copains. Mais aussi 4 mois de travail à Philadelphie avec une visite de la côte Est des États-Unis et des chutes du Niagara, une conférence en Chine (avant le Covid !), une semaine de Summer School en Sicile, beaucoup de changements de bureau à Gustave Roussy, un changement de labo à Centrale, une collocation dans le 5ème à côté de la rue Mouffetard (avec une victoire en coupe du monde), un autre appartement à Gentilly coincé entre le périphérique et l'A6, une traversée de la moitié Sud du GR20, des vacances pour faire du surf dans les Landes, mais aussi des cours de Python et de Deep Learning à donner à des étudiants, l'organisation d'un data challenge, des vacances d'escalade en Espagne, Turquie, Suisse et France et plein d'autres choses.

Pour commencer, je remercie mes directeurs de thèse, Nikos, Eric et Maria. Un énorme merci à Maria Vakalopoulou qui m'a accompagné pendant toute cette thèse, que ce soit pour finir les articles pour les conférences à 4h du matin, pour aller en Chine à Shenzhen ou pour me remotiver quand j'en avais besoin. J'aurais voulu écrire une phrase en grec dans mon manuscrit de thèse, mais malheureusement les seuls mots que je connais sont ευχαριστώ, Καλημερα et Χρόνια Πολλά, ce qui signifie respectivement *merci*, *bonjour* et *bon anniversaire*.

Un remerciement tout particulier à Paul-Henry Cournède pour l'accueil dans son laboratoire. Merci à tous les membres du jury pour avoir généreusement accepté d'évaluer mes travaux; mes deux rapporteurs Stéphanie Allassonière et Hervé Delingette pour la relecture de mon manuscrit et leurs remarques, et les autres évaluateurs, Mattias Heinrich, Christos Davatzikos et Marleen de Bruijne. Et pour finir sur la soutenance, merci à toutes les personnes présentes physiquement, en visio ou en pensée pendant ce jour particulier.

J'ai eu la chance de rencontrer beaucoup de doctorants pendant toute cette thèse. J'ai une pensée pour Mihir, Marie-Caroline, Arthur, Stergios, Maria et les autres doctorants du CVN avec qui j'ai commencé ma thèse. Je pense également à tous les doctorants et chercheurs étrangers que j'ai croisé en conférence à l'étranger. Merci aussi à tous les doctorants du laboratoire MICS (beaucoup trop nombreux pour être cité ici) pour leur accueil, la bonne ambiance studieuse et le bon esprit. Enfin merci à tous les membres de l'équipe de Gustave Roussy, Emilie, Alexandre, Stéphane, Sylvain, Angela, Roger, Jade, Nathan, et Charlotte. Et surtout, je remercie énormément les doctorants avec qui j'ai partagé plusieurs bureaux, enfermé au -1 de Gustave Roussy, sans aucune fenêtre, Enzo Battistella, Marvin Lerousseau, Théophraste Henry et Amaury Leroy. On a vécu beaucoup de choses dans ce sous-sol, ce fut un plaisir de partager les bons et mauvais moments avec vous. Bon courage pour vos fin de thèses et rendez vous à vos soutenances.

Merci aussi à Philippe de Vomécourt pour le temps passé ensemble à Gustave Roussy et l'organisation du data challenge 2018.

J'ai eu la chance d'aller passer 4 mois de ma thèse dans un laboratoire de l'université de Pennsylvanie à Philadelphie. Je remercie Christos Davatzikos pour l'accueil dans son laboratoire, Anahita pour l'encadrement et Vishnu pour nos conversation et nos repas du midi. J'ai profité de ces 4 mois pour visiter une partie de la côte Est américaine avec Aline, Valentin, Raphaëlle et Thibault. Merci pour les parties de dame de pique, les expéditions à Atlantic City, les brunchs, les rooftops, les célébrations du 4 juillet et toutes les autres choses que j'oublie.

