Novel Multi-Scale Architecture for Medical Image Registration

by

Vignesh Sivan

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2022

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Medical image registration is an integral component of many medical image analysis pipelines. While registration has conventionally been solved using optimization techniques, there is growing interest in the application of deep learning to medical image registration. Deep learning based image registration (DLIR) methods have shown promising results, outperforming or being competitive with optimization-based methods for some standard datasets. DLIR still lags behind conventional optimization-based methods in some settings, however.

In large-displacement settings, where there is large discrepancy between the anatomical structures being registered, DLIR still lags behind conventional optimization in performance. Recent advances in DLIR has demonstrated improved registration performance in large-displacement settings. A striking feature of these novel DLIR approaches is that they employ neural network architectures and architectural techniques that are fairly common in optical flow. This fact, coupled with the similarity in inputs and outputs between optical flow networks and DLIR networks begs the question: can adoption of optical flow network architectures improve generalization of DLIR.

This thesis addresses this question. One of the challenges with directly adopting optical flow networks is that 3D medical images tend to require much more GPU memory than traditional networks. In this thesis, a novel DLIR architecture for large displacement registration is developed. Multiple scales were incorporated as well as an explicit cost volume computation, both of which are commonplace in optical flow literature. At inference time, a finetuning step was also investigated to improve performance for a range of medical imaging datasets. The novel DLIR approach developed, Recurrence With Correlation Network (RWCNet), was able to achieve 1.48mm keypoint loss on the NLST dataset, competitive the previous state of the art. On the OASIS dataset, it is able to achieve a dice score of 80.7%, marginally higher than the state of the art, at 80%.

Ablation studies were performed to highlight the significance of these approaches. It was found that the computation of a cost volume greatly improves network performance and iterative refinement through a recurrent neural network both greatly improve performance in the large displacement setting. These architectural features do not impact registration in the small displacement setting. This suggests that different architectures are better suited to different datasets. Overall, the DLIR architecture developed represents continued progress towards improving medical image registration that will ultimately translate to improvements in clinical diagnosis and treatment.

Acknowledgements

First, I would like to sincerely thank my supervisors, Dr. Alex Wong and Dr. Stewart Mclachlin, for their support and guidance. It was a pleasure to work under such supportive and and inspiring researchers.

I would also like to thank Dr. Michael Hardisty from Sunnybrook Research Institute for his mentorship. He helped me get past countless technical roadblocks and was always generous with his time and insights. Also, a huge thanks to my collaborators for the MICCAI 2022 Learn2Reg competition, Teodora Vujovic and Raj Ranabhat.

I would like to thank my readers, Dr. David Clausi and Dr. Yue Hu for taking the time out of their busy schedules to review and provide valuable feedback for my thesis.

I am grateful to Synaptive Medical for giving me the opportunity to work on an extremely exciting clinical project. Thanks are also in order for Dr. Sadegh Raeisi and the team at Foqus Inc. for an extremely rewarding internship experience.

I am thankful for my friends at VIP lab. They were a constant source of joy and camaraderie and the work presented in this thesis would not have been possible without hundreds of impromptu table tennis matches.

Dedication

This thesis is dedicated to the three who raised me: Sivan, Muthulakshmi and Madhavi.

Table of Contents

List of Figures

List of Tables

Chapter 1

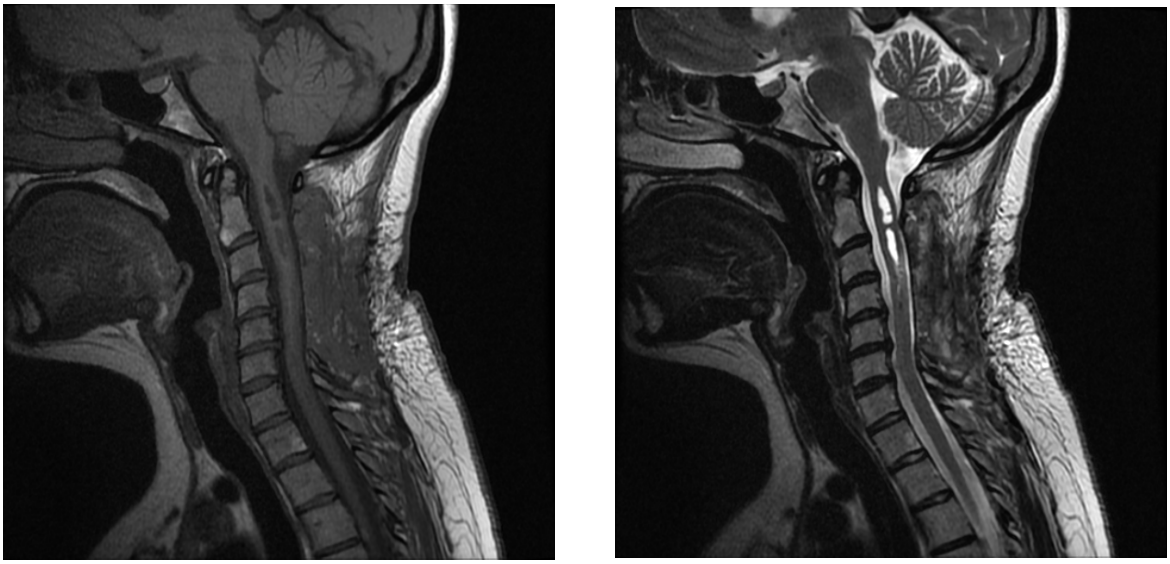
Introduction

Image registration is the task of spatially aligning two different images containing the same or similar objects. Registration of medical images is an especially challenging problem because the spatial statistics of objects in medical images can vary greatly. This chapter provides a high-level overview of the registration problem and its clinical applications.

1.1 Medical Image Registration

To understand the challenges of the image registration problem, it can be helpful to consider the characteristics of images generated by different imaging modalities. Computed Tomography (CT) derives contrast from the variable attenuation of X-ray beams as they pass through a heterogeneous object. For biological tissue, CT can provide excellent contrast in both hard and soft tissue. Magnetic Resonance Imaging (MRI) derives contrast from the variable resonance properties of protons in different tissue. While less ubiquitous than CT, it allows imaging with unrivalled soft tissue contrast. CT and MRI images of the same anatomical structure can visually look very different. The appearance of MRI scans can also vary greatly, based on the kind of tissue weighting they use. Figure 1.1 shows T1 and T2-weighted MRI scans of the spine of the same patient. Similar objects have extremely different intensities and structures in one image are not present in the other. The registration of images from different modalities is useful because it brings to the same coordinate frame features captured by different modalities and has exciting clinical applications such as pre-operative planning [1].

Another challenge in medical image registration is that the objects being imaged can vary greatly in appearance. For example, when registering images of the same anatomical



(a) T1-weighted spine

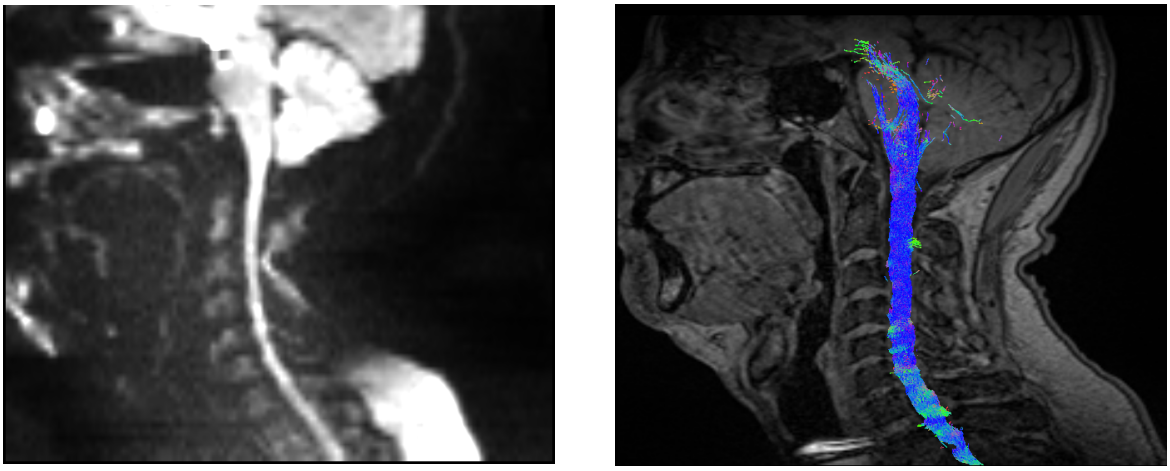
(b) T2-weighted spine

Figure 1.1: T1 and T2 weighted MRI scans of the same subject. Note the presence of, possibly, edema at C2-C3 in both volumes.

structure from different subjects, the size and shape of the anatomical structures being imaged can vary based on age, gender and other subject attributes. Yet another challenge for registration algorithms stems from the fact that medical image data is inherently 'long-tail'; this describes the phenomenon that any medical image dataset will overwhelmingly represent the most common diseases, while many uncommon diseases will either be sparse or not represented entirely. This poses a challenge to the design of registration algorithms, as the algorithm should generalize to data that wasn't observed during its development.

1.2 Applications of Medical Image Registration

Medical image registration has a myriad of applications. Registration of scans of the same subject with different modalities can highlight different anatomical features, and multi-modal registration in this context provides the ability to collocate these various anatomical features in the same frame of reference. This can be essential for image-guided pre-surgical planning [1], where features from different imaging modalities are needed for clinical decision-making. Registration of the same image at different points in time can be used for quantitative disease tracking [2], or for image-guided radiotherapy [3].



