EVALUATING VOXELMORPH, A LEARNING-BASED 3D NON-LINEAR REGISTRATION ALGORITHM, AGAINST THE NON-LINEAR SYMMETRIC NORMALIZATION TECHNIQUE FROM ANTS

Victoria Madge

Biological and Biomedical Engineering Program

McGill University

December 2020

A Thesis submitted to the Faculty of Graduate Studies and Research in partial fulfilment of the requirements for the degree of Master of Engineering

ABSTRACT

Medical image registration is the process of aligning two images of the same scene into the same image space and is a fundamental step in many image processing applications. For the registration of brain images between subjects, non-linear diffeomorphic registration is favoured since such techniques are capable of compensating for tissue deformation while maintaining brain topology. Recently, deep learning has shown success in a wide variety of medical image-analysis tasks, including image registration. VoxelMorph is a deep learning-based non-linear technique promising fast diffeomorphic registrations and claiming comparable results to Symmetric Normalization from Advanced Normalization Tools (ANTs SyN). However, the comparison between the two methods was based solely on Dice scores of automatically segmented labels. Using automatic segmentations could muddle results, and using Dice scores, an indirect evaluation measure, is an incomplete evaluation of goodness of fit. Additionally, the smoothness parameters of the ANTs SyN algorithm were altered to be more similar to those of VoxelMorph, thus restraining ANTs SyN's capacity to achieve a successful registration. This thesis presents an evaluation of VoxelMorph against the native, unaltered ANTs SyN, offering comparisons with direct and indirect evaluation metrics using data with manual gold standard segmentation labels. This evaluation was performed in experiments with three databases: a database of simulated deformations of the VoxelMorph atlas, BrainWeb20, and Neuromorphometrics. Results from the first experiment show ANTs SyN outperforms VoxelMorph in the presence of simulated deformation. Results from the second and third experiment show VoxelMorph produces inter-subject registration results comparable to those of ANTs SyN.

ABRÉGÉ

Le recalage d'images médicales est le processus d'alignement de deux images de la même scène dans le même espace d'image et est considéré comme une étape fondamentale dans de nombreuses applications de traitement d'image. Pour effectuer un recalage entre deux images cérébrales, le recalage difféomorphique non-linéaire est favorisé car il est capable de déformer le tissu cérébral tout en conservant sa topologie. Récemment, l'apprentissage en profondeur a connu du succès en plusieurs tâches d'analyse d'images médicales, y compris le recalage d'images. VoxelMorph est une technique non-linéaire basée sur l'apprentissage promettant des recalages difféomorphes rapides et revendiquant des résultats comparables à la normalisation symétrique des outils de normalisation avancés (ANTs SyN). Cependant, la comparaison entre les deux méthodes était basée uniquement sur les scores de Dice des étiquettes automatiquement segmentées. L'utilisation de ces étiquettes pourrait brouiller les résultats, et l'utilisation des scores de Dice, une mesure d'évaluation indirecte, est une évaluation incomplète de la qualité de l'ajustement. De plus, les paramètres de régularité de l'algorithme ANTs SyN ont été modifiés pour être plus similaires à ceux de VoxelMorph, limitant ainsi la capacité de ANTs SyN. Cette thèse vise à évaluer de manière complète VoxelMorph par rapport au ANTs SyN non modifié, offrant des comparaisons avec des mesures d'évaluation directe et indirecte à l'aide de données avec des segmentations manuelles. Cette évaluation sera réalisée en trois expériences avec trois bases de données: des déformations simulées, BrainWeb20 et Neuromorphometrics. Les résultats de la première expérience montrent que ANTs SyN surpasse VoxelMorph en présence de déformation simulée. Les résultats de la deuxième et troisième expérience montrent que VoxelMorph peut effectuer un recalage inter-sujet comparable à celui des ANTs SyN.

ACKNOWLEDGEMENTS

First of all, this thesis would not be possible without the guidance and support of my supervisor, Dr. D. Louis Collins. I would like to thank him for giving me the opportunity to work in his lab, providing me with great connections and resources, and for being exceptionally understanding when things became difficult. His compassion and patience have helped me thrive and persevere. I am honoured and proud to call myself his student and could not have asked for a better supervisor.

I would also like to thank all members of the lab, but especially Dr. Vladimir Fonov, Dr. Mahsa Dadar, and Dr. Phil Novosad for their generous advice and help with all things image processing. From small peculiarities with the Minc Toolkit or system environments, to suggestions for registration parameters and image deformation approaches, I could not have made it to the finish line without you.

I feel very fortunate to have loving family and friends who have been my personal cheerleaders throughout this entire process. I would especially like to thank my sister Bridget Robertson, and my friends Daniel A. DiGiovanni, Fares El Tin, Jai Hebel, and Myrna Megalla. You all have been so supportive and encouraging through the frustration, exhaustion, and many tears; but also, elation, success, and lots of laughter. I am indebted to you all.

Lastly, I would like to thank myself. This was a gruelling and arduous journey, and not at all like I had originally planned. The biggest obstacle I had to overcome was my own expectations. Though I am not sure I was completely successful in my struggle, I can at least find peace knowing that I have accomplished something big. I did it.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Motivation

In the scientific community, publishing novel work has many incentives. But with the growing number of publications, it is equally important to see the validation of such work in the literature [1]. Reproducibility in science is imperative to give credibility where it is due through replication and validation of another's work. Validation is essential to innovation: to evaluate progress, to improve efficiency, and to establish fairness in publications [2]. The Open Science Collaboration assessed the reproducibility of 100 experimental and correlational studies in physiological science by observing p-values, effect sizes, and other measures in replications of the original studies [1]. Of their many findings, only 35 of the replicated studies were statistically significant ($p < 0.05$), while 97 of the original studies claimed statistical significance. Analyses such as the one conducted by the Open Science Collaboration suggest the need to improve reproducibility in psychology, and this can be extended to all areas of science.

One field of science where collaboration among institutions has prospered, thanks to its infrastructure, is in the medical image-analysis community [3]. However, there is less success in software sharing and reproducibility among colleagues. It is important, especially in a field that directly affects the health of the public, to make such studies, software and data included, open source, in order to facilitate reproducibility, to ultimately accelerate innovation [2].

Recently, deep learning (DL) has shown success in a wide variety of medical image-analysis tasks, including image registration, but requires repeatability in their methodology and experimentation to consider the techniques credible and comparable against the current state of the art [4]. State-of-the-art automatic non-learning-based registration techniques have long been the solution for aligning medical images, improving over manual alignment strategies which heavily depend on user expertise and are subject to inter- and intra-rater variability. Some non-linear diffeomorphic techniques, referred to as "classic" techniques in this thesis, are considered the benchmark in the registration of brain images, and will be discussed in-depth alongside DL-based registration techniques in the Background Chapter of this thesis.

Among the newly published DL-based registration techniques is VoxelMorph [5], a non-linear diffeomorphic registration algorithm promising fast, topology-preserving registration of brain images. VoxelMorph has been claimed to have registration results comparable to the Symmetric image Normalization method by Advanced Normalization Tools (ANTs SyN) [6]; however, the evaluation performed by Dalca *et al.* between VoxelMorph and ANTs SyN could be improved. First, the training and testing of the VoxelMorph model was performed using fully automatic FreeSurfer-based [7] segmentations of magnetic resonance imaging (MRI) volumes [5], and not manual gold-standard segmentations [8]. Thus, errors in automatic segmentation confound registration quality metrics. Second, the smoothness parameters of the ANTs SyN algorithm were altered to be more similar to those of VoxelMorph [5], thus restraining ANTs SyN's capacity to achieve a successful registration. Finally, Dalca *et al.* based their comparison solely on Dice scores, an indirect measure, which alone is not a complete evaluation of goodness-of-fit [8].

The objective of this thesis is to compare the registration technique of VoxelMorph to ANTs SyN. It will be important to maintain the comparison of VoxelMorph, a newly published technique,

to ANTs SyN, published in 2009, as this is how Dalca *et al.* had performed their comparison, and this thesis aims to validate their results. Validating the results from Dalca *et al.* will be performed in three experiments; first, using simulated deformations of the target atlas from the VoxelMorph paper [5]; second, using a digital phantom database with manual gold standard labels – BrainWeb20 [9]; and third, using a clinical database with manual gold standard segmentations from Neuromorphometrics [10]. The methods will be compared using direct metrics, such as recovery error, as well as indirect evaluation metrics, such as Dice score, Cohen's Kappa and Hausdorff distance. Both whole brain and specific brain tissue and structure labels will be used for measuring direct and indirect metrics. Indirect whole brain metrics will use the EvaluateSegmentation tool [11] for the analysis, and the other metrics mentioned will rely on the Minc Toolkit [12]. It is hypothesized that the validation methods used to compare the registration results from VoxelMorph and ANTs SyN will be sufficient to discern whether one approach outperforms another.

1.2 Thesis Overview

This thesis is organized into four chapters. Chapter 2 presents an overview of medical image registration, including definitions of terms and techniques. Chapter 2 also includes a literature review of non-linear diffeomorphic registration "classic" techniques and non-linear DL-based techniques for brain image registration. Chapter 3 discusses the methods and results from the three experiments in a publication which has been submitted to *Medical Image Analysis*. Chapter 4 includes a discussion of the results, possible future work related to the thesis, and a conclusion.

1.3 Contributions

The contributions from this thesis are the following:

i. Simulated realistic deformations, using deformation warps obtained from the registration technique ANIMAL [13], to evaluate VoxelMorph's effectiveness in recovering deformed images and to measure the recovery error in comparison to ANTs SyN.

ii. Compared the performance of VoxelMorph to ANTs SyN with indirect evaluation metrics such as Dice score, Cohen's Kappa, and Hausdorff distance, using whole brain, grey matter, white matter, and cerebrospinal fluid labels from a digital phantom database – BrainWeb20 [14].

iii. Compared the performance of VoxelMorph to ANTs SyN with indirect evaluation metrics such as Dice score, Cohen's Kappa, and Hausdorff distance, using cortical and deep brain structures from the Neuromorphometrics database [10] to observe their movement in an atlas-based inter-subject registration task.

CHAPTER 2

Background

This chapter presents a comprehensive literature review of both “classic” non-linear diffeomorphic registration methods and deep learning (DL) based non-linear registration methods. In this thesis, “classic” methods refer to any non-linear diffeomorphic registration method that is not based on DL. But first, in order to understand both classic and DL-based methods, an overview of image registration is required. The overview will be discussed in terms of the importance and applications of registration, followed by a definition of image registration.

The background of this thesis was organized with the intent to first motivate the reader with applications of registration to highlight its importance. Once motivation is set, the reader can understand image registration in terms of the process and different variations of registration. If the reader so chooses, they can circle back to the motivation to further set the importance of image registration. Predominantly, the knowledge flow in the background section follows the same structure as review papers by Oliveira *et al.* [15] as well as Zitová *et al.* [16], with some references to Maintz *et al.* [17] and Crum *et al.* [18], although Maintz *et al.* followed a different overall knowledge flow than the former authors. For a more detailed understanding of image registration, please refer to these publications directly.

2.1 The Importance of Image Registration and Applications

Medical image registration is a fundamental step in many image processing applications [15]. Particularly, image registration plays important roles in the application of image segmentation and atlas formation, diagnostics, treatment, and monitoring of various pathologies, as well as image-guided surgical planning, surgical simulation, and intra-operative brain shift correction. To highlight the significance of image registration, some applications will be discussed.

2.1.1 Applications in Atlas Formation and Image Segmentation

Firstly, for the purpose of atlas formation – where an atlas is an anatomical template which summarizes structure and volume and is a useful tool when performing group analyses in image processing [19] – and anatomy segmentation, image registration is particularly essential. Leow *et al.* used registration to create three-dimensional (3D) atlases of intra-subject (i.e. within-subject) brain changes across different magnetic resonance (MR) scanning sequences as part of the Alzheimer's Disease Neuroimaging Initiative (ADNI) [20]. The utilization of registration here was crucial in the determination of stable scanning techniques with the least amount of intra-patient variability, ensuring quality data for public use [21]. Gooya *et al.* performed an atlas-based registration (i.e. registering images to a target atlas) of brain images with gliomas [22]. By registering with two different atlases, one of a normal (i.e. healthy) brain and one created from a model for tumour growth, Gooya *et al.* produced better tumour segmentations, a challenge in and of itself, compared to related works. So, both atlas formation and segmentation were feasible here due to registration. Another atlas-segmentation combination is observed with Automatic Nonlinear Image Matching and Anatomical Labeling (ANIMAL) [13], a segmentation-registration technique by Collins *et al.* who used a registration algorithm to register MR images to a labelled atlas in

order to anatomically segment the brain [13]. The brain is objectively segmented here by applying the inverse registration transformation to the atlas labels to segment the brain in its native space (i.e. patient acquisition coordinates).

2.1.2 Applications in Disease Diagnosis, Monitoring and Treatment

In disease diagnosis, treatment, and therapy, registration adds a level of precision and accuracy. This is seen in work by Staring *et al.*, who developed a registration technique, which uses both the intensity and local structure data within the image, to register T2-weighted cervical MR images of the same patient to previous scans for radiation therapy [23]. The registration technique enabled precise dose targeting to the necessary anatomy, while avoiding critical organs. Lavely *et al.* also demonstrated the accuracy of two registration techniques, one automatic and the other semi-automatic, in the application of radiotherapy treatment [24]. Both techniques registered positron emission tomography (PET) and computed tomography (CT) images and were validated using a brain phantom[1] as well as patient data. Foskey *et al.* improved accuracy in image-guided radiation therapy for prostate cancer using registration to compensate for regions of miscorrespondence due to organ motion in CT images [25]. The registration method facilitates accurate organ segmentation and image-guided radiation therapy, targeting specific organs while limiting dosage to other organs. More recently in monitoring and treatment, van der Hoorn *et al.* developed a two-stage semi-automatic registration technique to register MR scans from different time points to better monitor glioma patients before, during, and after treatment [26]. Demonstrating improved accuracy with disease diagnosis, Huang *et al.* developed a two-step registration technique to align real time 3D ultrasound images to dynamic MR and CT images for

[1] A brain phantom is an object that typically simulates the tissue and imaging behaviour of a real brain [105].

cardiac diagnosis, and as well as for surgical navigation, to improve the quality of cardiac procedures [27].

2.1.3 Applications in Surgical Simulation, Planning, and Image-Guided Surgery

Image registration can also be applied in many surgical simulation and surgical planning applications. In cardiac and orthopedic surgery, King *et al.* developed a technique to register real-time ultrasound images to their pre-operative MR images during minimally invasive cardiac procedures [28]. The technique increases the inherent signal-to-noise ratio in the ultrasound acquisition and corrects for bulk motion, such as respiratory motion, during cardiac surgery. Hurvitz *et al.* developed an intraoperative atlas-based registration technique for bone-surface reconstruction of X-ray images of the femur using a pre-operative CT intensity atlas to reduce miscorrespondences between images [29]. This application is important for image fusion of intraoperative images to preoperative CT scans during orthopedic surgery.

In efforts to improve neurosurgical planning, Nandish *et al.* used image registration and image fusion of CT and MR images to help localize brain lesions and determine skull incisions pre-surgery [30]. For image-guided neurosurgery, the registration technique by Maurer *et al.* using implantable fiducial markers was considered critical work for its time in 1997 [15][31]. Their technique uses the implantable fiducial markers as a set of points to register one image to the other, in this case, CT to MR images [32]. More recently, Drouin *et al.* have implemented an intraoperative ultrasound-MR registration technique, within their neuronavigation system – Intraoperative Brain Imaging System (IBIS), to correct for brain shift [33]. The technique uses image landmarks from preoperative MR data, instead of invasive implantable fiducial markers, to update the preoperative MR images during surgery to correct for brain shift using real-time

intraoperative ultrasound images. Use of this technique within IBIS improves patient-to-image mapping and increases the use of neuronavigation during surgery.

From the many applications discussed, the importance of image registration is evident; however, an understanding of image registration must be established before delving into related work in classic registration and DL-based non-linear registration.

2.2 Defining Image Registration

Medical image registration is the process of aligning two images of the same scene or anatomy into the same image space [15]. This concise definition of registration requires a lot of unpacking to understand (1) what is the process, (2) how is the image pair related and how do they differ, and (3) what is an image space. These concepts will be defined in the context of brain image registration; however, examples of image registration using other anatomical scenes may be used for illustrating specific terms. The concepts will be defined in reverse order.

2.2.1 Image Space: Coordinate Systems, Standardized Spaces, and Atlases

Here, the term image space refers to a system of coordinates unique to a specific image [31]. To illustrate, if one image, herein referred to as the "moving image", is registered to another image, herein referred to as the "fixed image", the features of the moving image (i.e. structures, boundaries, intensity values, etc.) would share the same position (i.e. coordinates) as the corresponding features from the fixed image within the coordinate system of the fixed image.

The coordinate system can vary from the native coordinate system of each image (i.e. patient image acquisition coordinates determined by the scanner geometry), to a standardized brain-based coordinate system such as the Talairach space [34] or the MNI space, which has evolved over the

years. The first iteration of the MNI space – the MNI305 – was proposed by Collins *et al.* [35]. This brain-based standardized space, based on the Talairach stereotaxic coordinate system, aligns images for easier localization of anatomical points of interest in the brain and facilitates voxel-to-voxel comparison between subjects [36]. The coordinate system defines its origin on three axes: one axis passes through the superior aspect of the anterior commissure (AC) and the inferior edge of the posterior commissure (PC), the second axis passes through the midline plane, perpendicular to the AC-PC axis and the third, perpendicular to the first two, in the left-right direction [35]. This thesis will use the latest iteration of the MNI space in its methodology – the International Consortium for Brain Mapping atlas, or ICBM152 [37], as well as other niche image spaces such as the VoxelMorph atlas space. The ICBM152 atlas is a brain template created from approximately 150 images from a normative young adult population and was collected from three different scanning locations [36]. This atlas provides improved contrast form the top of the brain to the bottom of the cerebellum and has been integrated into many brain mapping software.

2.2.2 Image Pair: Intra- and Inter-Subject Registration

Returning to the definition, registration is the process of alignment between an image pair, of which the pair of images to be registered can differ in many ways. The images can be from the same subject – intra-subject registration – or different subjects – inter-subject registration [16]. For example, Woods *et al.* demonstrates the performance of their automated intra-subject registration technique on the same patient with eight different PET scans [38]. The technique uses voxel-to-voxel intensity ratios and aligns each slice in order to minimize variance across all voxel ratios. An example of inter-subject registration is the work from Collins *et al.*, which registers MR images from different subjects to a standardized space [35]. Inter-subject registration poses a particular problem due to the dissimilarities in brain sizes and shapes, or inter-subject variability, and if

acquired from different imaging sites, image parameters such as slice thickness, pixel size, and resolution can also pose difficulties in performing the task [15][39].

2.2.3 Image Pair: Mono- and Multi-Modal Registration

The image pair can also differ in terms of the sensor used to perform each image acquisition. The image pair can be from the same image acquisition type – called mono-modal registration – or from different acquisition types – multi-modal registration. The previously mentioned method by Woods *et al.* – later to be developed as the technique AIR [38]– is an example of monomodal registration, but has been extended to multi-modal intra-subject registration as well. The method differs slightly, however, as corresponding voxels between MR and PET differ in intensity [38]. To accommodate the extra imaging modality, the method utilizes manual skull-stripping to remove all non-brain structures for easier correspondence, and then partitions MR pixels into tissue type to best match to the same tissue type found in the PET image using weighted averages of the normalized standard deviations of the MR pixel values. This is just one of the many ways that multi-modal image registration can be performed.

2.2.4 The Registration Process: Transformation Model

The process of image registration itself varies greatly, but can be fundamentally broken down into three main components: (1) determining the transformation model to register a moving image to a fixed image, (2) determining a measure of similarity between each image of the image pair, and (3) determining the search strategy or optimization technique to obtain the transform that maximizes similarity between the image pair [15] [16] [17] [40] [41]. Each of these components can be interpreted in many different ways and applied using many different techniques.

When discussing the first component of the registration process, a transformation between an image pair refers to a mapping of the voxels from the moving image to the fixed image space [16]. The transformation can be local or global, that is, it can be used to register an entire image or only a region of an image [15]. Transformations can be rigid or non-rigid. Rigid registration has transformations which only involve rotation or translation of the moving image to the fixed image space. For example, if you have a 3D moving image to be registered to a 3D fixed image, there would be a total of six degrees of freedom, or parameters, accounted for in registering the moving image: three parameters for rotation in the x, y, and z directions, and three parameters for translation in the x, y, and z directions. Non-rigid transformations encompass many different transformation types, including affine, projective, and curved.

Affine transformations include twelve parameters, and account for translation, rotation, scaling, and shearing in the x, y, and z directions of a 3D image. Affine transformations are categorized as linear transformations, along with rigid transformations [40]. If one can think of an affine transformation as a mapping of parallel straight lines onto parallel straight lines, projective transformations could be defined as mapping straight lines onto straight lines, but here parallel straight lines are not constrained to remain parallel [17]. Curved transformations are considered a higher order transformation which map straight lines onto curved lines [17] [41]. They are also commonly referred to as deformable, elastic, or fluid transformations and are categorized as non-linear registration in the literature [15] [16] [18].