De très nombreux amis ont partagé une partie de ma vie pendant ces presque 4 ans.

Jean Delbecque avec qui on a fait toutes les classes de la 6ème à la terminale.

Le groupe du Toddy et plus si affinités : Romain, Léa, Quentin, Willy, Arthur, David, Audrey, Jean, Fatma, avec qui j'ai partagé des nombreuses bières (ou cidre), des séances de ciné et de jeux de société, vacances sportives et week-ends.

Merci à tous les membres de la meilleure troupe d'improvisation de Paris i.e : Les Chutes Libres ! Dans cette joyeuse équipe, on trouve Nessy, Nassim, Lou, Arilès, Laurent, Mathilde, Gwen, Baptiste, Christine, Yannick, Hippolyte, Timothée, Thomas, Erika et Sylvain. Toutes nos improvisations à base de conspirations, de Luna Corp, d'autres jeux de mots, et nos AGs interminables ont été des moments de respirations très importants pour moi. D'ailleurs si quelqu'un lit ses lignes en 2022, on a un spectacle tous les mardis soirs à l'improvibar (et aussi une page Facebook).

Merci à tous les copains avec qui j'ai fait de l'escalade dans les salles à Paris, sur les rochers de Fontainebleau et sur les falaises de France et du monde. Dans un ordre aléatoire : Gabriel, Paul, Ivan, Félix, Adrien, Zénon, Henri, Giulio, Clément, Aurore, Hugo, Grégoire, Caro, Mathilde, Lara, beaucoup beaucoup d'autres. Je remercie grandement tous les colocataires du boulevard des Brotteaux à Lyon qui m'ont vu passé entre un train et une falaise et m'ont accueilli.

Merci aux différents collocataires anciens (NCW, Tournefort) et actuel (Gentilly). Un prix spécial à Arthur Chavignon, avec qui j'ai partagé plusieurs appartements mais aussi l'expérience de la thèse, de la recherche académique française et de la maturité scientifique. On aurait pu créer une série Youtube intitulé *Deux collocataires font une thèse, ça tourne mal !*. Finalement, on finit tous les deux docteurs.

Je remercie enfin toute ma famille pour ce qu'ils m'ont apporté depuis presque 28 ans et ce qu'ils m'apportent encore, mes parents Thierry et Isabelle, mon frère Samuel, ma sœur Pascaline, mes grands-parents, oncles, tantes et cousins. Je dédie particulièrement ma thèse à mon grand père, Roland Duval, né le 13 novembre 1934 et décédé d'un cancer en janvier 2016 pendant mes études d'ingénieur à Centrale. Je ne pense pas que mes travaux de thèse vont changer quelque chose pour le traitement des malades du cancer, mais j'ai essayé de faire le mieux que je pouvais.

Maintenant, une nouvelle aventure commence et elle promet beaucoup !

PS : Bravo à celui ou celle qui est arrivé jusqu'à la fin de ce manuscrit de thèse, j'espère que cette lecture fut agréable ou au moins intéressante.

Théo Estienne, Décembre 2021, Gentilly

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Méthodes d'apprentissage profond pour le recalage 3D d'images médicales

Mots clés: Imagerie Médicale, Recalage d'Images, Apprentissage Profond, Réseau de Neurones à Convolution, Radiothérapie

Résumé: Cette thèse se concentre sur des nouvelles approches d'apprentissage profond (aussi appelé deep learning) pour trouver le meilleur déplacement entre deux images médicales différentes. Ce domaine de recherche, appelé recalage d'images, a de nombreuses applications dans la prise en charge clinique, notamment la fusion de différents types d'imagerie ou le suivi temporel d'un patient. Ce domaine est étudié depuis de nombreuses années avec diverses méthodes, telles que les méthodes basées sur des difféomorphismes, sur des graphes ou sur des équations physiques. Récemment, des méthodes basées sur l'apprentissage profond ont été proposées en utilisant des réseaux de neurones convolutifs.