(a) (b)

Figure 1.2: a) sagittal view of cord tractography bundles registered to a T1-weighted MRI of the same patient b) three-dimensional view of streamlines intersecting with an axial slice of the T1 MRI

Another common application of medical image registration is in population studies, where images can be used to characterize physiological attributes of a population. Sometimes, it can be used to generate a template image, called an 'atlas', that characterizes typical features of a population [4]. Generating an atlas entails the deformable alignment of a representative sample of images from a population to yield an 'average' image, from which new subject images can be more easily registered to in reference to locating key anatomical features or regions. A common use of atlases is in brain studies, where the goal might be to understand neurological functioning of specific brain regions [5], or to understand how aging and various diseases quantitatively affect brain anatomy [6].

Finally, the application that spurred the contributions of this thesis is diffusion tractography imaging (DTI) analysis. DTI enables the visualization of diffusion in the body. It is underpinned by diffusion sensitive MRI, which derives contrast from the preferential diffusion of protons in the body. Tractography is a technique for visualizing neural tracts from DTI. It is often desirable to register DTI tracts to tissue-weighted MRI scans, so that structural anatomical features can be co-located with DTI tracts [7, 8]. This co-location has several clinical applications. For example, DTI tractography in the spinal cord can highlight regions where there is cord compression [9]. Figure 2.1a shows a diffusion weighted MRI scan and Figure 2.1b shows a T1-weighted MRI overlaid with DTI tracts generated from the diffusion weighted MRI scan.

1.3 Thesis Contributions and Overview

Conventional approaches to medical image registration formulate registration as an energy minimization problem, where the energy functional quantifies the acceptability of any given solution. There is growing interest in the application of deep learning to the task of medical image registration. An advantage of using deep learning neural networks over conventional optimization approaches is that these networks can often arrive at optimal solutions orders of magnitude faster. This speed can increase the usability of medical image registration in time-constrained environments such as those found in real-time medical care. Based on deep learning's success in the field of computer vision, it may also arrive at more accurate solutions than conventional optimization techniques. Indeed, recent work on deep learning image registration (DLIR) has shown its potential for yielding faster and more accurate transforms than conventional methods for some datasets.

There are three major architectural features that have become commonplace when designing neural network architectures for DLIR: multi-scale features [10], iterative refinement [11] and cost volume computation [12, 13, 14]. To the author's knowledge, there is not a single architecture that combines all of these components. This thesis explores this gap. The primary contribution is a novel architecture that combines iterative multi-scale recurrent networks with cost volume computation. This architecture is named *recurrence with correlation* and abbreviated as RWCNet, for its constituent components. This thesis benchmarks this architecture and compares it against other network architectures. Ablations are carried out to highlight architectural components that provide a strong implicit bias for the image registration problem.

This thesis is organized as follows:

- Chapter 2 introduces the image registration challenge from the perspective of an optimization problem. Literature that particularly influenced the work presented in this thesis is introduced.
- Chapter 3 details the primary contribution of this thesis: a novel deep learning based architecture for medical image registration. It borrows approaches that have worked well for 2D optical flow problems and applies them to the problem of 3D medical image registration.
- Chapter 4 summarizes results in the wider context of medical image analysis on real clinical data. Challenges to adoption are discussed, along with potential future work to address these challenges.

Chapter 2

Background

This chapter introduces topics and literature relevant to the main contributions of this thesis. This chapter is organized as follows: Section 2.1 provides the background for the optimization challenge faced with medical image registration. Section 2.2 provides a high-level overview of the different strategies for optimization-based registration. Finally, Section 2.3 provides a high-level overview of deep learning based approaches that have been used to solve this problem.

2.1 Image Registration Formulation

Let t be some spatial transform function that maps some moving image m to a target (or fixed) image, f. Both the moving and fixed images have the same number of dimensions n (i.e., $m, f \in \mathbb{R}^n$). In this section, the number of dimensions is assumed to be 3; thus the term for a position in a 3D image, voxel, is used. For simplicity, it can also be assumed that m and f have the same size. Given some similarity metric, $\mathcal{S}$ that describes the degree of spatial similarity between the moving and fixed images, a naive approach to finding an optimal registration would be to find a t that maximizes $\mathcal{S}$. That is,

$$t^* = \max_t \mathcal{S}(t(m), f) \tag{2.1}$$

where t^* is the optimal transform. By itself, the above formulation is ill-posed, since for each of the say, N, voxels in the image, there is an unrestricted number of possible sub-voxel locations ($\gg N$) that the transform can map each of the N voxels to.

Thus, some sort of regularization is needed to restrict the search space for the optimal solution. A common framework is variational optimization, where some regularization functional, $\mathcal{R}$, penalizes unlikely transforms based on prior belief. Smooth and small magnitude displacements are examples of prior beliefs that can be represented and penalized explicitly by regularizers. The objective function in this variational optimization scheme becomes:

$$t^* = \max_t(\mathcal{S}(t(m), f) + \mathcal{R}(t)) \tag{2.2}$$

Typically, t is a parametric function, belonging to a set of transformations, called the 'transform model', T. In the case of rigid registration, where only translation and rotation are permitted operations, T can be the set of functions parametrized by a homogeneous transform matrix with zero scaling and shearing. When the set of transforms is expanded to include affine transformations, the set of allowable functions might be parametrized as transform matrices with non-zero scaling and squaring.

In the case of deformable registration, the transform can be parametrized by a dense grid, ϕ. The transformation is then a resampling operation of the moving image onto this new grid. The objective function in this scenario becomes

$$\phi^* = \max_\phi(\mathcal{S}(m \circ \phi, f) + \mathcal{R}(\phi)) \tag{2.3}$$

where $m \circ \phi$ denotes the resampling operation. A common way to parameterize the displacement field, ϕ, is to use a dense deformation field, $D \in \mathbb{R}^{3\times H\times W\times D}$, that describes the displacement of each voxel or voxel in a moving image to a fixed image. With this parametrization, the objective function becomes:

$$\phi^* = \max_\phi(\mathcal{S}(m \circ (Id + D), f) + \mathcal{R}(D)) \tag{2.4}$$

where Id is the identity grid and $Id + D$ describes the (sub-) voxel coordinates of each moving voxel in the fixed coordinate frame. One strategy to solve such a problem would be through iterative gradient methods. One can perform simple gradient descent on the loss function described above, using:

$$\phi^t = \phi^{t-1} - \eta\nabla_\phi L \tag{2.5}$$

where L is the loss function, described by the expression inside the max in Equations (2.3) and (2.4), η is the learning rate, and ∇_ϕ is the jacobian operator.

2.1.1 Similarity and Regularization Functions

The objective function requires appropriate choices for the similarity function $\mathcal{S}$ and the regularization functional $\mathcal{R}$. The choice of regularization functional is usually more straightforward as it is not as dependent on the data. The L2-norm, for example, may be used as it promotes smaller displacement fields. For deformable registration, another common regularizer is the gradient of the transform function, which promotes smoothness of the spatial transform. A common way to represent this smoothness is to compute the sum of squared gradient of the displacement, $\sum(\nabla D)^2$.

The choice of similarity metric is dependent on the data in the registration problem. A simple choice is to directly compare voxel intensities. The mean squared error (MSE) for the moving and fixed image pair with size N can be expressed as:

$$MSE(m \circ \phi, f) = \frac{1}{N} \sum (m \circ \phi - f)^2 \tag{2.6}$$

There are several challenges with MSE. First, it only works when images are normalized and lie within the same intensity range. Moreover, it can be sensitive to outliers and noise. It can be especially problematic for modalities with especially poor signal to noise ratio, like ultrasound and diffusion-weighted MRI.

Another choice for images belonging to the same modality is the normalized cross correlation (NCC) metric. It measures the similarity between the moving and fixed images by measuring the correlation of patches extracted from the moving image and the fixed image. It is given by:

$$NCC(m', f) = \frac{1}{N} \sum_i \frac{(m'_i - \bar{m'})(f - \bar{f})}{\sigma_{m'} \sigma_f} \tag{2.7}$$

where m' is $\circ\phi$ and $\bar{k}$ and σ_k denote the mean and the standard deviation of the window in image k around voxel i. NCC might be advantageous when there might be some global intensity differences or higher localized noise, as the similarity is based on a neighborhood of points rather than relying on a single voxel.

In multi-modal registration, voxel intensities cannot be used directly, as dissimilar intensities might denote similar anatomical features. For example, consider Figure 1.1, which shows the same region imaged with two different MRI protocols: with T1-weighting and T2-weighting. It can be seen that the same anatomical features have completely different intensities. The dark cerebrospinal fluid in the T1 image is bright in T2 image.

A common metric in this case is mutual information (MI) or normalized mutual information (NMI). It can be computed by generating a 2D histogram of intensities between a pair of images (or sub-images) and measuring the joint distribution of the intensities. Like NCC, mutual information in the context of image registration is typically implemented using patches of the moved image.

Another similarity metric from the medical image registration literature is the discrepancy of the moved and fixed images in 'feature space'. Self similarity context (SSC) [15], for example, computes for each voxel p in image I and feature SC as follows:

$$SC(I, p) = \exp^{-\frac{(p-y)^2}{\sigma^2}}, \forall y \in \mathbb{N} \tag{2.8}$$

where $\mathbb{N}$ is the neighborhood of points around p and σ^2 is an estimate of the local or global noise in the image I. The similarity loss with SSC features then simply becomes $S(m) - S(t)$. The advantage of using metrics such as SSC over mutual information is that mutual information requires building a histogram of intensity values for each image patch that it compares, which can be expensive. In contrast, SSC is a simple kernel function applied to the image followed by a subtraction and is, thus, less computationally expensive.