Choosing the appropriate transformation type depends on which anatomical structures are being registered, as well as the computation power or time available to perform the task, since non-rigid registration consumes more computational resources than most rigid registration techniques [42]. An example where rigid transformation is preferred is when registering rigid anatomical

structures, such as bones. For example, Rasoulian *et al.* used a rigid registration method to individually register CT images of vertebrae to intraoperative ultrasound images for spine registration during surgery [43]. The use of rigid registration on individual vertebrae and amalgamating the information for the whole spine facilitates updated patient-to-image mapping during spine surgery and other image-guided spinal procedures, where there exist differences in patient position during scanning versus during the procedure. Besides the spine and other bones, most parts of the body are considered fluid, or viscoelastic, and thus require non-rigid forms of registration in order to compensate for tissue deformation [31] [40]. While rigid transformations are usually global transformations, non-rigid transformations can be local or global. Curved transformations are typically favoured in most registration task settings, since they are capable of mapping viscoelastic tissues while preserving viscoelastic tissue properties [15].

In general, there are two types of curved transformations: free-form transformations (usually referred to as free-form deformations or FFDs) or guided transformations [15]. FFDs allow any deformation, while guided transformations rely on models which take into account material properties of the tissue to control the deformations to the moving image. Guided transformations include flow-based models, such as fluid flow and optical flow. [15] [40].

Flow-based transformations model the movement between the moving and fixed images over time, and can be divided into two categories: fluid flow and optical flow [15] [16]. Fluid flow-based models model the fluid movement of one image to another, while optical flow uses intensity values in the images as a similarity function to drive the movement of one image to another.

FFDs solve for the mapping of the moving image to the fixed image space using a grid of control points [15]. These control points are capable of moving individually according to a given

similarity measure. These points within the transformation are interpolated from the moving to the fixed image using radial basis functions such as thin plate splines (TPS) or B-splines. Radial basis functions are capable of local geometric distortions and are thus beneficial in non-linear registration [16].

Regardless of the transformation model, be it rigid, affine, or FFD or flow-based, resampling methods or interpolators are required to estimate the similarity function. Common interpolating functions or resampling functions include cubic B-splines, TPS, tri-linear interpolation, and the nearest neighbour function, among others [16]. Resampling or interpolation can work in a forward direction, or a backward direction. A forward approach takes the voxels from the moving image and maps them according to the transformation model in the fixed image space [16]. This approach is risky, however, since it may produce holes or overlaps of voxels in the moved image (where the moved image is the moving image registered in the fixed image space). A backward approach takes into account the voxel coordinates of the fixed image and uses the inverse of the transformation to identify the location of the voxels from the moved image in the fixed image space [16].

Another important aspect of the transformation model is its invertibility and ability to preserve topology. One way to guarantee both is to use diffeomorphic transformation models. A diffeomorphic transformation is an invertible function that maps one manifold to another such that both the forward and inverse transformations are smooth and differentiable [44]. A diffeomorphic transformation, in theory, preserves topology. To illustrate, if one were given a task to physically align two real objects, and the only way to successfully complete this task is to tear or fold one of the objects to match the other, this would be a case where topology was not preserved [18]. The same can be applied to the registration of brain images. It is desirable to maintain image topology

(in terms of connectedness and boundaries) of both images during the registration process; hence, diffeomorphic registration techniques are preferred. Diffeomorphic registration techniques have invertible transformations, which implies that the transforms have a one-to-one mapping of the moving image voxels to the moved image voxels, and also preserve topology [15] [44]. Invertible transforms are transforms capable of "undoing a registration"; that is, reversing the direction of the transformation such that the moved image reverts to the original moving image. For inter-subject registration, topological differences between subjects may produce unrealistic transformations; for example in non-diffeomorphic registration methods, folding or tearing is possible when there is no correspondence in a given image location – like when the region of one gyrus of one subject is represented by two gyri in another [13] [44]. Diffeomorphic transformations are especially important here to ensure a continuous differential transformation which preserves topology. For reference in this thesis, both VoxelMorph and ANTs SyN claim diffeomorphic registrations. While in theory, diffeomorphic transformations are possible, in practice implications during implementation may not always results in diffeomorphic transformations.

2.2.5 The Registration Process: Similarity Measure

The second component of the registration process is determining a similarity measure between the two images. A similarity measure between an image pair is defined as a characteristic, metric, function, or quality that describes this similarity or dissimilarity of two images [15] [45]. Ideally, similarity measures would be capable of associating two images to one another down to the voxel level, to create a one-to-one mapping.

Many review papers on image registration classify similarity measures into two categories: intensity-based (or area-based) similarity measures, and feature-based (or geometric-based)

similarity measures [15] [16] [18] [41] [45]. A popular review paper on medical image registration by Maintz *et al.* classifies similarity measures differently, and includes similarity measures that are not image-based [17] such as physically invasive markers implanted or fixed onto the skull of the patient, like fiducial markers or a stereotactic frame. For the purpose of this Background Chapter, only image-based similarity measures are considered.

Intensity-based similarity measures are commonly computed in the regions which overlap between the fixed and moving images [15] [45]. They include comparison of raw intensity values of the voxels, intensity gradients, and statistical or mathematical criteria based on voxel intensity [15] [16] [17] [18] [45]. Statistical and mathematical criteria based on voxel intensity include measures such as cross-correlation, sum of squared differences (SSD), information theory, and Fourier domain-based measures [15] [16] [17] [18] [45].

Some similarity measures are more popular in particular registration tasks; for example, most cross-correlation and SSD similarity metrics are suitable for mono-modal registration tasks, but fail in most multi-modal registration tasks since the same tissue type in the fixed image may have very different intensity values in the moving image [18]. The SSD similarity metric assumes the fixed and moving images have identical image intensity properties apart from some Gaussian noise, and correlation-type similarity measures assume a linear relationship between the intensities in the fixed image and the intensities in the moving image, hence making it unsuitable for multi-modal registration tasks. However, correlation ratio and mutual information similarity measures work well for multi-modal registration tasks [18]. Generally, a correlation ratio similarity metric utilizes a ratio of the intensities from the same tissue type in both the moving and fixed images to guide the registration. Mutual information is an information theory-type similarity metric which typically uses Shannon entropy, computed using the joint probability distribution of the voxel

intensities in the moving and fixed images. A potential disadvantage of intensity-based similarity metrics used in registration tasks is that the anatomical data are not considered, that is, the structural or boundary information within the image [18].

Feature-based (or geometric-based) similarity metrics use explicit structural information from moving and fixed images [18]. Some feature-based similarity metrics require user input to define structural or anatomical regions to use for registration, but this is not always the case [17] [18]. Structural or anatomical features used as similarity measures can vary from points to lines to surfaces. Feature-based similarity metrics are generally used in rigid or affine registration tasks [18] [41], but can be used in some non-linear registration tasks as well [46]. Sometimes feature-based similarity metrics replace manual identification of structural landmarks with automatic segmentations to automatically (or semi-automatically) drive the registration, however these methods are limited by the quality of the segmentation [17]. Once anatomical or structural features are pulled from the image, distance measures, such as Euclidean distance, Mahalanobis distance, chamfer distance, and an SSD method called sum of squared distances, are used in the cost function to drive the registration [15] [17] [45]. Typically the features used are spread throughout the image so as to not bias the image by an alignment in one particular area [18]. Obviously, a disadvantage of using a handful of points, curves, or surfaces to drive the registration is that much of the image information is unused, unlike with intensity-based similarity measures where the whole volume can play a role in the registration. Some registration methods opt for a combination of intensity-based and feature-based similarity measures to improve registration accuracy [47] [48] [49].

The similarity measure, once selected, whether intensity-based or feature-based, is usually placed in a cost function which takes into account the parameters of the transformation (i.e. rigid, affine, FFD, flow-based, etc.) and, depending on the transformation type, the cost function can

also include smoothness constraints (i.e. for non-linear transformations: FFDs and flow-based) [15] [16] [17] [18] [41] [45]. Common smoothness constraints, or regularization terms, include bending energy or second order derivatives of the transformation model, and the Jacobian of the transformation model [15] [17] [44]. The parameters of the transformation are found using an optimization strategy as described in the next section.

2.2.6 The Registration Process: Optimization Strategy

The third component of the registration process is the optimization strategy. Optimization here refers to the way in which the best transformation between an image pair is found such that the similarity measure between the images is maximized [18]. Choosing an optimization strategy depends on the transformation type, the similarity measure, the cost function, and the desired accuracy of the registration.

There are typically two types of optimization strategies (besides brute force search of course) which are categorized into continuous and discrete methods [40] [44]. Continuous optimization strategies involve continuous variables, and require the cost function to be differentiable [40] [44]. Examples of continuous optimization techniques include gradient descent, conjugate gradient descent, Powell's optimization, downhill simplex method, Quazi-Newton optimization, Levenberg-Marquardt optimization, Newton-Raphson iteration, and stochastic gradient descent [15] [17] [40]. Discrete optimization strategies work the same way in that the strategy attempts to minimize the cost function while maximizing the image similarity, but does so using a set of discrete values [40]. Discrete optimization does have the advantage of being more efficient than continuous optimization but may not find the most optimal solution given its restrictions using a discrete set of parameters. Examples of discrete optimization methods include graph-based

methods, Markov Random Field, message passing methods, and linear-programming methods [40] [44].

Constraints in a cost function become very important especially when using an optimization strategy. The sole purpose of an optimization strategy is to find the transformation parameters that maximize the similarity between an image pair. So an optimization strategy may appear to do a good job because the similarity between a pair of images is near perfect, but this may not preserve topology in the case of non-linear transformations [18]. This is why it is important to have a transformation that is diffeomorphic, and a cost function (including the similarity metric) which constrains optimization.

Now that an understanding of registration and its various processes has been established, related work to the thesis must be discussed. To repeat, the purpose of this thesis is to compare a DL-based non-linear diffeomorphic registration technique against the classic technique ANTs SyN. Therefore, a review of classic registration techniques is required, to gain knowledge of what is considered the benchmark in brain image registration, and then later introduce emerging DL-based non-linear registration techniques.

2.3 Related Work in Non-Linear Diffeomorphic Registration

The eight classic techniques that will be discussed in this section are the top performing classic registration techniques which also promise diffeomorphic transformations, several of which are evaluated in the most recent evaluation of non-linear brain registration techniques by Klein *et al.* [50], while others are mentioned in several review papers [15] [16] [44]. The techniques are listed in chronological order of publication, and their methods are discussed in terms of the transformation model (and how the transformation is guaranteed to be diffeomorphic), the

similarity metric, and the optimization strategy employed, as well as how the method was evaluated and applied to a registration task. The list of classic methods comprises the following: Small-Deformation Inverse-Consistent Linear-Elastic Image Registration (SICLE) algorithm published in 1999 [46], Large deformation diffeomorphic metric mapping (LDDMM) published in 2004 [51], Jensen-Rényi Divergence (JRD) Fluid registration method published in 2006 [52], a diffeomorphic FFD algorithm published in 2006 [53], Diffeomorphic Anatomical Registration using Exponentiated Lie algebra (DARTEL) Toolbox published in 2007 [54], Diffeomorphic Demons published in 2007 [55], Advanced Normalization Tools (ANTs) Symmetric image Normalization method (SyN) published in 2008 [6], and Deformetrica published in 2014 [56].

2.3.1 SICLE

Small-deformation Inverse-Consistent Linear-Elastic image Registration, or SICLE, is a non-linear image registration algorithm developed by Johnson *et al.* [46] which builds on the very first non-linear diffeomorphic registration method by Christensen *et al.* [57]. The transformation model is an FFD which utilizes a thin-plate spline (TPS) interpolator, a radial basis function that is based on the energy needed to deform a thin metal plate. SICLE uses both intensity-based and feature-based measures as its similarity metric, making this a two-step process. The first step uses non-rigid landmark registration for large deformations to essentially align boundaries. Corresponding landmarks are chosen from both datasets in the image pair, and the transformation is optimized by minimizing the bending energy. The second step utilizes intensity-based registration to make small deformations, while maintaining feature-based correspondences from the first step. Intensity and landmark-based registration steps are alternated until an adequate solution is found. To achieve a diffeomorphic transformation, the forward transformation is averaged with the inverse of the reverse transformation, and vice versa. This claims to reduce the

effects of large inverse consistency errors, as forward and inverse transformations have more correspondence between one another. However, an evaluation of topology preservation was not mentioned, so it is questionable whether true diffeomorphic transformations occur for every image pair to be registered. The SICLE algorithm was applied to MRI data to demonstrate that using landmark and intensity-based similarity measures together achieve better results than using either measure alone.

2.3.2 LDDMM

Large Deformation Diffeomorphic Metric Mapping, or LDDMM, is a non-linear registration technique developed by Beg *et al.* [51]. The registration technique is based on derivations from the Euler-Lagrange equations proposed for the "image-matching problem" from two other publications [58] [59]. Euler-Lagrange Equations are partial derivatives used to solve for a stationary velocity field. The technique models the transformation as a velocity field or flow field [51]. The flow field is solved for using the Euler-Lagrange Equations, and by minimizing the intensity-based squared error norm dissimilarity metric between the images. Diffeomorphic transformations are guaranteed using the Jacobian of the transformation. A gradient descent optimization scheme is implemented to solve for the transformation. The algorithm demonstrated its capabilities by calculating geodesic distances between anatomical structures in databases of Schizophrenia and Alzheimer's patients, where data had been obtained from a collaborating academic institution.

2.3.3 JRD-Fluid

A fluid registration technique using the Jensen-Rényi Divergence (JRD) measure as a similarity metric was developed by Chiang *et al.*[52]. JRD is an information-based measure, much

like mutual information, but with more degrees of freedom. The transformation type is modelled as a simplified partial differential equation flow field governed by the Navier-Stokes fluid model – a common set of equations to describe the momentum of viscous fluid [60]. The partial differential equation contains constraints to ensure that large deformations are smooth [52]. The flow field is computed using kernel convolution, or Green's function of the linear differential operator, and is accelerated using fast Fourier transforms. To ensure a diffeomorphic transformation, only deformations with a Jacobian determinant of less than 0.5 are used. Of course, positive Jacobians are associated with smooth and continuous deformations fields, but specifically choosing a cut-off value of 0.5 was done to reduce precision errors which occur with larger positive Jacobian values [60]. This method was applied to detect brain shape changes in HIV/AIDS patients compared to normal controls in order to characterize cognitive impairment [52].

2.3.4 Diffeomorphic Registration using B-Splines

The paper this method is based off of initially proposed two methods, which are both diffeomorphic, to counter the comparisons at the time which showed that diffeomorphic algorithms outperform their non-diffeomorphic counterparts [53]. Only the more successful of two methods will be discussed.

This method implements multi-level FFDs in order to gain the advantage of larger and smaller control-point spacings [53]. Larger control-point spacings ensure smooth, global transformations, while smaller control-point spacings allow for more localized transformations but are computationally expensive. The control-point spacings are constrained using a "hard constraint" approach, where the maximum displacement of a control point is given by a predetermined bound [61]. Since the final deformation is built upon several concatenated FFDs, if

each FFD is diffeomorphic, the overall transformation is diffeomorphic [53]. All FFDs in this model are regularized by B-spline interpolation to ensure smooth deformations. The similarity measure used in the cost function is normalized mutual information, and the cost function includes a regularization term to ensure smoothness. The optimal registration between two images is found when the cost function is minimized, but no optimization strategy was mentioned which would accelerate this calculation and hence it is assumed that this method performs slower to other diffeomorphic registration methods. The method was used to register 20 T1-weighted MR images from normal subjects, and the registration performance was evaluated using manual anatomical segmentations. Results showed similar accuracy compared to non-diffeomorphic techniques but had the added benefit of preserving brain topology.

2.3.5 DARTEL

Diffeomorphic Anatomical Registration using Exponentiated Lie algebra (DARTEL), is a recursive non-linear diffeomorphic image registration technique developed by John Ashburner [54]. The technique uses a flow field transformation model and models the movement from the moving image to the fixed image using differential equations. The Euler method solves the differential equations by integrating over small steps in time [62], and a scaling and squaring method is used to enable faster integration [63]. In theory, to ensure a diffeomorphic transformation, the flow field is considered a Lie algebra [54]. Lie algebra is a bracket which contains three axioms, one of which is the Jacobi identity. The Jacobi identity is advantageous here since positive Jacobian determinants imply diffeomorphic properties, as seen in previously discussed methods in this chapter [51] [52]. The voxel-wise mean-square difference between the image pair was used as the similarity metric, which is beneficial in template-based inter-subject registration tasks [54]. The cost function is comprised of the prior probability of the flow field,

which ensures a realistic transformation, and the posterior probability of the flow field. The optimization function used to minimize the cost function is the Levenberg–Marquardt algorithm, which required the first and second derivatives of the flow field, found recursively using a brute force method – a full multigrid approach. A full multigrid approach was used here because they are computationally efficient at convergence using different spatial scales. The method was tested by creating a template from 471 T1-weighted MR images, comprised of both men and women from various ages [64]. The flow fields of each subject were then taken and used to predict the age and the sex of the subject [54]. Successful prediction results indicated that the deformations were precise enough to encode the necessary shape information to predict age and sex.

2.3.6 Diffeomorphic Demons

Diffeomorphic Demons, proposed by Vercauteren *et al.* [55], is a non-rigid registration algorithm which uses Thirion's Demons algorithm [65]. Thirion's Demons algorithm uses an analogue of Maxwell's Demons, a theory in which 'Demon forces' control a membrane between two chambers of gas. The Demon forces are capable of controlling the membrane by which fast and slow particles can permeate in only one direction, respectively. The result being that one chamber has only high energy gas particles and the other chamber has only low energy gas particles [65]. This theory contradicts the second law of thermodynamics due to the decrease in entropy from the final result. Demons forces are used in image registration by assuming that the boundaries of anatomical objects within the image are like this membrane, and the particles are scattered throughout the image and bound by their membranes. In order to match a fixed and moving image, the Demons forces work to move the particles across each of the membranes using a deformable grid approach. In addition, local anatomical characteristics are accounted for such that the

membranes match. The implementation of Thirion's Demons by Vercauteren *et al.* uses this 'diffusing model' approach for non-linear registration [55].

The transformation model of Vercauteren is an optical flow model, just as with Thirion's method [55] [65]. The transformation is expressed in the same way as with Thirion, using Demons forces, but additionally using the Jacobian determinant to determine if the transformation is diffeomorphic, and by assuming the deformation grid is a velocity vector field instead of a displacement field. The similarity metric used here is the voxel wise mean squared error, an intensity-based measure [55] with a Newton optimization strategy. The algorithm was compared with other methods and was shown to have been diffeomorphic and to perform smoother inter-subject registration with BrainWeb20 images [9] [55].

2.3.7 ANTs SyN

Advanced Normalization Tools (ANTs) Symmetric image Normalization method (SyN) is a non-linear diffeomorphic registration method developed by Avants *et al.* [6]. ANTs SyN, like LDDMM [51], uses Euler-Lagrange equations to solve the stationary velocity field which represents the transformation model of the moving image. This technique uses cross-correlation as a similarity metric in a diffeomorphic as well as "inverse consistent", or symmetric normalization, case [6]. Computing the cross-correlation Euler-Lagrange equations is symmetrized through a Jacobian operator, reducing the computation time significantly. The technique guarantees diffeomorphic transformations by including invertibility constraints during the optimization. The optimization itself is done by computing the two half velocity fields for each direction of image movement: moving to fixed image and fixed to moving image. Both Euler-Lagrange equations are computed in parallel to find the maximum cross-correlation between the

two images, and then their inverses are computed in order to generate a full solution in the direction of the moving to fixed image. ANTs SyN was evaluated against two other methods: Thirion's Demons algorithm [66], and an elastic cross-correlation optimizer[2] [67]. All three methods were to perform an inter-subject registration of twenty T1 MR images from ten normal elderly and ten Frontotemporal Dementia (FTD) patients in order to characterize patients for the disease based on volume measurements of different brain structures [6]. The Dice score calculated from the images registered by ANTs SyN most closely matched those of the gold standard volume measurements compared to the other two methods, indicating that ANTs SyN is reliable in normalization and volumetric measurement tasks. ANTs SyN has shown better performance in terms of accuracy and computation time compared to all other classic methods listed in this Chapter [50]. Because of its success in the literature, and despite its age, ANTs SyN will be compared to VoxelMorph in the validation study, discussed in Chapter 3.

2.3.8 Deformetrica

Deformetrica is a template-based non-linear registration technique developed by Durrleman *et al.* [56]. This algorithm was used to study shape differences between normal controls and patients with Down Syndrome. The registration works in a semi-automatic manner, where the user specifies the number of control points (in the tens or low hundreds) on a template that will parametrize the deformation field. The template is a series of labelled meshes, with initial control points located at the most anatomically variable points. Each image that is to be registered to the template is matched to the template using a similarity measure known as a "varifold metric", adapted from Charon and Trouvé, which are manifolds that contain unoriented tangent vectors,

[2] Note that this method is an inter-subject registration method but does not promise diffeomorphic transformations and hence was not listed as a classic method in this thesis.

and are useful in representing non-oriented shapes such as in brain images [68]. The varifolds drive the movement of the control points, which are modelled as a velocity field over time and are ensured to follow a diffeomorphic transformation so long as they are differentiable [56]. The registration process is optimized using Nesterov's gradient descent method. Deformetrica is applied to eight controls and eight Down Syndrome patients to assess the shape differences in complex deep brain structures between the groups. Results show statistically significant anatomical differences between the groups despite the low number of subjects.

