Longitudinal VoxelMorph: Spatiotemporal Modeling of Medical Images

by

Adelaide Woods Chambers

B.S., Electrical Engineering and Computer Science, M.I.T., 2019

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ENGINEERING
IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MAY 2020

Signature of Author: ______________________________________

Department of Electrical Engineering and Computer Science
May 12, 2020

Certified by: ______________________________________

John Guttag
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by: ______________________________________

Adrian Dalca
Assistant Professor, Harvard Medical School
Thesis Supervisor

Accepted by: ______________________________________

Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Longitudinal VoxelMorph: Spatiotemporal Modeling of Medical Images

by
Adelaide Woods Chambers

Submitted to the Department of Electrical Engineering and Computer Science
on May 12th, 2020 in partial fulfillment of the requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Medical image registration is an important initial step in many downstream clinical tasks. Repeated imaging for diagnostic, therapeutic, or scientific discovery is common. Classical longitudinal image registration systems are too slow to be useful in practice and are often only designed for a highly specific type of image data. Efficient pairwise image registration models are limited by not accounting for the temporal nature of the data. We present Longitudinal VoxelMorph, a novel machine-learning-based model for efficient and scalable spatiotemporal medical image registration. We also define a new evaluation metric to quantify the temporal smoothness of a longitudinal deformation field. We evaluate the model on cardiac cine-MRI data and echocardiography data, and find that Longitudinal VoxelMorph is more temporally consistent than state-of-the-art pairwise models, and achieves comparable or improved anatomical accuracy. Longitudinal VoxelMorph has the potential to be incorporated in downstream medical image tasks, such as image prediction and diagnosis, facilitating better clinical outcomes.

Thesis Supervisor: John Guttag
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Adrian Dalca
Title: Assistant Professor, Harvard Medical School

Acknowledgments

This thesis would not have been possible without the support and guidance of many people. First, I would like to thank my wonderful advisors, Prof. John Guttag and Prof. Adrian Dalca, for all of their mentorship. They have both eased the transition from undergraduate to graduate studies, and helped me grow as a researcher, thinker, and communicator.

I would also like to thank all the members of my lab, the Data Driven Inference Group at CSAIL. All of their feedback helped provide direction and refinement throughout the year, and I have really enjoyed hearing about the wide range of work done in the lab. I will especially miss the coffee breaks and impromptu book-club discussions next year.

My aunt, Ann McLaughlin, also helped this work by serving as a sounding board for all of my ideas. I had lots of fun at our Friday evening dinner and drinks sessions, excitedly discussing the week's successes and joking about the disasters.

I thank Matt Woicik, my amazing boyfriend, for always being there to celebrate the highs and support me in the lows. I'm grateful to him for making MIT infinitely more rewarding and full of laughter.

Finally, I'd like to thank my family. My awesome dad, Craig Chambers, and best-big-sister-ever, Caitlin Chambers, have been pillars of stability and love throughout my whole life. I'm very grateful for everything they've done for me and for their belief that, no matter what I set my mind to, I can accomplish it. And most especially, I'd like to thank my mom-by-blood and best-friend-by-choice, Sylvia Chambers, for having always been there for me. Even after reading this hundreds of times, I'm sure she would have really loved it.

Contents

List of Figures

Chapter 1

Introduction

MEDICAL imaging has become widespread over the past two decades [77]. Physicians use medical images both as diagnostic tests and to monitor ongoing treatment, and the images have become an important component of medical care.

Medical images can be 2-D, such as x-rays, or 3-D, such as Magnetic Resonance Imaging (MRI) scans. A 2-D image is made up of *pixels*; a 3-D scan is made up of *voxels*. A 3-D scan can be decomposed into 2-D *slices* by only considering voxels at a specific location along one of the image axes.

Image registration is a common pre-processing step for many image analysis tasks [38, 43, 60, 85]. In general, image registration determines an alignment between two or more images to place them in the same reference frame or measure correspondences between them. In medical imaging, registration often aligns an image for a specific patient to an atlas (reference) image, which usually represents a prototypical image for that patient's population [3, 14, 15, 31, 38, 80]. In the case of a brain image, for example, the registration could measure how the size of the patient's hippocampus differs from what is considered normal [84].

Recently, machine learning models have been employed to automate medical image registration [11, 14, 19, 23, 31, 40, 72, 90, 91]. To register images, these models usually compute a deformation field that maps between pixels in the images of interest [14, 22]. Different anatomical structures or tissue types are usually manifested as different pixel intensities, so a mapping between pixels of similar intensities across images is a common proxy for a mapping between structures across images [14, 22, 34].

Patients also often undergo repeated medical imaging. Repeated imaging of the same body part can be done during the diagnosis or treatment phase of an illness or injury, as repeated screenings during preventative care checkups, or as part of a medical study. The temporal medical data can be interpreted either as images taken from the same subject at distinct time points, or as frames in a video of that patient's anatomical activity.

This presents the need for image registration over time, also known as *spatiotemporal image registration* [22, 25, 31, 46, 50, 86]. Spatiotemporal image registration aligns multiple images taken from the same subject over time to each other. For n-D images, registration can be viewed as an $(n+2)$-D deformation field estimation problem, where the additional dimensions are space and time. In the case of repeated brain images, the registration could, for example, now measure the expansion of the patient's ventricles over time [28].

We present Longitudinal VoxelMorph, a novel machine-learning model for spatiotemporal image registration. In this work, we will use 2-D image examples. We therefore refer to images and pixels when discussing medical imaging artifacts and their constituents. Whenever data were originally 3-D scans, we consider 2-D slices of the scans instead. The modeling ideas, however, extend to higher dimensions for scans and voxels. We evaluate the model on several real-world medical image datasets, and find that the model achieves a more temporally smooth deformation field than state-of-the-art pairwise image registration models, while maintaining or improving accuracy on each individual image.

■ 1.1 Spatiotemporal Modeling Applications

An efficient and accurate model for spatiotemporal medical image registration has several important applications. First, registration between images from the same subject is useful for *segmentation propagation* [39, 49, 92]. Image segmentations identify specific regions of an image. In cardiac imaging, for example, the left ventricle is often segmented. The volume of the left ventricle at different phases of the cardiac cycle is used to compute the ejection fraction, which is an indicator for a patient's risk of cardiac failure, and is therefore a quantity of clinical interest [37].

However, it is costly for experts to segment these images. Particularly in the case of repeated images for the same subject, manually segmenting all images in the time series, of which there might be hundreds, is not practical. Instead, it would be ideal if a registration model could, given a segmentation for a single image in that time series, propagate the segmentation to all other images.

Second, there is clinical meaning in a deformation that measures changes in a specific structure over time. Although all information content about how a structure changes over time is present in the images themselves, an automatically computed deformation field can draw attention to important parts of the image and quantify changes.

Third, closely measuring development has medical significance. Brain development progression, for example, is important in both neurodegenerative and developmental studies. Tracking brain structures over time is common in both neuroscience and clinical neurology [4, 51, 52]. Previous work estimates an Alzheimer's disease (AD) patient's future symptom class trajectory by using two Magnetic Resonance Imaging (MRI) scans from that patient, as well as some genetic and clinical information gathered at the time of the MRIs [17]. Other work predicts future MRI scans for an AD patient from a single MRI and accompanying genetic and clinical information [13, 66].

Fourth, there is benefit in modeling the short-term changes of a lesion. For example, targeted radiation therapy for non-small cell lung cancer (NSCLC) patients is made more difficult by patient respiration. While the patient breathes, the lung tumor moves, making it hard to accurately target the cancerous cells while minimizing radiation to the surrounding healthy cells [58]. Models that track tumor movement over the course of a patient's breath cycle help mitigate the possibility of error during radiation, and have been introduced for this specific use case [33, 58].

Finally, image registration is a useful precursor to other important clinical tasks, such as prediction. In the example of symptom-trajectory prediction for AD patients, we believe that using as input to the prediction model a time series of more than two MRIs per patient will enable the model to achieve higher accuracy. Although this work will not explore prediction given a time series of images, it is a natural extension and an interesting avenue for future work.

■ 1.2 Current Medical Image Registration Techniques

Medical image registration between an image taken from a patient and a synthesized atlas image, often created to define a prototypical image from a population of interest, is a well-explored problem. Classical medical image registration models have optimized a loss function combining an image similarity loss and regularization penalty term for each pair of images. Classical methods are discussed more in Section 2.2.

Existing longitudinal image registration pipelines are often optimization-based, so run slowly over datasets of large images, or are specialized to a specific kind of image data. Current longitudinal medical image registration pipelines like FreeSurfer end up using neither optimization-based longitudinal registration models nor pairwise learning-based models, and instead create a template image for each subject before performing local optimizations to align the template image to the original input images [67].

Efficient and accurate learning-based models that are designed for use across different kinds of medical imaging data exist, but are often pairwise [14, 22]. To efficiently construct a deformation between the patient's image and the atlas, these pairwise methods commonly use a deep neural network to estimate a mapping between pixels in the two images. The mapping can be interpreted as a flow field, moving pixels in the patient's image to the location in the atlas image to which they align [14, 18, 22, 23, 60, 85, 90].

Although these subject-to-atlas registration models were not designed for longitudinal data, the idea of a deformation between images can be extended to image time series. Rather than registering a patient's image to an atlas, the same models can be converted to map between two images taken from the same patient at different points in time, creating a pairwise deformation. An overall deformation can be computed as a piecewise combination of the pairwise deformations for that patient. When pairwise models are used for image time series, the deformations they produce cannot benefit from information contained in any images apart from the two they consider at any given time. Even though pairwise models do not take advantage of the temporal aspect of the data, they are a reasonable baseline method for spatiotemporal registration.

Recently, registration of temporal medical data has also been investigated. One approach uses Gaussian kernels to model intensity changes over time for both inter- and intra-subject n-D registration [34]. Others use the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework [22]. These compose a spatial deformation and a temporal deformation. In a diseased population, for example, the average patient trajectory can be modeled as the temporal deformation, while a specific patient's deviation for the mean is modeled as the spatial deformation [22, 31]. Most of these models are infeasible in practice given the size of modern medical images and the number of images that might be included in a time series [31].

Some models seek to overcome this efficiency limitation by re-framing the image registration task as a longitudinal regression problem [34]. They extract low-dimensional characteristics of the image, such as the brain's posterior thalamic radiation, and estimate its development over time to address a specific clinical task. This is an effective solution to the estimation speed problem, but limits the utility of the model. Such a model is only useful when the task can be re-framed using a low-dimensional representation of the image.

Another recent approach for spatiotemporal image data uses deep-learning to efficiently estimate a deformation field between images. This method is similar to frame prediction for short-term temporal data, when the images can easily be interpreted as frames in a video. In this case, frames can be registered using a conditional variational autoencoder (CVAE) to create deformations, where the structure of previous and future frames inform the deformation for the current frame [46]. This approach, however, was designed for medical data where time between images is under a second, and therefore might not scale well to data that have larger time gaps.

1.3 Longitudinal VoxelMorph

In this work, we present Longitudinal VoxelMorph, a model for n-D spatiotemporal image registration across a time series of medical images taken from the same subject. Our main contributions are:

- **Longitudinal VoxelMorph:** We define a novel spatiotemporal image registration model with a b-spline representation that enables scalability. We implement a learning-based approximation network to efficiently register a time series of medical images.

- **Longitudinal Metric:** We propose a new evaluation metric, D*, to extend the existing pairwise Dice score for image registration to longitudinal data.

- **Real-world Evaluation:** We perform experiments on two medical image datasets, and find that Longitudinal VoxelMorph produces a deformation that is smoother in time than baseline state-of-the-art pairwise models, while achieving comparable or improved registration accuracy.

This thesis is structured as follows. Chapter 2 will supply background on medical imaging, image registration, and b-splines that will be useful in understanding the constraints and advantages of the model. Chapter 3 will define Longitudinal VoxelMorph. Chapter 4 will introduce useful metrics to quantify performance of spatiotemporal registration models, and Chapter 5 will present the results of the model in several experimental settings. Chapter 6 concludes the work, and outlines future research directions.

Chapter 2

Background

IN this chapter, we explore several medical imaging modalities and mathematical concepts relevant to this work, and provide some background on classical medical image registration. Together, they help motivate and clarify the model defined in the subsequent chapter.

■ 2.1 Medical Imaging Background

Medical imaging is frequently used as the principal method of diagnosing a wide range of conditions and monitoring ongoing treatment. Nearly all of us will undergo medical imaging many times over the course of our lives, varying from dental x-rays on a yearly basis to fetal ultrasounds during pregnancy. In this work, we use magnetic resonance images (MRIs), cine imaging data, and ultrasound.

All medical images are somewhat noisy, meaning the image they produce does not perfectly capture the underlying anatomy. There are many possible sources of this noise. Noise can be introduced by variation between machine settings, patient movement, or imperfect measurements. In the case of longitudinal modeling, where the images of interest come from the same subject, such variability can sometimes represent a substantial portion of the differences between two images [68]. A good longitudinal model should therefore be prepared to handle noisy data.

■ 2.1.1 Magnetic Resonance Imaging

Magnetic Resonance Images (MRIs) are 3-D grayscale images, often capturing specific body parts. MRI is a non-invasive technology that does not require harmful ionizing radiation, unlike other common medical imaging techniques like x-rays or computed tomography (CT/CAT) scans. MRI technology is based on nuclear magnetic resonance (NMR) technology, which was initially used to analyze molecule properties [29]. To produce an image, an MRI machine first creates a strong magnetic field, causing the protons in the body's tissue to align to the field. A set of radio frequency (RF) pulses are then sent through the patient [2, 29]. These RF pulses temporarily cause the protons to spin out of their alignment. The speed at which a molecule's protons return to alignment with the magnetic field depends on the chemical properties of the tissue. The speed can be measured to microsecond accuracy by the MRI machine [29]. An

MRI machine is programmed to identify different tissue types based on the speed at which the protons re-align to the magnetic field and then to reconstruct an image of the patient's underlying anatomy [2].

MRI is a widespread and powerful medical imaging technique, and is used for many interesting research datasets. The RF pulses can pass through a patient's bone structures without degrading, creating accurate images of tissues inside of or in close proximity to bony areas of the body [29]. The accurate images and lack of ionizing radiation make MRI the modality of choice for brain and fetal diagnostic imaging, or when repeated imaging is needed for either treatment or diagnosis [2]. This lends itself especially well to longitudinal medical imaging studies. It also enables clinical studies that include healthy patients by removing ethical concerns surrounding radiation exposure [53]. Despite their benefits, MRIs are more expensive to obtain than x-ray or CT scans and require more expertise to interpret, so are not always used [2, 53].