Besides image similarity, auxiliary data can be used to guide registration. Segmented regions of interest can be used to delineate regions of interest and maximizing overlap between these segmentations could be incorporated into the objective function. A common registration goal is then to maximize the Dice coefficient (DC) between the two sets of segmentations. It is given by:

$$DC(s_m \circ \phi, s_f) = 2 * \frac{s_m \circ \phi \cap s_f}{s_m \circ \phi \cup s_f} \tag{2.9}$$

Another common auxiliary data that is useful for image registration is the coordinates of landmarks in both the fixed and moving image. In voxel space, the discrepancy, LM, in the landmark coordinates can be expressed as

$$LM(D, L_m, L_f) = (L_f - (L_m + D))^2 \tag{2.10}$$

where L_f and L_m are the fixed and moving coordinates respectively.

Caveat Emptor

Registration differs from other image analysis tasks such as segmentation due to the fact that it is inherently ill-posed, which implies the existence of multiple solutions to maxi-

mize some similarity constraint. A side effect of this ill-posedness is that there can exist solutions that maximize some similarity metric while being, in some cases, wildly inaccurate. Moreover, Rohlfing [16] identified that intensity-based similarity metrics cannot distinguish between accurate and inaccurate registrations, even when multiple intensity-based similarity metrics are used in conjunction. They are able to create a *Completely Useless Registration Tool* that is able to generate good tissue overlap similarity while also generating inaccurate registration. They found that only metrics derived from auxiliary data such as dense anatomical landmarks and segmentation overlap were able to accurately distinguish between accurate and inaccurate registrations reliably.

2.1.2 Diffeomorphic Registration

Spatial transformations in the context of medical image registration should be diffeomorphic. Diffeomorphic maps have the elegant property of preserving the topology of the image by forcing a one-one mapping between the moving and transformed image, which prevents 'folding', which is the artifact caused by multiple voxels being mapped to the same location. To achieve an approximate diffeomorphic displacement, the displacement field ϕ is parametrized using a stationary velocity field, described using the differential equation $\frac{d\phi}{dt} = v(\phi^t)$ with initial condition $\phi^{(0)}$ being the identity mapping Id. The diffeomorphic displacement field, ϕ, is obtained through integration of the velocity field over a unit time step ($t \in [0, 1]$) through scaling and squaring, as implemented in DARTEL [17]. The final transform, ϕ^1 is approximated as e^v. This transformation has a positive jacobian determinant, guaranteeing locally invertible transforms (i.e. one-one mappings).

2.2 Optimization-based Registration

Broadly, there are two major categories of optimization-based registration. The first is gradient-based iterative optimization. The second approach is discrete optimization by labelling vertices of a markov random field (MRF), using a finite set of discrete displacements.

2.2.1 Continuous Optimization

A particularly influential gradient-based optimization approach is Demons [18, 19], based on Maxwell's famous thought experiment [20]. In Maxwell's thought experiment, demons

act as gate-keepers for gates that allow molecules to pass through. In the image registration setting, the demons can be thought of as gate-keeping the boundaries of voxels of the fixed image. Then, the optimization can be thought of as iteratively diffusing the voxels of the moving image through these fixed-image gates. At each iteration, n, each voxel in the moving image is displaced by a vector d, such that:

$$d^{(n+1)} = \frac{(m^{(n)} - f^{(0)})\nabla f^{(0)}}{(m^{(n)} - f^{(0)})^2 + |\nabla f^{(0)}|^2} \tag{2.11}$$

where $m^{(n)})$ is the moving image at the nth iteration and $f^{(0)}$ is the static image. To promote smoothness, a Gaussian filter is applied to the displacement field at every iteration. Many variants of the Demons algorithm have since been proposed. For example, note that (2.11) directly uses the intensities of the moving and fixed images. It can be modified to use features for multi-modal registration. One of the limitations of gradient-based approaches is that they can readily fall into local optima, especially in large deformation settings [21].

2.2.2 Discrete Optimization

Discrete optimization via MRF is the second family of methods for image registration. In this optimization setting, the displacement of each voxel is restricted to some finite set, $\mathbb{D} = \{l^1, ..., l^i\}$. The optimization then becomes an MRF labelling problem, where each node in the MRF is assigned a specific element from $\mathbb{D}$ [22, 23]. Similar to the continuous domain, the optimization process entails maximizing a similarity term while minimizing some regularization penalty associated with the displacement. This can be expressed as the energy of the MRF:

$$E_{MRF} = \sum_{i \in \mathcal{V}} \mathcal{S}(d^i) + \sum_{(i,j) \in \mathcal{E}} \mathcal{R}(d^i, d^j) \tag{2.12}$$

where $\mathcal{V}$ is the set of vertices of the MRF and $\mathcal{E}$ is the clique to which vertex i belongs. Finally, d^i is the displacement assigned to vertex i. The first term is the data similarity term and the second term is the regularization term. One approach to optimizing the energy is to use a dynamic programming strategy, such as Viterbi optimization. Dynamic programming entails the traversal of the MRF graph and recording of the most probable displacements given by the energy function for each node. Computation can be sped up by restricting the space of possible displacements; the transform model is usually restricted to a parametric set of transformations, such as a B-spline transform [23].

2.2.3 Iterative Discrete Optimization

It is possible to combine the benefits of continuous and discrete-domain approaches. For example, Hutchinson et al. [21] describes a gradient-based discrete optimization algorithm that is able to overcome local optima without requiring computationally expensive dynamic programming. Given fixed and moving features, a cost volume describing the goodness of fit can be computed for a discrete set of voxel-level displacements (i.e, the set of displacements is described by the set: $\mathbb{D} = \{0, \pm 1, \pm 2, ..., \pm l_{max}\}$).

$$C(x, d) = \mathcal{S}(F_f(x), F_m(x + d)) \tag{2.13}$$

where C is the cost volume, F_f and F_m describing fixed and moving image features respectively, x is a coordinate location and $d \in \mathbb{D}$. The flow field can then be iteratively optimized, as described by Algorithm 1. It requires a discrete set of update coefficients, $\mathcal{Q}$, describing the step size at each iteration. Note that at each iteration a smoothing is performed, similar to Demons. Unlike the continuous-domain case, however, the updates are not derived from the gradients, but from the best possible displacement, which is similar to the discrete MRF optimization scheme.

Algorithm 1 Non-parametric discrete optimization

Require: $\mathcal{Q}$, the set of update coefficients
Require: C, the cost volume
Require: S, a function that performs smoothing
  $d_{min} \leftarrow \arg\min C$
  for each $q \in \mathcal{Q}$ **do**
    $d_{smooth} \leftarrow S(d_{min})$
    $d_{min} \leftarrow \arg\min(C + q \times d_{smooth})$
  end for
  $d_{final} \leftarrow S(d_{min})$

ConvexAdam [24] combines the iterative discrete optimization described by Algorithm 1 with continuous-domain optimization. It works by first computing MIND (Modality Invariant Neighborhood Descriptor) features [25] of the fixed and moving image. First discrete optimization, as per Algorithm 1 is performed, followed by continuous domain gradient optimization as shown in (2.5). An ADAM [26] optimizer is used instead of pure gradient descent, however. ConvexAdam has been able to achieve state-of-the-art performance in several large displacement datasets.

2.3 Learned Image Registration

This section details recent efforts to apply deep neural networks to the problem of medical image registration. These methods, can take advantage of GPUs to drastically improve the speed of registration, when compared to traditional optimization methods, which have typically been CPU bound. The problem can also be viewed as amortized optimization, where the cost of the optimization is paid during training time, enabling faster registration at inference time than conventional methods.

2.3.1 Spatial Transformer Networks

One of the requirements for learned registration algorithms is that the application of the mapping function to the moving image, $\Phi(m)$, needs to be differentiable, in order to propagate loss back to a neural network. Spatial Transformer Networks (STN)s [27] provide a way to to apply this mapping differentiably. Using the notation from [27], D-dimensional voxel coordinates $p \in \mathbb{Z}^D$ are displaced by a flow field $u \in \mathbb{R}^D$. Here, $\mathbb{Z}^D$ and $\mathbb{R}^D$ are used to indicate the fact that voxel coordinates are integer valued, while the flow-field can be any real number. Then, the shifted voxel coordinates $p' \leftarrow p + u, p' \in \mathbb{R}^n$ describe the sub-voxel location of the new voxel. To obtain intensity values at integer coordinates, linear interpolation is carried out. In the case of bilinear interpolation, this warping operation can be expressed as:

$$\Phi(p) = \sum_{q \in \mathbb{N}p'} \Pi_{d \in D}(1 - |p'_d - q_d|) \tag{2.14}$$

where $\mathbb{N}$ is the neighborhood voxels around p'. The equation above consists solely of additions, multiplications and absolute values, for which gradients and sub-gradients can be readily computed. This enables the backpropagation of gradients for deep learning based registration.