In the comparison by Klein *et al.*, ANTs SyN and Diffeomorphic Demons topped the comparison against DARTEL, JRD-fluid, and SICLE, with SyN having a slight advantage in some cases such as volume and surface overlap results [50]. Other techniques in this section: LDDMM, Diffeomorphic registration using B-splines, and Deformetrica were not compared to other techniques in their respective publications. However, the reader could select a registration method based on the application to which each method was used; for example, using LDDMM on Schizophrenia or Alzheimer's data [51], or Deformetrica for data of patients with Down Syndrome [56].

2.4 Related Work in Registration Based on Deep Learning

Before delving into the literature, a brief definition of deep learning is required. Deep learning is a branch of Artificial Intelligence (AI) comprising intelligent software which learns concepts in a hierarchical manner and is able to predict complicated problems by breaking the problem down into simpler concepts [69]. Building the simple concepts one on top of another makes this type of AI "deep" learning, since many layers are required (many more so than, say, an artificial neural net). Briefly, deep learning layers typically include convolutional layers, pooling

layers, and drop-out layers to name a few, Convolutional layers use kernels with activation functions to convolve with the image and produce lower level features [69]. Pooling layers provide a summary of the outputs from the convolutional layers and allow invariance in the presence of small changes to the input of the deep learning architecture [69]. Dropout layers are used as computationally inexpensive method for regularizing models [69]. Together, these layers can provide a basic Convolutional Neural Network, a commonly employed deep learning architecture [69]. However, there are other architectures that can be employed, even in the case of image registration as will be seen in this section. For a more in-depth understanding of deep learning, please refer the book titled, "Deep Learning" written by Ian Goodfellow *et al.* [69].

Search criteria for related work in DL-based registration included the following: The words "deep learning" "registration" were used in PubMed. Papers were excluded if DL-based registration was not applied to brain images, or was not non-linear, or if the method was only used to supplement part of a registration pipeline, where the bulk of the registration was done by a classic method. Resulting from this search are six methods, which are discussed in the following six subsections, in chronological order. The methods are discussed similarly to the related work in Section 2.3, in terms of the transformation model, similarity metric, optimization strategy employed, as well as the evaluation and application of the method. However, since the DL-based methods may not follow classic methods of how their deformation fields are derived, details such as DL architectures used, inputs, outputs, training and testing employed, and other DL related topics necessary for a comprehensive summary, are noted.

2.4.1 Self-Supervised Fully Convolutional Network

This registration algorithm by Li and Fan proposes “deep self-supervision” in their Fully Convolutional Network (FCN) to register an image pair [70], where what they call “deep self-supervision” is advantageous to other methods discussed later in this section. An FFD transformation model type is predicted using the FCN, shown in Figure 1. The FCN architecture contains several layers which replace conventional registration approaches from classic methods. For example, regression layers create image pairs at various resolutions, which are then used to predict deformation fields at these resolutions. Deconvolutional layers then upsample these deformation fields, essentially interpolating the displacement vectors in the deformation field, thus replacing a B-spline or TPS interpolation. The FCN contains a loss function with regularization parameters for smoothing, as well as a similarity measure of normalized cross-correlation, which is used to predict the deformation fields at each resolution level. This multi-resolution prediction is what they call “deep self-supervision”. Conventional optimization strategies are replaced with forward and backpropagation during training. The network has a two-channel input which takes the fixed and moving images, and outputs a single deformation field. What was not discussed in this publication was how the deformation field is applied to the moving image.

The authors used the publicly available ADNI databases [71], as well as the LONI LBPA40 database [72] in two separate experiments. The ADNI data was used for an inter-subject registration task for the hippocampus. All volumes from ADNI GO and ADNI 2 were used for training, leaving 9600 image pairs from ADNI 1 left for testing. The LBPA40 data was used for an inter-subject registration of 54 different brain structures, where the ROIs had publicly available manual gold standard segmentations. Thirty volumes were used or training with the remaining 10 volumes saved for testing. The proposed method, in both experiments, was compared against

ANTs SyN, where performance was measured by Dice overlap of the 54 brain structures. Results from the ADNI experiment and the LBPA40 experiment showed slightly better Dice scores for the FCN-proposed method than for ANTs SYN. Although this method presents itself like VoxelMorph, ultimately VoxelMorph does present itself as a superior method since it promises diffeomorphic transformations.

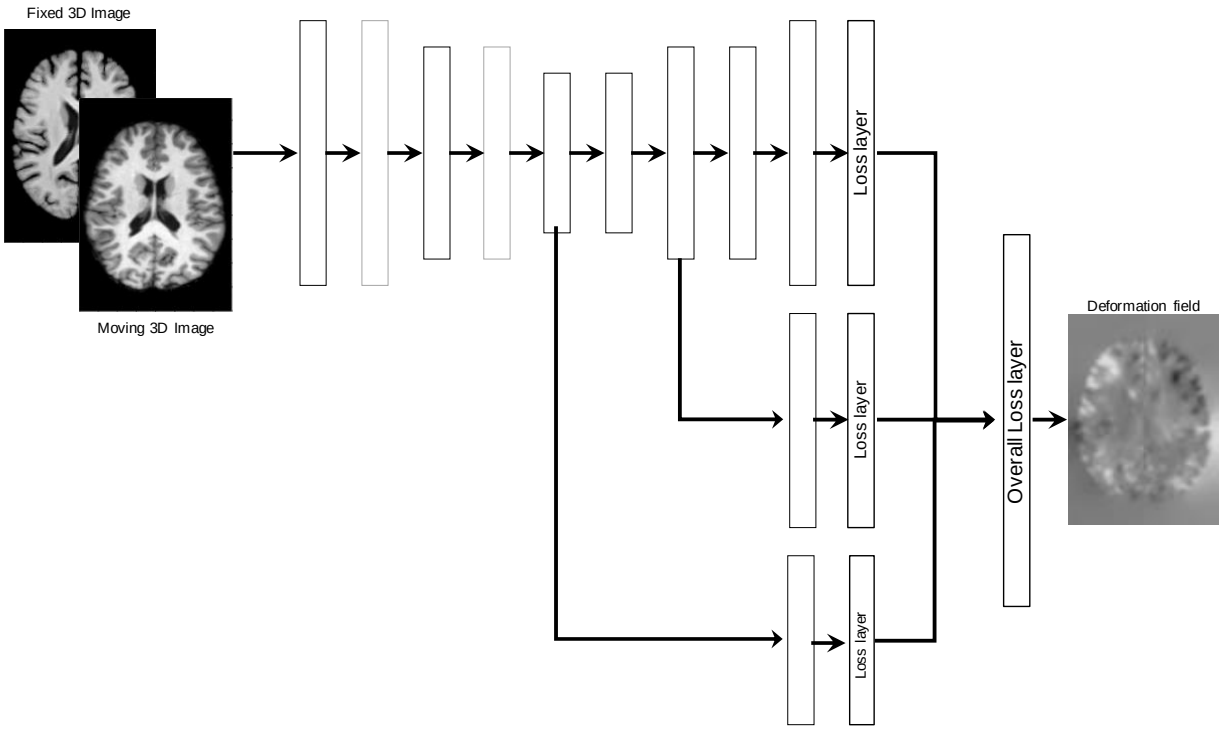

Figure 1: Fully Convolutional Network architecture (inspired by Li *et al.* [70]), showing the 2-input channel with fixed and moving images, three regression layers which extract multi-resolution deformation fields, and deconvolution layers which interpolate each deformation field through upsampling. An overall loss function controls for the resulting predicted deformation field and contains regularization parameters for smoothing and normalized cross correlation as a similarity metric.

2.4.2 PCANet

This registration method uses a Principal Component Analysis (PCA) based DL framework to learn features from a pair of images to be registered and uses these features to perform a non-rigid registration, making it advantageous in multimodal registration tasks. Briefly, PCA is an orthogonal linear transformation which is used to reduce the dimensionality while maximizing variability of a given dataset [73]. It works by solving for the largest eigenvalues and corresponding eigenvectors of a given image. PCANet uses an FFD-based transformation model, with B-spline interpolation [74]. The transformation model uses an objective function which comprises the similarity metric and smoothness term. The feature-based similarity metric calculates the Euclidean distance between the structural representations (as described below) of the image pair, which are predicted by the DL network. The network architecture, shown in Figure 2, contains two stages. The first stage takes image patches from the image pair, subtracts the patch mean from each image patch, and vectorizes the remaining information. PCA is then used to reduce the dimensionality of the vectors and produce eigenvectors (sorted in decreasing order of eigenvalue) which are used as convolution kernels. These convolutional kernels are then used to convolve the image patch pair to produce feature information which is used as input to the second stage in the network. The second stage takes the output from the first stage, removes its mean, is and vectorizes and condenses it in dimensionality using PCA to produce eigenvector convolution kernels, which are used to convolve the second stage input. The output from the second stage and the first stage are used to produce the fused feature image, or structural representation, which drives the registration. The FFD transformation is obtained by applying the cost function, which includes the Euclidean distance of this structural representation as well as a smoothness constraint.

The cost function is used in the Limited memory Broyden–Fletcher–Goldfarb–Shannon (L-BFGS) optimization algorithm.

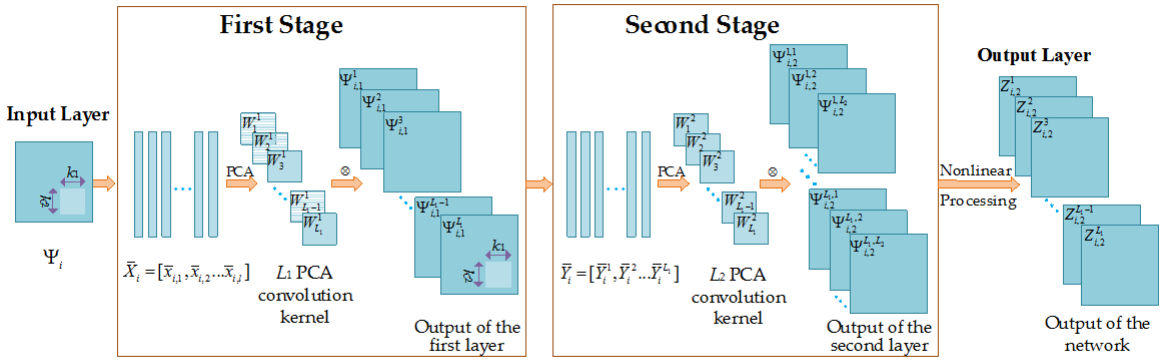

Figure 2: PCANet architecture (with permission) from Zhu *et al.* [74], showing the first and second stages of the network which act to pull higher level and low level feature information. Each stage shows the same process of vectorization and calculating PCA eigenvector kernels on the image patch pair. The input is shown as the image patch and the output is shown as the structural representation map which is used to drive the registration.

The network is trained in order to determine the weights of the convolution kernels in both stages 1 and 2. Five hundred images were used from a real brain database called Atlas [75] to train the network. BrainWeb20 [9], and a CT and MR image database called RIRE [76], were used for testing. The proposed method was evaluated using Target Registration Error (TRE) (i.e. the Euclidean distance between fixed and moving points) against four feature-based state-of-the-art methods: modality-independent neighborhood descriptor (MIND) [77], normalized mutual information (NMI) [78], Weber local descriptor (WLD) [79], and the sum of squared differences on entropy images (ESSD) [80]. Results show better TRE results using PCANet than with MIND, NMI, WLD and ESSD; however, these methods are not as well-known as the classic methods

listed in Section 2.3, and comparison to Diffeomorphic Demons or ANTs SyN would make the method more credible.

2.4.3 Cue-aware Deep Regression Network

Cao *et al.* developed this registration algorithm to successfully perform inter-subject registration between multiple databases and even databases with diseased populations [81]. The algorithm contains two networks which predict the similarity metric and displacement vectors for the deformation field, respectively, as shown in Figure 3. The similarity metric is a feature-based contextual "cue", or map, which is generated from multiscale information obtained from the first network in the architecture, the Context Driven Network. The Context Driven Network contains both convolution layers and pooling layers for feature extraction and assembly at multiple image scales. The first network takes the image patch pair as inputs and outputs the contextual cue map. The contextual cue, along with the same image patch pair, are inputs to the second network in the architecture, the Deep Regression Network. This network predicts and outputs the displacement vectors for the image patch pair. These displacement vectors for all patches in one image pair are then interpolated using block-wise TPS to produce the final FFD-type deformation field.

The network is trained on image patches, rather than an entire image, and therefore does not have the ease of end-to-end registration that some methods do. However, a novel sampling strategy called Key Point Truncated Balance sampling helps to reduce redundant feature and patch selection. The "key point" sampling, in the image space, obtains image points with large gradients to ensure patches with the most information are obtained, while the "truncated balance" sampling, in the displacement space, ensures maximum distribution between displacement vectors.

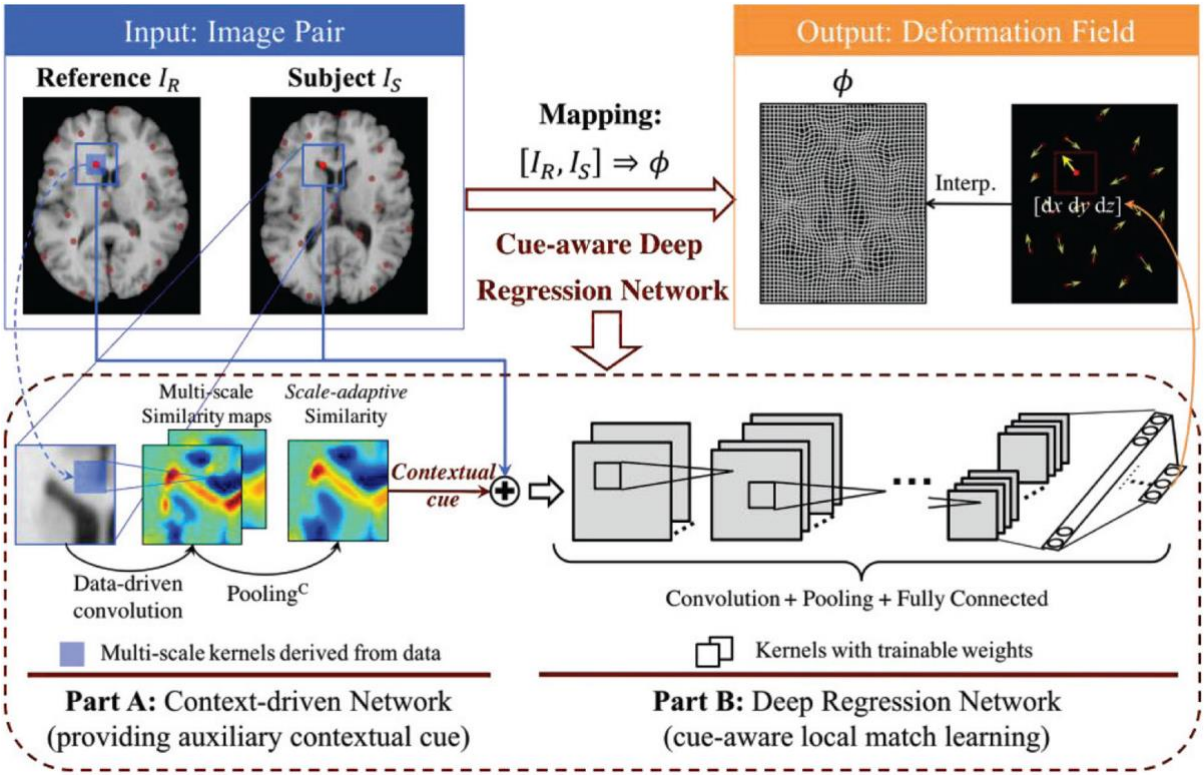

Figure 3: Cue aware deep regression network architecture (with permission) from Cao *et al.* [81], showing in the blue box patches of input image pair as inputs to both networks (labelled Part A and B in figure). The second network also takes the output from the first network, the contextual cue, as an input. The output of the second network is the deformation field of the image patch inputs (interpolated with TPS), shown in the orange box.

The network was trained using 40 image patch pairs from 20 images from the LONI LBPA40 database [72], and was tested using three databases in three different experiments. The first test used 15 LONI images, the second used images from the Information eXtraction from Images (IXI) database [82], and the third performed an inter-subject registration across the LONI and ADNI [20] databases. The experiments were used to compare the proposed method against three state of the art methods: ANTs SyN [6], as well as two Demons methods, SSD-Demons (i.e. the method mentioned earlier in this chapter) [55] and LCC-Demons (a Demons algorithm which uses a symmetric local correlation coefficient (LCC) as its similarity metric) [83]. The evaluation was

performed by measuring Dice scores between fixed and moving images, as well as by measuring the average surface distance. Grey and white matter segmentations were used, which had been produced from an automatic segmentation with manual corrections. Results from all three experiments showed comparable if not slightly better performance from the proposed deep leaning method. However, the three methods that this method is compared to create diffeomorphic registrations, while it is assumed that this method does not produce diffeomorphic transformations since they are not mentioned.

2.4.4 Adversarial Similarity Network

This registration technique is based on a General Adversarial Network (GAN), in which the similarity metric and deformation field are both predicted in this generator-discriminator network [84]. Generator and discriminator networks are trained using unsupervised adversarial training which is an advantage for this method. Adversarial training is also advantageous since it can improve the robustness of the network [85]. The discrimination network acts as the similarity metric, since it learns to discriminate between good and bad alignment of an image pair, represented by a similarity probability between 1 and 0 (good and bad, respectively) [84]. The discrimination network takes as inputs the deformation field, that is, the output from the registration network, and outputs the similarity probability. The registration network is a regression Unet, supervised by the similarity probability from the discrimination network. The registration network takes as inputs the image patch pair, and outputs the deformation field. The registration network is trained to convince the discrimination network that a good alignment has been made, so its cost function comprises both a smoothness term and a term to approximate a similarity probability of 1 between the image patch pair. After the registration network, a deformable transformation layer interpolates new voxel locations in the moving image using the

displacement vectors in the deformation field. The entire architecture can be seen in Figure 4. Of course, when testing occurs the registration network is used on its own.

The networks were trained using 30 images from the LONI LBPA40 database [72], totalling 26,000 image patch samples. The networks were tested using four public datasets [50]: the remaining 10 images from LONI LBPA40, Internet Brain Segmentation Repository (IBSR18), Columbia University Medical Center (CUMC12), and the MGH/MIT/HMS Athinoula A. Martinos Center for Biomedical Imaging (MGH10). The method was compared to Diffeomorphic Demons [55] and ANTs SyN [6]. Evaluation was performed by measuring the Dice score of 54 ROIs, in which the proposed method performed the best in 42 structures and were comparable in the other 12 structures [84]. As with the previous DL-based method, this method is compared to techniques which produce diffeomorphic transformations, while it is assumed that this method does not produce such transformations as diffeomorphisms are never mentioned in the methodology.

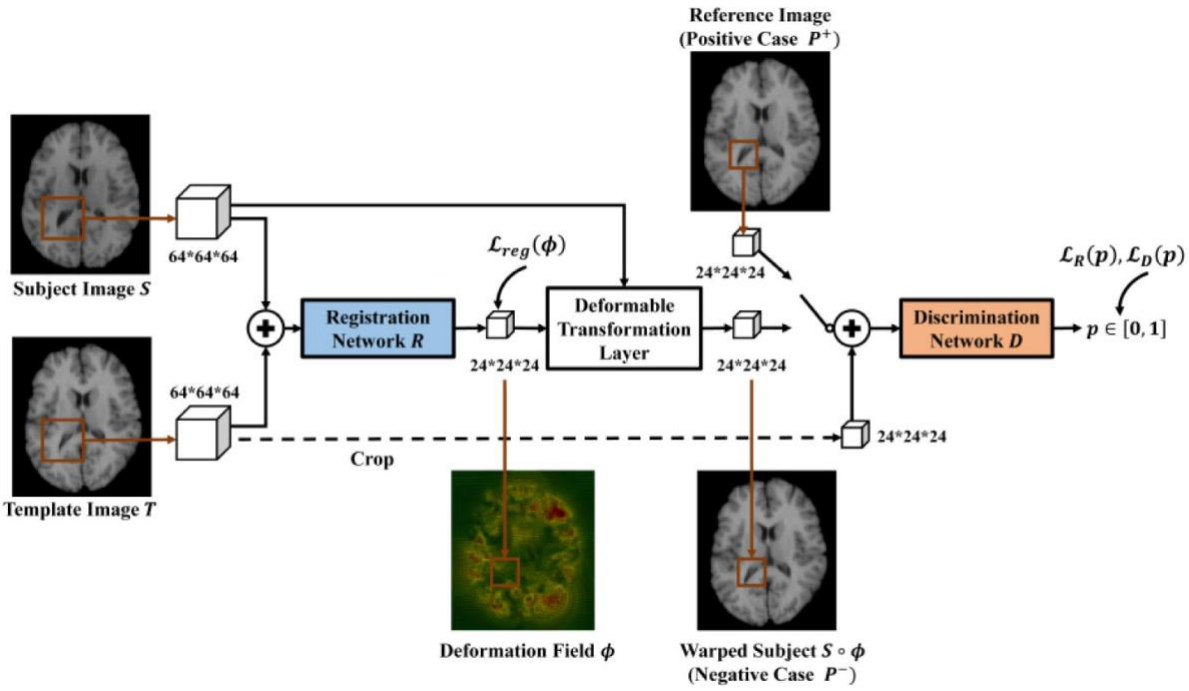

Figure 4: Adversarial Similarity Network architecture and training strategy (with permission) from Fan *et al.* [84], showing 64x64x64 patches of fixed (template) and moving (subject) images as inputs to the registration network. The registration network outputs the deformation field of the image patch, which is used to deform the moving image (subject) to match the fixed image (template). The figure shows how the discriminator network is trained using positive and negative misalignment cases of registrations, acting as the similarity metric to determine goodness-of-fit.