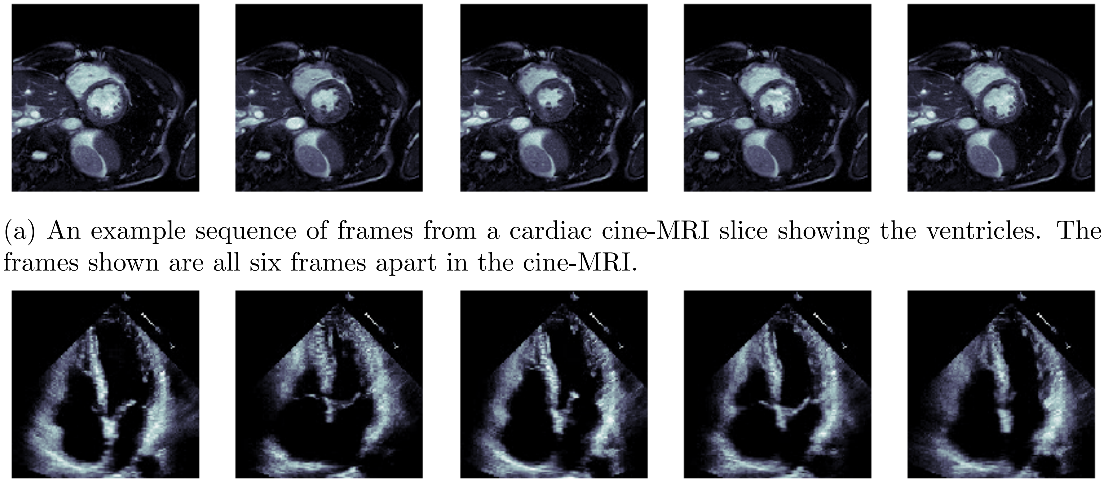
■ 2.1.2 Cine-MRI

Cine images are a set of images that, together, can be interpreted as frames in a video. Cine-MRIs are a set of repeated MRIs taken in close succession that are then combined to form a brief video. Cardiac cine-MRIs, which show the tissues in the heart over the course of a heartbeat, are a common type of cine-MRI. They are often constructed in conjunction with electrocardiograms (ECGs) that, for each frame, measure its location in the cardiac cycle [5].

For a normal MRI, multiple RF pulses are emitted to measure the *average* re-alignment speed, as any individual measurement might be incorrect. Similarly, cine-MRIs are usually averaged given multiple cycles of measurement. In the case of the heart, each frame is constructed from RF pulses taken over several non-ectopic heartbeats. Combined with ECG data measuring the phase in the cardiac cycle, the re-alignment speeds for a specific phase are averaged, creating re-alignment data for a synthetic heartbeat. This synthetic heartbeat data is then used to produce the cine-MRI frames. Figure 2.1a shows an example frame from a cardiac cine-MRI slice. Collecting data across multiple heartbeats can require patients to hold their breath for ten to twenty seconds (the duration of a single resting heartbeat often ranges between 0.7 and 1.25 seconds) introducing another source of variability between the cine-MRI frames if patients are unable or forget to do so [5, 10].

■ 2.1.3 Ultrasound Imaging

Diagnostic ultrasound imaging is a non-invasive imaging technique that uses sound waves to produce 2-D images of tissues and organs. These sound waves have higher than 20KHz frequencies, above the human hearing range (hence the name *ultra*sound), but modern ultrasound machines usually have MHz frequencies. These ultrasound waves are sent into the human body using a transducer, and are reflected back towards the transducer when the waves hit a tissue boundary. The transducer can measure the speed at which the wave is traveling and the time it took to return. Similar to



(a) An example sequence of frames from a cardiac cine-MRI slice showing the ventricles. The frames shown are all six frames apart in the cine-MRI.

(b) An example sequence of frames from an echocardiogram showing the ventricles. The frames shown are all 19 frames apart in the echocardiogram.

Figure 2.1: Basic example frames of a cardiac cine-MRI and echocardiogram, each showing the ventricles of the heart from a different view. Although the cardiac cine-MRI frames show more detail and much sharper boundaries between structures, it is also much more expensive to obtain and the added detail is not always clinically necessary [64].

echolocation, the ultrasound scanner can calculate the distance between the transducer and the tissue boundary, and use this information to produce an image [6].

Ultrasound machines enable non-invasive imaging of internal organs, and do not require ionizing radiation [6]. They are not, however, generally accurate in areas with bones or with air pockets [6]. Only when the surrounding area is fully or partially filled with fluid can ultrasounds sometimes image bones, most notably for fetal ultrasounds [6]. A cardiac ultrasound is also called an echocardiogram.

Although the quality and resolution of an echocardiogram is not as high as that of a cardiac MRI, as compared in Figure 2.1, a cardiac MRI costs over 5.5 times more than an echocardiogram, as of 2005 [64]. It is therefore valuable to create an image registration model that works well for both data modalities.

■ 2.2 Classical Image Registration Overview

Medical image registration is a well-explored field [8, 12, 16, 24, 35, 79, 88, 89]. Classical image registration techniques optimized a deformation field by minimizing the cost function

$$L(\phi, x, y) = L_{\mathsf{sim}}(\phi, x, y) + \lambda L_{\mathsf{smooth}}(\phi), \tag{2.1}$$

where ϕ is the deformation field estimated by the model and x and y are the two images being registered [14, 78]. L_{sim} is some image similarity loss, often mean squared error based on pixel intensity [14, 31], mutual information [65], or cross-correlation [11]. L_{smooth} enforces some smoothness constraint on the estimate of ϕ. The λ parameter conveys the tradeoff between an accurate end-to-end deformation field and an anatomical desire for smoothness.

Several well-established classical medical image registration models, including elastic-type models [12, 26, 75], statistical parametric mapping [9], b-spline free-form deformations (FFDs) [70], discrete methods [24, 35], and Demons [63, 79] interpret the deformation ϕ as a displacement vector field between two images. Methods that produce diffeomorphic deformation fields, thereby enforcing smooth changes over space and time, are particularly popular for medical image registration since the diffeomorphic constraint mirrors medical understanding of anatomical development and motion. Models that produce diffeomorphic displacement vector fields include Large Diffeomorphic Distance Metric Mapping (LDDMM) models [16, 20, 41, 42, 54, 87, 89], DARTEL [8], diffeomorphic demons [82], and standard symmetric normalization (SyN) [11].

These classical approaches are not learning-based, however, and instead optimize the cost function given in Equation (2.1) for every pair of new images. Even though new algorithms that adapt these methods for GPUs can solve this optimization problem for each pair of images on the order of minutes [56, 57], this does not scale well to studies over large population sizes or time series including many images [14].

2.3 Basis-spline Background

We use basis-splines (b-splines) as an efficient and scalable spatiotemporal representation for the estimated deformation field in Longitudinal VoxelMorph. B-splines are a common mathematical modeling formulation, first suggested by Isaac Schoenberg in 1946 [73]. B-splines are used to define a continuous curve or surface given a relatively small set of points, so can serve as a sparse parameterization. They are widely used because of their smoothness, generalizability, and efficiency [34, 36, 44, 62, 70, 71]. A simple example of a b-spline curve is given in Figure 2.2.

A b-spline of order k is defined by a set of control points, ϕ_{CP}, a set of knots $T = \{t_0, t_1, \ldots, t_n : t_0 \leq t_1 \leq \textbf{...} \leq t_n\}$ where n is the number of knots, and a set of basis functions $\{N_{i,j} : 1 \leq j \leq k, i \in \{0, 1, \ldots, n+k\}\}$. The b-spline is a piecewise combination of n polynomials of degree $k-1$, which are joined at the knot locations. Between two adjacent knots, the b-spline is therefore C^{∞} continuous. At a knot of multiplicity m, the b-spline is C^{k-m-1} continuous [62].

The basis functions are defined using the knot vector T. They can be recursively written as

$$N_{i,1}(t) = \begin{cases} 1, & \text{if } t_i \leq t \leq t_{i+1} \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

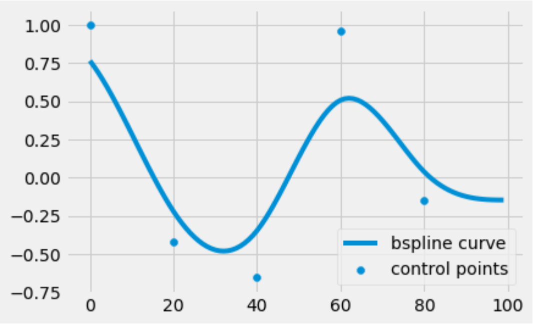


Figure 2.2: An example of a b-spline curve. Given simplifying assumptions, the curve is fully parameterized by the five control points, shown as blue dots.

and

$$N_{i,k}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} N_{i,k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} N_{i+1,k-1}(t). \tag{2.3}$$

This recurrence relation guarantees that $N_{i,k}(t) > 0$ for $t_i < t < t_i + k$, and that $N_{i,k}(t)$ is C^{k-2} continuous at each knot of multiplicity 1. B-splines have *local support*, i.e., $\forall t \notin [t_i, t_{i+k}] : N_{i,k}(t) = 0$ [62].

The b-spline curve is then defined using the basis functions $N_{i,k}$ and the control points $\phi_{CP} = \{\phi_0, ..., \phi_n\}$ as

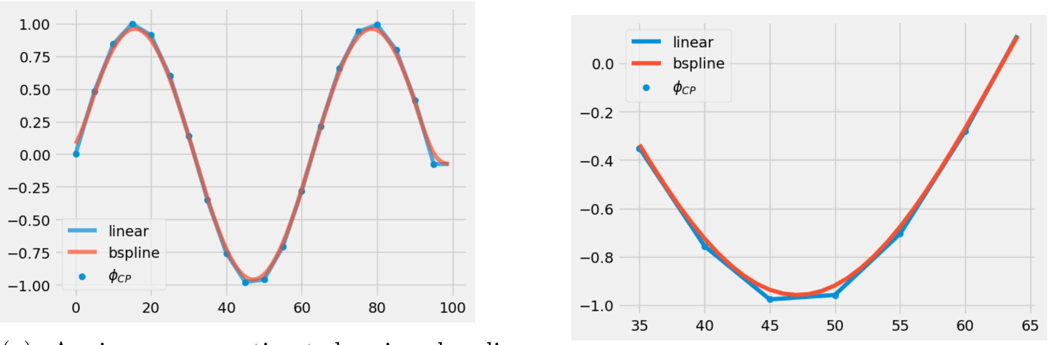
$$r(t) = \sum_{i=0}^{n} \phi_i N_{i,k}(t), \;\; n \geq k-1, t \in [t_{k-1}, t_{n+1}], \tag{2.4}$$

following de Boor's formulation [27, 62]. The local support property of the basis functions extends to the b-spline curve. Importantly, a single span of the b-spline curve is fully determined by only k control points and, conversely, a single control point only influences k spans. This is in contrast to other mathematical curve parameterizations, such as Bézier curves [62]. Local support provides a significant modeling advantage, since it enables us to tune only a few control points to adjust a specific part of the curve, and avoid affecting parts of the curve that are far away from those control points. It also decreases the computational resources required to calculate $r(t)$, since only the nearest control points are relevant [70].

The formulation of b-spline curves in higher-dimensional cases is analogous to two dimensions [62]. For example, in 3-D the b-spline surface is defined as

$$r(u, v) = \sum_{i=0}^{m} \sum_{j=0}^{n} \phi_{ij} N_{i,k}(u) N_{j,l}(v). \tag{2.5}$$

To simplify calculations and computation time, specific b-spline parameterizations can be written in a closed form. Cubic b-splines (i.e., b-splines of order $k = 4$), are commonly used for modeling problems [34, 36, 44, 70, 71], and we use them in the remainder



(a) A sine curve estimated using b-spline and linear interpolation, with control points placed 5 units apart.

(b) The estimated sine curve from Fig. 2.3a over a smaller window.

Figure 2.3: The result of b-spline and linear interpolation given a set of control points lying along a sine curve. As can be seen, the b-spline curve is smooth and does not directly pass through control point set ϕ_{CP}, unlike the linear curve.

of this work. When the knots T are placed uniformly, such that $\forall i \in [0, ..., n-1]$: $|t_{i+1} - t_i| = \Delta$, and all control points are spaced uniformly, separated by δ_d along the d-axis, the 4-D b-spline surface can be written in closed form as

$$r(x, y, z) = \sum_{l=0}^{3} \sum_{m=0}^{3} \sum_{n=0}^{3} B_l(\frac{x}{\delta_x} - \lfloor \frac{x}{\delta_x} \rfloor) B_m(\frac{y}{\delta_y} - \lfloor \frac{y}{\delta_y} \rfloor) B_n(\frac{z}{\delta_z} - \lfloor \frac{z}{\delta_z} \rfloor) \phi_{i+l, j+m, k+n}, \quad (2.6)$$

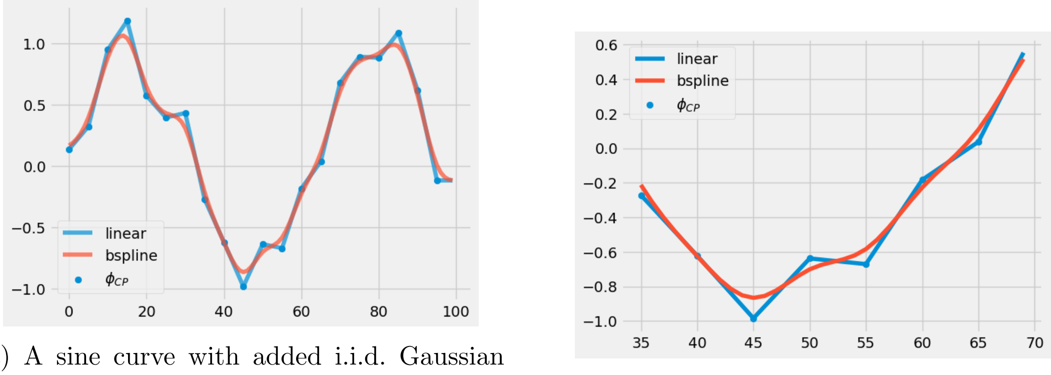
where $i = \lfloor \frac{x}{\delta_x} \rfloor$, $j = \lfloor \frac{y}{\delta_y} \rfloor$, and $k = \lfloor \frac{z}{\delta_z} \rfloor$ [70]. The basis functions B_l are defined by

$$B_l(u) = \begin{cases} \frac{(1-u)^3}{6} & \text{if } l = 0 \\ \frac{3u^3 - 6u^2 + 4}{6} & \text{if } l = 1 \\ \frac{-3u^3 + 3u^2 + 3u + 1}{6} & \text{if } l = 2 \\ \frac{u^3}{6} & \text{if } l = 3 \end{cases}. \quad (2.7)$$

The closed form given in Equation (2.6) exhibits the local support property, where only the four control points closest to point (x, y, z) in each dimension impact the value of $r(x, y, z)$. The work presented for the remainder of this thesis is based on the closed form b-spline parameterization given in Equations (2.6) and (2.7), using uniformly spaced control points and uniformly placed knot vectors.