Les méthodes utilisant l'apprentissage profond ont obtenu des résultats similaires aux méthodes classiques tout en réduisant considérablement le temps de calcul et en permettant une prédiction en temps réel. Cette amélioration provient de l'utilisation de processeurs graphiques (GPU) et d'une phase de prédiction où aucune optimisation n'est requise. Cependant, les méthodes utilisant l'apprentissage profond ont plusieurs limites, telles que le besoin de grandes bases de données pour entraîner le réseau ou le choix des bons hyperparamètres pour éviter des transformations trop irrégulières.

Dans ce manuscrit, nous proposons diverses modifications apportées aux algorithmes de recalage à l'aide de deep learning, en travaillant sur différentes types d'imagerie et de parties du corps. Nous étudions dans un premier temps la combinaison des tâches de segmentation et de recalage proposant une nouvelle architecture conjointe. Nous nous appliquons à des jeux de données d'IRM cérébrales, en explorant différents cas : des cerveaux sans et avec tumeurs. Notre architecture comprend un encodeur et deux décodeurs et le couplage est renforcé par l'introduction d'une fonction de coût supplémentaire. Dans le cas de la présence d'une tumeur, la fonction de similarité est modifiée tel que l'entraînement se concentre uniquement sur la partie saine du cerveau, ignorant ainsi la tumeur. Ensuite, nous passons au scanner abdominal, une localisation plus difficile, à cause des mouvements et des déformations naturelles des organes. Nous améliorons les performances d'apprentissage grâce à l'utilisation de pré-apprentissage et de pseudo segmentations, l'ajout de nouvelles fonction de coût pour permettre une meilleure régularisation et une stratégie multi-étapes. Enfin, nous analysons l'explicabilité des réseaux d'enregistrement en utilisant une décomposition linéaire et en s'appliquant à l'IRM pulmonaire et l'hippocampe cérébrale. Grâce à notre stratégie de fusion tardive, nous projetons des images dans l'espace latent et calculons une nouvelle base. Cette base correspond à la transformation élémentaire que nous étudions qualitativement.

Title: Deep learning-based methods for 3D medical image registration

Keywords: Medical Imaging, Image registration, Deep Learning, Convolutional Neural Networks, Radiotherapy

Abstract: This thesis focuses on new deep learning approaches to find the best displacement between two different medical images. This research area, called image registration, have many applications in the clinical pipeline, including the fusion of different imaging types or the temporal follow-up of a patient. This field is studied for many years with various methods, such as diffeomorphic, graph-based or physical-based methods. Recently, deep learning-based methods were proposed using convolutional neural networks.

These methods obtained similar results to non-deep learning methods while greatly reducing the computation time and enabling real-time prediction. This improvement comes from the use of graphics processing units (GPU) and a prediction phase where no optimisation is required. However, deep learning-based registration has several limitations, such as the need for large databases to train the network or tuning regularisation hyperparameters to prevent too noisy transformations.

In this manuscript, we investigate diverse modifications to deep learning algorithms, working on various imaging types and body parts. We study first the combination of segmentation and registration tasks proposing a new joint architecture. We apply to brain MRI datasets, exploring different cases : brain without and with tumours. Our architecture comprises one encoder and two decoders and the coupling is reinforced by the introduction of a supplementary loss. In the presence of tumour, the similarity loss is modified such as the registration focus only on healthy part ignoring the tumour. Then, we shift to abdominal CT, a more challenging localisation, as there are natural organ's movement and deformation. We improve registration performances thanks to the use of pre-training and pseudo segmentations, the addition of new losses to provide a better regularisation and a multi-steps strategy. Finally, we analyse the explainability of registration networks using a linear decomposition and applying to lung and hippocampus MR. Thanks to our late fusion strategy, we project images to the latent space and calculate a new basis. This basis correspond to elementary transformation witch we study qualitatively.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France