2.4.5 BIRNet

This registration method offers a supervised method with ground truths obtained from classic methods [86]. BIRNet is a hierarchical dual-supervised Fully Convolutional Network (FCN) capable of end-to-end registration. That is, the network takes as inputs image patches, and outputs the displacement vectors for that image patch pair and is capable of predicting the entire deformation field from patches. What makes the network "dual-supervised" is contained within its loss function, which acts as the similarity metric. The loss function comprises the displacement differences between the predicted and ground truth deformations, where ground truth deformations

originate from both ANTs SyN [6], and Diffeomorphic Demons [55]; and the intensity differences between the moved and fixed images [86]. Both differences are calculated as normalized gradients and drive the registration process. The network architecture is a typical U-Net shape, but with additional layers in between to help make connections between the high and low-level features to which the authors refer to as "gap-filling" (See Figure 5).

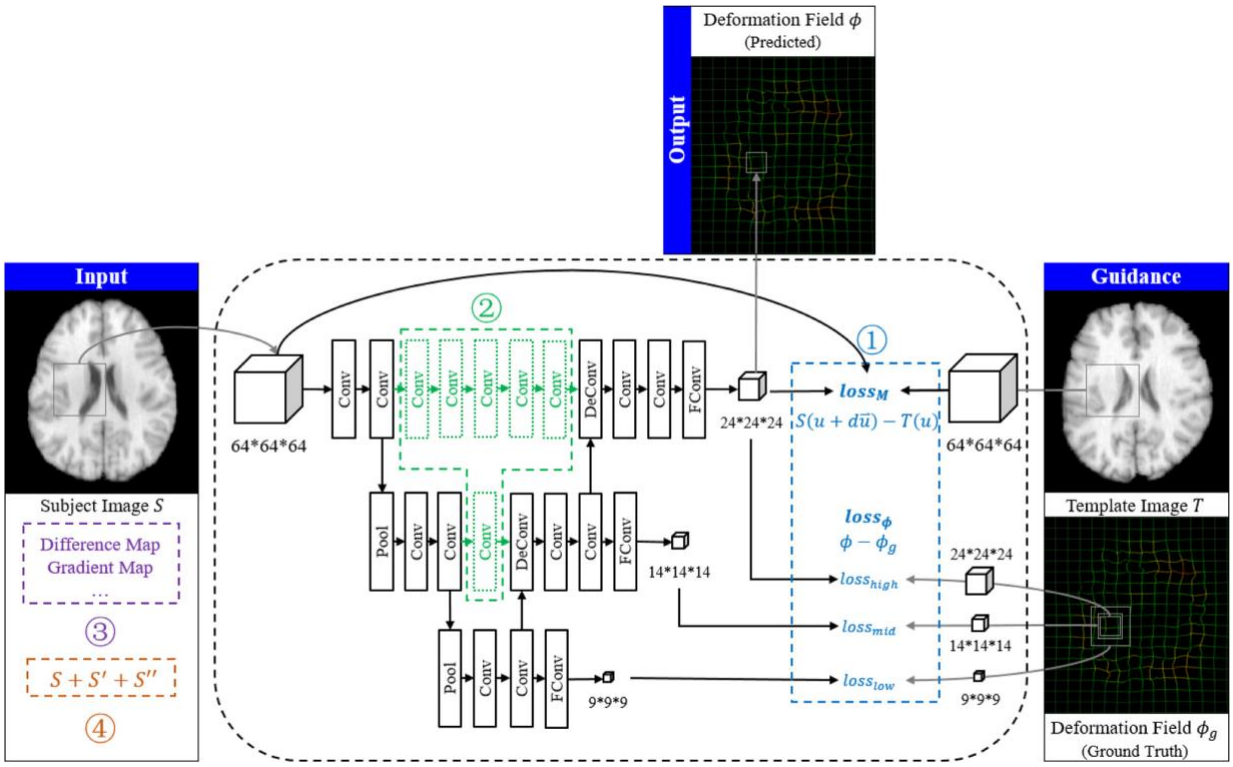

Figure 5: BIRNet architecture and training strategy (with permission) from Fan *et al.* [86]. The network takes a 64x64x64 image patch pair as input which undergoes convolutions and deconvolutions in the Unet architecture. The Unet architecture contains "gap-filling" layers, which are additional layers between the Unet which bridge connections between high and low level features. BIRNet is trained with ground truth deformation fields from classic methods which is used in the loss function along with the predicted deformation field to optimize for goodness-of-fit.

The network was trained on 54,000 image patches from 30 subjects of the LONI LPBA40 database [72], with data augmentation performed through warping the images with percentages of

their ground truth deformations to increase robustness of the network. BIRNet was evaluated in an atlas-based registration, using the first image from LONI LBPA40 as the template, against Diffeomorphic Demons, LCC-Demons [83] and ANTs SyN among others. Testing was done using four public datasets: Internet Brain Segmentation Repository (IBSR18), Columbia University Medical Center (CUMC12), and the MGH/MIT/HMS Athinoula A. Martinos Center for Biomedical Imaging (MGH10), and the Information Extraction from Images database (IXI) [82]. Dice scores on ROIs of cortical grey structures indicate BIRNet has the best performance; however dice score alone is not an accurate measure of goodness-of-fit [8]. BIRNet was compared against two diffeomorphic registration algorithms but fails to produce diffeomorphic transformations itself.

2.4.6 VoxelMorph

VoxelMorph is a DL-based non-linear diffeomorphic registration algorithm developed by Dalca *et al.* [5]. The advantage of VoxelMorph over other learning-based registration methods is its end-to-end unsupervised framework; therefore, no ground truth is required..

Considering only the basic model shown in Figure 6, VoxelMorph uses a stationary velocity field as its transformation model, as introduced by John Ashburner in DARTEL [54]. The convolutional neural network, in this case a Unet, learns the stationary velocity field and spatial transformation layers perform the diffeomorphic integration to obtain the deformation field [5]. Specifically, the Unet takes as inputs the moving and fixed image which both go through a series of convolutions and deconvolutions to arrive at a prediction of the mean and covariance for the probabilistic model of the stationary velocity field. This posterior registration probability is estimated by the Unet in order to obtain the most likely registration field given the cost function.

The cost function acts as the similarity metric between the fixed and moving images. The cost function is a combination of the Expectation of the similarity between the fixed and moved image, and Kullback-Leibler (KL) divergence. KL-divergence is a measure of difference in probability distributions [87] and is used to guide the posterior registration probability closer in similarity to the prior probability of the stationary velocity field [5]. The network is optimized by stochastic gradient descent methods, which acts as the optimization strategy in comparison to classic methods. Once the network has predicted the stationary velocity field via the mean and covariance of the posterior registration probability, the deformation field is computed through seven scaling and squaring layers, which are similar to Euler integration as performed in DARTEL. The deformation field is then used to warp the moving image using a learned spatial transformer which resamples the moving image. The technique guarantees diffeomorphic registrations since each step in its framework is differentiable.

The data used for training and testing were 3731 T1-weighted brain MRI scans from eight datasets [5]. These includes: OASIS [88], ABIDE [89], ADHD200 [90], MCIC [91], PPMI [92], HABS [93], and Harvard GSP [94]. Images were skull stripped using FreeSurfer [7], resampled to 1 mm isotropic sampling, intensity normalized, and affine registered to the atlas. For training, 3231 volumes were used, leaving 250 volumes for validation and 250 volumes for testing. Segmentation labels were automatically generated by FreeSurfer [7], and labels were compared against the atlas using Dice scores to evaluate their registration performance against ANTs SyN. It is important to note that the transformation parameters for their version of ANTs SyN were altered to fit the "performance standards" of the proposed method. Briefly, this means the smoothness parameters of ANTs SyN were altered to produce results which would then be comparable to VoxelMorph,

but this comes to the detriment of ANTs SyN's performance. VoxelMorph performed comparably to the modified ANTs SyN when evaluating Dice scores of 29 segmented brain structures.

It is noted that VoxelMorph is one of few techniques mentioned here which have provided fully open source access to their methods. As stressed in the Introduction Chapter of this thesis, open source publications facilitate reproducibility in science and help drive innovation forward [1] [2]. Since its publication, VoxelMorph has been used as a baseline in another method called FAIM [95] by Kuang *et al*.

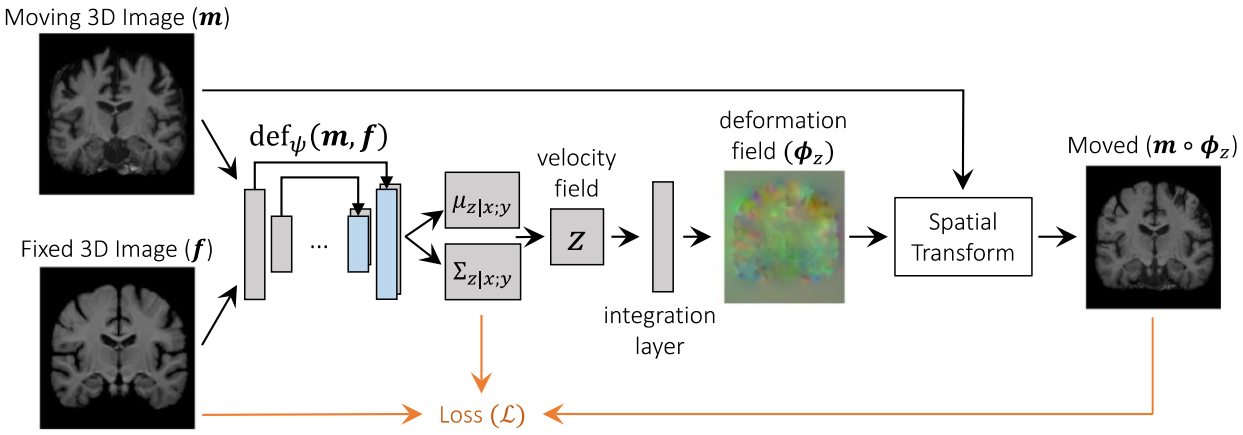

Figure 6: VoxelMorph architecture (with permission) from Dalca *et al.* [5], showing inputs: atlas and moving image, as well as the Unet which predicts the mean and covariance of the posterior registration probability, p(z), to give the stationary velocity field, z. Seven scaling and squaring layers calculate the deformation field, which is then resampled to create the moved image.

2.5 Evaluation Metrics

Common evaluation metrics for image registration can be categorized into direct and indirect metrics. Direct metrics judge the misalignment between two images through measurements of distance, whereas indirect measurements judge misalignment, for example, using segmentations

of the registered images. Examples of metrics used in this thesis include direct metrics such as recovery error and indirect metrics such as Dice score, Cohen's Kappa, and Hausdorff distance.

The recovery error is used to compare two transformations, in this case, the recovered transformation and the simulated deformations in Experiment A. The recovery error is defined as the root mean square (rms) difference in position of a set of points transformed forward through the known transformation, and then back through the inverse of the recovered transformation. Equation 1 shows the recover error:

$$rec\ error\ = RMS\big((x,y,z) - T_c(x,y,z)\big) \tag{1}$$

where T_c is the concatenation of the forward transformation and the inverse of the recovered transformation. Dice score and Cohen's Kappa are examples of overlap metrics, where two registered images have their segmentations overlain to see what percentage of the labels intersect [11]. The formula for Dice score is shown in Equation 2.

$$DICE = \frac{2\,|S_g \cap S_t|}{|S_g| + |S_t|} \tag{2}$$

where S_g is the ground truth segmentation, S_t is the test or experimental segmentation, and $|\ |$ represents the number of voxels [11]. Cohen's Kappa tends to provide a more robust calculation of overlap, since Cohen's Kappa takes into account the probability that overlap occurred by chance [11]. The formula for Cohen's Kappa is shown in Equation 3:

$$KAPPA = \frac{P_a - P_c}{1 - P_c} \tag{3}$$

where P_a represents when the two segmentations overlap, and P_c is the hypothetical probability that the segmentation overlap occurred by chance. Last of the indirect metrics is Hausdorff

distance, which measures the shortest distance between two furthest voxels in overlain segmentations of registered images. Hausdorff distance can be sensitive to outliers, so usually a 95^{th} percentile Hausdorff distance is used [11].

2.6 Moving Forward

Proceeding to Chapter 3, the chosen DL-based registration method, VoxelMorph, will be evaluated against the original unmodified ANTs SyN in a paper whose contents have been submitted for publication in *Medical Image Analysis*. The paper is titled, "Evaluating VoxelMorph, a learning-based 3D non-linear registration algorithm, against the non-linear Symmetric Normalization technique from ANTs," with authors Victoria Madge, Philip Novosad, Daniel A. DiGiovanni, and D. Louis Collins.

CHAPTER 3

Results

This thesis has discussed the importance of registration in medical image-analysis tasks. Recently, DL has found success in many image processing tasks, including registration, which have been reviewed in the previous chapter. Of the DL techniques studied, VoxelMorph appears as the first of its kind in end-to-end registration of whole images, producing a deformation field for inter-subject registration. It was directly compared to a non-learning state-of-the-art method, Advanced Normalization Tools Symmetric Normalization (ANTs SyN) published by Avants *et al.* in 2008 [6].

In this chapter, VoxelMorph will be compared against ANTs SyN in an evaluation which includes both direct and indirect metrics, manual gold standard segmentations, and by using unaltered versions of both methods. The contents of the following paper have been submitted for publication in *Medical Image Analysis*.

Evaluating VoxelMorph, a learning-based 3D non-linear registration algorithm against the non-linear Symmetric Normalization technique from ANTs

Victoria Madge[a, b], Philip Novosad[a, b], Daniel A. Di Giovanni[a, c], D. Louis Collins[a, b, c]

[a] *McConnell Brain Imaging Centre, Montreal Neurological Institute and Hospital, Montreal, QC, Canada*
[b] *Department of Biomedical Engineering, McGill University, Montreal QC, Canada*
[c] *Department of Neurology and Neurosurgery, McGill University Montreal QC, Canada*

Abstract

Medical image registration is the process of aligning two images of the same scene into the same image space and is considered a fundamental step in many image processing applications. Recently, deep learning has shown success in a wide variety of medical image analysis tasks, including image registration. VoxelMorph is a learning-based non-linear technique promising fast diffeomorphic registrations while claiming comparable results to ANTs SyN. However, the indirect comparison between the two methods was based solely on Dice scores of automatically segmented labels, and the smoothness parameters of the ANTs SyN algorithm were altered to be more similar to those of VoxelMorph. This paper aims to compare VoxelMorph against the native, unaltered ANTs SyN using both direct evaluation with simulated deformations and indirect metrics based on manual gold standard segmentation labels. Results from the first experiment show ANTs SyN significantly outperforms VoxelMorph in the presence of simulated deformation. Experiments with real data demonstrate VoxelMorph has a small but significant advantage in inter-subject registration compared to ANTs SyN.

3.1 Introduction

Medical image registration is the process of aligning two images of the same scene into the same image space. This is considered a fundamental step in many image processing applications [15] [16]. In particular, image registration plays important roles in the application of brain segmentation and atlas formation [13] [21] [22], diagnostics, treatment, and monitoring of various pathologies [23] [24] [25] [26] [27], as well as image-guided surgery and surgical planning [30] [28] [29], surgical simulation [96], and intra-operative brain shift correction [32] [33].

For the registration of brain images between subjects, non-linear forms of registration are favoured since they are capable of compensating for tissue deformation [15] [31] [37]. However, a common problem among the use of non-linear transformations is the inability to maintain brain topology. Diffeomorphic transformations guarantee image topology preservation as well as transform invertibility in a one-to-one mapping between images [44].

Recently, DL has shown success in a wide variety of medical image-analysis tasks, including image registration, but requires repeatability and validation to deem the techniques credible against the current state-of-the-art [4]. Reproducibility in science is imperative to give credibility where it is due, through replication and validation of another's work [2]. This paper aims to evaluate the registration technique of VoxelMorph [5] to the state-of-the-art ANTs SyN [6].

3.1.1 Related Work

Several state-of-the-art non-linear diffeomorphic registration techniques, termed "classic" methods here, have found success in various brain registration applications. The techniques mentioned here use different transformation models to deform the "moving" image to the "fixed"

image. Some techniques use Free Form Deformation (FFD) models, such as Small-deformation Inverse-Consistent Linear-Elastic image Registration (SICLE) developed by Johnson *et al.* [46]. SICLE uses an FFD model and thin-plate spline (TPS) interpolation, with intensity-based and feature-based similarity measures to achieve alignment [46]. Another technique developed by Rueckert *et al.* uses multi-level hierarchical FFDs in order to gain the advantage of larger and then smaller constrained control-point spacings, which ensures both smooth local and global transformations [53]. B-spline interpolations are then used to regularize the diffeomorphic registration of the moving image. Although not necessarily in all cases, the use of FFDs and their interpolators can pose problems in the form of holes or folds in the recovered image [16], which would not make the registration diffeomorphic. Johnson et al. have overcome this issue in their technique by constraining their FFD transformations using diffeomorphic fluidic properties [46], and Rueckert et al. ensures diffeomorphic transformations by limiting the maximum number of control point displacements for each FFD [53].

Diffeomorphic Demons, proposed by Vercauteren *et al.* [50], uses Thirion's Demons forces [65] to express an optical flow transformation model for diffeomorphic registration guaranteed using the Jacobian determinant [97]. The intensity-based voxel wise mean square error is used as the similarity metric, with a Newton optimization strategy. This technique has reasonable computation times compared to other non-linear techniques, but is outperformed in terms of accuracy by these very same techniques [50].

Other diffeomorphic techniques rely on flow fields or velocity fields to model the movement of one image to another, including Large Deformation Diffeomorphic Metric Mapping (LDDMM) [51], JRD-fluid [52], Diffeomorphic Anatomical Registration using Exponentiated Lie algebra (DARTEL) [54], Deformetrica [56], and Advanced Normalization Tools (ANTs) Symmetric

image Normalization method (SyN) [6]. In the model, the ordinary differential equation which represents the velocity field of the moving image can be solved using Euler-Lagrange Equations [6] [52], or similarly, Euler's method [52][54]. The methods differ, however, in their approach to make the registration diffeomorphic and in the similarity metric used in the method. For example, LDDMM guarantees diffeomorphisms using smoothness constraints [51], while DARTEL uses the Jacobi identity in Lie Algebra [54]; similarly, JRD-fluid uses a positive Jacobian determinant threshold [52]. For similarity metrics, the JRD-fluid method uses the Jensen-Rényi Divergence information-based measure as a similarity function, Deformetrica uses a manifold of unoriented tangent vectors called varifolds as a similarity measure, as well as a template to drive the registration [56], whereas DARTEL uses a voxel-wise mean square difference measure [54].

However, of these velocity field-based techniques, ANTs SyN consistently ranked one of the best in diffeomorphic registration according to the most recent study evaluating fourteen non-linear deformation algorithms applied to human brain MR images [50]. The Klein study [50] was cited by Dalca *et al.* as justification to determine which classic registration method to compare to VoxelMorph [5].

ANTs SyN uses cross-correlation (among other similarity metrics) in a diffeomorphic and "inverse consistent" process known as symmetric normalization [6]. Computing the Euler-Lagrange cross-correlation equations to solve for the velocity field is symmetrized through a Jacobian operator. The technique guarantees diffeomorphic transformations by including invertibility constraints during the optimization. The optimization is done by computing the two half velocity fields for each direction of image movement: moving-to-fixed and fixed-to-moving. Velocity field equations are solved in parallel to find the maximum cross-correlation between the two images. Their inverses are then computed in order to generate a full solution in the moving-

to-fixed image direction. ANTs SyN has shown to be reliable in normalization and volumetric measurement tasks with Alzheimer's Disease and Frontotemporal Dementia patients [98].

Classic registration methods, while successful in their respective applications, often suffer from long computation times. Recently, deep learning has shown success in a wide variety of medical image analysis tasks, including image registration [4]. Learning-based registration offers potentially faster registration times with arguably comparable results. However, where the "learning" takes place in these non-linear registration techniques, comparable to classic brain registration techniques, can vary.

PCAN net, for example, uses a deep learning framework to learn features from a pair of images to be registered and then proceeds with a more classic form of non-linear registration of two images [74]. The two-stage network takes image patches as inputs, which are then vectorized and passed through Principal Component Analysis (PCA) convolution kernels to extract pertinent high-level and low-level structural information. This information is used as a similarity metric in an FFD-type registration technique using B-spline interpolation.

Other techniques use deep learning frameworks to predict the deformation field itself. Some use image patches as inputs to address memory limitation issues, and thus cannot predict an entire deformation field in one go [70][81][84]. In their methods, Cao *et al.* are able to construct a final deformation field using TPS interpolation to interpolate between image patches [81]. Interpolation approaches are also utilized for Fan *et al.* [84], while Li *et al.* use deconvolutional operators to upsample, in order to interpolate the final deformation field [70].