■ 2.3.1 B-spline interpolation examples

A small set of control points can fully define b-spline curves with uniformly distributed knots and uniform control-point spacing. For example, a set of control points can be used to smoothly estimate a sine curve, as in Figure 2.3.



(a) A sine curve with added i.i.d. Gaussian noise $Y \sim \mathcal{N}(0, 0.25)$ estimated using b-spline and linear interpolation, with control points placed 5 units apart.

(b) The estimated sine curve with noise from Fig. 2.4a over a smaller window.

Figure 2.4: The result of b-spline and linear interpolation given a set of control points lying along a sine curve when i.i.d. Gaussian noise $Y \sim \mathcal{N}(0, 0.25)$ is introduced. In this case, the b-spline curve is better able to reflect the underlying sine curve since it is not constrained to pass directly through the control points.

This example illustrates two important differences between linear interpolation given ϕ_{CP} and b-spline interpolation given ϕ_{CP}. First, unlike linear interpolation, a b-spline curve need *not* directly pass through the control points. If we wanted a curve that passed through the point set ϕ_{CP}, as the linear interpolation curve does, a b-spline could still achieve this. It would simply require a different set of control points ϕ'_{CP}. Not passing directly through ϕ_{CP} can be a significant advantage of the b-spline curve. If, for instance, we have reason to believe that the control points may be noisy, the curve that follows the overall trend of ϕ_{CP} without adhering too closely to any single point can produce a more robust overall curve. A synthetic example of this phenomenon is shown in Figure 2.4.

Second, the b-spline curve is differentiable, unlike linear interpolation, as shown in figures 2.3 and 2.4. Smoothness, reflecting the clinical understanding of the smooth change of anatomical structures over space and time, is a central theme throughout this work. As described in Section 2.3, every point in a cubic b-spline curve with uniformly distributed knots and control points has at least C^2 continuity [62].

B-spline interpolation offers a sparse representation of a smooth curve in n-D, and can be optimized more easily than other differentiable n-D curves given its local support property [70]. It is more robust to independent and identically distributed (i.i.d.) noise than simpler interpolation schemes, such as linear interpolation, but is still faithful to the overall trend of ϕ_{CP} that parameterizes the curve. For these reasons, b-splines lend themselves well to modeling anatomical changes over time and space.

Chapter 3

Model

IN this chapter we introduce our main contribution, Longitudinal VoxelMorph, an efficient and scalable spatiotemporal medical image registration model. We define the model and an accompanying learning-based approximation network architecture. Finally, we provide an analysis of tradeoffs considered in the model design.

We aim to register a time series of medical images to each other. At a high level, one goal of an image registration method is to transform all input images to a shared coordinate system. This shared coordinate system is usually the coordinate system of one of the images, often referred to as the *fixed* or *target* image. The transformation defines a deformation that maps pixels in the other images, sometimes called the *moving* or *source* images, back to the fixed image. Applying the deformation to a moving image therefore produces an estimate of the fixed image [11, 14, 19, 38, 40, 72, 74, 85, 90, 91]. For spatiotemporal registration, the deformation field can be interpreted anatomically as quantifying the structural changes captured in the medical images at different points in time [4, 28, 39, 49, 92].

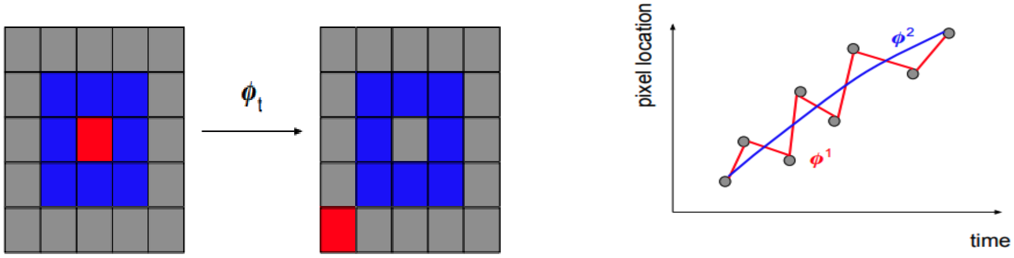
Longitudinal VoxelMorph registers a time series X of medical images from a single patient. To define a shared coordinate system, we choose the patient's first image, x_0, as the fixed image. Based off of the classical image registration cost function given in Equation (2.1), we estimate a deformation field ϕ that minimizes

$$\hat{\phi} = \min_{\phi} \sum_{i=1}^{|X|-1} L_{\mathsf{sim}}(\phi, x_0, x_i) + \lambda L_{\mathsf{smooth}}(\phi), \tag{3.1}$$

where L_{sim} is an image similarity loss and L_{smooth} enforces some smoothness constraint on the deformation ϕ. We build a learning-based approximation network to estimate the parameters of the model, as outlined in Section 3.2.

We use a sparse representation of ϕ, parameterized by b-splines, to represent and estimate the deformation field in an efficient way. Since a single medical image, and particularly a 3-D scan, is often very large, supporting time series with more than two images can exceed memory constraints, even with access to modern GPUs [76]. (Using CPUs is infeasible, since they run on the order of 55 times slower for pairwise 3-D image registration tasks [59].)

We also want the deformation field ϕ to be smooth. Anatomical activity is medically



(a) An example deformation field ϕ_t that is *not* anatomically consistent. ϕ_t warps the pixels in the first grid to produce the second, where the pixel coloration identifies the deformation of individual pixels. In this case, ϕ_t is not anatomically consistent, since the red pixel should not be able to exit all of its encircling neighbors (shown in blue) given our understanding of anatomical structures' activity.

(b) Two example deformation fields, ϕ^1 (shown in red) and ϕ^2 (shown in blue), with an example pixel's deformed location across the images shown as gray points. Even though ϕ^1 passes through the points exactly, it is possible that the deformed locations are noisy because of noise in the input images, and ϕ^2 is more temporally consistent, since we expect anatomical activity to be relatively smooth over time. Medically, ϕ^2 is more useful.

Figure 3.1: Basic examples to illustrate the concepts of (a) anatomical consistency and (b) temporal consistency.

understood to change smoothly over space and time. The hippocampus does not tear as it atrophies, for example [4]. Nor does the lining of the ventricles expand and contract in a jagged motion during the cardiac cycle. Instead, the ventricles move smoothly [81]. Mathematically, ϕ is diffeomorphic, meaning that ϕ is invertible and both ϕ and ϕ^{-1} are differentiable functions.

Spatially, this guarantees *anatomical consistency*, so underlying structures in the images move smoothly and obey normal physical laws, such as preventing tissues from being cut out from the center of an image or moved into an entirely different part of the image [55]. An example of an anatomically inconsistent ϕ is shown in Figure 3.1a. The temporal analog of anatomical consistency is *temporal consistency*, which requires that changes are smooth over time. A temporally inconsistent model, for example, might optimize a deformation field for each individual image, potentially resulting in a non-differentiable ϕ, which would not reflect clinical understanding of anatomical changes over time, as shown in Figure 3.1b [47].

■ 3.1 Model Definition

In this section, we define the variables and important hyperparameters used in Longitudinal VoxelMorph.

First, the model defines the hyperparameter F as the number of images in each patient's time series and the hyperparameter K as the maximum time gap between the first and last image from any of the patients. That is, images within a time series must all occur within K time of each other.

The model also defines the following variables.

- $X = [x_0, x_1, \ldots, x_{F-1}]$. These are the set of images that make up the time series for a patient. Each image can be n-dimensional, so $x_i \in \mathbb{R}^n$. The sequence X is sorted by the time at which the image was taken, so that x_0 is the first image and x_{F-1} is the last available image for that patient. As required by the hyperparameter, we have $|X| = F$ for all patients.
- $T = [t_0, t_1, \ldots, t_{F-1}]$. These are the set of times at which each image was taken. We define $t_0 = 0$, so that all times are relative to the baseline image. That is, image x_i was taken t_i time after x_0 was taken. The hyperparameter K requires that $t_{F-1} \leq K$.

We model ϕ_v as the b-spline control-point representation for the vector velocity field v of the deformation ϕ, using uniformly spaced control points and knot vector placement. That is, ϕ_v defines the spatial velocity at each control point location. Given 2-D images, where initially $x_i \in \mathbb{R}^{W \times H}$, the control-point representation has shape

$$\phi_v \in \mathbb{R}^{\frac{W}{\delta_x} \times \frac{H}{\delta_y} \times 2 \times \frac{F}{\delta_T}}, \tag{3.2}$$

where δ_x, δ_y, and δ_T are spatial hyperparameters defined by the model. These determine the level of sparsity in our representation of the deformation field ϕ. We define δ_d as the spatial-sparsity ratio along the dth dimension of the input image. In the case of a 2-D image, for example, we would define δ_x and δ_y. Similarly, we define δ_T as the temporal-sparsity ratio. This determines the sparsity of the model's representation over time.

We let ϕ_d be the control point parameterization for the deformation field. Therefore,

$$\phi_d = \int \phi_v dt. \tag{3.3}$$

The integration does not change the dimensions of the control point grid from those of ϕ_v.

Finally, we model the dense deformation field ϕ using b-spline interpolation of the control point grid ϕ_d at every pixel location. Given the uniformly spaced control points and knot vector placement, this interpolation is defined in Equation (2.6). For a 2-D image, this therefore produces

$$\phi \in \mathbb{R}^{W \times H \times 2 \times F-1}. \tag{3.4}$$

We define ϕ_t as the slice of the vector field ϕ when fixing time t, so that

$$\phi_t \in \mathbb{R}^{W \times H \times 2}. \tag{3.5}$$

Based on the model, we aim to learn an approximation of the deformation ϕ that minimizes

$$\hat{\phi} = \min_{\phi} \sum_{i=1}^{F-1} \|x_0 - \phi_i \circ x_i\| + \lambda \|\phi\|. \tag{3.6}$$

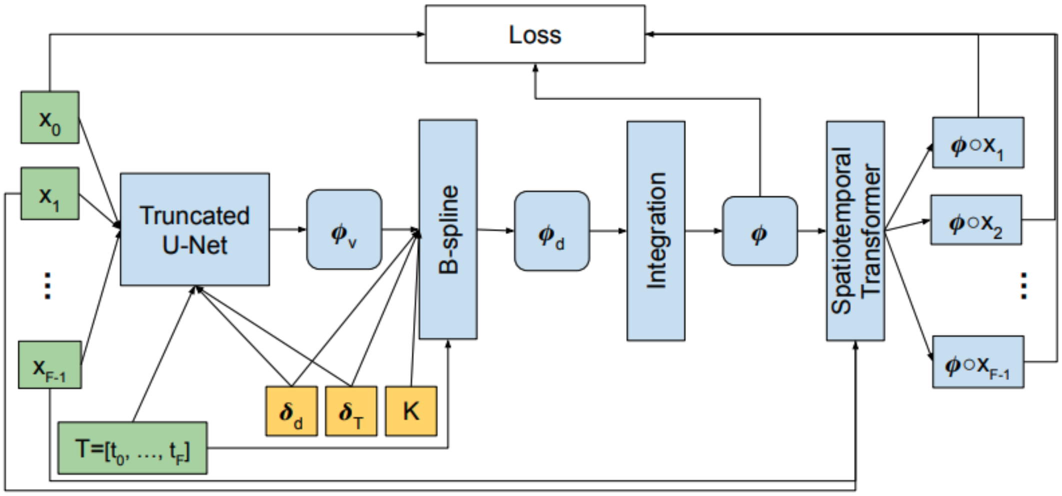


Figure 3.2: An overview of the proposed network approximating the spatiotemporal deformation ϕ. The network inputs X and T are shown in green, while the hyperparameters to the model, δ_d, δ_T, and K, are shown in yellow. All model layers and computed values are shown in blue.

■ 3.2 Learning-based Approximation

Rather than approximating ϕ independently for each new time series of images, we use a network $g_{F,K,\delta_d,\delta_T}(X,T) = \phi$. The network optimizes

$$g_{F,K,\delta_d,\delta_T}(X,T) = \min_{g} \sum_{i=1}^{F-1} \|x_0 - g(X,T)_i \circ x_i\| + \lambda \|g(X,T)\| \,. \tag{3.7}$$

After training, the network g can be used to estimate ϕ for new example time series more efficiently than a classical approach that optimizes all examples independently.

We define our network g as follows. Figure 3.2 visually depicts the network architecture.

- We estimate ϕ_v using a truncated U-Net [69]. The U-Net architecture is a commonly used convolutional neural network (CNN) specifically designed for medical images [14, 47, 69]. We define $h(X,T) = \mathsf{UNet}[l]$ to be the output of the lth layer in the upsampling path of the U-Net. The value of l is based off of the hyperparameters δ_d, such that the spatial dimensions of $h(X,T)$ are as close as possible to the desired spatial dimensions of ϕ_v. Any additional required reshaping is done immediately following $h(X,T)$ to produce the correct dimensions of ϕ_v, and we refer to this final reshaping as f. Therefore, $\phi_v = f(h(X,T))$.

- We estimate ϕ_d using a scaling and squaring layer [23]. In a fully-resolved velocity vector field, we can imagine taking small steps along a vector field and summing up the displacement at each location from each step to produce the overall displacement field.

  We define Δ as the number of steps we want to take along the velocity field v. Define the displacement at each pixel after i steps as the vector field $d^{(i)}$. We compute $d^{(1)} = \frac{v}{\Delta}$. Then, to improve the model efficiency, we approximate $d^{(2)}$ as $d^{(1)}+d^{(1)}$, $d^{(4)} \approx d^{(2)}+d^{(2)}$, and so on until $d^{(\Delta)} \approx d^{(\frac{\Delta}{2})}+d^{(\frac{\Delta}{2})}$. In each summation, we use linear interpolation to calculate the displacement at each pixel location in the vector field. The higher the Δ, the higher the accuracy of the approximated integration, but the slower the computation time. For this work, we chose $\Delta = 2^7$.

  Although the intuition is simpler with a fully-resolved vector field, the same can be done with b-spline control points. Specifically, this approach requires that for some control-point parameterization ϕ_{CP},

$$\mathsf{bspline}(\tfrac{\phi_{CP}}{2}) + \mathsf{bspline}(\tfrac{\phi_{CP}}{2}) = \mathsf{bspline}(\phi_{CP}),$$

  where $\mathsf{bspline}(\cdot)$ denotes b-spline interpolation using the given set of control points. This would guarantee that summing the fully-resolved displacement fields parameterized by control points for two smaller steps is the same as resolving the displacement field parameterized by the summation of the control points for each step.