2.3.2 Neural Networks for Image Registration

Voxelmorph is one of the earliest and best-known works that combines DLIR with an unsupervised or semi-supervised loss [28]. It consists of a single 3D U-net [29] that outputs a flow field for a pair of fixed and moving images, as well as a spatial transformer network for warping the moving image. It also uses scaling and squaring layers to ensure that the spatial mapping is diffeomorphic. Voxelmorph was found to achieve comparable results

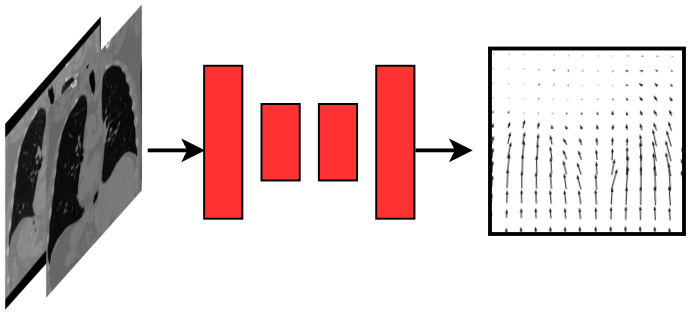
with conventional optimization techniques for brain registration tasks, while requiring much less time to generate the transformation. Unfortunately, its performance was found to be worse for large displacement datasets. [30].

Several works have since been proposed to address Voxelmorph's performance in large displacement settings. One of these networks is LAPIRN [10], which was found to achieve improved performance through multi-scale refinement. It builds a feature pyramid by learning features at high resolution and downsamples these features to lower resolutions. It applies coarse to fine refinement of the flow field, by first training a network to learn to register downsampled inputs, before learning to register progressively upsampled inputs. The intuition behind coarse to fine refinement is that at low resolutions, the high level features disappear and only the low-frequency shape of the image remains. This is an 'easier' problem than matching both the shape and the higher frequency features, and it is less sensitive to noise.

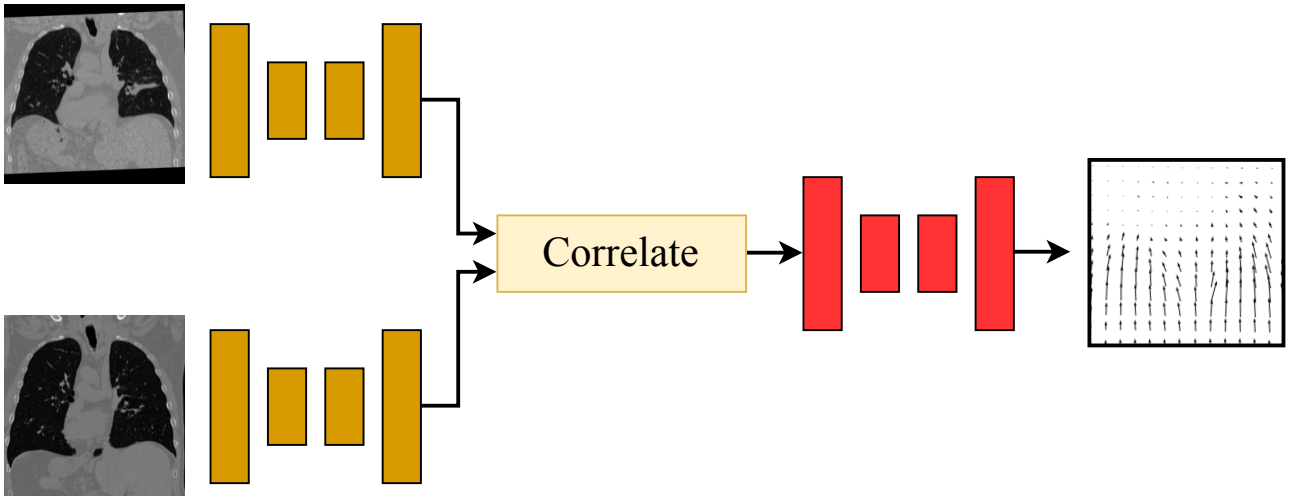
A common approach in optical flow literature (as discussed in the next section) is the computation of a cost volume between features in the fixed and moving network. This cost volume computes the similarity between each pair of voxels in the moving and fixed images, providing a measure of 'goodness of fit' for each displacement. It can be computed using Equation (2.13). It is common for there to be a maximum displacement, such that $|d| < d_{max}$. This d_{max} value is sometimes called the 'search radius' and can aid in conserving space. Some approaches compute the cost volume for all locations [32]. Moreover, the cost function can also vary. Dot products are used in [33] instead of the mean squared error.

Instead of directly passing the features directly into the network that outputs a flow field, the cost volume is passed into the network that outputs the flow field. Figure 2.1 illustrates how this approach differs from the approach used by Voxelmorph, which simply stacks the input images. This additional layer of indirection can serve as a way to force the network to learn useful feature representations of the input images. It has been found to improve performance in many 2D optical-flow networks. Probabilistic Dense Displacement Network (PDDNet) [14], uses this technique to compute 6D cost volume between learned features in the three-dimensional moving and fixed images. ConvexAdam [24] is another approach that computes a cost volume. Unlike PDDNet, however, it is a discrete-optimization based method that uses the cost volume to provide a goodness-of-fit value for each of the displacement in the discrete set of possible displacements.

Another similarity between PDDNet and ConvexAdam is an 'instance optimization' step, that fine tunes the flow field generated by the network by directly optimizing the loss objective using iterative gradient-based optimizer as shown in Equation. This optimization is performed for a very few number of steps (< 100, for example) and is performed directly on the GPU. Thus it is significantly faster than traditional optimizer packages



(a)



(b)

Figure 2.1: Two different neural network architectures for image registration. a) Input images are stacked and passed through one or more networks that output a flow field. This architecture is used by Voxelmorph [28], CascadeNet [31] and LapIRN [10]. b) Cost-volume based architecture where features are explicitly learned by the network for each image and the cost volume that computes some goodness of fit tensor. PDDNet [14] implements this architecture for medical image registration.

which implement slower, but more sophisticated optimizers.

2.3.3 Optical Flow Networks

Optical flow is the general problem of computing the motion of an object from images. Often, this motion can be described using a flow field. Optical flow is an important problem in computer vision and has a range of applications including action recognition and pose estimation. It differs from the image registration problem in that the images are typically of the same object and the consistency of the intensity of objects between both objects is expected. The general image registration problem does not have these constraints. Nevertheless, the two problems are similar. In deep learning settings, the inputs to the network are two images between which to compute flow and the output is a flow field. As such, optical flow architectures can be used directly for medical image registration.

Flownet [12] is one of the earliest and best known works that applied deep learning to the optical flow problem. It introduced two architectures: FlownetCorr and FlownetSimple, abbreviated as FlownetC and FlownetS. These architectures differ in one fundamental aspect: the computation of a cost volume. In FlownetS, the input images are stacked together and fed to the network as input, which outputs a flow field. FlownetC, in contrast, has a feature-extractor network that first computes features for both the images. Next, these features are used to compute a cost volume, which is then fed into a second network that outputs a flow field. While the original Flownet paper found that both network architectures had similar performance, later works [13] have shown that computing an explicit correlation between fixed and moving features results in improved performance. A follow-up work, Flownet2 [13], demonstrates this result, and also improves the original Flownet architecture's results by showing that iterative refinement through cascading of networks yields better results.

PWCNet [33] implements and builds on key concepts that have become prevalent in optical flow literature: feature pyramids, iterative refinement and explicit cost volume computation. Feature pyramids are built by generating features, typically at full resolution, and downsampling them to lower resolutions. Intuitively, a flow field that achieves good resolution at high resolution also achieves one at a lower resolution. Optical flow for lower resolution features might be simpler since the range of displacement is smaller. PWCNet iteratively refines the flow fields from a coarse to fine manner, by feeding in warped images at higher resolutions and combining flow fields generated at higher resolutions with those generated at lower resolutions.

Finally, Recurrent All Pairs Field Transform (RAFT) [32], shares many of the key ideas that drive PWCNet's performance: multi-scale features and explicit cost volume computation. Rather than using a cost volume between a subset of voxels in the fixed and moving features, RAFT computes correlation between all pairs, thereby removing the inherent limitation of restricting flow to a certain radius. It also uses a lightweight RNN to iteratively update a displacement field, and on each iteration it samples from this cost volume. RAFT is able to achieve state-of-the-art optical flow performance and its architecture inspired the one developed in this thesis.

2.4 Chapter Summary

Image registration can be posed as an optimization problem consisting of a data fidelity or image similarity term as well as a smoothness-based regularization term. Moreover, it is common to parameterize the spatial transformation as a flow field that describes the amount of displacement each spatial location experiences in 3D. This formulation is not dissimilar to optical flow in computer vision, where the objective is to find a flow field that describes the motion of objects between two images. Interestingly, neural network architectural patterns such as multi-scale and iterative refinement have found success in both domains. A challenge with extending optical flow architectural patterns to medical images is that medical images tend to be larger, partly owing to the fact that medical images can be 3D.

The next section explores a novel architecture for medical image registration that is heavily influenced by architectures from optical flow literature. Like RAFT and PWCNet, multiple scales are used and an explicit cost volume is used as input to the model. The architecture overcomes the challenge of adopting optical flow architectures to large medical image data. This work explores whether the combination of explicit cost volume with multi-scale iterative refinement brings improvements to medical image registration.

Chapter 3

Recurrence With Correlation Network for Medical Image Registration

This chapter outlines the main contribution of this thesis: the *Recurrence With Correlation Network* (RWCNet) which combines explicit cost volume computation, common in optical flow networks, with multi-scale refinement for medical image registration. It is organized as follows: Section 3.1 provides some background and motivation for the architecture presented, Section 3.2 outlines the specifics of the architecture developed, Section 3.3 describes the experiments carried out, Section 3.4 outlines and discusses the significance of the results and, finally, Section 3.5 summarizes the main contributions of RWCNet.