Two deep learning techniques capable of full deformation field predictions include BIRNet [86] and VoxelMorph [5]. BIRNet is a hierarchical Fully Convolutional Network (FCN) which

uses ground truth deformations from both ANTs SyN and Diffeomorphic Demons, as well as the difference between the fixed and moved images in a dual-supervision registration technique [86]. The FCN comprises a Unet with additional "gap-filling" layers to help make connections between the high and low-level features. However, the supervised nature of this technique is a disadvantage, since the registration results from classic methods are required as inputs to train the model. However, the second of the two methods – VoxelMorph – requires no ground truth.

VoxelMorph has an advantage over other deep learning-based registration methods due to its end-to-end unsupervised framework [5]. VoxelMorph uses a stationary velocity field as its transformation model, with a similar theoretical background as DARTEL [54]. Here, the Unet learns parameters to form the stationary velocity field and spatial transformation layers perform the diffeomorphic integration to obtain the deformation field [5]. The Unet takes as inputs the moving and fixed image and outputs the prediction of the mean and covariance for the posterior registration probability. Once the network has predicted the stationary velocity field via the mean and covariance of the posterior registration probability, the deformation field is computed through seven scaling and squaring layers, which are similar to Euler integration as performed in DARTEL [54]. The deformation field is then used to warp the moving image using a learned spatial transformer which resamples the moving image. The technique guarantees diffeomorphic registrations since each step in its framework is differentiable.

The authors of VoxelMorph claim to have diffeomorphic registration results comparable to ANTs SyN [6]; however, there are several points within the methodology and evaluation performed by Dalca *et al.* which could be improved. First, the training and testing of the VoxelMorph model was performed using FreeSurfer-based [7] segmentations of MRI volumes [5], and not manual gold standard segmentations [8]. Thus, segmentation errors in the

automatically extracted labels could confound registration quality metrics. Second, the smoothness parameters of the ANTs SyN algorithm were altered to be more similar to those of VoxelMorph [5], thus restraining ANTs SyN's potential capacity to achieve a successful registration. Finally, the methods comparison was based solely on Dice scores, an indirect metric, which alone is not a complete evaluation of goodness-of-fit [8].

3.2 Methods

The objective of this paper is to comprehensively evaluate ANTs SyN and VoxelMorph. This evaluation will be performed in three experiments, each serving a different evaluation purpose.

3.2.1 Data

Experiment A – Deformed VoxelMorph Atlas

Experiment A evaluates how well both VoxelMorph and ANTS SyN can recover simulated deformations of a single image. While one could generate random deformations to test each method's performance, such simulated transformations may not be realistic. Realistic deformations were estimated with Automatic Nonlinear Image Matching and Anatomical Labeling (ANIMAL) [13] to map individual subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [20] to the VoxelMorph atlas. ANIMAL deformations were chosen here since they do not advantage either VoxelMorph or ANTs SyN, but give smooth, realistic estimates of the required deformations. The VoxelMorph atlas was chosen here since the VoxelMorph model has been trained to register moving images to this target atlas, and thus would not be at a disadvantage compared to using a different template. The VoxelMorph atlas was borrowed from

another study published from the co-authors which provided an analysis framework for multimodal image studies with applications in stroke [99]. The atlas contains left and right hemisphere labels as well as deep grey structure segmentations including basal ganglia.

The ADNI database is a publicly available database compiled of multicentre imaging data designed to advance research in Alzheimer's Disease detection and tracking [20]. Outside of this experiment, ANIMAL [13] was used to register images from the ADNI database, of varying disease severities, to a normal template. Briefly, ANIMAL estimates displacement vectors between a subject's MRI and a template brain at various resolution levels, building a deformation field from these vectors [13]. The disease severity levels selected for this experiment include normal controls, early Mild Cognitive Impairment (MCI), late MCI, and Alzheimer's Disease.

Thirty randomly selected ADNI-ANIMAL deformations for each disease severity were applied to the VoxelMorph atlas and its corresponding label, building a dataset of 120 different MRI-label pairs. The data for Experiment A is used to test the performance of both VoxelMorph and ANTs SyN to recover these simulated deformations. Since the VoxelMorph atlases are still in its own space after the deformations are applied, and the intensity of the image is unaltered from the original atlas, no pre-processing is required.

Experiment B – BrainWeb20

Experiment B uses the BrainWeb20 database, which comprises twenty T1-weighted digital MR phantoms created from twenty normal adults [9]. For each of the 20 subjects, ten manually corrected fuzzy mask brain structure volumes (e.g., cerebrospinal fluid (CSF), grey matter, white matter, fat, muscle, skin, skull vessels, connective tissue, dura and marrow) were used to simulate T1w MRI volumes. The simulator takes these tissue masks and changes them to intensity values

using first principles block equations. Thus the data are digital phantoms which represent a physical distribution of human head scans and thus can be considered truth, and can help understand the accuracy in the performance of these two techniques [8].

The BrainWeb20 MRI volumes underwent pre-processing similar to the VoxelMorph methodology [5]: skull stripping, affine registration to the original VoxelMorph atlas, and intensity normalization. However, pre-processing was completed using algorithms from the Minc Toolkit [12], a collection of image processing software released by the McConnell Brain Imaging Centre at the Montreal Neurological Institute.

Experiment C - Neuromorphometrics

Experiment C uses a database with gold standard manual segmentations from Neuromorphometrics [10]. The database comprises 93 unique T1-weighted raw MR volumes with labels of manually segmented neuroanatomical structures, inspected and certified by a neuroanatomical expert [10]. This database was chosen in order to evaluate both VoxelMorph and ANTs SyN on real MRI data with manual gold standard segmentations, an important step in validation [8]. Specifically, in this experiment, the performance of each method will be evaluated using labels from selected anatomical structures. Due to computation constraints, twenty subjects were chosen from the 96 subjects at random. Computation constraints also restricted structure inclusion, whereby three cortical and three deep grey structures were chosen at random. Cortical and deep grey structures were chosen specifically to see how the registration methods would behave when performing a non-linear registration to these regions. The three cortical structures that were randomly selected include: the middle temporal gyrus, the frontal pole, and the post-

central gyrus; and the three deep grey structures that were randomly selected include: the amygdala, the hippocampus, and the thalamus.

Similar to the data in Experiment B, the twenty selected Neuromorphometrics MRI volumes underwent skull stripping, affine registration to the original VoxelMorph atlas, and intensity normalization for pre-processing. All pre-processing was performed using the Minc Toolkit [12].

3.2.2 Metrics

The evaluation metrics used in the experiments include both indirect and direct measurements. Indirect metrics are measured using the EvaluateSegmentation tool, an open source implementation of common and well-defined evaluation metrics for medical image segmentation [11]. The tool includes overlap-based metrics, volume-based metrics, pair-counting-based metrics, and distance metrics, which have all been implemented for simple one-label cases, and some metrics are applicable to multi-label cases. The EvaluateSegmentation tool [11] is used here to measure Dice score, Cohen's Kappa, and Hausdorff distance . Dice score was selected to directly compare results published by Dalca *et al.* Cohen's Kappa was selected as another overlap metric to provide a robust calculation of overlap, since Cohen's Kappa takes into account the probability that overlap occurred by chance [11]. The Hausdorff distance metric was chosen in order to calculate the maximum error for each method. Since Hausdorff distance can be sensitive to outliers, a 95^{th} percentile Hausdorff distance (hereby referred to as H95) is used to avoid potential outliers [11]. Other evaluation metrics from the EvaluateSegmentation tool [11] were considered redundant or were not applicable in the evaluation of registration methods.

Direct measurements are possible with simulated data as the answer is known. The recovery error is used to compare two transformations, in this case the recovered transformation and the

simulated deformations in Experiment A. The recovery error is defined as the root mean square (rms) difference in position of a set of points transformed forward through the known transformation, and then back through the inverse of the recovered transformation. Equation 4 shows how the recovery error is calculated:

$$rec\ error\ = RMS\big((x,y,z) - T_c(x,y,z)\big) \tag{4}$$

where T_c is the concatenation of the forward transformation and the inverse of the recovered transformation. The recovery error is implemented using *xfmconcat* and *transformtags* algorithms from the Minc Toolkit [12] to combine deformation and moving to fixed transformation metrics and create a grid of points by which the NumPy library from Python [100] is used to subtract the grid of points of one image from another in order to get a physical distance between the points. The distances are averaged in each direction (x, y and z) and the rms of the distances is calculated as well, using Python.

3.2.3 Experiment A

The purpose of Experiment A is to evaluate the performance of both VoxelMorph and ANTs SyN in the presence of simulated deformations using both Dice, Cohen's Kappa, H95, and recovery error.

The pre-trained VoxelMorph model (https://github.com/voxelmorph/voxelmorph) was applied to the 120 deformed volumes generated for Experiment A to register each volume to the original VoxelMorph atlas. The algorithm was configured to output the transformation along with the recovered MRI volume and warp parameters. Each volume took approximately 30 seconds to 1 minute to register on an NVIDIA TitanX GPU. Included in the VoxelMorph package was a

separate script which resampled the labels with the recovered transformation before calculating Dice scores. The VoxelMorph package from Dalca contained a separate script to apply the recovered transformation to resample the labels (using nearest neighbour) to calculate Dice scores internally. This script was modified to output the resampled labels so that the evaluation can be performed with EvaluateSegmentation.

The ANTs SyN algorithm was applied to estimate 120 3D non-linear registrations to the original VoxelMorph atlas. The updated field includes 0.1, 3 and 0 which represent the gradient step at each fluid iteration, the update field variance limit at each fluid iteration, and the total field variance limit at elastic iteration (in this case, there is no elastic regularization) [101]. With three hierarchical levels, the maximum number of iterations is 70, 50 and 20 respectively, with a CC threshold (1e-6) which allows the algorithm to move on to the next level if CC does not improve within the convergence window of 10 iterations. The three levels have resolutions of 4, 2 and 1 with default smoothing sigmas 2, 1 and 0. The algorithm output the recovered transformation, which was later used to resample the volumes and the corresponding labels using the resampling tool from the Minc Toolkit [12] [102]. Tri-linear interpolation was used to resample the MRI data, and nearest neighbour interpolation was used for labels [102]. The ANTs SyN technique took between 37 to 43 minutes to register each volume on a Xeon E3-1275 V6 CPU with 3.80 GHz clock speed.

The recovered labels from both VoxelMorph and ANTs SyN were compared against the original VoxelMorph atlas labels using indirect metrics from the EvaluateSegmentation tool [11]. Brain masks (the union of all brain labels, i.e., grey matter, white matter and CSF) were used in this evaluation since some measurements in the EvaluateSegmentation tool, including H95 and

Cohen's Kappa, cannot evaluate on multi-label files. The recovered transforms from both VoxelMorph and ANTs SyN were used to calculate the recovery error.

3.2.4 Experiment B

The purpose of Experiment B is to perform inter-subject atlas-based registration with both VoxelMorph and ANTs SyN using known MR volumes and their segmentations via a phantom database – BrainWeb20 [9]. The goal will be to see how well the subjects are aligned in the atlas template space by doing pair-wise comparisons of anatomical structures between different subjects.

As in Experiment A, atlas-based registration using VoxelMorph and ANTs SyN was performed for Experiment B. While VoxelMorph resampled the MRI data internally, the ANTs SyN output was recovered using the resampling tool from the Minc Toolkit [12] [102] just as in Experiment A. Recovered labels were also resampled as in Experiment A; the label for the VoxelMorph output was resampled using the separate script mentioned in Section 3.2.3, and the label for the ANTs SyN output was resampled using Minc Toolkit [12] [102]. Once again, both labels were resampled using nearest neighbour interpolation into the template space. If the methods recovered the necessary transformations without error, the anatomy of all subjects would be perfectly aligned to the template and structure comparisons between subjects would achieve Dice and Cohen kappa scores of 1.0 and H95 distances of 0.0 mm.

The original VoxelMorph label set was limited to left and right hemisphere segmentations, as well as some deep brain structural segmentations, which did not match the segmentations of the BrainWeb20 database [9]. Therefore, whole brain, grey and white matter, as well as CSF segmentations from BrainWeb20 [9] were used in pairwise comparisons between subjects. The

labels, transformed with simulated and recovered transformations were compared against one another within the target space and were evaluated using the indirect metrics from the EvaluateSegmentation tool [11]: Dice score, Cohen's Kappa, and H95.

3.2.5 Experiment C

The purpose of Experiment C is to perform inter-subject atlas-based registration using both VoxelMorph and ANTs SyN using a database with manual gold standard segmentations from Neuromorphometrics [10]. In particular, the methods are evaluated on their ability to register anatomical structures. Twenty subjects and six anatomical structures were chosen to measure goodness-of-fit for both methods.

Atlas-based registration using the pre-trained VoxelMorph model and ANTs SyN was performed for Experiment C as for Experiments A and B. Outputs and labels were recovered using the same methods as for Experiments A and B. As with Experiment B, recovered labels (middle temporal gyrus, frontal pole, and post-central gyrus, amygdala, hippocampus, and thalamus) were compared against one another within the target space and were evaluated using the indirect metrics from the EvaluateSegmentation tool [11]: Dice score, Cohen's Kappa, and H95.

3.3 Results

3.3.1 Experiment A

Figures 7, 8 and 9 compare the whole brain Dice scores, Cohen's Kappa, and H95, respectively, for VoxelMorph and ANTs SyN in Experiment A. The boxplots are arranged on the x-axis by increasing disease severity, with the assumption that increasing disease severity increases the voxel displacement in the deformation. Disease severity increases from normal controls (nc), to early Mild Cognitive Impairment (emci), late MCI (lmci), and Alzheimer's Disease (ad). Results from a two-way analysis of variance (ANOVA) show there is little difference in Dice score between disease severities; however, the difference between methods is statistically significant ($p < 0.001$) for both Dice scores (Fig. 7) and Cohen's Kappa (Fig. 8), meaning ANTs SyN significantly outperforms VoxelMorph. Results from the two-way ANOVA show H95 results were not significantly different for method or severity (Fig. 9).

Table 1 shows the results of the recovery error for both ANTs SyN and VoxelMorph in Experiment A. The table also shows average directional biases for x, y and z directions across all images, for each disease severity. From the table, results indicate that ANTs SyN significantly outperforms VoxelMorph ($p < 0.001$).

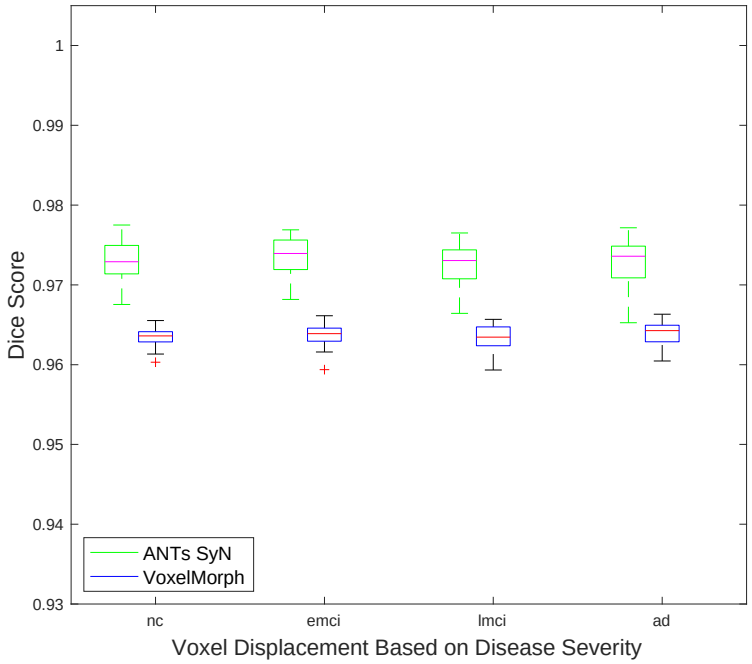

Figure 7: Dice score results of ANTs SyN (green) and VoxelMorph (blue) with increasing voxel displacement represented by increasing disease severity: nc, emci, lmci, and ad.

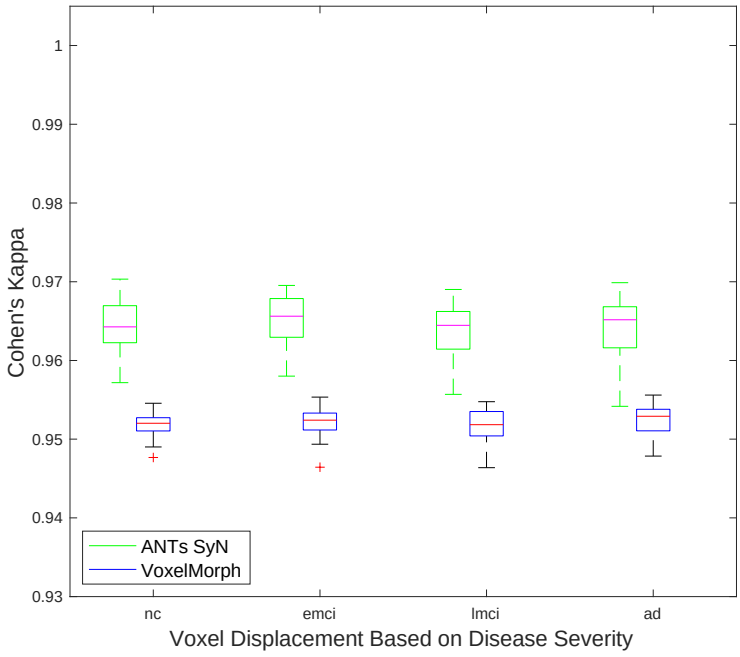

Figure 8: Cohen's Kappa of ANTs SyN (green) and VoxelMorph (blue) with increasing voxel displacement represented by increasing disease severity: nc, emci, lmci, and ad.

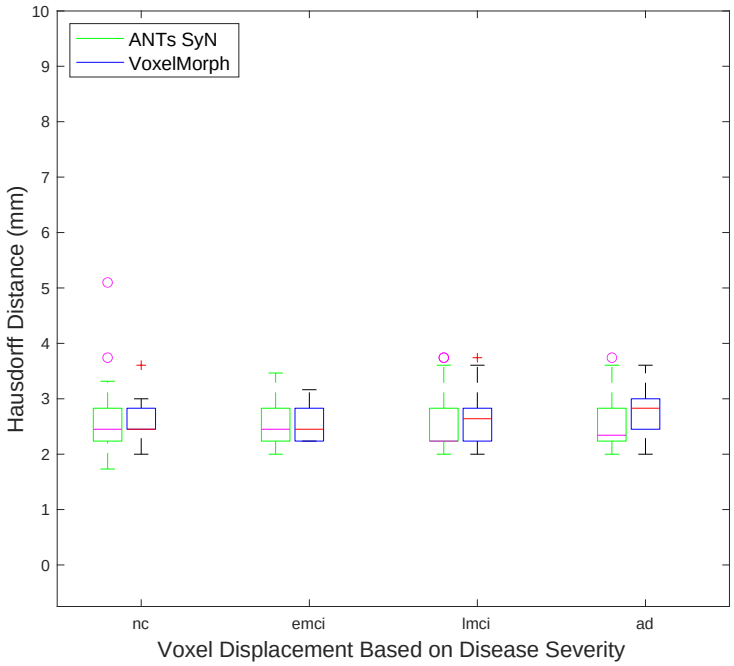

Figure 9: H95 results of ANTs SyN (green) and VoxelMorph (blue) with increasing voxel displacement represented by increasing disease severity: nc, emci, lmci, and ad.

Table 1: Directional biases in x, y and z, and recovery error of VoxelMorph and ANTs SyN from Experiment A, averaged per disease severity: nc, emci, lmci, and ad. Statistically significant results from the better performing method are bolded.

Method	Disease Severity	Bias in x (mm)	Bias in y (mm)	Bias in z (mm)	Recovery error (mm)
VoxelMorph	nc	-0.48	-1.17	0.39	2.173
	emci	-0.49	-0.78	0.40	2.099
	lmci	-0.48	-1.00	0.45	2.111
	ad	-0.36	-1.15	0.43	2.099
ANTs SyN	nc	**-0.10**	**-0.31**	**0.0072**	**0.669**
	emci	**-0.10**	**-0.21**	**-0.0018**	**0.638**
	lmci	**-0.094**	**-0.27**	**0.015**	**0.668**
	ad	**-0.050**	**-0.30**	**0.014**	**0.659**

3.3.2 Experiment B

Figure 10 depicts the whole brain, grey matter, white matter, and CSF Dice score results for both VoxelMorph and ANTs SyN from Experiment B. Results are depicted as a colour map for each label. Dark red indicates a perfect Dice score of 1.0, while dark blue indicates a Dice score

of 0.0. All pairwise comparisons between subjects are done. With all subjects ordered the same way from top to bottom on the vertical axis, and left to right on the horizontal axis, a diagonal of perfect Dice scores is seen between the same volumes. Two-way repeated measures ANOVA reveals that differences in Dice Scores and Cohen's Kappa between methods are statistically significant ($p < 0.001$). The mean Dice score for each method, and tissue type are shown in Table 2, where VoxelMorph performs better than ANTs SyN.

Table 2: Average Dice scores, Cohen's Kappa, and H95 for Experiment B. Averages are listed per method and tissue type. Statistically significant mean differences are in bold for the better performing method.