  To see that this is true, consider the 1-D b-spline curve analogous to the 3-D parameterization given in Equation (2.6). Specifically,

$$r(p) = \textstyle\sum_{l=0}^{3} B_l(\frac{p}{\delta_x} - \lfloor \frac{p}{\delta_x} \rfloor)\phi_{CP_{\lfloor \frac{p}{\delta_x} \rfloor + l}},$$

  where B_l is defined as in Equation (2.7). The value of p defines the location to interpolate at, and ϕ_{CP} is the sparse control-point representation. Computing $\frac{\phi_{CP}}{2}$ does not impact p, so let $c_l := B_l(\frac{p}{\delta_x} - \lfloor \frac{p}{\delta_x} \rfloor)$, where c_l is a constant in terms of ϕ_{CP}. Similarly, let $i_l = \lfloor \frac{p}{\delta_x} \rfloor + l$, which is again a constant in terms of p.

  Then, for some location p,

$$\mathsf{bspline}(\tfrac{\phi_{CP}}{2}) + \mathsf{bspline}(\tfrac{\phi_{CP}}{2}) = \textstyle\sum_{l=0}^{3} c_l(\frac{\phi_{CP}}{2})_{i_l} + c_l(\frac{\phi_{CP}}{2})_{i_l} = \sum_{l=0}^{3} c_l \times 2(\frac{\phi_{CP}}{2})_{i_l} = \sum_{l=0}^{3} c_l(\phi_{CP})_{i_l} = \mathsf{bspline}(\phi_{CP}).$$

  This analysis extends to higher-dimensional b-spline parameterizations. Therefore, fully-resolving summed control-point representations of step-sized displacements is equivalent to summing fully-resolved step-sized displacements. It is more efficient to sum the sparse control-point representation, so this is what is implemented in the network g.

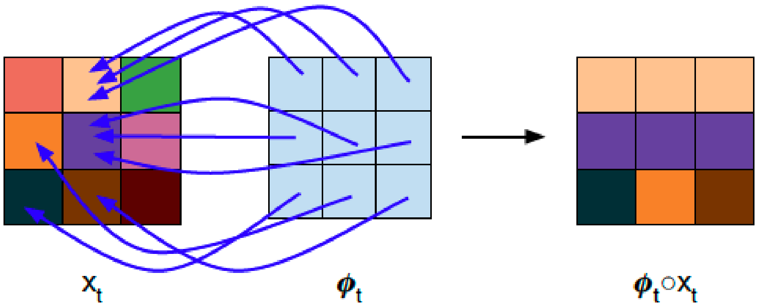


Figure 3.3: A 2-D example deformation, as applied by the spatiotemporal transformer layer. Each index (i, j) in ϕ_t contains an 2-D vector that specifies the pixel in x_t that should be used for the (i, j)th pixel in the warped image. The field ϕ_t is a surjective mapping that covers the estimated warped image $\hat{x}_t$, producing an estimated image that is well-defined and simple to compute. It is *not* necessary for the vectors in ϕ to map to index pixel locations. Values corresponding to non-integer locations within x_t are calculated with linear interpolation using the location's neighboring pixels.

We therefore use this layer to approximate the control-point representation for the displacement vector field according to the recursive relation

$$\phi_d^{(1)} = \frac{\phi_v}{2^7}, \tag{3.8}$$

and

$$\phi_d^{(s)} = \phi_d^{(\log_2 s)} + \phi_d^{(\log_2 s)}, \tag{3.9}$$

with the final estimate $\phi_d = \phi_d^{(2^7)}$.

- We estimate the full-resolution deformation field ϕ using a b-spline interpolation layer. We scale the image times relative to the maximum time frame supported by the model, so that the final control points in time refer to the longest supported time gaps. For each pixel location in an image and each time in the input vector T (scaled relative to K), we interpolate the displacement vector at that pixel for that time. We compute $\phi = \mathsf{bspline}(\phi_d, \frac{T}{K})$.

- Finally, we use a spatiotemporal transformer layer to warp x_t back to the reference frame of x_0 for all $t > 0$. If we define the warped version of x_t as $\hat{x}_t$, this layer implements the surjective function $\hat{x}_t = \phi_t \circ x_t$ for $0 < t < F$, as shown in Figure 3.3. For each pixel in $\hat{x}_t$, the deformation defines the location in x_t that corresponds to that pixel. The field therefore covers the set of pixels in $\hat{x}_t$, producing an estimated deformation that is well-defined and simple to compute, as shown in Figure 3.3.

■ 3.3 Fixed Image Choice

To define a shared coordinate system for a patient's image data, we selected x_0 as the fixed image. We chose to use an image from the input time series as the fixed image rather than constructing a separate prototypical image to use as an atlas. We made this design decision mindful of applications like segmentation propagation, as discussed in Chapter 1, where we only have a segmentation available for the baseline image (*i.e.*, x_0 has an accompanying segmentation, and the task is to estimate a segmentation for all other images in the time series). For applications like this, constructing a separate prototypical image, mapping the segmentation to that space, and then mapping the prototypical segmentation to all of the other images would introduce more errors to the estimated segmentations with each mapping. These errors can be avoided by using the existing baseline image as the fixed image.

■ 3.4 Analysis of Error Propagation

Different spatiotemporal registration models could choose to learn a slightly different interpretation of the deformation field ϕ, and each interpretation choice would lead to different behavior when exposed to noisy input data. The input data might be noisy because of changes to specific scanner settings between images, patient movement, or imperfect imaging technology [68]. Ideally, noise in one part of the input data or other errors in the model's estimated deformation field should not be amplified or create errors in other parts of the model.

In this section, we perform a theoretical analysis of the Longitudinal VoxelMorph deformation's response when exposed to noisy data. We also compare it to two other deformation field interpretations, and show that the ϕ computed by Longitudinal VoxelMorph is the most robust to noise.

To facilitate different interpretations of ϕ, in this section we will refer to the warped version of x_t in the fixed reference frame as $\hat{x}_t$. Furthermore, we will analyze the behavior of each interpretation using the inverse of ϕ (ϕ^{-1}, which is guaranteed to exist since ϕ is diffeomorphic) to adhere to a more intuitive understanding of x_i mapping to x_j, where $i < j$.

The three analyzed interpretations of ϕ are as follows.

1. In Longitudinal VoxelMorph, the deformation field ϕ_t^{-1} represents the mapping from the baseline image in the sequence to the estimate of the next image. Mathematically,
$$\hat{x}_{t+1} = \phi_t^{-1}(x_0), 0 \leq t \leq T-2. \tag{3.10}$$
This formulation will be called the *baseline* deformation interpretation throughout this section.

2. In our first alternative, the deformation field ϕ_t^{-1} represents the deformation from the previous image in the *input* sequence to the estimate of the next image. That

is, we could interpret ϕ^{-1} according to

$$\hat{x}_{t+1} = \phi_t^{-1}(x_t), 0 \leq t \leq T-2. \tag{3.11}$$

For the remainder of this section, we will refer to this interpretation as the *input timestep* deformation interpretation. To distinguish this timestep interpretation of ϕ^{-1}, we will use the notation $\phi^{-1}_{t \to t+1} \coloneqq \phi_t^{-1}$.

3. In our second alternative, the deformation field ϕ_t^{-1} represents the deformation that warps the previous image in the *estimated* sequence, i.e.

$$\hat{x}_{t+1} = \phi_t^{-1} \circ \phi_{t-1}^{-1} \circ \phi_{t-2}^{-1} \circ ... \circ \phi_1^{-1} \circ \phi_0^{-1}(x_0), 0 \leq t \leq T-2. \tag{3.12}$$

This will be referred to as the *estimated timestep* deformation interpretation for the remainder of this section. Again, in later parts of this section we will use the notation $\phi^{-1}_{t \to t+1} \coloneqq \phi_t^{-1}$ to convey that this is a timestep-based interpretation.

3.4.1 Introducing Noise to the Data

Consider introducing noise to the input data sequence. For a theoretical analysis, let $\bar{x}_t$ be a 'perfect' image, i.e., let it noiselessly capture the patient's anatomy at time t. Let $x_j = \bar{x}_j + \gamma_\sigma$ define a noisy image, for $j > 0$ and $\gamma_\sigma = \mathcal{N}(0, \sigma)$. To simplify the analysis, we constrain the baseline image x_0 to always be perfect.

Define $\bar{\phi}_t^{-1}$ to be the 'perfect' deformation field for time point t, i.e. $\bar{x}_{t+1} = \bar{\phi}^{-1}_{t \to t+1}(\bar{x}_t), 0 \leq t \leq T-2$ for in the timestep interpretations listed above, or $\bar{x}_t = \bar{\phi}_t^{-1}(x_0), 0 \leq t \leq T-1$ in the baseline interpretation. Then, let $\phi_i^{-1} = \bar{\phi}_i^{-1} + \epsilon_\sigma, 0 < i \leq T-1$, where $\epsilon_\sigma = \mathcal{N}(0, \sigma)$, and let $\phi_k^{-1} = \bar{\phi}_k^{-1}, \forall k \neq i, 0 \leq k \leq T-1$. That is, all deformations except for the ith are 'perfect', and we have introduced Gaussian noise to the ith deformation. This mimics the case where there is an error in the deformation field that is not necessarily caused by noise in an input image. It might be impossible, for instance, to perfectly capture the deformation with a sufficiently sparse representation of ϕ_v.

This is an improbable set of assumptions for a real-world application. First, there are no noiseless medical images, because of limitations in medical imaging. Second, if $\phi_i^{-1} \neq \bar{\phi}_i^{-1}$, then likely $\phi_{i+1}^{-1} \neq \bar{\phi}_{i+1}^{-1}$. In Longitudinal VoxelMorph's model formulation, for instance, in order to introduce noise into ϕ_i^{-1}, there must be noise introduced into ϕ_v, since the mapping $\phi_v \xrightarrow{\int \phi_v dt, \mathsf{bspline}(\phi_d)} \phi$ is deterministic. If control point $\phi_v[x, y, t]$ has been shifted by the introduction of noise, the cubic b-spline interpolation will slightly adjust the result of the interpolated locations nearest that control point in each dimension. Therefore, if noise is introduced to ϕ_i^{-1} then a control point that caused at least part of the noise in ϕ_i^{-1} also influences ϕ_{i+1}^{-1}, and therefore likely introduces noise to ϕ_{i+1}^{-1} as well.

However, these assumptions simplify the analysis and still convey the advantages and disadvantages of different interpretations of ϕ^{-1}. In some cases, the subsequent sections comment on steps where a different, more realistic set of assumptions might change the relative value of different interpretations, but on the whole this basic framework facilitates analysis.

We consider each of the possible model interpretations for the deformation field individually and analyze how the noise changes the network behavior.

■ 3.4.2 Baseline Deformation Interpretation with Noise

In this interpretation, the model uses the deformation field to estimate an image according to $\hat{x}_{t+1} = \phi_t^{-1}(x_0)$, as in Equation (3.10). This presents the following cases:

1. $t \neq i : \hat{x}_{t+1} = \phi_t^{-1}(x_0) = \bar{\phi}_t^{-1}(\bar{x}_0) = \bar{x}_{t+1}$. In this case, the estimated image for time $t+1$ is perfect. Under the assumption that x_0 is perfect, this does not depend on j.

2. $t = i : \hat{x}_{t+1} = \phi_t^{-1}(x_0) = \bar{\phi}_t^{-1} \circ \epsilon_\sigma(\bar{x}_0) = \epsilon_\sigma(\bar{x}_{t+1})$. Here, an additional deformation of ϵ_σ is applied to the true value. This analysis is still independent of j.

These results are compared with those of the two alternative deformation field interpretations in Section 3.4.5.

■ 3.4.3 Input Timestep Deformation Interpretation with Noise

In this interpretation, the model uses the deformation field to estimate an image according to $\hat{x}_{t+1} = \phi^{-1}_{t\to t+1}(x_t)$, as in Equation (3.11). We consider the following cases:

1. $i \neq t, j \neq t : \hat{x}_{t+1} = \phi^{-1}_{t\to t+1}(x_t) = \bar{\phi}^{-1}_{t\to t+1}(\bar{x}_t) = \bar{x}_{t+1}$. In this case, the estimated image at time $t+1$ is perfect.

2. $i = t, j \neq t : \hat{x}_{t+1} = \phi^{-1}_{t\to t+1}(x_t) = \bar{\phi}^{-1}_{t\to t+1} \circ \epsilon_\sigma(\bar{x}_t) = \epsilon_\sigma \circ \bar{\phi}^{-1}_{t\to t+1}(\bar{x}_t) = \epsilon_\sigma(\bar{x}_{t+1})$. ϵ_σ and $\bar{\phi}_t^{-1}$ commute, since each deformation can be thought of as a vector field, and the composition of two displacement vector fields is their sum. Therefore, this estimation is noisy, applying an unwanted additional deformation of ϵ_σ.

3. $i \neq t, j = t : \hat{x}_{t+1} = \phi^{-1}_{t\to t+1}(x_t) = \bar{\phi}^{-1}_{t\to t+1}(\bar{x}_t + \gamma_\sigma) = \bar{\phi}^{-1}_{t\to t+1}(\bar{x}_t) + \bar{\phi}^{-1}_{t\to t+1}(\gamma_\sigma) = \bar{x}_{t+1} + \bar{\phi}^{-1}_{t\to t+1}(\gamma_\sigma)$, since the deformation field ϕ^{-1} acts as a location index into x_t, and is therefore distributive. This estimation is again noisy, adding the warped value $\bar{\phi}^{-1}_{t\to t+1}(\gamma_\sigma)$ to the true value.

4. $i = t, j = t : \hat{x}_{t+1} = \phi^{-1}_{t\to t+1}(x_t) = \epsilon_\sigma \circ \bar{\phi}^{-1}_{t\to t+1}(\bar{x}_t + \gamma_\epsilon) = \epsilon_\sigma(\bar{x}_{t+1} + \bar{\phi}^{-1}_{t\to t+1}(\gamma_\sigma)) = \epsilon_\sigma(\bar{x}_{t+1}) + \epsilon_\sigma \circ \bar{\phi}^{-1}_{t\to t+1}(\gamma_\sigma)$. This estimation is noisy as well, applying an additional unwanted warp to the true $\bar{x}_{t+1}$ and then adding $\epsilon_\sigma \circ \bar{\phi}(\gamma_\epsilon)$ to that term as well.

Section 3.4.5 compares these findings with the other potential deformation field interpretations listed in Section 3.4.

3.4.4 Estimated Timestep Deformation Interpretation with Noise

In this interpretation, the model uses the deformation field to estimate an image according to $\hat{x}_{t+1} = \phi^{-1}_{t\to t+1} \circ \phi^{-1}_{t-1\to t} \circ ... \circ \phi^{-1}_{1\to 2} \circ \phi^{-1}_{0\to 1}(x_0)$, as in Equation (3.12). This leads us to consider the following cases:

1. $t < i : \hat{x}_{t+1} = \phi^{-1}_{t\to t+1} \circ ... \circ \phi^{-1}_{0\to 1}(x_0) = \bar{\phi}^{-1}_{t\to t+1} \circ ... \circ \bar{\phi}^{-1}_{0\to 1}(\bar{x}_0) = \bar{\phi}^{-1}_{t\to t+1}(\bar{x}_t) = \bar{x}_{t+1}$. This is a perfect estimation of x_{t+1}. This does not depend on j, unlike the analysis for the input timestep deformation interpretation given in Section 3.4.3. Under the assumption that $x_0 = \bar{x}_0$, any noise in the sequence's subsequent images does not directly impact the model's estimations (*i.e.*, it only impacts the model through its effects on ϕ^{-1}).