3.1 Background for RWCNet

While DLIR has shown a lot of promise by outperforming conventional optimization methods on some standard datasets, challenges to adoption remain. Until recently, DLIR lagged behind conventional approaches when a large deformation field was required [30]. In this setting, the solution space that achieves good similarity between fixed and moving images is large and it can be difficult to learn an optimal solution for unseen data.

Several strategies have been explored to improve DLIR performance for large displacements. One strategy is iterative refinement of a flow field, where the network learns to iteratively augment a displacement field with smaller displacements [10, 11]. At each iteration, the space of possible solutions is smaller and, thus, becomes an easier optimization

problem. In optical flow literature, it is common to combine iterative models with explicit cost volume computation for continuous-domain optimization [33, 32]. These methods have demonstrated excellent performance for 2D images, often achieving state-of-the-art. The computation of an explicit cost volume is not as common for 3D medical image registration, since it can be prohibitively large to store in GPU memory. To the author's knowledge, there is no work that incorporates multi-scale refinement with an explicit cost volume computation.

Another desired property of DLIR is patient-specific fine tuning, as the distribution of medical images is long-tail and unlikely to be fully represented by the training dataset. There are several strategies to patient-specific fine-tuning. ConvexAdam [24] and PDDNet [14] contain an 'instance optimization' layer where a flow field is refined using iterative gradient-based optimization (e.g. Adam optimization).

In this work, the objective of the DLIR architecture development was to combine multi-scale iterative learning with an explicit cost volume computation. Potential memory issues are addressed by feeding corresponding fixed and moving patches at higher resolutions. A modification was also introduced to the commonly found instance optimization layer that is able to improve performance on a large displacement dataset. This combination of approaches, to the author's knowledge, is novel in DLIR. The technical contributions are listed as follows:

- A novel network architecture, *Recurrence with Correlation* (RWCNet), for DLIR that outperforms other coarse to fine networks in some datasets.
- Ablations of architectural components as well as the fine-tuning step to demonstrate the performance of the various architectural features of the registration network.

3.2 Methods

Sub-network Architecture.

The architecture for the recurrent neural network (RNN) sub-network is shown in Figure 3.1. Given a fixed and moving image pair (f and m, respectively), RWCNet obtains fixed and moving features (F_f and $F_{m,0}$) by feeding both images through a feature extractor network. A voxel-wise correlation between the fixed and moving features is computed, C. Due to the large number of dimensions, the correlation is restricted so that only voxels within a certain 'search range', r of the moving voxel are considered. Additionally, the

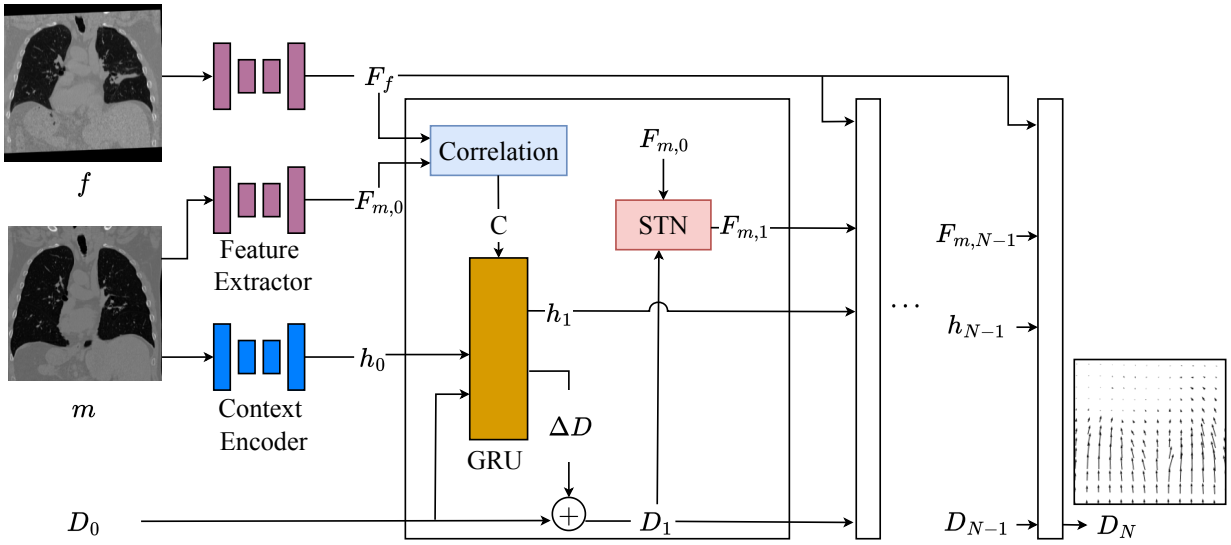


Figure 3.1: RAFT-inspired [32] RNN network architecture. It has four main components: 1) a feature extractor network for extracting fixed and moving features from the images, as well as a context encoder network for extracting moving image context 2) computation of a cost volume through correlation of the input features 3) update aggregate displacement and hidden state using a GRU. 4) spatial transformer network (STN) that warps the moving features using the aggregate flow field.

moving image is fed through a context network that extracts contextual information for the hidden network. The output of the contextual network is used as the initial hidden state of the RNN, h_0. Finally, a displacement field with zero displacement, D_0 is initialized.

The hidden state, flow and cost volume are fed into an update block that is a modified gated recurrent unit (GRU) [34]. The GRU is almost identical in implementation to the one used by RAFT [32], and it outputs a new hidden state and a new displacement field, h_1 and ΔD. The new displacement field is used to update the aggregate displacement, i.e, $D_1 = \Delta D + D_0$. This new displacement field can subsequently be used to warp the moving features using a spatial transformer network, denoted STN. The warped features will be used in the next RNN iteration for cost volume computation. This process of updating displacement field with GRU cell and updating features is repeated for N RNN time steps.

Coarse to Fine Registration.

Like LAPIRN and many optical flow networks, RWCNet adopts a course to fine approach to image registration. For each resolution, s, a new RNN is trained using inputs from the previous resolution. The model is not trained end to end. The weights from previous resolutions are frozen at higher resolutions. At high resolutions, computing the cost volume for the whole volumes becomes prohibitively expensive; as such, the approach is to divide the input images into uniform, non-overlapping windows or patches. The size of the patches at each resolution is parameterized by the 'patch factor', $p^s \in [0, 1]$. The size of the patches at resolution s is computed as $p^s \times S$ where S is the size of the full image at $1\times$ resolution. At higher resolutions, the flow field is used to warp the initial moving image (or patch) using flow fields computed at lower resolutions. Furthermore, the final hidden state is cached at lower resolutions and added to the initial hidden state at higher resolutions, increasing the non-linearity of the network and providing additional context to the network.

Instance Optimization.

At inference time, the flow field generated by the network is fine tuned using gradient-based iterative optimization. An ADAM-based optimizer is used in favour of the simple gradient-descent approach described by Equation (2.5), as implemented in PDD-Net and ConvexAdam.

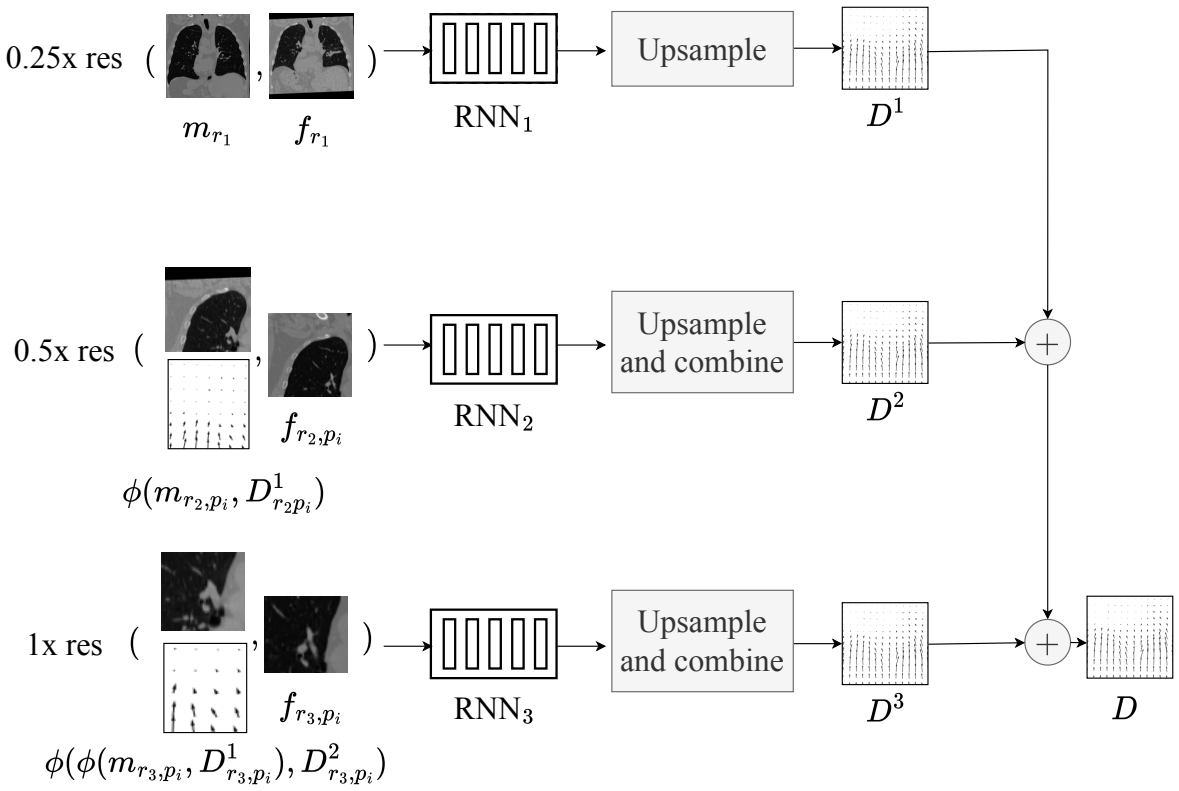


Figure 3.2: Multi-scale refinement of flow field using separate RNNs for each resolution. Note, at higher resolutions, patches of the input image are fed into the network, rather than the entire image. Not shown is the fact the final hidden layer is upsampled and concatenated with the first hidden state of the next resolution.