Evaluation Metric	Tissue Type	$\boldsymbol{\mu}$ VoxelMorph	$\boldsymbol{\mu}$ ANTs SyN	$\Delta\boldsymbol{\mu}$
Dice score	Whole brain	**0.98 ± 0.0075**	0.96 ± 0.015	**0.02**
	Grey matter	**0.62 ± 0.091**	0.58 ± 0.10	**0.04**
	White matter	**0.75 ± 0.060**	0.71 ± 0.060	**0.04**
	CSF	**0.50 ± 0.12**	0.43 ± 0.14	**0.07**
Cohen's Kappa	Whole brain	**0.97 ± 0.011**	0.95 ± 0.021	**0.02**
	Grey matter	**0.59 ± 0.099**	0.54 ± 0.11	**0.05**
	White matter	**0.73 ± 0.065**	0.68 ± 0.079	**0.05**
	CSF	**0.49 ± 0.12**	0.41 ± 0.14	**0.08**
H95 (mm)	Whole brain	7.64 ± 2.77	8.07 ± 2.84	0.43
	Grey matter	11.73 ± 4.04	12.28 ± 4.37	0.55
	White matter	12.75 ± 3.55	12.75 ± 3.65	0.00
	CSF	16.36 ± 4.78	15.88 ± 4.25	0.48

Figure 11 depicts the whole brain, grey matter, white matter, and CSF Cohen's Kappa for both VoxelMorph and ANTs SyN for Experiment B. Colour scales are the same as for Dice scores in Figure 10. Two-way repeated measures ANOVA reveals that differences between methods are statistically significant ($p < 0.001$). The mean Cohen's Kappa for each method, and tissue type are shown in Table 2, where VoxelMorph performs better than ANTs SyN.

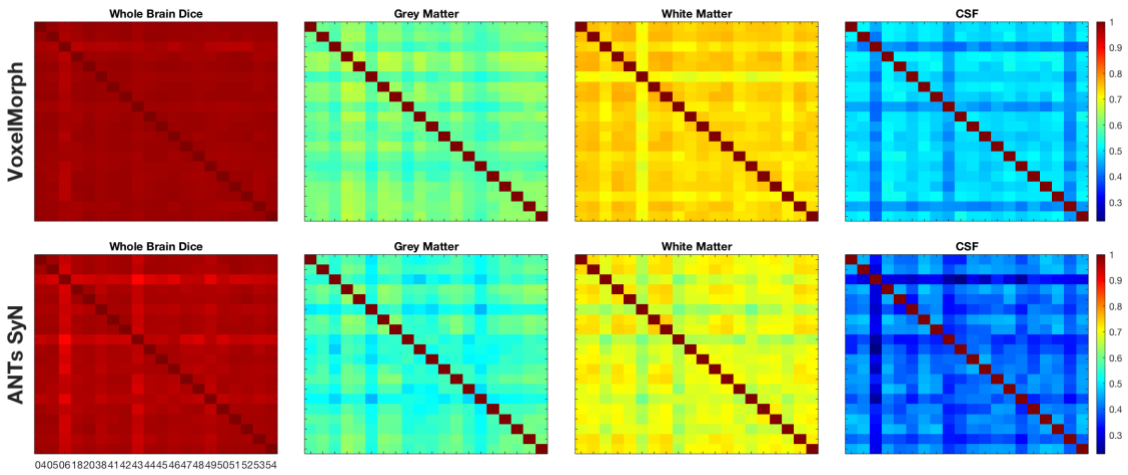

Figure 10: Dice Score of BrainWeb20 Images Registered by VoxelMorph (top) and ANTs SyN (bottom). Dark red values indicate a Dice score of 1, indicating perfect overlap, while dark blue is a score of 0, indicating no overlap.

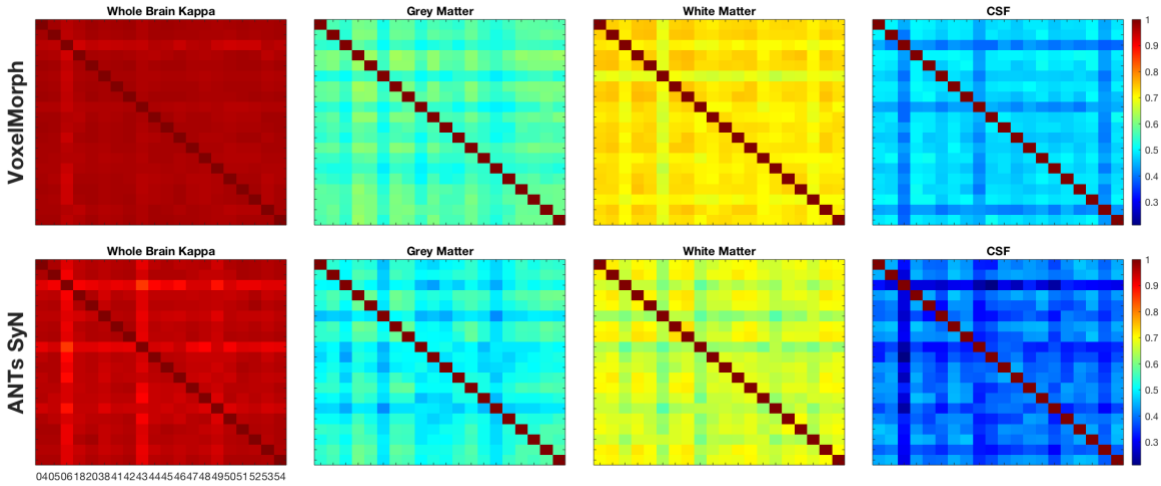

Figure 11: Cohen's Kappa of BrainWeb20 images registered by VoxelMorph (top) and ANTs SyN (bottom). Dark red values indicate a Cohen's Kappa value of 1, indicating perfect overlap, while dark blue is a value of 0, indicating no overlap.

Figure 12 depicts the whole brain, grey matter, white matter, and CSF H95 results for both VoxelMorph and ANTs SyN for Experiment B. Here, Dark blue, or a value of 0 mm, is best. Differences in label type and method are significant from a two-way repeated measures ANOVA ($p < 0.05$); however, while differences between tissue types is significant, a t-test reveals

differences between methods is not significant. The mean H95 for each method, and tissue type are shown in Table 2.

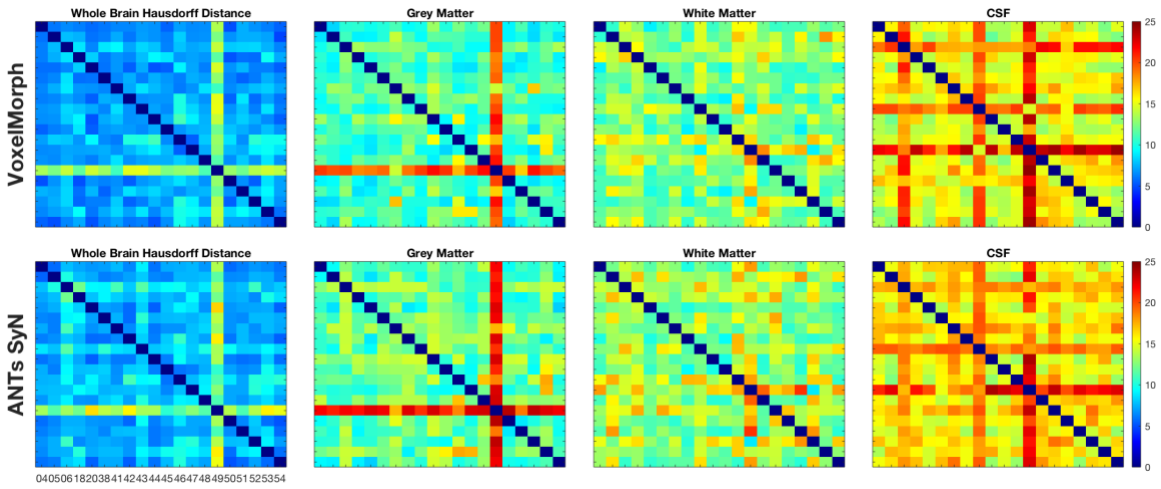

Figure 12: H95 results of BrainWeb20 images registered by VoxelMorph (top) and ANTs SyN (bottom). Dark blue indicates perfect overlap and a maximum error of 0.

3.3.3 Experiment C

Figure 13 depicts the Dice scores of the three cortical structures and three deep grey structures from twenty subjects registered by both VoxelMorph and ANTs SyN from Experiment C. Results are depicted in the same manner as in Figures 10 and 11 from Experiment B. All pairwise comparisons between subjects are done. Two-way repeated measures ANOVA reveals that differences in both structure and method are statistically significant ($p < 0.05$). The mean Dice score for each method, and tissue type are shown in Table 3. It is noted that given the values in Table 3, there appears to be no advantage of one method over the other. Cohen's Kappa results were very similar and are not shown here.

Figure 14 depicts the H95 of the three cortical structures and three deep grey structures for both VoxelMorph and ANTs SyN from Experiment C. Results are depicted in the same manner as

in Figure 12, although the deep grey structures have a smaller scale compared to the scale of the cortical structures (Figure 14). The mean H95 for each method, and tissue type are shown in Table 3. Differences in H95 are significant between structures, but there is no significant difference between methods according to a two-way repeated measures ANOVA.

Table 3: Average Dice scores and H95 for Experiment C. Averages are listed per method and tissue type. Statistically significant mean differences for the better performing method are in bold.

Evaluation Metric	Tissue Type	$\boldsymbol{\mu}$ VoxelMorph	$\boldsymbol{\mu}$ ANTs SyN	$\Delta\boldsymbol{\mu}$
Dice score	Middle temporal gyrus	**0.58 ± 0.12**	0.55 ± 0.12	**0.03**
	Frontal pole	0.43 ± 0.21	**0.44 ± 0.20**	**0.01**
	Post-central gyrus	**0.42 ± 0.16**	0.40 ± 0.17	**0.02**
	Amygdala	0.58 ± 0.14	**0.66 ± 0.10**	**0.06**
	Hippocampus	0.69 ± 0.09	**0.72 ± .08**	**0.03**
	Thalamus	**0.84 ± 0.05**	0.83 ± 0.05	**0.01**
H95 (mm)	Middle temporal gyrus	13.42 ± 3.89	13.28 ± 3.76	0.14
	Frontal pole	9.27 ± 4.23	9.40 ± 4.19	0.13
	Post-central gyrus	14.13 ± 4.59	15.20 ± 5.36	1.07
	Amygdala	5.50 ± 2.32	4.63 ± 2.23	0.87
	Hippocampus	6.41 ± 2.37	5.74 ± 2.21	0.67
	Thalamus	4.77 ± 1.69	4.62 ± 1.48	0.15

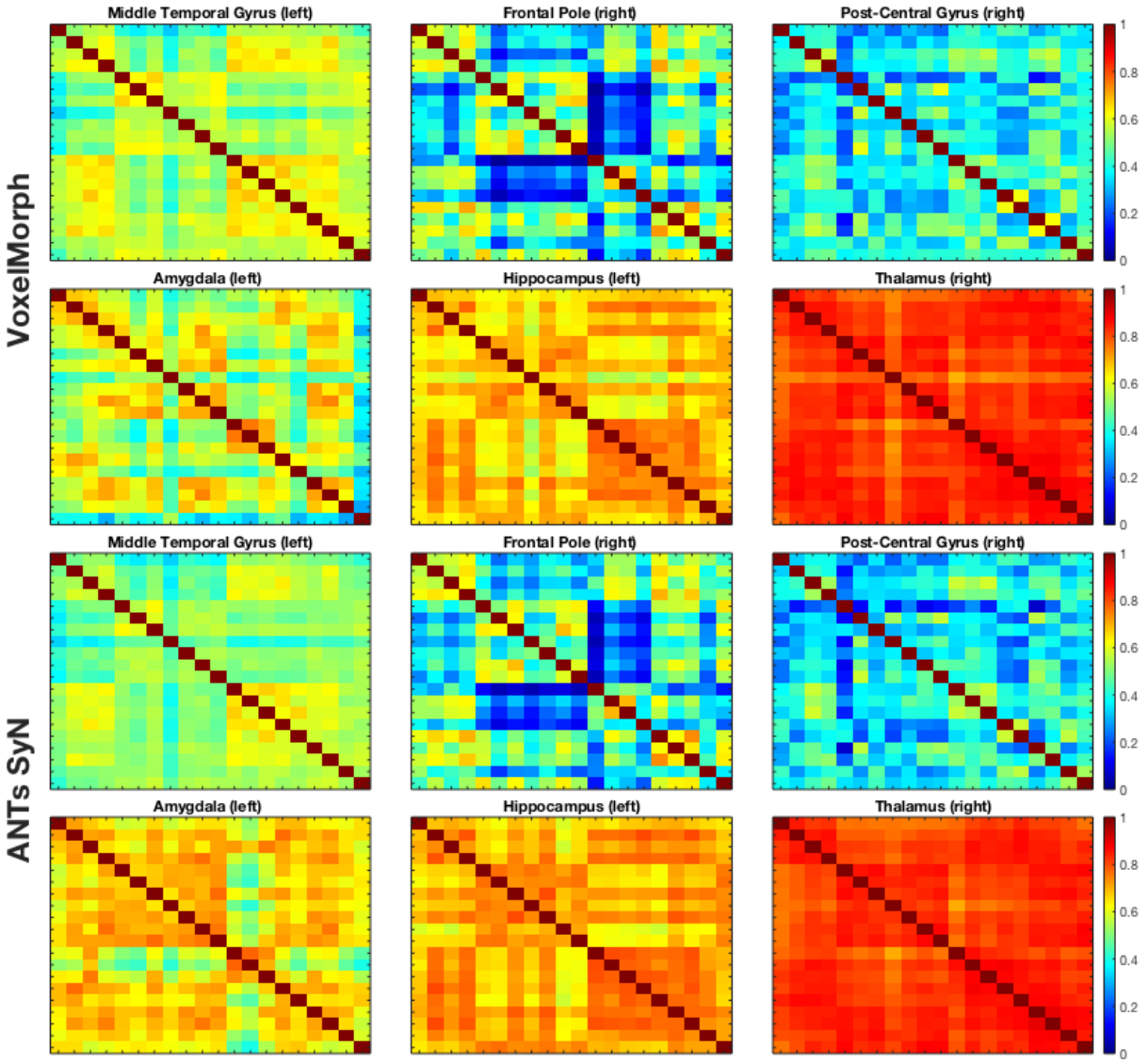

Figure 13: Dice Score of three cortical structures and three deep grey structures from twenty randomly selected Neuromorphometrics volumes registered by VoxelMorph and ANTs SyN.

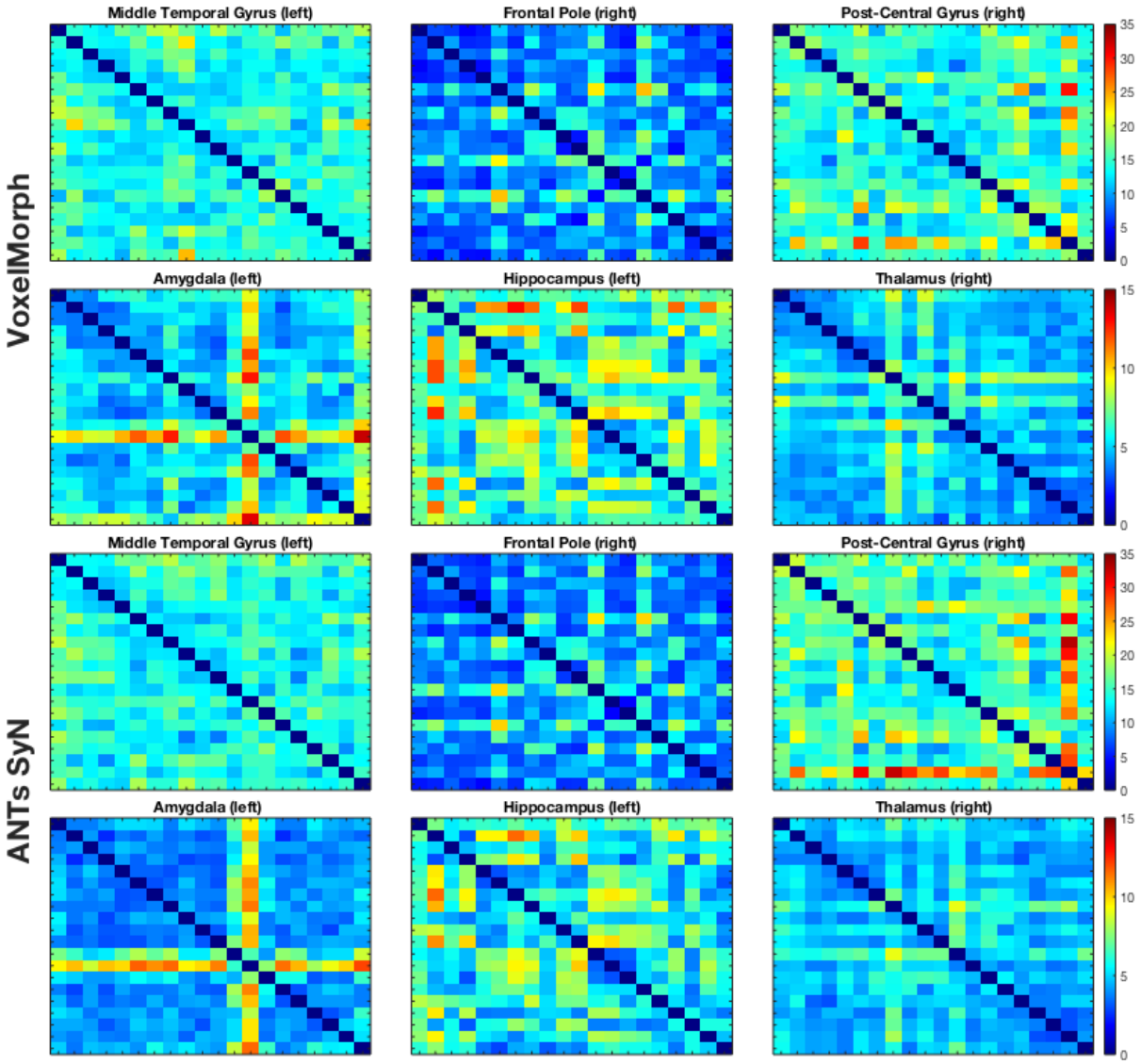

Figure 14: H95 of three cortical structures and three deep grey structures from twenty randomly selected Neuromorphometrics volumes registered by VoxelMorph and ANTs SyN. Note that the cortical structures have H95 results ranging from 0 to 35 mm while deep grey structures have H95 results ranging from 0 to 15 mm.

3.4 Discussion

3.4.1 Experiment A

ANTs SyN demonstrates superior performance in both Dice score and Cohen's Kappa results (Figures 7 and 8, respectively), as well as in recovery error results (Table 1) in Experiment A. The results for Dice score and Cohen's Kappa imply better structure overlap between the original atlas labels when recovering the simulated deformations with ANTs SyN.

Further inspection of the recovered images revealed fewer holes or overlaps from ANTs SyN compared to those from VoxelMorph, and ANTs SyN tends to keep the general shape of gyri and sulci better than VoxelMorph does. In theory, the methods compared should be diffeomorphic, but in practice, approximations in the implementation of each method can give non-diffeomorphic results. Qualitative findings such as those seen in Figure 15, such as the cortical folds and discontinuities in topology, seen in all three subjects, could indicate that VoxelMorph did not produce diffeomorphic registrations in all of the registration tasks. Figure 15 shows examples of cases where sulci do not take on a sulci-like shape and gyri do not take on a gyri-like shape (examples are bound in red boxes within the figure).

Interestingly, the results from Dice score, Cohen's Kappa, and recovery error were statistically significant when comparing methods, but not when taking into account disease severity according to the two-way ANOVAs calculated. This shows that both methods are capable of recovering images with simulated voxel displacements equally, according to their respective performance levels.

Recovery error results from Table 1 show much smaller average error and directional bias with ANTs SyN compared to VoxelMorph. H95 results (Figure 9) were not statistically significant from a two-way repeated measures ANOVA. This implies that errors of ANTs SyN and VoxelMorph are similar.

A further inspection of H95 results from ANTs SyN does reveal greater variability among images with smaller deformations (see Figure 9). One might expect the worst performing H95 measurements to be of volumes deformed with AD transformations since recovered transformation would have to make larger local deformations in order to compensate for the large ventricles in AD patients. But this was not the case as seen in Figures 7, 8 and 9. It is suspected that variability among results here were mostly due to anatomical differences shown in ADNI-ANIMAL deformations, such as larger protruding temporal regions compared to the original VoxelMorph atlas, which would produce such results.