2. $t = i : \hat{x}_{t+1} = \phi^{-1}_{t\to t+1} \circ ... \circ \phi^{-1}_{0\to 1}(x_0) = \bar{\phi}^{-1}_{t\to t+1} \circ \epsilon_\sigma \circ \bar{\phi}^{-1}_{t-1\to t} \circ ... \circ \bar{\phi}^{-1}_{0\to 1}(\bar{x}_0) = \epsilon_\sigma \circ \bar{\phi}^{-1}_{t\to t+1}(\bar{x}_t) = \epsilon_\sigma(\bar{x}_{t+1})$. This estimated image is noisy, applying an additional, unwanted deformation of ϵ_σ.

3. $t > i : \hat{x}_{t+1} = \phi^{-1}_{t\to t+1} \circ ... \circ \phi^{-1}_{i\to i+1} \circ ... \circ \phi^{-1}_{0\to 1}(x_0) = \bar{\phi}^{-1}_{t\to t+1} \circ ... \circ \bar{\phi}^{-1}_{i\to i+1} \circ \epsilon_\sigma \circ ... \circ \bar{\phi}^{-1}_{0\to 1}(\bar{x}_0) = \epsilon_\sigma \circ \bar{\phi}^{-1}_{t\to t+1} \circ ... \circ \bar{\phi}^{-1}_{0\to 1}(\bar{x}_0) = \epsilon_\sigma \circ \bar{\phi}^{-1}_{t\to t+1}(\bar{x}_t) = \epsilon_\sigma(\bar{x}_{t+1})$. We observe that the deformation ϵ_σ is propagated to $\hat{x}_t$ even when $t > i$. Again, this analysis is independent of j.

The following section will compare these results to the other deformation field interpretations considered in Section 3.4.

3.4.5 Interpretation with Noise Analysis Comparison

The analysis presented earlier in this section demonstrates some of the theoretical tradeoffs between the possible interpretations of the deformation field ϕ^{-1}.

In the analysis of the input timestep deformation interpretation defined in Equation (3.11), the error in $\hat{x}_{t+1}$ depends on both i and j, meaning it is impacted by noise in the input data and noise in the deformation field, even when these two sources of noise are not assumed to be correlated. This is the only interpretation considered in this section that is directly impacted by noise in the data itself. In some cases, it is possible that the noise in the image and the noise in the deformation field help mitigate each other for future image estimations ($i.e.\ \epsilon_\sigma(\bar{x}_{t+1}) + \epsilon_\sigma \circ \Phi_{t\to t+1}(\gamma_\sigma) < \epsilon_\sigma(\bar{x}_{t+1}), t = i = j$), but in general we would prefer for the deformation field estimation to be robust to noise in the data. It is also worth noting that this interpretation would prevent the model from serving the needs of some important temporal registration applications. In the case of segmentation propagation problems, for example, the model would not have access to an input segmentation at every time point, creating the need for automated segmentation propagation in the first place.

When analysing the estimated timestep deformation interpretation given in Equation (3.12), the noise in the deformation field for a specific time point is propagated through all subsequent estimated images. In the more realistic case of noise at multiple values of t in ϕ_t^{-1}, it is again possible that the noise introduced by each deformation could counteract the others, but noise propagation is not desirable in general.

Finally, when considering the baseline deformation interpretation defined in Equation (3.10) and used in Longitudinal VoxelMorph, the analysis showed that noise only appeared in the estimation where the noisy deformation field ϕ_i^{-1} was applied. No noise was introduced directly from noisy images, assuming x_0 is perfect, and no noise was propagated to estimations of later time points. Although this prevents errors in multiple ϕ_t^{-1} from counteracting each other, it also prevents errors from compounding each other.

Overall, this theoretical analysis leads us to believe that the baseline deformation interpretation for ϕ^{-1} produces the most noise-tolerant results and, by extension, deformation field ϕ. It is therefore the deformation interpretation used in Longitudinal VoxelMorph.

Chapter 4

Longitudinal Evaluation Metrics

IN this chapter, we present several metrics that we use to evaluate spatiotemporal registration. Some metrics, such as Dice score [30] and surface distance, are commonly used to evaluate pairwise registration performance, while new metrics proposed in this work take into account the temporal aspect of the data.

■ 4.1 Dice Score

Many registration evaluation metrics are based on *segmentations*. A segmentation labels an image region as belonging to a specific structure. In the case of cardiac images, for example, a segmentation might show the area of an image that belongs to the left ventricle. For a brain MRI, a segmentation could label the hippocampus.

Given a set of segmentations $S = \{S_0, ..., S_{F-1}\}$ that correspond to input images $X = \{X_0, ..., X_{F-1}\}$, a registration model can use the estimated deformation field ϕ to warp S_i to the reference frame of the fixed segmentation S_0 according to $\phi \circ S_i$ for all $i > 0$. In the case of a perfect registration model and perfectly segmented data, $\phi \circ S_i = S_0$ for all i.

Given two segmentations, S_A and S_B, the Dice score [30] is given by

$$\mathsf{Dice}(S_A, S_B) = \frac{2 \times |S_A \cap S_B|}{|S_A| + |S_B|}. \tag{4.1}$$

Therefore, the Dice score is equal to 1 if the two segmentations align perfectly, and 0 if they are disjoint.

We can extend Dice scores to the case where $|S| > 2$. For each $i > 1$, $\mathsf{Dice}(S_0, \phi_i \circ S_i)$ can help quantify the deformation field ϕ's anatomical accuracy. Specifically, it can measure how faithfully ϕ deforms anatomical structures over time, which is an important metric for applications such as segmentation propagation.

It does not, however, measure other desirable quantities such as temporal consistency, which a registration model may or may not directly enforce. In this case, the pairwise nature of the Dice score rewards a registration model that optimizes ϕ_i for each time point individually, without considering the overall trajectory. This sacrifices temporal consistency for anatomical accuracy [47].

To better visualize temporal consistency, consider the example application of segmentation propagation. A video of the resulting segmentations would ideally be smooth, reflecting temporally consistent anatomical movement, rather than a jagged, frame-by-frame representation of local optima. This tradeoff between temporal consistency and intensity matching motivates the longitudinal Dice score proposed in Section 4.3.

■ 4.2 Surface Distance

Surface distance between two segmentations in the same reference frame is another common way to quantify the performance of a registration model. In this case, the segmentations contain the border of an anatomical structure, so $B_A = \{p_0^A, ..., p_{n-1}^A\}$, where p_j^A is the jth point that lies on the boundary of segmentation A. If $B_B = \{p_0^B, \ldots, p_{m-1}^B\}$, then the mean surface distance for S_A and S_B is computed according to

$$\mathsf{MSD}(S_A, S_B) = \frac{1}{n+m}(\sum_{j=0}^{n-1} \min_{0 \le k < m} \|p_j^A - p_k^B\| + \sum_{k=0}^{m-1} \min_{0 \le j < n} \|p_k^B - p_j^A\|). \tag{4.2}$$

This yields $\mathsf{MSD}(S_A, S_B) = 0$ if S_A and S_B line up perfectly, and the surface distance continues to increase as S_A and S_B diverge.

MSD measures a slightly different relationship between segmentations than Dice score, so it is useful to evaluate registration models using both metrics. For example, consider a segmented anatomical structure that has a large volume to surface area ratio. In this case, the Dice score remains high even as edges of the segmentation diverge, since the overlap in the center of the structure is still large, while surface distance can more clearly capture differences between the segmentations.

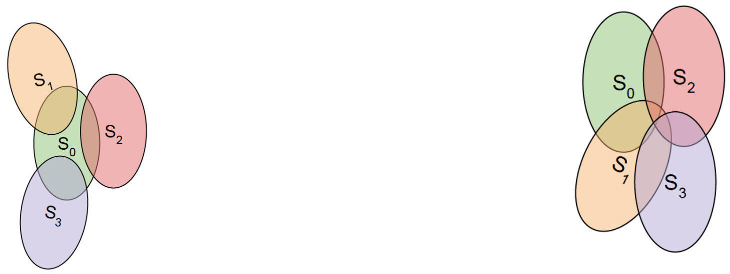
In other respects, surface distance is still very similar to the Dice score metric. It is also inherently pairwise, and will reward models that optimize local accuracy at each time point at the cost of temporal consistency.

■ 4.3 D* Score

We want the spatiotemporal registration model to estimate a temporally consistent deformation field. As defined in Chapter 3, a temporally consistent deformation encodes smooth changes over time. In general, we aim to reward a smooth, globally optimized temporal trajectory, rather than a jagged, locally optimized trajectory.

In order to quantify temporal consistency as well as anatomical accuracy in a model's deformation field, we propose a new, longitudinal Dice coefficient, referred to as D*. D* is intended to highlight the differences between a temporally consistent deformation field and one that locally optimizes the result for each individual time point at the expense of smoothness, as shown in Figure 4.1.

Let S_0 be the segmentation for the fixed image. Then, assuming all segmentations are perfect, there is some optimal deformation field $\bar{\phi}$ for that subject. The field $\bar{\phi}$



(a) Example warped segmentations of a model locally optimized for the segmentation overlap between S_i and S_0.

(b) Example warped segmentations of a model optimized for the segmentation overlap between S_i and S_0, but with global information that can used to impose temporal consistency.

Figure 4.1: Two examples of three warped segmentations, S_1, S_2, and S_3, compared to the segmentation S_0 of the fixed image. These examples highlight the differences between a model that optimizes ϕ_t for each time point individually, and a model that optimizes ϕ_t with temporal smoothness constraints placed on ϕ. In this example, the warped segmentations depicted in Fig. 4.1a achieve a better Dice score when only compared to S_0, and are therefore more optimal locally than the segmentations in Fig. 4.1b, but the warped segmentations in Fig. 4.1b are more consistently located, and are therefore more globally optimal than the segmentations in Fig. 4.1a. Interpreted temporally, even if the segmentations in Fig. 4.1a better align to S_0, they lead to a more jagged time series of segmentations than those in Fig. 4.1b.

should be temporally smooth, since a deformation can be interpreted as the change of anatomical structures over time, which is scientifically understood to be temporally consistent [47].

Slight errors ϵ in the deformation field field, introduced by limitations of the sparse representation or estimation inaccuracy, produce $\phi = \bar{\phi} + \epsilon$. A model that optimizes locally for each such i would independently select the ϵ that minimized the error at each time point; a model that optimizes globally to include longitudinal consistency would select a smooth ϵ.

If the error term ϵ is smooth, we would expect $\|\epsilon_i - \epsilon_{i+1}\| < \|\epsilon_i - \epsilon_j\|$ for $j \gg i$. This may not be true for every value of j, particularly in the case of cyclic motion, but we would expect it to hold in general. Intuitively, this means the error of time-adjacent deformations should be similar if the computed ϕ is temporally consistent.

For any $0 < i, j < F$, we know that

$$\begin{gathered}\|\phi_i(S_i) - \phi_j(S_j)\| = \left\|\bar{\phi}_i(S_i) + \epsilon_i(S_i) - (\bar{\phi}_j(S_j) + \epsilon_j(S_j))\right\| = \\ \|S_0 + \epsilon_i(S_i) - S_0 - \epsilon_j(S_j)\| = \|\epsilon_i(S_i) - \epsilon_j(S_j)\|.\end{gathered}$$

Medically, we would expect the change in anatomy between two images taken closer in time to be smaller than the changes between images taken further apart in time. This may not be true for every possible pair of images, but we aim to optimize this in general. Mathematically, we can write this as $\|S_{i+1} - S_i\| < \|S_j - S_i\|$ for most values of $j \gg i$. Furthermore, if ϵ is not too large (meaning the deformations are relatively accurate) and is smooth over time, then we have a small $\|\epsilon_{i+1} - \epsilon_i\|$, relative to $\|\epsilon_j - \epsilon_i\|$, for most values of $j \gg i$.

If $\|S_{i+1} - S_i\|$ is small and $\|\epsilon_{i+1} - \epsilon_i\|$ is small, we would expect $\|\epsilon_{i+1}(S_{i+1}) - \epsilon_i(S_i)\|$ to be small as well. That is, we expect warped segmentations from time-adjacent images to be relatively close to each other, compared to segmentations from distant time images, if ϕ is temporally consistent. To capture this intuition mathematically, let $D_{ij} = \mathsf{Dice}(\phi_i \circ S_i, \phi_j \circ S_j)$, from Equation 4.1. (Define $D_{0j} = \mathsf{Dice}(S_0, \phi_j \circ S_j)$.) Then, we let

$$D_i^*(\alpha) = \frac{\sum_{j=i+1}^{F-1} (1-\alpha)^{j-i-1} D_{ij}}{\sum_{j=0}^{F-j-2} (1-\alpha)^j},\ 0 \le i < F-1, \tag{4.3}$$

where $\alpha \in [0, 1]$ is a hyperparameter. If the image data is inherently cyclic, such as cardiac imaging over the course of a heartbeat, D_{F-1}^* can also be computed as $D_{F-1}^*(\alpha) = D_{0,F-1}$, thereby considering the image at time 0 and the image at time $F-1$ to be neighbors. For the remainder of this section, we will consider the case of non-cyclic motion, but this can easily be extended.

Next, we compute D* according to

$$D^*(\alpha) = \frac{1}{F-1} \sum_{i=0}^{F-2} D_i^*(\alpha). \tag{4.4}$$

Therefore, $D^*(\alpha = 0)$ computes the average of all of the pairwise Dice scores. This rewards a model for which $\|\epsilon_i - \epsilon_j\|$ is small for all values of i and j. It does not, however, distinguish pairs i and j that are close to each other versus far away. Conversely, $D^*(\alpha = 1)$ computes the average of the pairwise Dice scores between warped segmentations of time-adjacent images. That is, it rewards a model that computes ϕ such that $\|\epsilon_{i+1} - \epsilon_i\|$ is small, but does not directly measure $\|\epsilon_i - \epsilon_j\|$ for $j > i + 1$. The benefits and pitfalls of setting α to either 0 or 1 are illustrated in Figure 4.2.

Values of α between 0 and 1 highlight the tradeoff between a temporal consistency with neighboring images and a consistent warped segmentation from all of the images. Mathematically, values of α close to 0 reward models that minimize $\|\epsilon\|$, while values of α close to 1 minimize $\frac{1}{F-2}\sum_{i=0}^{F-2}\|\epsilon_{i+1} - \epsilon_i\|$. The precise value of α that is best for evaluating a model is likely application-specific.