3.3 Experiments

3.3.1 Datasets

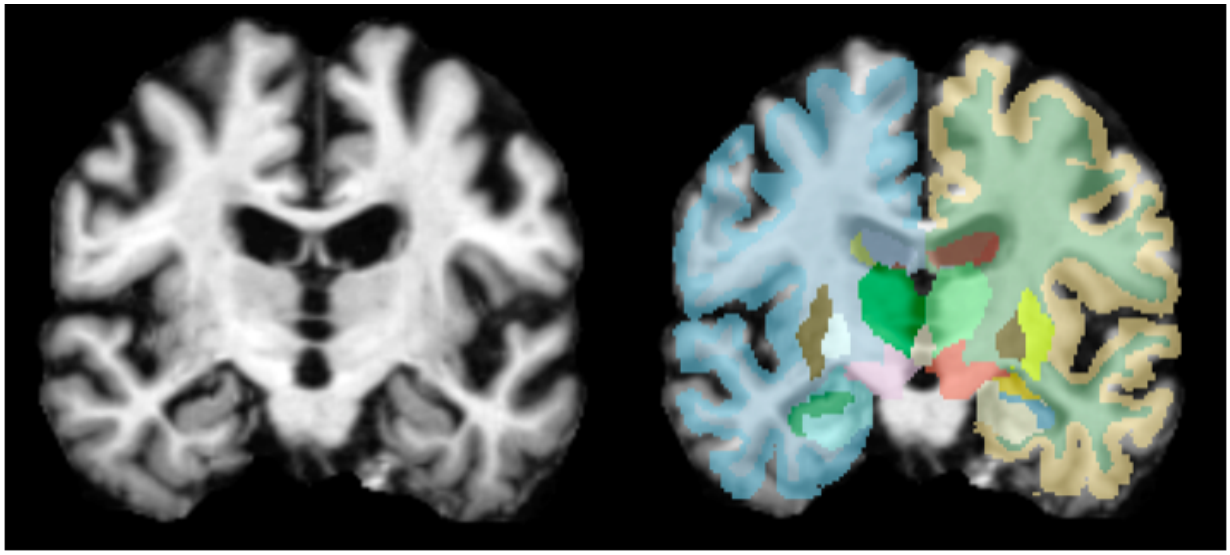
RWCNet is trained and validated with the OASIS [35] and NLST datasets [36]. Figure 3.3 shows samples from the OASIS and NLST datasets. The OASIS dataset consists of 414 T1-weighted MRI scans of individuals from ages 18-96 with mild to severe Alzheimer's. The scans are skull-stripped and resampled onto an isotropic grid and cropped to a uniform size. 35 segmentation labels are provided for important brain regions. The dataset is split into 395 images for training and 19 for validation. Inter-subject registration in this context could be used for constructing a sub-population brain atlas or for analysing intensity changes in consistent brain regions that are linked to disease progression.

NLST is a lung-CT dataset with pairs of inhale/exhale scans; keypoints and masks are provided by the 'Learn2Reg' (L2R) 2022 competition [37] for semi-supervised training. A subset of the image pairs (100 out of 150) of the NLST dataset released by the L2R competition is used for both training and validation, with a 90:10 training/validation split. Corresponding keypoints indicating locations of lung nodules in the inhalation and exhalation images are provided as auxiliary information. Since respiration is accompanied by a large change in lung volume, registration of the NLST dataset represents a large-displacement setting.

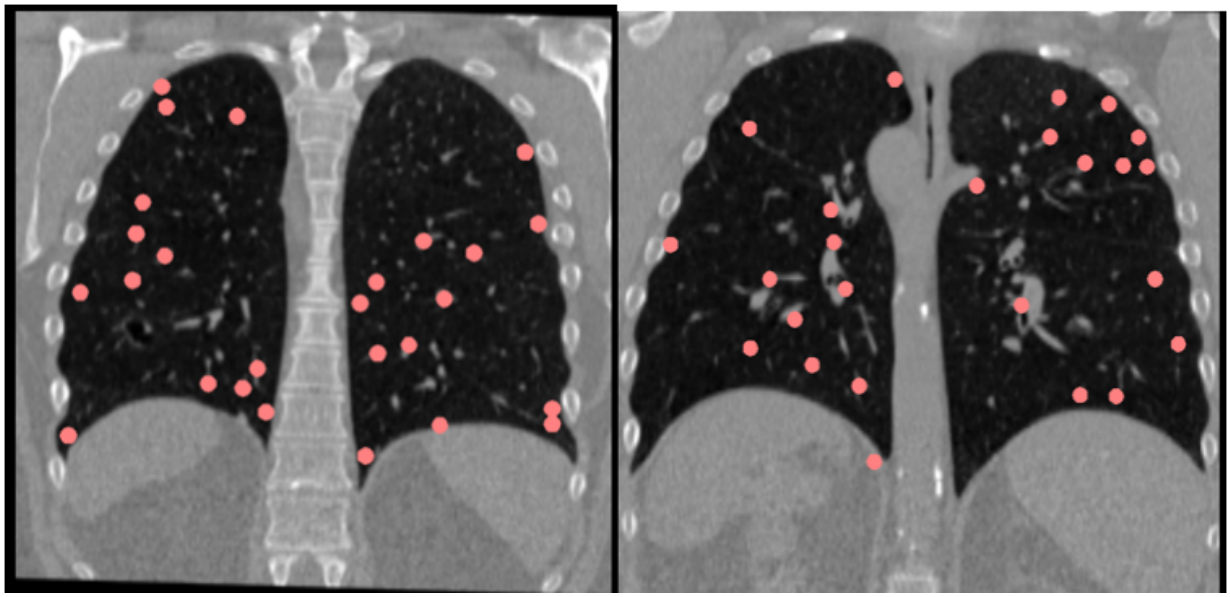
3.3.2 Training Parameters

Each sub-network is trained separately and in a coarse to fine manner. Additionally, the inputs at higher resolutions are patched to decrease GPU memory requirements. Table 3.1 shows the number of steps and the sizes of the inputs at each resolution. To address over fitting, dropout with probability 0.5 is used for the feature networks. The network takes about 30 hours to train both the NLST and OASIS datasets on an NVIDIA A100 with 32GB of RAM.

For OASIS, the similarity component of the loss function was a summation of the mean squared error (MSE) between the warped moving image and the fixed image intensities as well as the Dice loss between the warped segmentation and the fixed segmentation. The regularization loss was the average gradient of the flow field, as used in Voxelmorph. For NLST, the data was range normalized to between 0 and 1, with -4000 serving as the minimum value and 16000 serving as the maximum value. The loss function was a combination of the MSE and the keypoint discrepancy loss The mean gradient of the flow



(a) OASIS



(b) NLST

Figure 3.3: Sample images from the OASIS (a) and NLST (b) datasets. Note, segmentations are available for the OASIS dataset and keypoints indicating the positions of nodules are available for the NLST dataset.

Table 3.1: Resolution-specific Parameters.

Resolution	RNN Steps	Patches Per Image	Patch Size[1]	Training Steps
0.25	12	1	$0.25 \times S$	30000
0.5	12	8	$0.25 \times S$	45000
1	4	8	$0.5 \times S$	60000

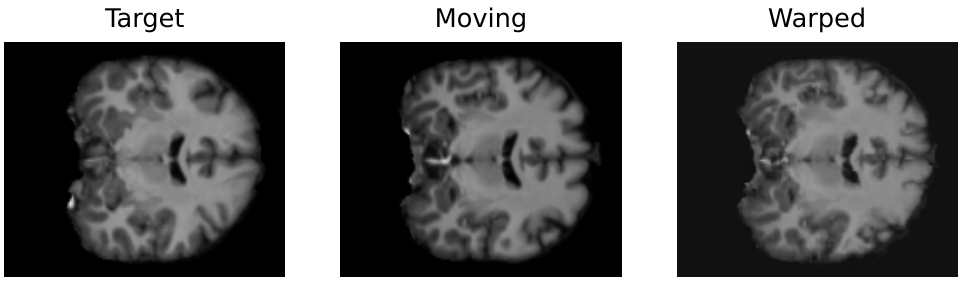
field was used as the regularization loss on the flow field. For both datasets, an ADAM optimizer with learning rate of 3×10^{-4} was used.

LAPIRN, another multi-resolution model, was trained on OASIS using training parameters from [10]. For NLST, training was carried out with a supervised discrepancy loss between the fixed and moving keypoints once the displacement is applied to the moving keypoints. MSE was used as the image-space loss for NLST, as it was found that it improved the performance on the validation dataset as when compared to normalized cross correlation (NCC). ConvexAdam, an optimization-based method, was found to outperform LAPIRN for the large displacement dataset and was also tested. Since the approach also incorporates instance-specific fine tuning, ablation tests were performed on ConvexAdam without instance-specific fine tuning to study performance of discrete optimization step on the dataset.