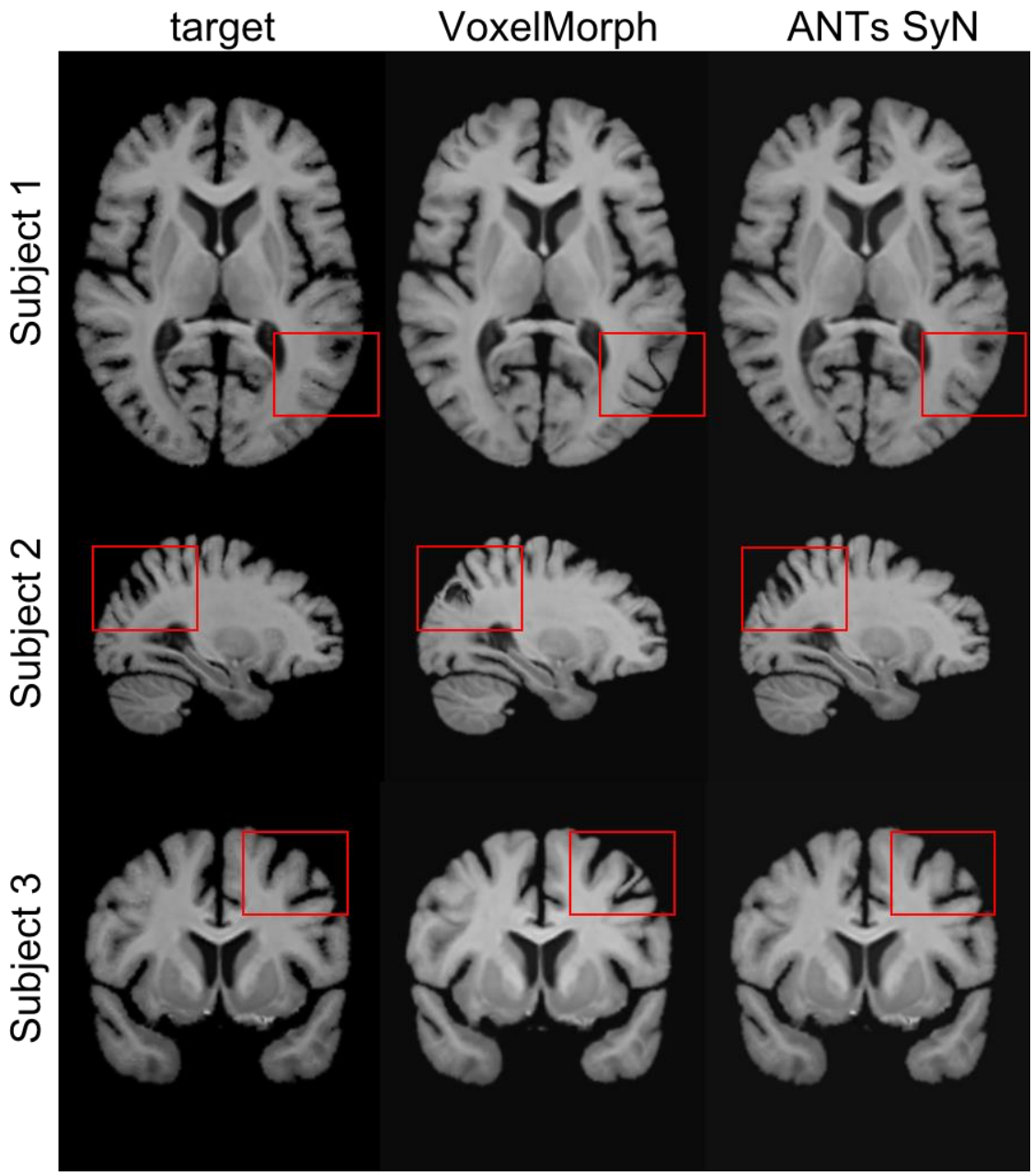

Figure 15: Comparison of three different recovered MRI volumes (rows 1, 2 and 3) from Experiment A, where red boxes show discontinuities in VoxelMorph recovered volumes (column 2) compared to ANTs SyN recovered volumes (column 3) and the target atlas (column 1).

3.4.2 Experiment B

VoxelMorph appears to significantly outperform ANTs SyN, as shown in Dice scores and Cohen's Kappa results on whole brain, grey matter, white matter, and CSF labels in Experiment B (Figures 10 and 11, respectively). The mean differences between the two methods for both metrics can be seen in Table 2. While the differences between methods are significant ($p < 0.001$), the metrics on whole brain, grey and white matter labels do not differ greatly, therefore it can be said that both methods perform at similar levels, with VoxelMorph having a small but statistically significant advantage. Results comparing H95 metrics between methods and labels found no significance in a two-way repeated measures ANOVA. Although results were visually indistinguishable from H95 results, they were only significant between tissue type and not between methods.

Notable discrepancies, where results for some subjects appeared worse than others, seemed to be due to larger inter-subject variability for both methods in all metrics (Figures 10, 11, and 12). Anatomical variabilities such as crooked midlines, smaller total white matter area, and asymmetrical right and left hemispheres, for subjects 06, 38 and 49, can explain the poorer results observed. Finally, it is noted that the worst results for all three indirect evaluation metrics are seen for CSF labels, since these labels generally show the most inter-subject variability.

3.4.3 Experiment C

Experiment C shows mixed performances for both VoxelMorph and ANTs SyN on Dice, Cohen's Kappa and H95 results; however, two 2-way repeated measures ANOVAs reveal results are statistically significant for Dice and Cohen's Kappa results ($p < 0.05$). It is noted that as with Experiment B, H95 results (Figure 14) differences are driven by structure type and not by methods.

To aid in deciphering which method outperformed the other, mean Dice scores are reported in Table 3 where ANTs SyN outperforms VoxelMorph in three of the six structures; namely, the right frontal pole, the left amygdala, and the left hippocampus (Figure 13). On the other hand, VoxelMorph reports better Dice scores on the left middle temporal gyrus, the right post-central gyrus, and the right thalamus. Differences in mean Dice scores of each structure can help to discern which method performs best, and since both mean and median values are similar, indicating an unskewed distribution, these results are trustworthy. For structures with mean Dice score differences that are small, for example 0.03 or smaller, it could be said that these results are comparable since the difference in structure overlap is so small. However, a closer look at differences in recovered amygdala structures, where the mean difference was larger – 0.06, indicate ANTs SyN is more capable of maintaining accurate structure alignment.

For the H95 results in Table 3, mean differences show that ANTs SyN outperforms VoxelMorph in four structures: the left middle temporal gyrus, left amygdala, left hippocampus, and right thalamus (Figure 14). VoxelMorph outperformed ANTs SyN in the registration of the right frontal pole and right post-central gyrus, with post-central gyrus having a difference of 1.07 mm. Differences below 1 mm (i.e., the voxel size) are arguably very small, and it can be said that the performance of the methods is comparable. A closer look at differences in the right post-central gyrus, which exceeded 1 mm, indicate that VoxelMorph normalizes inter-subject variability more than ANTs SyN. Again, structure type drove the variability among results for H95, and thus these results do not aid in choosing the better registration method.

Figures 13 and 14 also highlight inter-database variability in Dice scores for middle temporal gyrus and hippocampus, seen as "blocks" along their respective colour maps. For context, Neuromorphometrics is comprised of volumes from four different databases: ADNI [20], a 20

Repeats dataset of 20 nondemented subjects scanned on two visits within 90 days and labelled by two raters; the Child and Adolescent NeuroDevelopment Initiative (CANDI) database [103]; and the Open Access Series of Imaging Studies (OASIS) database [88]. Of the twenty randomly selected subjects from the Neuromorphometrics database, four were from the 20 Repeats dataset, seven were from ADNI [20], three were from CANDI [103], and six were from OASIS [88]. Most notably in this "inter-database" variability are the results from CANDI volumes [103], seen in frontal pole Dice scores, and amygdala Dice scores and H95. These discrepancies are to be expected since the volumes from the CANDI dataset [103] are from adolescents and children, who have different ratios of grey to white matter compared to adults [104]. Other discrepancies in both Dice scores and H95 for both methods appear to be due to variability between subjects which could not be completely normalized by either method.

One understandable limitation present in all three experiments was the use of the pretrained VoxelMorph model. Had a subset of this data been used to train VoxelMorph, it is possible that registrations would have improved performance from indirect and direct metrics in all three experiments. However, it must be said that the model was trained on 3731 T1-weighted brain MRI scans from eight datasets, including OASIS [88]. Therefore, one would assume that the model is capable of registration from a wide range of normal subjects. This could explain VoxelMorph's slightly better performance in Experiments B and C; and since Experiment A contained "non-normal" deformations, this could explain VoxelMorph's poor performance. Ultimately, it can be said that for either method further parameter tuning, or training, could yield improved results.

3.5 Conclusion

The objective of this paper was to compare VoxelMorph to the state-of-the-art ANTs SyN. While Experiment A showed that ANTs SyN outperforms VoxelMorph when recovering simulated deformations, with recovery errors approximately 1.5 mm less than VoxelMorph, Experiments B and C demonstrate that VoxelMorph normalizes anatomical variability between subjects better than ANTs SYN. VoxelMorph achieves comparable results to ANTs SyN in a small fraction of the time, but results have suggested that transformations may not always be diffeomorphic; however, if the given task relies heavily on topological preservation, taking the time to perform a diffeomorphic registration from ANTs SyN may be preferred. Future work on quantitative evaluation of diffeomorphisms could effectively determine if one method outperforms another.

3.6 Acknowledgements

This study was supported by grants from CIHR, NSERC, and The Healthy Brains for Healthy Lives Foundation.

CHAPTER 4

Discussion, Future Work, and Conclusions

4.1 Discussion

This thesis compared the registration performance of VoxelMorph to ANTs SyN. The previous chapter contained what may be published in a paper submitted to *Medical Image Analysis*, which explains the methods used to achieve this evaluation and presented key findings. This chapter will review the methods and key findings of the previous chapter in terms of the contributions, strengths, and limitations of the methods.

4.1.1 Experiment A

Key findings from Experiment A include better performance from ANTs SyN, shown in both Dice score and Cohen's Kappa coefficient (Figures 7 and 8, respectively), as well as recovery error results (Table 1), with statistical significance between methods ($p < 0.001$). Firstly, the results for Dice score and Cohen's Kappa coefficient imply better structure overlap with the original VoxelMorph atlas labels when the recovered deformation comes from ANTs SyN.

From inspection, it is noted that both the Dice scores and Cohen's Kappa results show significantly better performance for ANTs SyN since it appears that ANTs SyN makes fewer holes or overlaps in the recovered labels and tends to keep the general shape of the gyri and sulci compared to VoxelMorph. Examples of this can be seen from Figure 15. This could indicate that

VoxelMorph did not produce diffeomorphic registrations in all of the registration tasks. Figure 15 shows examples of cases where sulci do not take on a sulci-like shape and gyri do not take on a gyri-like shape (examples are bound in red boxes within the figure).

Interestingly, the results from Dice score, Cohen's Kappa, and recovery error were statistically significant when comparing method to method, but not when taking into account disease severity according to the two-way ANOVAs calculated. Lack of variability in disease severity shows that both methods are capable of recovering images of varying voxel displacements to their respective performance levels equally.

H95 results (Figure 9) were not statistically significant from a two-way repeated measures ANOVA. This implies that the maximum error of ANTs SyN and VoxelMorph are similar. Additionally, recovery error results show smaller error and directional bias from ANTs SyN, where the recovery error is calculated from the average error across the entire image. So, although both methods have similar maximum error, overall ANTs SyN performs better in terms of average error, structure overlap, as well as directional bias in the presence of deformation.

However, it is noted that ANTs SyN appears to have greater variability in H95 results (see Figure 9). One might expect the worst performing H95 measurements to be of volumes deformed with AD transformations. This would be because the recovered transformation would have to make larger local deformations in order to compensate for the large ventricles in AD patients. But upon further inspection, it was noted that this was not the case, as previously discussed. Further investigation reveals the worst performing H95 cases are due to inter-subject variability, with brain shapes from ADNI-ANIMAL transformations recovering larger protruding temporal regions, and thus producing larger voxel displacement in the simulated database. It is possible that the

resampling technique used with ANTs SyN (i.e. the Minc Toolkit [12]) did not have enough iterations in its default settings in order to resample the voxels correctly. However, there is no sign of mirroring or repetition[3] in all of the recovered MRI or label volumes.

4.1.2 Experiment B

Key findings from Experiment B include better performance from VoxelMorph compared to ANTs SyN, as shown in Dice scores and Cohen's Kappa results on whole brain, grey matter, white matter, and CSF labels (see Figures 10 and 11). The mean differences between both methods for both metrics can be seen in Table 2. While the differences between methods are significant ($p < 0.001$), the errors on whole brain, grey and white matter labels do not differ greatly, therefore it can be said that both methods perform well, with VoxelMorph having a small but statistically significant advantage. Results comparing H95 metrics between methods and labels found no significance in a two-way repeated measures ANOVA. Although results were visually indistinguishable from H95 results, they were only significant between tissue type alone, however this has no bearing on the comparison between methods (see Figure 12).

Furthermore, discrepancies in the Dice, Cohen's Kappa, and H95 results were investigated. Most notable discrepancies where results stood out as being poorer than the rest of the results seemed to be due to (extreme) inter-subject variability. For example, subject 06 has a crooked midline which affected results seen in Dice and Cohen's Kappa results for whole brain and CSF labels for both methods (see Figures 10 and 11). Subject 38 has a smaller total white matter area compared to other subjects and hence results are noticeable for both methods in both Dice and

[3] Mirroring or repetition is observed in images where fewer resampling iterations result in a mirrored image of the brain or part of the brain across the x y or z axis.

Cohen's Kappa results (see Figures 10 and 11). Finally, subject 49 has highly asymmetrical right and left hemispheres, and thus results are noticeably worse for H95 results of both methods in particular (see Figure 12). Finally, it is noted that in general the worst results for all three indirect evaluation metrics are seen for CSF labels, since these labels will show the most inter-subject variability.

Despite the discrepancies mentioned above, overall VoxelMorph does slightly outperform ANTs SyN. But it is noted that these results are quite different than those observed from Experiment A. From Experiment A, it was observed that ANTs SyN outperformed VoxelMorph in the presence of deformation. This could be due to the fact that a one-to-one mapping exists and is known, since deformations were simulated using transformations from ANIMAL [13], but between subjects a one-to-one homology is not guaranteed. A lack of a guaranteed one-to-one homology between subjects also explains why results from the same evaluation metrics were worse in Experiment B compared to Experiment A. Specifically with the results of the H95 results, the average whole brain H95 is approximately 3 mm in Experiment A, but is 7-8 mm in Experiment B. Worse performance could be due to stray individual voxels detached from the main structure despite using 95^{th} percentile of H95 results.

Experiments A and B are both limited to evaluating each method using whole brain labels. Results from Experiment B confound conclusions on which method performs best for an inter-subject registration task, despite using grey and white matter as well as CSF labels in its evaluation. Perhaps smaller brain structure labels would help in the evaluation of the performance of each method, which can be observed in Experiment C.

4.1.3 Experiment C

Key findings from Experiment C shows mixed performances for both VoxelMorph and ANTs SyN on Dice, Cohen's Kappa and H95 results; however, two 2-way repeated measures ANOVAs reveal results are statistically significant for Dice and Cohen's Kappa results ($p < 0.05$). Results from Cohen's Kappa were not shown as they are very similar to those of Dice. It is noted that as with Experiment B, H95 results have variability driven mostly by structure type and not between methods, as there is no significant variability between methods alone (see Figure 14).

To aid in deciphering which method outperformed the other, mean Dice scores are reported in Table 3. From the table, ANTs SyN outperforms VoxelMorph in three of the six structures; namely, the right frontal pole, the left amygdala, and the left hippocampus (see Figure 13). Otherwise, VoxelMorph reports better dice scores on the left middle temporal gyrus, the right post-central gyrus, and the right thalamus (see Figure 13). The differences between the mean Dice scores of each structure can give indication of the performance of each method since both mean and median values are similar, indicating an unskewed distribution. For structures with mean Dice score differences of 0.03 and lower, it could be said that these results are comparable since the difference in structure overlap is so small. However, a closer look at differences in recovered amygdala structures, where the mean difference was 0.06, indicate ANTs SyN is more capable of maintaining structure topology, and can also explain the poorer results observed from VoxelMorph.

For H95 results in Table 3, mean differences show ANTs SyN outperforms VoxelMorph in four structures: the left middle temporal gyrus, left amygdala, left hippocampus, and right thalamus (see Figure 14). VoxelMorph outperformed ANTs SyN in the registration of the right frontal pole

and right post-central gyrus, with post-central gyrus having a difference of 1.07 mm. Differences below 1 mm are arguably very small, and it can be said that the performance of the methods is comparable. A closer look at differences in the right post-central gyrus, which exceeded 1 mm, indicate that VoxelMorph normalizes inter-subject variability more than ANTs SyN. Again, structure type drove the variability among results for H95, and thus results may not be useful to select the best method for registration.

Further investigation into visual discrepancies in Figures 13 and 14 was performed. Specifically, it was noted that Dice scores for middle temporal gyrus and hippocampus have similar Dice scores which can be seen in blocks along their respective colour maps. To understand this, it must be said that Neuromorphometrics is comprised of volumes from four different databases: ADNI [20], 20 Repeats dataset of 20 nondemented subjects scanned on two visits within 90 days and labelled by two raters, the Child and Adolescent NeuroDevelopment Initiative (CANDI) database [103], and the Open Access Series of Imaging Studies (OASIS) database [88]. Of the twenty randomly selected subjects from the Neuromorphometrics database, four were from the 20 Repeats dataset, seven were from ADNI, three were from CANDI [103], and six were from OASIS [88]. It appears that some structures tend to have similar Dice scores within these datasets. Notably, CANDI [103] seems to perform poorly in comparison to other datasets, of which these findings can be seen from the frontal pole Dice scores, and the amygdala Dice scores and H95 results. These discrepancies are to be expected since the volumes from the CANDI [103] dataset are from adolescents and children, who have different ratios of grey to white matter compared to adults [104]. The fact that CANDI [103] datasets perform slightly better for VoxelMorph then with ANTs SyN indicates again that VoxelMorph is better at normalizing inter-subject variability. Other

discrepancies in both Dice scores and H95 results for both methods appear to be due to variability between subjects which was unable to be completely normalized by either method.

Comparing Experiment C to Experiments A and B, results are worse for both methods; however, this is expected since the labels used in this experiment are of smaller structures, and therefore overlap of these structures is worse than with larger areas such as white matter or whole brain labels. It is also noted that H95 results of deep grey structures from this method compared well against those from Experiment B, but cortical grey structures performed poorly. This can be expected since there may be larger deformations to normalize cortical structure between subjects, but less so in some deep grey structures, and from these larger deformations maximum error can occur from individual voxels which are not connected to the main cortical structure.

The implications of the registration time per subject also needs to be considered when determining which method is better. ANTs SyN registers a single subject in roughly 37 to 43 minutes on a Xeon E3-1275 V6 CPU with 3.80 GHz clock speed. While VoxelMorph requires several days of training [5], but is able to register a single subject in 30 seconds to 1 minute on an NVIDIA TitanX GPU. In applications where fast computations times are important, such as with image guided neurosurgery [33], VoxelMorph is advantageous over ANTs SyN. But where precise alignment is important, ANTs SyN is more favourable.

One understandable limitation present in all three experiments was the use of the pretrained VoxelMorph model. Had a subset of the data been used to train VoxelMorph, it is possible that registrations would appear diffeomorphic, and there would be improved performance from indirect and direct metrics in all three experiments. Although the model was trained on 3731 T1-weighted brain MRI scans from eight datasets, including OASIS [88], one would assume that the model is

capable of registration of normal subjects. This would explain VoxelMorph's slightly better performance in Experiments B and C; and since Experiment A contained "non-normal" deformations, this would explain VoxelMorph's poor performance. Ultimately, it can be said for either method, that further parameter tuning, or training, could yield improved results.

Evaluating reproducibility requires several measurements, not only the ability to repeat a study [1]. The intention of this thesis is to stress the importance of reproducibility in science, since the validation of novel publications drives research forward in a pragmatic way. The hypothesis of this thesis was that the validation methods used to compare the performance VoxelMorph and ANTs SyN would be enough to determine which method performs best. The results which have been discussed do show that the validation methods were sufficient to determine which method is better depending on the registration task. The results suggest a trade-off between methods in terms of computation speed and quality of results.

4.2 Future Work

The evaluation performed in this thesis can hopefully highlight the capabilities and drawbacks of both VoxelMorph and ANTs SyN. In the future, further quantitative evaluation on the diffeomorphic properties of the transformations will be performed in order to assess whether both methods are producing diffeomorphic registrations for all experiments. For example, calculating the Jacobian determinant over the entire deformation field can assess whether the registrations are in fact diffeomorphic, as discussed in Chapter 2 of this thesis. Additionally, perhaps a VoxelMorph model, trained using subsets of data from experiments, could provide improved results, and fine-tuning parameters of ANTs SyN without hampering its capabilities, are also possible for expected improvements.

Finally, VoxelMorph provides a solid building block for potential applications in intraoperative image registration to correct for brain shift during image guided surgery tasks. It would be interesting to observe the capabilities of deep learning registration when given imaging data with brain tumours, since tumour location is patient specific. The advantages of fast (i.e. in seconds) non-linear registration of brain images during surgery reveals the advantages of deep learning while potentially maintaining the quality of classic methods.

4.3 Conclusions

The objective of this thesis was to compare VoxelMorph to the state-of-the-art ANTs SyN. The validation techniques employed to evaluate both methods were sufficient in determining which method outperforms the other. Specifically, ANTs SyN is a preferable method when dealing with registration tasks which have deformations present. Otherwise, there is a trade-off in terms of registration quality and computation time depending on if ANTs SyN or VoxelMorph is chosen, respectively. In applications where fast computations times are important, VoxelMorph may be considered; but where precise alignment is important, ANTs SyN is more favourable. It is recommended that the reader reference the Background and Results Chapters of this thesis as a guide in selecting an appropriate method for a given registration task. Additionally, the reader can take away the validation methods employed in this thesis and apply them to other comparisons of medical image processing tools.