The D* score should not fully replace the Dice score as an evaluation technique. Specifically, when used in registration tasks, the Dice score computes deformation accuracy relative to the ground-truth segmentation, S_0. A deformation field can achieve a high D* score without any anatomical accuracy, as shown in Figure 4.3. It is, however, useful in distinguishing between models that achieve a comparable Dice score, or other similar metric measuring anatomical accuracy. Therefore, we will evaluate performance on both Dice score and D* score in order to capture both anatomical accuracy and temporal consistency.

This formulation can be expanded analogously to other traditionally pairwise metrics, including surface distance, by replacing D_{ij} with MSD_{ij}.

■ 4.4 Consistency Metrics

In registration models, we might have access to the deformation field as well as the warped images during evaluation time. In this case, we can directly analyze the deformation field ϕ to check for anatomical and longitudinal consistency.

■ 4.4.1 Anatomical Consistency

To determine whether the model is anatomically consistent, we compute the Jacobian determinant of ϕ, $|J(\phi)|$. The Jacobian matrix $J(\phi)$ computes the first order partial derivative of ϕ with respect to all of the inputs, and can therefore be used to identify diffeomorphic transforms and capture the amount of local expansion or contraction in part of the field [8, 11, 14, 28, 40, 79, 88]. For each pixel p in the input images, if $\phi^{(p)}$ represents the deformation at this pixel and $|J(\phi^{(p)})| \leq 0$, then the deformation is non-invertible at p, and therefore not diffeomorphic. If $|J(\phi^{(p)})| < 0$, then the deformation at p reverses its orientation, causing the pixel to fold on itself. This is also not diffeomorphic. We can therefore quantify the level of anatomical consistency for a deformation ϕ by the proportion of pixels for which $|J(\phi^{(p)})| > 0$.

■ 4.4.2 Temporal Consistency

We propose a new metric, $C_{\textsf{temp}}$, to quantify temporal consistency. Given a set of deformations that all achieve an anatomical accuracy within ϵ of each other, we prefer

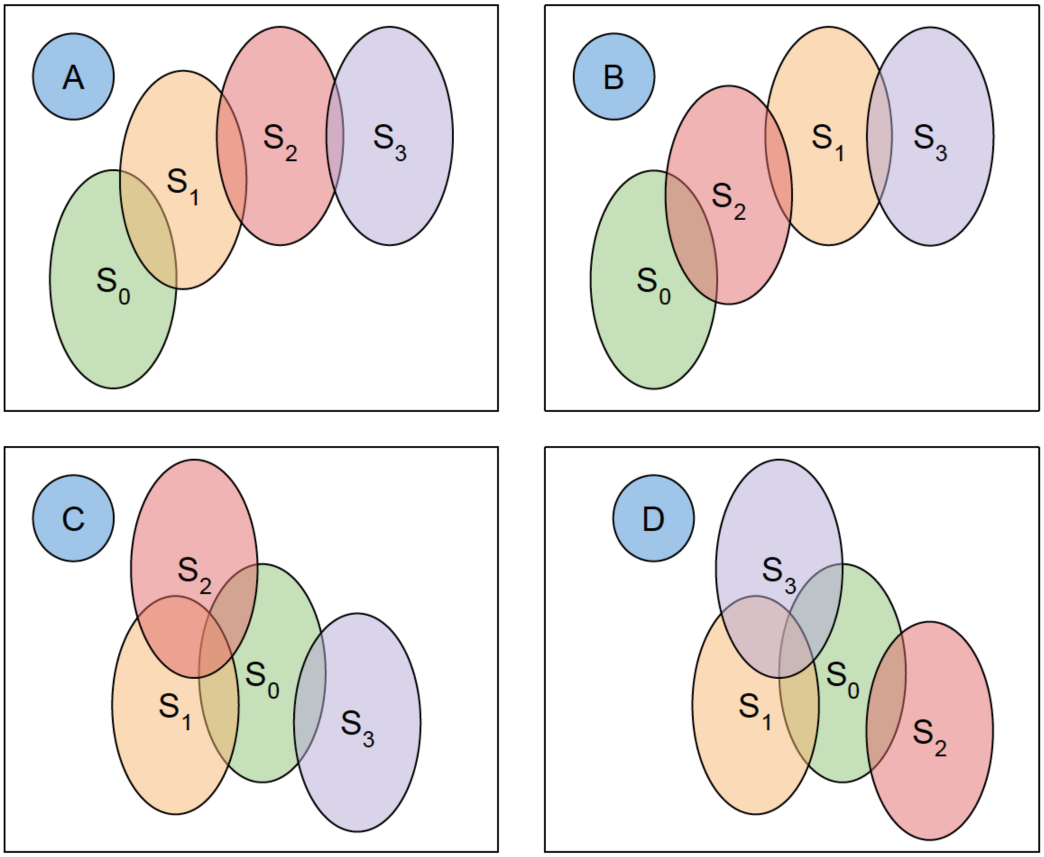


Figure 4.2: Four example scenarios of three warped segmentations, $S_1, S_2,$ and S_3 where S_i is the warped segmentation at time i, and a fixed segmentation S_0. In scenarios A and B, the $D^*(\alpha = 0)$ score would be the same, since the only difference between the two is the ordering of segmentations, while the $D^*(\alpha = 1)$ score would favor case A, where S_1 is closer to S_0 and S_3 is closer to S_2, compared to scenario B. Similarly, the $D^*(\alpha = 0)$ score is the same in scenarios C and D, while $D^*(\alpha = 1)$ favors C. In these comparisons, $D^*(\alpha = 1)$ successfully rewards temporal consistency. However, $D^*(\alpha = 1)$ prefers scenario A to scenario D, even though the warped segmentations in A continue to drift off in a direction that is not anatomically correct, while $D^*(\alpha = 0)$ scores scenario D more highly than scenario A, rewarding the clustered segmentations for all time points. Although these are exaggerated examples, the same method can quantify subtler differences between real segmentations.

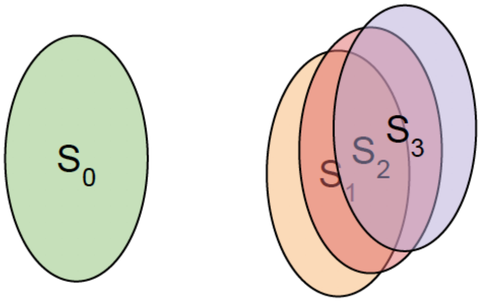


Figure 4.3: In this example with warped segmentations S_1, S_2, and S_3, and a fixed segmentation S_0, the model could achieve a relatively high D^* score due to the overlap between S_1, S_2 and S_3 for all values of α, but have no accuracy with regard to where the segmentations should be located anatomically. Meanwhile, their Dice score would be 0. D* is therefore intended to be used as a supplement to Dice score for spatiotemporal image registration evaluation.

the simplest of these deformations. A function from a simpler function class is less likely to overfit to the data than a more complex function. To quantify the temporal simplicity of a deformation, we consider the third-order derivative of the deformation with respect to time.

We propose using the change in acceleration implied by the deformation ϕ as a method to capture some temporal consistency information. In most cases, the registration model does not provide a continuous velocity field. Instead, this work proposes using $|\phi_{t+1} - \phi_t| \coloneqq \dot{\phi}_t$ as a piecewise estimate for the velocity at time t, since ϕ_t can be interpreted as the displacement between time 0 and time t. Similarly, we define $|\dot{\phi}_{t+1} - \dot{\phi}_t| \coloneqq \ddot{\phi}_t$ as an estimate for the acceleration at time t, and $|\ddot{\phi}_{t+1} - \ddot{\phi}_t| \coloneqq \dddot{\phi}_t$ as an estimate for the change in acceleration at time t. Then, we propose quantifying temporal consistency directly from ϕ according to $C_{\mathsf{temp}} = \left\| \dddot{\phi} \right\|$, where lower values of C_{temp} imply that ϕ is more temporally consistent.

This temporal consistency score C_{temp} should always be accompanied by an anatomical correctness metric. Many deformations would have $C_{\mathsf{temp}} = 0$ without learning any meaningful registration between images. Furthermore, anatomical structures *do* have a non-zero change in acceleration. Instead, C_{temp} can help distinguish between models that achieve ϵ-similar results on anatomical correctness metrics, such as the Dice score from Section 4.1, to determine which model best adheres to scientific understanding of anatomical activity.

Chapter 5

Evaluation

In this chapter, we evaluate Longitudinal VoxelMorph against several pairwise baselines on real-word medical images data. First, we evaluate the model on cardiac cine-MRI slices, which are high-quality, detailed images. Second, we evaluate Longitudinal VoxelMorph on a set of echocardiograms, which contain much less detail but are naturally 2-D. For a comparison between frames of a cardiac cine-MRI and frames of an echocardiogram, see Figure 2.1.

■ 5.1 Baseline Models

We compare the performance of Longitudinal VoxelMorph against several baselines. First, we consider the *naive model* that assumes no change between images. That is, the deformation field ϕ that the naive model estimates is the identity transformation. Although this is very simplistic, it is a reasonable baseline when registering images from the same subject, and performs well on images where development is slow over time, such as brain MRIs [66].

Second, we compare to VoxelMorph [23]. VoxelMorph is an efficient model for medical image registration, estimating a deformation field $\phi_{\mathsf{vxm\_pair}}(x, y)$ for each pair of images (x, y) that it is provided. To adapt VoxelMorph to our longitudinal setting, for a time series of images $X = [x_0, \ldots, x_{F-1}]$, we compute

$$\phi_{\mathsf{vxm}} = [\phi_{\mathsf{vxm\_pair}}(x_0, x_1), \phi_{\mathsf{vxm\_pair}}(x_0, x_2), \ldots, \phi_{\mathsf{vxm\_pair}}(x_0, x_{F-1})],$$

where each $\phi_{\mathsf{vxm\_pair}}$ is computed independently.

VoxelMorph does not estimate $\phi_{\mathsf{vxm\_pair}}$ at full resolution because of memory constraints with large input data, but instead uses linear interpolation as a final step fully resolve the deformation field [23]. Later sections will refer to this model as *linear VoxelMorph*. We also implemented a pairwise VoxelMorph model that uses b-spline interpolation to fully resolve $\phi_{\mathsf{vxm\_pair}}$. We will refer to this model as *b-spline VoxelMorph*.

Finally, we implemented a version of Longitudinal VoxelMorph using linear interpolation. We will refer to this version as *linear Longitudinal VoxelMorph*. For clarity, this chapter will refer the the Longitudinal VoxelMorph model presented in section 3.1 as *b-spline Longitudinal VoxelMorph*. The inclusion of linear Longitudinal VoxelMorph and b-spline VoxelMorph facilitate a comparison between linear and b-spline interpolation in addition to the comparison between longitudinal and pairwise model formulations.

We built the models using Keras [21], running on top of Tensorflow [7], and trained and evaluated the models using NVIDIA GPUs.

5.2 Cardiac cine-MRI experiment

Our first evaluation of Longitudinal VoxelMorph is on a set of cardiac cine-MRI images.

5.2.1 Cardiac Atlas Project Data

The Cardiac Atlas Project (CAP) compiled a set of cine-MRI slices from the Defibrillators to Reduce Risk by Magnetic Resonance Imaging Evaluation (DETERMINE) clinical trial [1, 32]. The data include cine-MRIs from 450 subjects, of which we randomly chose 100 to use for the experiment. The CAP DETERMINE dataset is made up of subjects with coronary artery diseases and mild-to-moderate left ventricular dysfunction. Each cardiac cine-MRI is 4-D (3-D image plus time), and made up of a set of slices, which are 3-D (2-D image plus time). We use the sequence of frames from each slice as the input sequence for each model to register.

Each slice of the cine-MRI is usually under 10 mm thick, with a gap of under 2 mm between slices. Each cine-MRI required a patient breath hold of 8-15 seconds [48]. Figure 5.1 shows the number of slices included in each subject's cardiac cine-MRI, the number of frames in each slice, and the number of pixels in a single frame of each slice. Although the size of the frames vary, we reshape every frame to 256x256 pixels using scipy's cubic spline interpolation as a pre-processing step [83]. We define the time at which each cine-MRI frame was taken to be the frame number in the sequence, so that there is one time unit between adjacent frames.

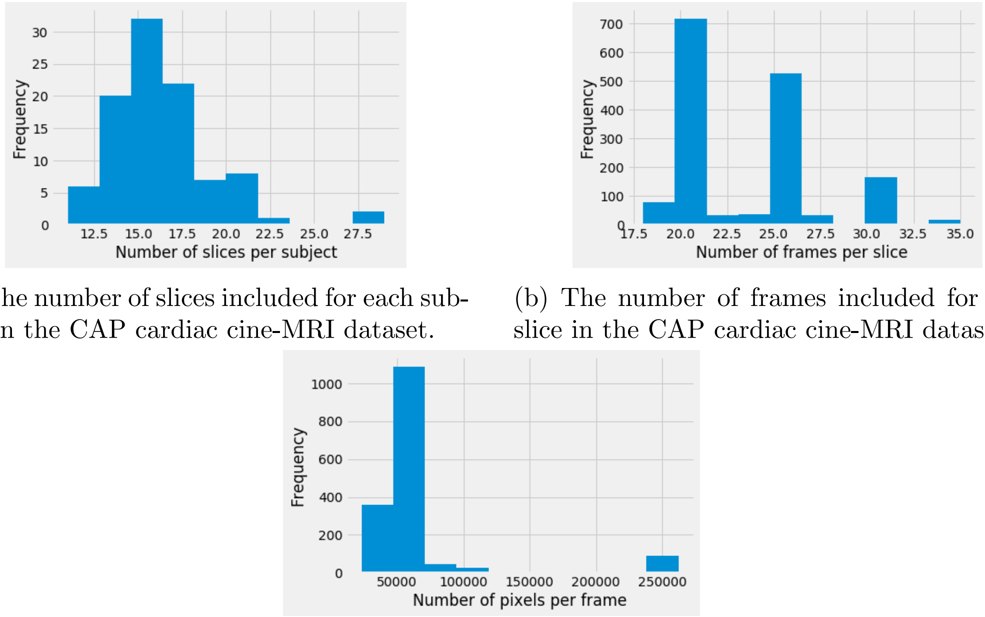
CAP also provides left-ventricular wall segmentations for each of the subjects. These segmentations delineate the boundary of the left ventricle, as shown in Figure 5.2 [32]. We use these segmentations to evaluate the models' performance.

We randomly divided the 100 subjects into training, validation, and test subjects. We use 78 training subjects, 11 validation subjects, and 11 test subjects.

5.2.2 Training and Hyperparameters

We ran the b-spline and linear Longitudinal VoxelMorph models for 750 epochs of training, and the b-spline and linear VoxelMorph models for 1500 epochs of training, using the Adam optimizer [45] and a learning rate of 1×10^{-4}.