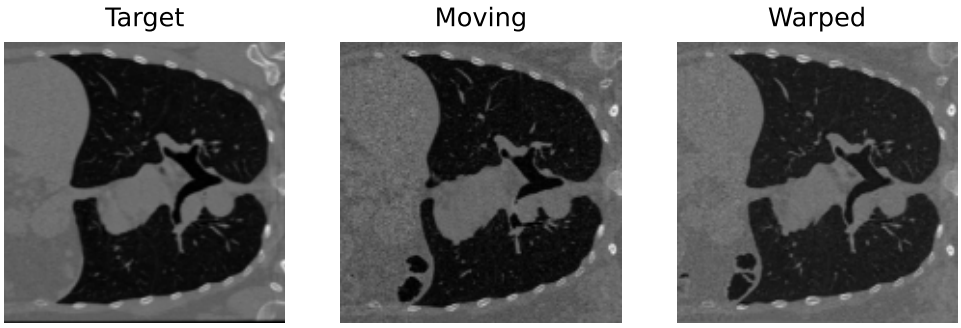
3.4 Results

Figure 3.4 shows qualitative results generated for the NLST and OASIS datasets. Table 3.2 summarizes the results when comparing RWCNet with LAPIRN and ConvexAdam on the NLST and OASIS datasets. Note that RWCNet without fine-tuning outperforms LAPIRN and simple convex optimization without instance optimization on the NLST dataset. With Adam optimization, both methods converge to the same point. RWCNet is also able to outperform both methods on the validation split of OASIS. Fine-tuning with Adam actually worsens the performance of the model on the OASIS dataset, as it converges to the same local optima as ConvexAdam.

Table 3.3 shows results from the architectural ablation tests. These ablation results 3.3 provide interesting insights into the role that architecture plays in registration accuracy in different datasets. Unsurprisingly, multi-resolution registration plays a crucial role in the accuracy of RWCNet; registering at only 4× downsampling on the NLST dataset yields a keypoint discrepancy of 5.52 mm, whereas registering at multiple resolutions yields a



(a) OASIS sample results



(b) NLST sample results

Figure 3.4: Qualitative results, showing the target, moving and warped images for OASIS (a) and NLST (b).

Table 3.2: Experiment Results on NLST and OASIS Validation

Experiment	NLST Keypoint Accuracy (mm) ↓	OASIS Dice (%) ↑
Zero Displacement	9.73	52.4
ConvexAdam [24]	1.48	76.4
ConvexAdam w/o IO [24]	2.78	75.6
Direct optimization with ADAM	2.03	76.2
LAPIRN [10]	5.51	80.0
RWCNet	**2.11**	**80.7**
RWCNet with ADAM IO	**1.48**	**76.4**

discrepancy of 2.11mm. OASIS Dice, likewise, drops from 80.7% to 74.0%.

The impact of correlation and number of RNN time steps is markedly different for both datasets. In the OASIS datasets, replacing correlation with stacking of the input feature tensors does not drastically impact the registration performance. The Dice score drops by 0.7, which might be negligible and might be explainable by the randomness inherent in neural network training. Likewise, when only 2 RNN time steps are used in RWCNet, the drop in accuracy is even lower in the OASIS dataset; the Dice score only drops by 0.6%.

This is in contrast to NLST, where decreasing the number of RNN time steps and removing correlation drastically decrease performance. The keypoint accuracy decreases

Table 3.3: Ablation Experiment Results on NLST and OASIS Datasets

Experiment	NLST Keypoint Accuracy (mm) ↓	OASIS Dice (%) ↑
RWCNet	2.11	80.7
RWCNet with single resolution (4x)	5.52	74
RWCNet without correlation	4.10	80.0
RWCNet with 2-timestep GRU	5.17	80.1

Table 3.4: Inference times for LAPIRN, RWCNet and ConvexAdam

Registration Method	Inference Time (seconds)
RWCNet	**26.1**
ConvexAdam	7.8
LAPIRN	0.33

to 4.10mm when correlation is not computed. Likewise, when only 2 time steps are used, the accuracy decreases to 5.17mm. As mentioned, NLST is a large displacement dataset, whereas the OASIS dataset requires small displacements for registration. The results suggest that cost volume computation and recurrent networks provide better support for generalization on large displacement datasets.

Table 3.4 shows the inference speeds of the three methods. These speeds were measured on an A100 GPU with 32GB GPU RAM. Note, that RWCNet takes considerably longer. This is mostly due to the need for patching and unifying tensors, which is performed on the CPU to conserve memory.

3.5 Chapter Summary

The DLIR network developed in this thesis, RWCNet, borrows key ideas from optical flow literature; specifically that of cost volume computation and iterative refinement, to improve medical image registration performance. For the large displacement NLST dataset, it was found that the model does not outperform variational optimization methods, although it is more performant than comparable deep learning methods. In the small displacement setting, RWCNet has comparable performance to another multi-scale model, LAPIRN. Instance optimization, which entails the fine tuning of the flow field using an ADAM optimizer, does improve the performance of RWCNet so that it is similar to the performance of ConvexAdam. Unfortunately, directly adopting instance optimization to the network is problematic, as it worsens the performance of the network on the OASIS dataset. Future work should investigate whether the instance optimization step can be augmented with features learned by the network to maintain accuracy on the small-displacement task while still maintaining gains for the large-displacement registration task.

The key takeaway of this work stems from the ablation studies; it shows that architectural features such as explicit cost volume computation and iterative refinement via

RNN improves the ability of the network to generalize in the large displacement setting, while having negligible impact on the performance in the small displacement setting. Future work might benefit from this finding and, in practice, it might justify the need for dataset-specific registration architectures.

Chapter 4

Conclusion

This chapter summarizes the contributions from Chapter 3 in the broader context of medical image registration and image analysis. Next steps for better understanding the results of this work and for improving the accuracy and usability of the architecture are discussed.

4.1 Summary of Thesis and Contributions

DLIR techniques proposed in recent years have typically been outperformed by direct optimization methods in large displacement settings. In this work, it was shown that performance can be improved on a large displacement Lung CT dataset by combining explicit cost-volume computation and multi-scale iterative refinement. With instance optimization, RWCNet is competitive with ConvexAdam, achieving the same keypoint discrepancy error of 1.48mm. Without fine tuning, the registration accuracy in the lung dataset is worse, by about 1mm. In small displacement settings, RWCNet achieves competitive performance with LAPIRN, with marginally higher Dice loss of 80.7%, compared to 80.1%. With fine tuning, however, the Dice loss drops to around 74.0%.

Ablations were performed to illustrate the benefits of the adopted methods. It was shown that multi-scale iterative refinement plays a key role; performing registration at a single resolution drastically decreases the accuracy of the model. RNN steps and explicit cost volume computation are also significant, however they mostly play a role in the large displacement dataset. On OASIS, the Dice score does not change drastically without these features. This suggests that these architectural features are valuable in large displacement settings. This is a significant finding, as it can inform future work for large displacement

settings. The next section more concretely details future work that needs to be carried out.

4.2 Future Work

4.2.1 Future Experiments

The network was trained and validated against inter-subject brain MRI data and lung CT data. While these datasets offer their own unique challenges, they do not wholly represent the variety of datasets that need to make use of registration. Future work should test this model against multimodal datasets, such as CT-MR or even MR-Ultrasound. Another setting where this model hasn't been tested is that of hard-tissue registration. For example, the spinal cord does not deform in a manner similar to soft tissue. Performing registration on such a dataset can yield an interesting point of comparison to the results presented in this paper.

4.2.2 Automated Hyperparameter Tuning

The findings in this paper highlight the need for dataset-specific hyperparameter selection. In the small displacement setting RWCNet yields almost identical Dice accuracy as LapIRN. In the larger displacement setting, RWCNet drastically outperforms LapIRN. Unfortunately, RWCNet is more resource intensive than LapIRN and its inference speed is significantly slower than LapIRN. Thus, there is no advantage to using RWCNet in small displacement settings. Automatic self-configuring network based on the attributes of the dataset, similar to nnU-net [38] for medical image segmentation, would be a desirable property for the usability of DLIR. Such a tool can help clinical researchers use this network without having to worry about the minutiae of registration, since registration is often a means to an end in many clinical analysis pipelines, such as tractography.

4.2.3 Strategies for Improving Generalization

Lastly, the ability of the model to generalize to data outside of the training set is not well explored. Even in situations where large datasets can be obtained for training, the distribution of images in clinical practice can be long-tail. In practice, the image might be captured using a different imaging protocol than those in the training dataset, or the image might reflect some disease or anomaly that is not present in the training dataset.