In the evaluation of VoxelMorph and ANTs SyN, it was demonstrated that ANTs SyN outperforms VoxelMorph when recovering deformed images, and more specifically, ANTs SyN achieves smaller average error, structure overlap, as well as directional bias. Furthermore, both methods achieve comparable structure overlap and maximum error for inter-subject registration

tasks. While VoxelMorph appears to normalize anatomical variability between subjects, qualitative assessments reveal potential non-diffeomorphisms which were comparatively not seen in the same subjects recovered from ANTs SyN. Thus, for inter-subject registration tasks, it appears there is a trade-off between methods. At the risk of recovering some non-diffeomorphic registrations, VoxelMorph achieves comparable results to ANTs SyN in a fraction of the time; however, if the registration task relies heavily on topological preservation, taking the time to perform a diffeomorphic registration from ANTs SyN may be preferred.

Future work on quantitative evaluation of diffeomorphisms or spending more time on improving results through highly specific fine-tuning of parameters as well as training, could effectively determine if one method stands out from the other and improve the evaluation of reproducibility of these techniques.

References

[1] Open Science Collaboration, “Estimating the reproducibility of psychological science,” *Science (80-.).*, vol. 349, no. 6251, pp. 943-aac4716-9, 2015.

[2] L. Ibáñez, R. Avila, and S. Aylward, “Open source and open science: How it is changing the medical imaging community,” in *2006 3rd IEEE International Symposium on Biomedical Imaging: From Nano to Macro - Proceedings*, 2006, vol. 2006, pp. 690–693.

[3] T. S. Yoo and D. N. Metaxas, “Open science - Combining open data and open source software: Medical image analysis with the Insight Toolkit,” *Medical Image Analysis*. 2005.

[4] G. Haskins, U. Kruger, and P. Yan, “Deep learning in medical image registration: a survey,” *Mach. Vis. Appl.*, vol. 31, no. 8, pp. 1–18, 2020.

[5] A. V Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, “Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces,” *Med. Image Anal.*, vol. 57, pp. 226–236, 2019.

[6] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, “Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain,” *Med. Image Anal.*, vol. 12, no. 1, pp. 26–41, 2009.

[7] FreeSurfer, “FreeSurfer,” 2013. [Online]. Available: https://surfer.nmr.mgh.harvard.edu/. [Accessed: 29-Aug-2020].

[8] A. Fenster and B. Chiu, “Evaluation of segmentation algorithms for medical imaging,”

Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc., vol. 7, pp. 7186–7189, 2005.

[9] Berengere Aubert-Broche, M. Griffin, G. B. Pike, A. C. Evans, and D. L. Collins, "Twenty New Digital Brain Phantoms for Creation of Validation Image Data Bases," *IEEE Trans. Med. Imaging*, vol. 25, no. 11, pp. 1410–1416, Nov. 2006.

[10] V. S. Caviness, N. Theodore Lange, N. Makris, M. Reed Herbert, and D. Nelson Kennedy, "MRI-based brain volumetrics: Emergence of a developmental brain science," *Brain and Development*, vol. 21, no. 5. pp. 289–295, 01-Jul-1999.

[11] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Med. Imaging*, vol. 15, p. 29, 2015.

[12] R. D. Vincent *et al.*, "MINC 2.0: A Flexible Format for Multi-Modal Images," *Front Neuroinform*, vol. 10, p. 35, 2016.

[13] D. L. Collins and A. C. Evans, "ANIMAL: validation and applications of non-linear registration-based segmentation," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 11, no. 8, pp. 94–116, 1997.

[14] B. Aubert-Broche, M. Griffin, G. B. Pike, A. C. Evans, and D. L. Collins, "Twenty new digital brain phantoms for creation of validation image data bases," *IEEE Trans. Med. Imaging*, vol. 25, no. 11, pp. 1410–1416, 2006.

[15] F. P. M. Oliveira, C. Foundation, and J. M. R. S. Tavares, "Medical image registration : A review," *Comput. Methods Biomech. Biomed. Eng. ISS*, vol. 5842, no. December 2013, 2014.

[16] B. Zitová and J. Flusser, “Image registration methods: A survey,” *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, Oct. 2003.

[17] J. B. A. Maintz and M. A. Viergever, “A survey of medical image registration,” *Med. Image Anal.*, vol. 2, no. 1, pp. 1–36, 1998.

[18] W. R. Crum and D. L. G. Hill, “Non-rigid image registration: theory and practice,” *Br. J. Radiol.*, vol. 77, pp. S140–S153, 2004.

[19] V. Beliveau *et al.*, “A high-resolution in vivo atlas of the human brain’s serotonin system,” *J. Neurosci.*, vol. 37, no. 1, pp. 120–128, 2017.

[20] A. D. N. Initiative, “ADNI | Alzheimer’s Disease Neuroimaging Initiative,” 2003. [Online]. Available: http://adni.loni.usc.edu/. [Accessed: 22-May-2020].

[21] A. D. Leow *et al.*, “Longitudinal stability of MRI for mapping brain change using tensor-based morphometry,” *Neuroimage*, vol. 31, no. 2, pp. 627–640, 2006.

[22] A. Gooya, G. Biros, and C. Davatzikos, “Deformable registration of glioma images using em algorithm and diffusion reaction modeling,” *IEEE Trans. Med. Imaging*, vol. 30, no. 2, pp. 375–390, 2011.

[23] M. Staring, S. Klein, M. A. Viergever, J. P. W. Pluim, and U. A. van der Heide, “Registration of Cervical MRI Using Multifeature Mutual Information,” *IEEE Trans. Med. Imaging*, vol. 28, no. 9, pp. 1412–1421, 2009.

[24] W. C. Lavely *et al.*, “Phantom validation of coregistration of PET and CT for image-guided radiotherapy,” *Med. Phys.*, vol. 31, no. 5, pp. 1083–1092, 2004.

[25] M. Foskey *et al.*, “Large deformation three-dimensional image registration in image-guided radiation therapy,” *Phys. Med. Biol.*, vol. 50, no. 24, pp. 5869–5892, Dec. 2005.

[26] A. van der Hoorn, J. L. Yan, T. J. Larkin, N. R. Boonzaier, T. Matys, and S. J. Price, “Validation of a semi-automatic co-registration of MRI scans in patients with brain tumors during treatment follow-up,” *NMR Biomed.*, vol. 29, no. 7, pp. 882–889, Jul. 2016.

[27] X. Huang, J. Ren, G. Guiraudon, D. Boughner, and T. M. Peters, “Rapid dynamic image registration of the beating heart for diagnosis and surgical navigation,” *IEEE Trans. Med. Imaging*, vol. 28, no. 11, pp. 1802–1814, Nov. 2009.

[28] A. P. King *et al.*, “Registering preprocedure volumetric images with intraprocedure 3-D ultrasound using an ultrasound imaging model,” *IEEE Trans. Med. Imaging*, vol. 29, no. 3, pp. 924–937, Mar. 2010.

[29] A. Hurvitz and L. Joskowicz, “Registration of a CT-like atlas to fluoroscopic X-ray images using intensity correspondences,” *Int. J. Comput. Assist. Radiol. Surg.*, vol. 3, no. 6, pp. 493–504, 2008.

[30] S. Nandish, G. Prabhu, and K. V Rajagopal, “Multiresolution image registration for multimodal brain images and fusion for better neurosurgical planning,” *Biomed. J.*, vol. 40, pp. 329–338, 2017.

[31] J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes, *Medical image registration*, vol. 46. 2001.

[32] C. R. Maurer, J. Michael Fitzpatrick, M. Y. Wang, R. L. Galloway, R. J. Maciunas, and G. S. Allen, “Registration of head volume images using implantable fiducial markers,” *IEEE*

Trans. Med. Imaging, vol. 16, no. 4, pp. 447–462, 1997.

[33] S. Drouin *et al.*, "IBIS: an OR ready open-source platform for image-guided neurosurgery," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 3, pp. 363–378, 2017.

[34] J. Talairach and P. Tournoux, *Co-Planar Stereotaxic Atlas of the Human Brain: 3-D Proportional System: An Approach to Cerebral Imaging*. 1988.

[35] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans, "Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space," *J. Comput. Assist. Tomogr.*, vol. 18, no. 2, pp. 192–205, 1994.

[36] A. C. Evans, A. L. Janke, D. L. Collins, and S. Baillet, "Brain templates and atlases," *Neuroimage*, vol. 62, no. 2, pp. 911–922, 2012.

[37] J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, and J. Lancaster, "A probabilistic atlas of the human brain: Theory and rationale for its development," *Neuroimage*, vol. 2, no. 2, pp. 89–101, 1995.

[38] R. P. Woods, J. C. Mazziotta, and S. R. Cherry, "Mri-pet registration with automated algorithm," *J. Comput. Assist. Tomogr.*, vol. 17, no. 4, pp. 536–546, 1993.

[39] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans, "Automatic 3D intersuject Registration fo MR Volumetric Data in Standardized Talairach Space," *Journal of Computer Assisted Tomography*, vol. 18, no. 2. pp. 192–205, 1994.

[40] G. Song, J. Han, Y. Zhao, Z. Wang, and H. Du, "A Review on Medical Image Registration as an Optimization Problem," *Curr. Med. Imaging Rev.*, vol. 13, no. 3, pp. 274–283, 2017.

[41] A. P. Keszei, B. Berkels, and T. M. Deserno, “Survey of Non-Rigid Registration Tools in Medicine,” *Journal of Digital Imaging*, vol. 30, no. 1. pp. 102–116, 2017.

[42] M. Wang and P. Li, “A Review of Deformation Models in Medical Image Registration,” *J. Med. Biol. Eng.*, vol. 39, no. 1, pp. 1–17, 2019.

[43] A. Rasoulian, P. Abolmaesumi, and P. Mousavi, “Feature-based multibody rigid registration of CT and ultrasound images of lumbar spine,” *Med. Phys.*, vol. 39, no. 6, pp. 3154–3166, May 2012.

[44] A. Sotiras, C. Davatzikos, and N. Paragios, “Deformable medical image registration: A survey,” *IEEE Trans. Med. Imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[45] G. P. Penney, J. Weese, J. A. Little, P. Desmedt, D. L. G. Hill, and D. J. Hawkes, “A comparison of similarity measures for use in 2-D-3-D medical image registration,” *IEEE Trans. Med. Imaging*, vol. 17, no. 4, pp. 586–595, 1998.

[46] H. J. Johnson and G. E. Christensen, “Consistent landmark and intensity-based image registration,” in *IEEE Transactions on Medical Imaging*, 2002, vol. 21, no. 5, pp. 450–461.

[47] F. P. M. Oliveira, A. Sousa, R. Santos, and J. M. R. S. Tavares, “Spatio-temporal alignment of pedobarographic image sequences,” *Med. Biol. Eng. Comput.*, vol. 49, no. 7, pp. 843–850, 2011.

[48] X. Papademetris, A. P. Jackowski, R. T. Schultz, L. H. Staib, and J. S. Duncan, “Integrated intensity and point-feature nonrigid registration,” in *Lecture Notes in Computer Science*, 2004, vol. 3216, no. PART 1, pp. 763–770.

[49] D. D. B. Carvalho *et al.*, “Joint intensity-and-point based registration of free-hand B-mode ultrasound and MRI of the carotid artery,” *Med. Phys.*, vol. 41, no. 5, pp. 52904–52905, 2014.

[50] A. Klein *et al.*, “Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration,” *Neuroimage*, vol. 46, no. 3, pp. 786–802, 2009.

[51] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, “Computing large deformation metric mappings via geodesic flows of diffeomorphisms,” *Int. J. Comput. Vis.*, vol. 61, no. 2, pp. 139–157, 2005.

[52] M. C. Chiang *et al.*, “Fluid registration of medical images using Jensen-Rényi Divergence reveals 3D profile of brain atrophy in HIV/AIDS,” in *2006 3rd IEEE International Symposium on Biomedical Imaging: From Nano to Macro - Proceedings*, 2006, vol. 2006, pp. 193–196.

[53] D. Rueckert, P. Aljabar, R. A. Heckemann, J. V. Hajnal, and A. Hammers, “Diffeomorphic registration using B-splines,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2006, vol. 4191 LNCS, pp. 702–709.

[54] J. Ashburner, “A fast diffeomorphic image registration algorithm,” *Neuroimage*, vol. 38, pp. 95–113, 2007.

[55] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, “Diffeomorphic demons: efficient non-parametric image registration.,” *Neuroimage*, vol. 45, pp. S61–S72, 2009.

[56] S. Durrleman *et al.*, “Morphometry of anatomical shape complexes with dense deformations and sparse parameters,” *Neuroimage*, vol. 101, pp. 35–49, 2014.

[57] G. E. Christensen, S. C. Joshi, and M. I. Miller, “Volumetric transformation of brain anatomy,” *IEEE Trans. Med. Imaging*, vol. 16, no. 6, pp. 864–877, 1997.

[58] A. Trouvé, “An Infinite Dimensional Group Approach for Physics Based Models in Patterns Recognition,” *Math. Subj. Classif.*, vol. 1313, pp. 1–36, 1991.

[59] P. Dupuis, U. Grenander, and M. I. Miller, “Variational problems on flows of diffeomorphisms for image matching,” *Q. Appl. Math.*, vol. 56, no. 3, pp. 587–600, 1998.

[60] G. E. Christensen, R. D. Rabbitt, and M. I. Miller, “Deformable templates using large deformation kinematics,” *IEEE Trans. Image Process.*, vol. 5, no. 10, pp. 1435–1447, 1996.

[61] Y. Choi and S. Lee, “Injectivity conditions of 2D and 3D uniform cubic B-spline functions,” *Graph. Models*, vol. 62, no. 6, pp. 411–427, 2000.

[62] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*. Chichester, UK: John Wiley & Sons, Ltd, 2016.

[63] N. J. Higham, “The scaling and squaring method for the matrix exponential revisited,” *SIAM J. Matrix Anal. Appl.*, vol. 26, no. 4, pp. 1179–1193, 2005.

[64] C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. A. Henson, K. J. Friston, and R. S. J. Frackowiak, “A voxel-based morphometric study of ageing in 465 normal adult human brains,” *Neuroimage*, vol. 14, no. 1 I, pp. 21–36, 2001.

[65] J. P. Thirion, “Image matching as a diffusion process: An analogy with Maxwell’s demons,”

Med. Image Anal., vol. 2, no. 3, pp. 243–260, 1998.

[66] J. P. Thirion, "Non-Rigid Matching Using Demons," *IEEE Comput. Vis. Pattern Recognit.*, pp. 245–251, 1996.

[67] P. Hellier *et al.*, "Retrospective evaluation of inter-subject brain registration," *IEEE Trans. Med. Imaging*, vol. 22, no. 9, pp. 1120–1130, 2003.

[68] N. Charon and A. Trouvé, "The varifold representation of nonoriented shapes for diffeomorphic registration," *SIAM J. Imaging Sci.*, vol. 6, no. 4, pp. 2547–2580, 2013.

[69] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[70] H. Li and Y. Fan, "Non-rigid image registration using self-supervised fully convolutional networks without training data," in *Proceedings - International Symposium on Biomedical Imaging*, 2018, vol. 2018-April, pp. 1075–1078.

[71] S. G. Mueller *et al.*, "The Alzheimer's Disease Neuroimaging Initiative," 2008.

[72] D. W. Shattuck *et al.*, "Construction of a 3D probabilistic atlas of human cortical structures," *Neuroimage*, vol. 39, no. 3, pp. 1064–1080, Feb. 2008.

[73] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, 2016.

[74] X. Zhu, M. Ding, T. Huang, X. Jin, and X. Zhang, "PCANet-based structural representation for nonrigid multimodal medical image registration," *Sensors (Switzerland)*, vol. 18, no. 5, 2018.

[75] E. D. Vidoni, “The Whole Brain Atlas,” *J. Neurol. Phys. Ther.*, vol. 36, no. 2, p. 108, 2012.

[76] National Institutes of Health, “Retrospective Image Registration Evaluation.” 2003.

[77] M. P. Heinrich *et al.*, “MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration,” *Med. Image Anal.*, vol. 16, no. 7, pp. 1423–1435, Oct. 2012.

[78] Z. F. Knops, J. B. A. Maintz, M. A. Viergever, and J. P. W. Pluim, “Normalized mutual information based registration using k-means clustering and shading correction,” *Med. Image Anal.*, vol. 10, no. 3 SPEC. ISS., pp. 432–439, Jun. 2006.

[79] J. Chen *et al.*, “WLD: A robust local image descriptor,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1705–1720, 2010.

[80] C. Wachinger and N. Navab, “Entropy and Laplacian images: Structural representations for multi-modal registration,” *Med. Image Anal.*, vol. 16, no. 1, pp. 1–17, Jan. 2012.

[81] X. Cao, J. Yang, J. Zhang, Q. Wang, P. T. Yap, and D. Shen, “Deformable image registration using a cue-aware deep regression network,” *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1900–1911, 2018.

[82] Imperial College London, “IXI Dataset – Brain Development,” *IXI - Extracción de información de imágenes (EPSRC GR / S21533 / 02)*, 2016. [Online]. Available: https://brain-development.org/ixi-dataset/. [Accessed: 23-May-2020].

[83] M. Lorenzi, N. Ayache, G. B. Frisoni, and X. Pennec, “LCC-Demons: A robust and accurate symmetric diffeomorphic registration algorithm,” *Neuroimage*, vol. 81, pp. 470–483, Nov.

2013.

[84] J. Fan, X. Cao, Z. Xue, P. T. Yap, and D. Shen, “Adversarial similarity network for evaluating image alignment in deep learning based registration,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11070 LNCS, pp. 739–746.

[85] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2019.

[86] J. Fan, X. Cao, P.-T. Yap, and D. Shen, “BIRNet: Brain image registration using dual-supervised fully convolutional networks,” *Med. Image Anal.*, vol. 54, pp. 193–206, 2019.

[87] S. Kullback and R. A. Leibler, “On Information and Sufficiency,” *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951.

[88] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner, “Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults,” *J. Cogn. Neurosci.*, vol. 22, no. 12, pp. 2677–2684, 2010.

[89] A. Di Martino *et al.*, “The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism,” *Mol. Psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[90] P. M. Milham, F. Damien, M. Maarten, and H. M. Stewart, “The ADHD-200 Consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience,”

Front. Syst. Neurosci., vol. 6, no. SEPTEMBER, pp. 1–5, 2012.

[91] R. L. Gollub *et al.*, "The MCIC collection: A shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia," *Neuroinformatics*, vol. 11, no. 3, pp. 367–388, 2013.

[92] K. Marek *et al.*, "The Parkinson Progression Marker Initiative (PPMI)," *Prog. Neurobiol.*, vol. 95, no. 4, pp. 629–635, 2011.

[93] A. Dagley *et al.*, "Harvard Aging Brain Study: Dataset and accessibility," *Neuroimage*, vol. 144, pp. 255–258, 2017.

[94] A. J. Holmes *et al.*, "Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures," *Sci. Data*, vol. 2, 2015.

[95] D. Kuang and T. Schmah, "FAIM – A ConvNet Method for Unsupervised 3D Medical Image Registration," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11861 LNCS, pp. 646–654.

[96] K. Miller *et al.*, "Modelling brain deformations for computer-integrated neurosurgery," *Int. j. numer. method. biomed. eng.*, vol. 26, no. 1, pp. 117–138, 2010.

[97] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: efficient non-parametric image registration.," *Neuroimage*, vol. 45, no. 1 Suppl, pp. S61–S72, Mar. 2009.

[98] A. BB, G. M, and G. JC, "Symmetric diffeomorphic image registration: Evaluating

automated labeling of elderly and neurodegenerative cortex and frontal lobe," *LNCS*, vol. 4057, pp. 50–57, 2006.

[99] R. Sridharan *et al.*, "Quantification and analysis of large multimodal clinical image studies: Application to stroke," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8159 LNCS, pp. 18–30.

[100] Numpy, "NumPy," 2019. [Online]. Available: https://numpy.org/. [Accessed: 03-Aug-2020].

[101] D. Pustina and P. Cook, "Anatomy of an antsRegistration call," 2017. [Online]. Available: https://github.com/ANTsX/ANTs/wiki/Anatomy-of-an-antsRegistration-call. [Accessed: 23-Nov-2020].

[102] P. Neelin, "mincresample," 1993. [Online]. Available: http://bic-mni.github.io/man-pages/man/mincresample.html. [Accessed: 25-Jul-2020].

[103] NITRC, "NITRC: CANDI Neuroimaging Access Point: Tool/Resource Info," 2007. [Online]. Available: https://www.nitrc.org/projects/candi_share. [Accessed: 27-Aug-2020].

[104] C. Lebel and C. Beaulieu, "Longitudinal development of human brain wiring continues from childhood into adulthood," *J. Neurosci.*, vol. 31, no. 30, pp. 10937–10947, 2011.

[105] H. Iida *et al.*, "Three-dimensional brain phantom containing bone and grey matter structures with a realistic head contour," *Ann. Nucl. Med.*, vol. 27, no. 1, pp. 25–36, 2013.