We used mean squared error for the image similarity loss, and the L2 norm of the deformation field for the regularization penalty. We trained models with different regularization parameters (λ), and empirically found that values $\lambda = 0.01$ achieved the highest pairwise Dice scores for the baseline pairwise VoxelMorph models on the validation set. The average Dice scores on the validation set were highest when $\lambda = 0.05$ for the Longitudinal VoxelMorph models.

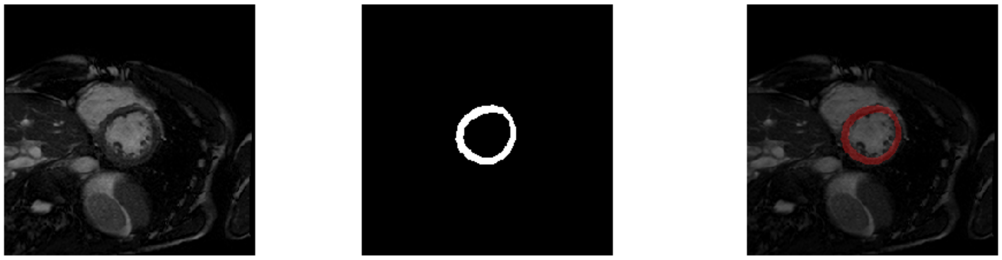


(a) The number of slices included for each subject in the CAP cardiac cine-MRI dataset.

(b) The number of frames included for each slice in the CAP cardiac cine-MRI dataset.

(c) The number of pixels included for each frame in each slice in the CAP cardiac cine-MRI dataset, before they are resized to 256x256.

Figure 5.1: CAP cardiac cine-MRI data distribution for the number of slices in each subject, the number of frames in each slice, and the number of pixels in each frame.



(a) Example frame from a cardiac cine-MRI slice showing the left ventricle.

(b) Example segmentation of the left ventricle for the frame shown in Figure 5.2a.

(c) The left ventricle segmentation 5.2b, shown in red, overlaid on the cine-MRI slice 5.2a.

Figure 5.2: Each frame of each slice in the CAP dataset has an accompanying left ventricle segmentation. Each segmentation outlines the left-ventricular wall, and does not include tissue inside the ventricle, as shown in Fig. 5.2c. These segmentations are used to evaluate the model's performance.

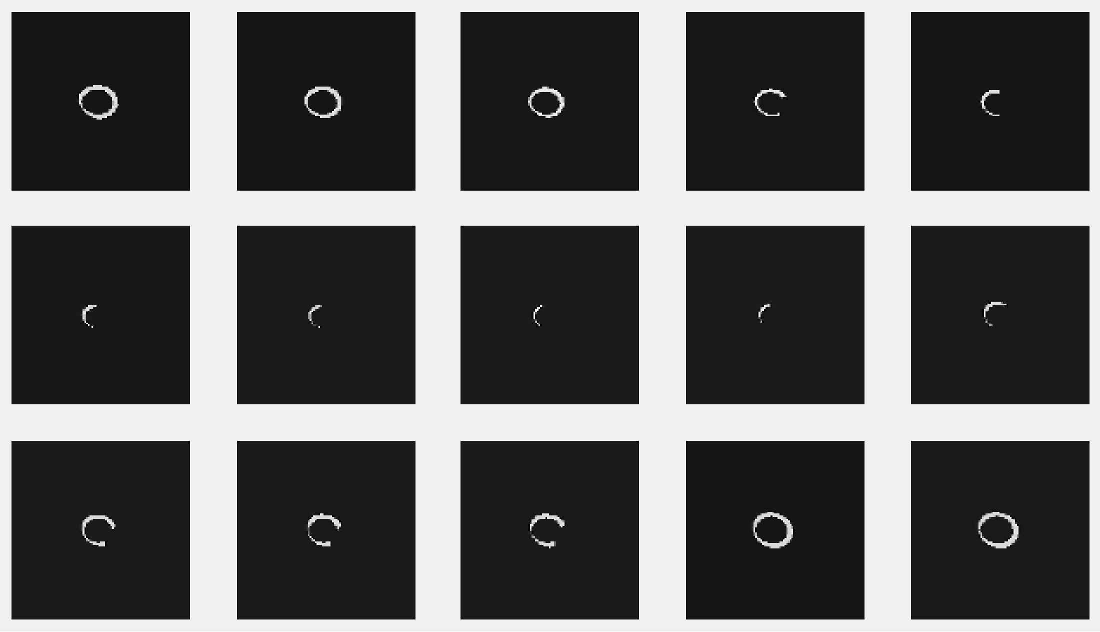


Figure 5.3: The accompanying segmentation for 15 frames of a cardiac cine-MRI slice. Some cine-MRI slices show slightly different anatomy in different frames, as cardiac structures expand and contract over the time. This is reflected in the segmentations. Here, the border of the left ventricle partially exits and the re-enters a slice. The model is exposed to some examples such as this one during training, validation, and testing time. This violates the model assumption that the true deformation is diffeomorphic since each image does not contain the same anatomy. We would expect worse performance on this kind of example.

For this cardiac cine-MRI experiment, the models were given $F = 15$ frames from a single slice for a patient to register. We set the maximum time value to $K = 40$, meaning we never try to model any frame later than the 40th in the sequence.

We trained models with different values of the spatial- and temporal-sparsity ratios, *i.e.*, with different values of δ_x, δ_y, and, in the case of the longitudinal models, δ_T. We used $\delta_x = \delta_y$, and refer their value as δ in the subsequent section.

■ 5.2.3 Analysis

We trained different versions of Longitudinal VoxelMorph with spatial-sparsity ratios of $\delta = 4$ and $\delta = 8$, and with $\delta_T = 3$ and $\delta_T = 5$. Figure 5.4a shows the training losses for the Longitudinal VoxelMorph models we implemented. On our validation dataset, we empirically found that the best b-spline Longitudinal VoxelMorph model and the best linear Longitudinal VoxelMorph model both used $\delta = 4$ and $\delta_T = 3$. Here we defined *best* by the highest average Dice scores across 10 randomly selected validation examples.

We also trained different versions of the baseline pairwise VoxelMorph models with different sparsity levels (training loss shown in Figure 5.4b). The best b-spline VoxelMorph baseline model used $\delta = 2$ and the best linear VoxelMorph baseline model used $\delta = 4$.

We compared the top-performing model of each type on the test set.

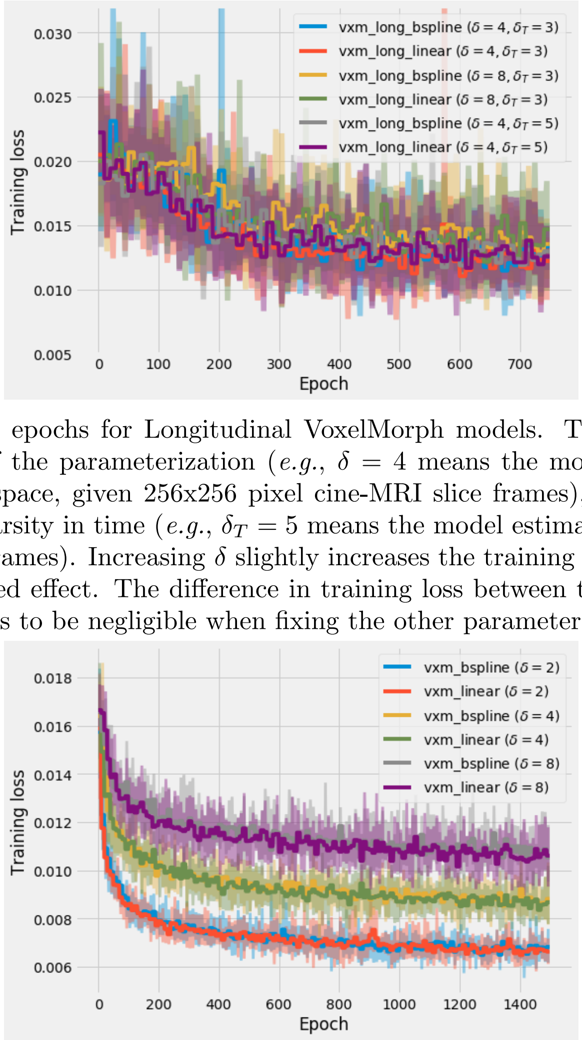
■ 5.2.4 Results

Overall, our results show that the baseline pairwise VoxelMorph models slightly outperform the Longitudinal VoxelMorph models in anatomical-correctness metrics on this dataset, but the Longitudinal VoxelMorph models are more temporally consistent. We find that the Longitudinal VoxelMorph models achieve a Dice score **2.715% lower** than the pairwise VoxelMorph models on average over our test set. The Longitudinal VoxelMorph models achieve a $D^*(\alpha = 1)$ value **0.462% higher** than the pairwise VoxelMorph models. The MSD for the Longitudinal VoxelMorph models was **8.710%** higher than the pairwise models.

All models have high spatial consistency on the test set. With the optimal spatial- and temporal-sparsity values, we did not observe a significant difference between models using b-spline versus linear interpolation.

Figure 5.5a shows the pairwise Dice scores, where every warped segmentation is compared to the fixed segmentation of the 0th frame in the cine-MRI slice. Figure 5.5b shows the D* score, as defined in Equation (4.4), across varying values of the parameter α.

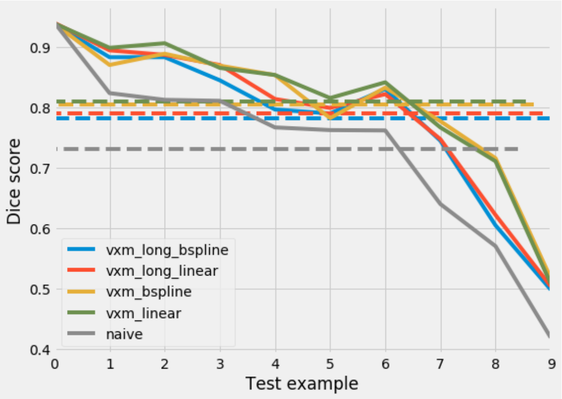
We found that the pairwise VoxelMorph models performed best in terms of Dice score, followed by both Longitudinal VoxelMorph models, which perform very similarly. For low values of α, the pairwise VoxelMorph models achieved better $D^*(\alpha)$ scores than the Longitudinal VoxelMorph models, suggesting that the pairwise models had a lower overall error in the deformation field. As α increased, though, the Longitudinal



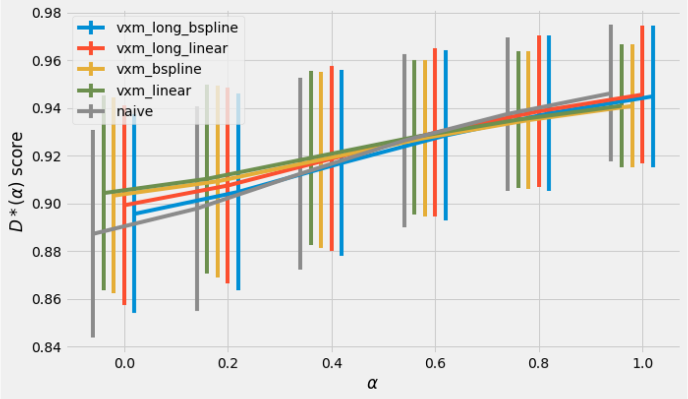
(a) Loss over training epochs for Longitudinal VoxelMorph models. The labeled δ indicates the spatial sparsity of the parameterization (*e.g.*, $\delta = 4$ means the model estimates a 64x64 control-point grid in space, given 256x256 pixel cine-MRI slice frames), and δ_T indicates the parameterization's sparsity in time (*e.g.*, $\delta_T = 5$ means the model estimates 3 control points in time, given 15 input frames). Increasing δ slightly increases the training loss, but increasing δ_T does not have a marked effect. The difference in training loss between the b-spline and linear model versions appears to be negligible when fixing the other parameters.

(b) Training loss over epochs for baseline VoxelMorph models. The $\phi_{\mathsf{vxm\_pair}}$ for each model is calculate at a sparsity factor of δ (*e.g.*, with $\delta = 4$, $\phi_{\mathsf{vxm\_pair}}$ is a 64x64 vector field, given a 256x256 cine-MRI frame).

Figure 5.4: The loss over training epochs for the Longitudinal VoxelMorph (Fig. 5.4a) and baseline VoxelMorph (Fig. 5.4b) models on the cardiac cine-MRI dataset. The solid line gives the average loss across a window of 10 training epochs, and the surrounding band of the same color shows the minimum and maximum epoch losses over that same period.



(a) The average left-ventricular wall Dice scores the for 10 testing examples. The average Dice score for each subject is computed according to the average of $[\mathsf{Dice}(x_0, x_1), \ldots, \mathsf{Dice}(x_0, x_{F-1})]$, as in Equation (4.1). The average pairwise Dice scores across all 10 subjects is shown as a dotted line. The b-spline and linear VoxelMorph models achieve the highest average Dice score, followed by the b-spline and linear Longitudinal VoxelMorph models. All learned models outperform the naive model baseline. Different interpolation schemes for different models perform similarly. The poor Dice score of the worst examples can be explained by changing anatomy present in the cine-MRI slice, as shown in Figure 5.3.



(b) The average left-ventricular wall D* scores for the 10 testing examples across different values of α, as defined in Equation (4.4). The standard deviations across testing examples for the fixed α values are also shown. On average, the pairwise VoxelMorph models slightly outperform the longitudinal models for low values of α, and the longitudinal models outperform the pairwise models for higher values of α. This suggests that the deformations estimated by the longitudinal models are more temporally consistent.

Figure 5.5: The average pairwise Dice scores, relative to the baseline image, and the D* scores across different values of α for each of the models for 10 randomly selected testing examples. All test examples were held out during training and validation stages. The pairwise Dice scores shown in Fig. 5.5a best capture the anatomical accuracy of the warped segmentations, while the D* scores better measure the deformation's temporal consistency.

VoxelMorph models performed best, suggesting the error in their deformation fields was most temporally consistent. An explicit comparison of the models' average performance on Dice and D* scores can be found in Table 5.1.

We compare the models' average MSD in Figure 5.6a. We found that the pairwise VoxelMorph models slightly outperform the Longitudinal VoxelMorph models, and all outperform the naive baseline. However, if we exclude the final two test slices, both of which contain a vanishing left ventricular wall (as in Figure 5.3), the difference in MSD between Longitudinal VoxelMorph and pairwise VoxelMorph models decreases, as shown in Figure 5.6b. An explicit comparison of the models' average MSD can be found in Table 5.2.

We also quantified the models' spatial and temporal consistency using only the deformation fields. Overall, all models estimate spatially consistent deformations, but the Longitudinal VoxelMorph models are much more temporally consistent. The evaluation of our top-performing models on each of these metrics is given in Table 5.3.