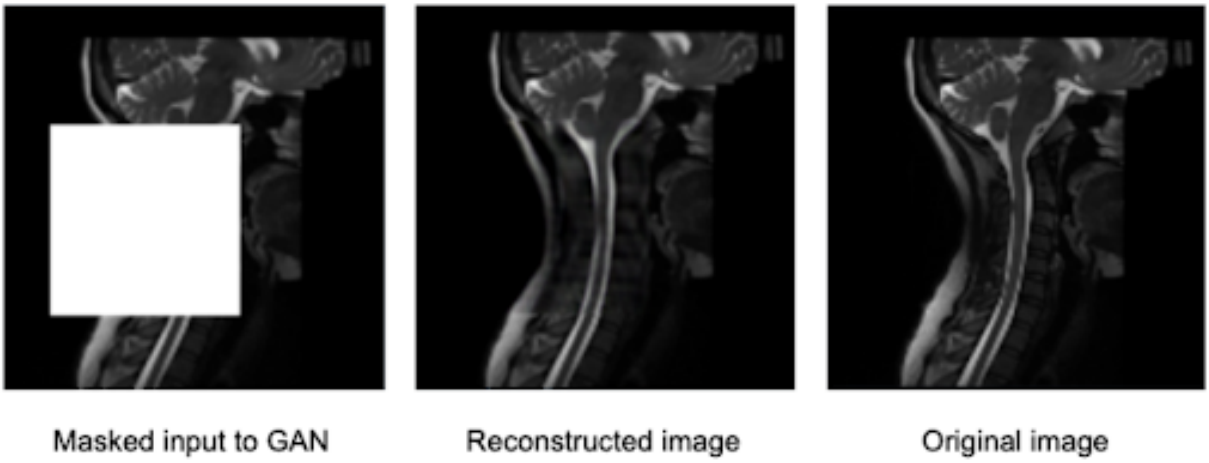


Figure 4.1: Qualitative comparison of image inpainting result (center) using a 2D generative network with input (left) and original image (right)

There are several strategies for improving generalization accuracy. Data augmentation to artificially introduce noise or other anomalous behaviour can, in practice, improve generalization accuracy of DLIR [39]. Self-supervised training is another approach to improving generalization accuracy. To the author's knowledge, self-supervised training has not been widely studied for image registration. Inpainting, which is the task of generating an image from masked data, is a common pretext task. Figure shows an example of inpainting, using a network trained by the author of this thesis, applied to the spine. Future work should study the effect of using inpainting and other self-supervised learning strategies to initialize weights for image registration.

Another avenue for improved registration is to generate synthetic displacement fields during training time and to modify training so that the network tries to learn the displacement field in a 'supervised' fashion. Such an approach has been studied widely in literature, but suffers from the limitation that it is not easy to generate physiologically accurate displacements at training time [40]. For the spinal cord tractography problem, there might be a way to combine finite element modelling of the spine to provide a way to generate physiologically realistic deformations for this problem space. While there is some prior work [41] in this space, there is space for future work to further explore this problem.

References

[1] Petter Risholm, Alexandra J. Golby, and William M. Wells. Multi-Modal Image Registration for Pre-Operative planning and Image Guided Neurosurgical Procedures. *Neurosurgery clinics of North America*, 22(2):197–206, April 2011.

[2] Vladlena Gorbunova, Pechin Lol, Haseem Ashraf, Asger Dirksen, Mads Nielsen, and Marleen de Bruijne. Weight preserving image registration for monitoring disease progression in lung CT. *Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 11(Pt 2):863–870, 2008.

[3] Fengxiang Li, Jianbin Li, Zhifang Ma, Yingjie Zhang, Jun Xing, Huanpeng Qi, and Dongping Shang. Comparison of internal target volumes defined on 3-dimensional, 4-dimensonal, and cone-beam CT images of non-small-cell lung cancer. *OncoTargets and Therapy*, 9:6945–6951, 2016.

[4] A.W. Toga and P.M. Thompson. The role of image registration in brain mapping. *Image and vision computing*, 19(1-2):3–24, January 2001.

[5] Susan M. Sunkin, Lydia Ng, Chris Lau, Tim Dolbeare, Terri L. Gilbert, Carol L. Thompson, Michael Hawrylycz, and Chinh Dang. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Research*, 41(Database issue):D996–D1008, January 2013.

[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. *Medical Image Analysis*, 12(1):26–41, February 2008.

[7] Stewart McLachlin, Jason Leung, Vignesh Sivan, Pierre-Olivier Quirion, Phoenix Wilkie, Julien Cohen-Adad, Cari Marisa Whyne, and Michael Raymond Hardisty.

Spatial correspondence of spinal cord white matter tracts using diffusion tensor imaging, fibre tractography, and atlas-based segmentation. *Neuroradiology*, 63(3):373–380, March 2021.

[8] Jose Soares, Paulo Marques, Victor Alves, and Nuno Sousa. A hitchhiker's guide to diffusion tensor imaging. *Frontiers in Neuroscience*, 7, 2013. Publisher: Frontiers.

[9] Benjamin M. Ellingson, Noriko Salamon, Davis C. Woodworth, and Langston T. Holly. Degree of subvoxel spinal cord compression measured with super-resolution tract density imaging (TDI) correlates with neurological impairment in cervical spondylotic myelopathy. *Journal of neurosurgery. Spine*, 22(6):631–638, June 2015.

[10] Tony C. W. Mok and Albert C. S. Chung. Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks, June 2020. arXiv:2006.16148 [cs, eess].

[11] Shengyu Zhao, Yue Dong, Eric I.-Chao Chang, and Yan Xu. Recursive Cascaded Networks for Unsupervised Medical Image Registration. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10599–10609, October 2019. arXiv:1907.12353 [cs].

[12] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning Optical Flow with Convolutional Networks, May 2015. arXiv:1504.06852 [cs].

[13] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks, December 2016. arXiv:1612.01925 [cs].

[14] Mattias P. Heinrich. Closing the Gap between Deep and Conventional Image Registration using Probabilistic Dense Displacement Networks, July 2019. arXiv:1907.10931 [cs].

[15] Mattias Paul Heinrich, Mark Jenkinson, Bartlomiej W. Papież, Sir Michael Brady, and Julia A. Schnabel. Towards Realtime Multimodal Fusion for Image-Guided Interventions Using Self-similarities. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Camille Salinesi, Moira C. Norrie, and Óscar Pastor, editors, *Advanced Information Systems Engineering*, volume 7908,

pages 187–194. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. Series Title: Lecture Notes in Computer Science.

[16] Torsten Rohlfing. Image Similarity and Tissue Overlaps as Surrogates for Image Registration Accuracy: Widely Used but Unreliable. *IEEE Transactions on Medical Imaging*, 31(2):153–163, February 2012. Conference Name: IEEE Transactions on Medical Imaging.

[17] John Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95–113, October 2007.

[18] J.-P. Thirion. Non-rigid matching using demons. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 245–251, June 1996. ISSN: 1063-6919.

[19] J. P. Thirion. Image matching as a diffusion process: an analogy with Maxwell's demons. *Medical Image Analysis*, 2(3):243–260, September 1998.

[20] Harvey S. Leff and Andrew F. Rex, editors. *Maxwell's demon 2: entropy, classical and quantum information, computing*. Institute of Physics, Bristol ; Philadelphia, 2003. OCLC: ocm51569169.

[21] Mattias P. Heinrich, Bartlomiej W. Papież, Julia A. Schnabel, and Heinz Handels. Non-parametric Discrete Registration with Convex Optimisation. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Alfred Kobsa, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Demetri Terzopoulos, Doug Tygar, Gerhard Weikum, Sébastien Ourselin, and Marc Modat, editors, *Biomedical Image Registration*, volume 8545, pages 51–61. Springer International Publishing, Cham, 2014. Series Title: Lecture Notes in Computer Science.

[22] Benjamin M Glocker. Random Fields for Image Registration. page 116.

[23] Ben Glocker, A. Sotiras, N. Komodakis, and N. Paragios. Deformable medical image registration: setting the state of the art with discrete methods. *Annual review of biomedical engineering*, 2011.

[24] Hanna Siebert, Lasse Hansen, and Mattias P. Heinrich. Fast 3D Registration with Accurate Optimisation and Little Learning for Learn2Reg 2021. In Marc Aubreville, David Zimmerer, and Mattias Heinrich, editors, *Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis*, volume 13166, pages 174–179.

Springer International Publishing, Cham, 2022. Series Title: Lecture Notes in Computer Science.

[25] Mattias P. Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V. Gleeson, Sir Michael Brady, and Julia A. Schnabel. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical Image Analysis*, 16(7):1423–1435, October 2012.

[26] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017. arXiv: 1412.6980.

[27] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks, February 2016. arXiv:1506.02025 [cs].

[28] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, August 2019. arXiv: 1809.05231.

[29] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, June 2016. arXiv:1606.06650 [cs].

[30] Mattias P. Heinrich and Lasse Hansen. Voxelmorph++ Going beyond the cranial vault with keypoint supervision and multi-channel instance optimisation, February 2022. arXiv:2203.00046 [cs].

[31] Jo Schlemper, Jose Caballero, Joseph V. Hajnal, Anthony Price, and Daniel Rueckert. A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction. *arXiv:1704.02422 [cs]*, November 2017. arXiv: 1704.02422.

[32] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow, August 2020. arXiv:2003.12039 [cs].

[33] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume, June 2018. arXiv:1709.02371 [cs].

[34] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches, October 2014. arXiv:1409.1259 [cs, stat].

[35] Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, September 2007.

[36] National Lung Screening Trial Research Team. The National Lung Screening Trial: Overview and Study Design. *Radiology*, 258(1):243–253, January 2011.

[37] Learn2Reg - Grand Challenge.

[38] nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation | Nature Methods.

[39] Kh Tohidul Islam, Sudanthi Wijewickrema, and Stephen O'Leary. A deep learning based framework for the registration of three dimensional multi-modal medical images of the head. *Scientific Reports*, 11(1):1860, January 2021. Number: 1 Publisher: Nature Publishing Group.

[40] Yabo Fu, Yang Lei, Tonghe Wang, Walter J Curran, Tian Liu, and Xiaofeng Yang. Deep learning in medical image registration: a review. *Physics in Medicine & Biology*, 65(20):20TR01, October 2020.

[41] Jihun Kim, Kazuhiro Saitou, Martha M. Matuszak, and James M. Balter. A finite element head and neck model as a supportive tool for deformable image registration. *International Journal of Computer Assisted Radiology and Surgery*, 11(7):1311–1317, July 2016.