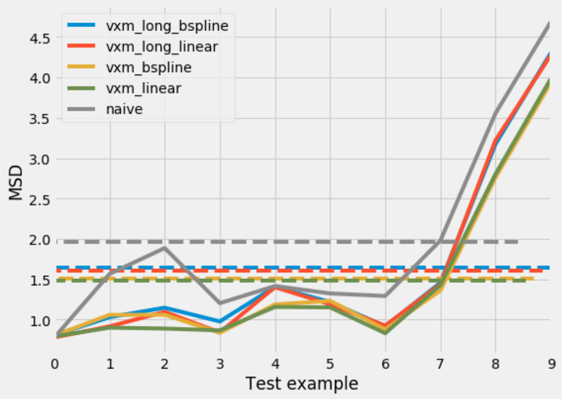
The performance for all models on this dataset is slightly impaired by the change in anatomy between different frames from the same slice, as presented in Figure 5.3. With small differences between the models' evaluation performance, it is difficult to determine the impact of the frames' variable anatomy. Although the echocardiogram

Model	Mean Dice Score	Mean $D^*(\alpha = 0)$	Mean $D^*(\alpha = 1)$
vxm_long_bspline	0.781 (0.185)	0.896 (0.042)	0.945 (0.030)
vxm_long_linear	0.790 (0.183)	0.899 (0.183)	0.946 (0.029)
vxm_bspline	0.805 (0.169)	0.903 (0.041)	0.941 (0.026)
vxm_linear	0.811 (0.169)	0.904 (0.049)	0.941 (0.026)
naive_model	0.731 (0.224)	0.887 (0.044)	0.946 (0.029)

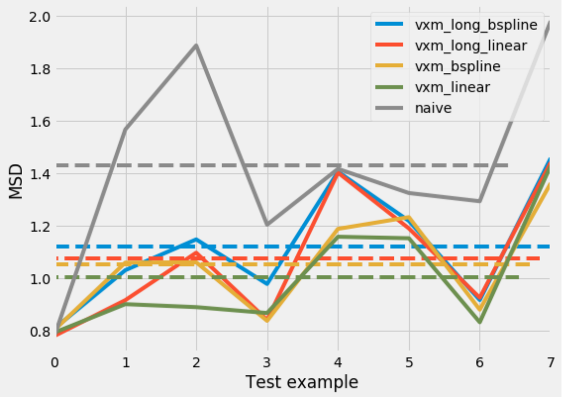
Table 5.1: Comparison of the models' Dice-based evaluation results. We show the average Dice, $D^*(\alpha = 0)$, and $D^*(\alpha = 1)$ scores across the ten test subjects from Fig. 5.5. Higher values are better. Standard deviations are given parenthetically.

Model	MSD ($n = 10$)	MSD ($n = 8$)
vxm_long_bspline	1.644 (2.053)	1.121 (0.506)
vxm_long_linear	1.609 (2.097)	1.074 (0.499)
vxm_bspline	1.512 (1.871)	1.053 (0.486)
vxm_linear	1.481 (1.876)	1.003 (0.487)
naive_model	1.969 (2.184)	1.432 (0.817)

Table 5.2: Comparison of the models' MSD evaluation results. We show the average MSD across the ten test slices shown in Fig. 5.6a, and the average MSD across the eight test slices shown in Fig. 5.6b, where we exclude the two slices where the left-ventricular wall significantly exits the slice before re-entering (see Fig. 5.3). Lower values are better. Standard deviations are given parenthetically.



(a) The average MSD, relative to the baseline image, for each of the models across the same 10 randomly selected testing examples as Fig. 5.5. The average MSD for each subject is computed according to the average of $[\mathsf{MSD}(x_0, x_1), \ldots, \mathsf{MSD}(x_0, x_{F-1})]$, as in Equation (4.2). All of the learned models consistently outperform the naive model, and the pairwise VoxelMorph models slightly outperform the Longitudinal VoxelMorph models overall.



(b) The models' average MSD, relative to the baseline images for the first 8 testing examples in Fig. 5.6a. This excludes the two testing slices that contained a dramatically vanishing left-ventricular wall, as in Fig. 5.3. When only considering these examples, the Longitudinal VoxelMorph models perform much more similarly to the pairwise baselines.

Figure 5.6: The average MSD for the models. Fig. 5.6a shows the MSD for the same ten randomly selected testing examples shown in Fig. 5.5. Fig. 5.6b shows the MSD over the first 8 of these examples, excluding the two examples where the segmentation for the left-ventricular wall significantly exited the slice before re-entering it, as shown in Fig. 5.3.

Model	% of $\lvert J(\phi_t)\rvert \leq 0$	$C_{\mathsf{long}} = \left\lVert \dddot{\phi} \right\rVert$
vxm_long_bspline	0.0 (0.0)	55.171 (24.951)
vxm_long_linear	$2.223 \times 10^{-5} (5.401 \times 10^{-5})$	55.710 (27.328)
vxm_bspline	0.0 (0.0)	119.736 (47.003)
vxm_linear	0.0 (0.0)	124.867 (52.662)
naive_model	0.0 (0.0)	0.0 (0.0)

Table 5.3: Comparison of the models on spatial- and temporal-consistency metrics on 10 test examples. First, we show the percentage of pixels for which the Jacobian of the deformation has a non-positive determinant (lower is better), meaning the deformation is not spatially diffeomorphic at that location (see Chapter 4). Second, we show the L2 norm of the estimated acceleration (lower is better) implied by the deformation field. The estimated flow field for the Longitudinal VoxelMorph models have a smaller change in acceleration, suggesting that they are more temporally consistent. The standard deviations are given parenthetically.

data is not as precise as the cine-MRI data, the experiment presented in the next section does not have the limitation of inconsistent anatomy between frames.

5.3 Echocardiogram experiment

Our second evaluation of Longitudinal VoxelMorph is on a set of echocardiography videos.

5.3.1 EchoNet Data

We use the EchoNet-Dynamic dataset [61], which includes 10,030 echocardiograms. Each echocardiogram is 3-D (2-D images plus time). The echocardiograms are from the 4-chamber apical view, although taken using varying angles, positions, and image acquisition techniques. These differences are intended to mimic the normal variation in clinical echocardiograms. The videos in the EchoNet dataset have been cropped and downsampled to 112x112 pixel frames using cubic interpolation [61]. Figure 5.7 shows the distribution of the number of frames in each echocardiography video.

The EchoNet dataset also includes expert tracings for the heart's left ventricle [61]. These tracings, an example of which is given in Figure 5.8b, use a set of lines to define the area of interest. We then converted the tracing to an area segmentation, as in Figure 5.8c, using linear interpolation between the endpoints of adjacent lines.

The left-ventricle expert tracings are only available for two points in the cardiac cycle: the end-systolic phase (ES, the moment at which the left ventricle finishes contracting) and end-diastolic phase (ED, the moment at which the left ventricle finishes expanding) frames. The end-systolic volume (ESV) and the end-diastolic volume (EDV) quantify the volume of blood remaining in the ventricle at their respective phases, and are used to compute the cardiac ejection fraction (EF). EF is a common indicator of

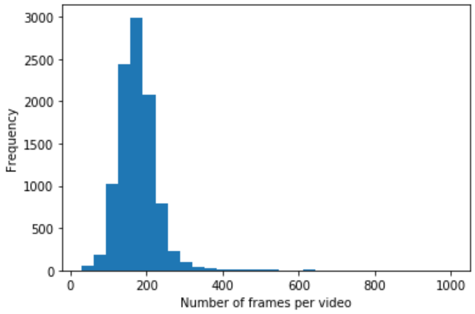
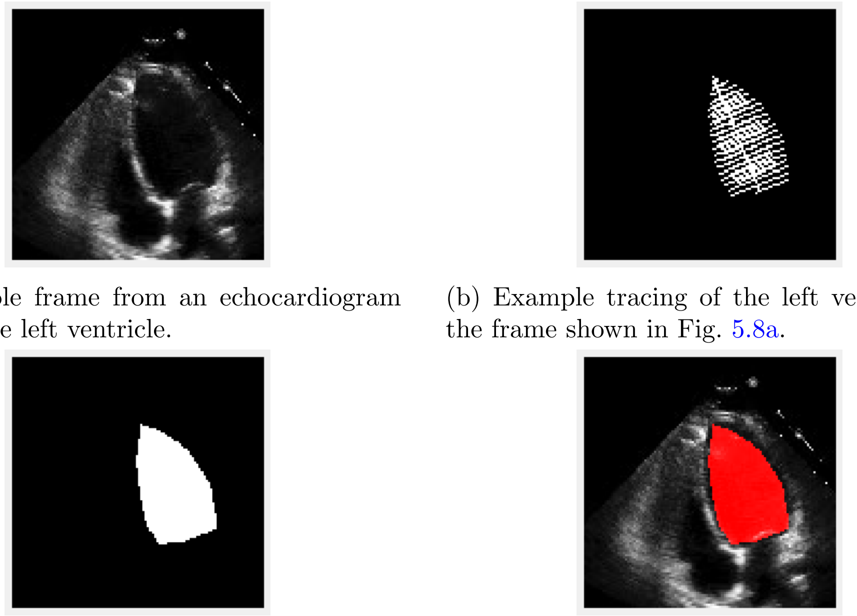


Figure 5.7: The number of frames included in each video of the EchoNet dataset.



(a) Example frame from an echocardiogram showing the left ventricle.

(b) Example tracing of the left ventricle for the frame shown in Fig. 5.8a.

(c) The left ventricle segmentation created from the tracing in Fig. 5.8b using linear interpolation between endpoints of adjacent lines.

(d) The left ventricle segmentation in Fig. 5.8c, shown in red, overlaid on the echocardiogram frame in Fig. 5.8a.

Figure 5.8: Each frame in the EchoNet dataset has an accompanying left ventricle tracing, which we convert to a segmented area as a pre-processing step. These segmentations are used to evaluate the models' performance.

risk, and is therefore a clinical value of interest [37]. Since only these two tracings are included for each subject, we will only use these two frames for the segmentation-based evaluation metrics presented in Chapter 4. We are unable to evaluate the models using D*, since ground-truth segmentations are not available for the other frames.

We divided the dataset into 7460 training examples, 1288 validation examples, and 1288 testing examples. During validation- and test-time, we ignored any examples where the end-diastolic and end-systolic phase frames did not both occur in the first F frames.

5.3.2 Training and Hyperparameters

We trained the b-spline and linear Longitudinal VoxelMorph models for 750 epochs, and the baseline b-spline and linear VoxelMorph models for 1500 epochs. We used the Adam optimizer [45] and a learning rate of 1×10^{-4}. We again used mean squared error for the image similarity loss and the L2 norm of the deformation field for the regularization penalty. Empirically, we found an optimal regularization parameter $\lambda = 0.01$ for the baseline VoxelMorph models and $\lambda = 0.05$ for the Longitudinal VoxelMorph models on the validation set.

In the echocardiography video experiment, the models are given $F = 25$ video frames to register, where each adjacent frame is again considered to be one time unit apart. We set the maximum time to be $K = 150$, so we never model any frames past the 150th in a sequence.

As in the cardiac cine-MRI experiment, we trained models with varying spatial- and temporal-sparsity parameters, δ and δ_T respectively.

5.3.3 Analysis

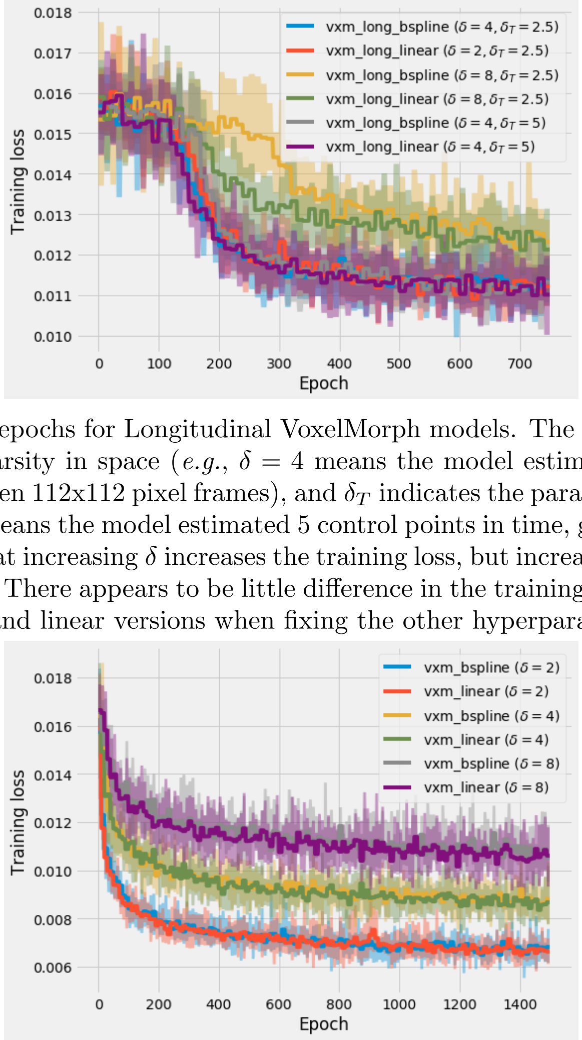
We trained different versions of Longitudinal VoxelMorph with spatial-sparsity ratios of $\delta = 4$ and $\delta = 8$, and with $\delta_T = 2.5$ and $\delta_T = 5$. Figure 5.9a shows the training losses for the Longitudinal VoxelMorph models that we trained. On our validation dataset, we empirically found that the best b-spline Longitudinal VoxelMorph model and the best linear Longitudinal VoxelMorph model both used $\delta = 4$ and $\delta_T = 5$. We defined *best* based on the Dice score and MSD across 10 randomly selected validation examples.

We also trained different versions of the baseline pairwise VoxelMorph models with varying sparsity levels (training loss shown in Figure 5.9b). The best b-spline and linear VoxelMorph baseline models both used $\delta = 2$.

We compared the top-performing model of each type on the test set.

5.3.4 Results

Overall, we found that the Longitudinal VoxelMorph models generally outperformed the baseline pairwise VoxelMorph models on segmentation-based metrics, with both types of models consistently outperforming the naive baseline. When compared in the reference frame of the 0th echocardiogram frame, the Longitudinal VoxelMorph Models achieved



(a) Loss over training epochs for Longitudinal VoxelMorph models. The labeled δ indicates the parameterization's sparsity in space (*e.g.*, $\delta = 4$ means the model estimated an 28x28 control point grid in space, given 112x112 pixel frames), and δ_T indicates the parameterization's sparsity in time (*e.g.*, $\delta_T = 5$ means the model estimated 5 control points in time, given 25 input frames). Here, we can notice that increasing δ increases the training loss, but increasing δ_T by a factor of 2 has a negligible effect. There appears to be little difference in the training loss after convergence between the b-spline and linear versions when fixing the other hyperparameters.

(b) Training loss over epochs for baseline VoxelMorph models. The $\phi_{\mathsf{vxm\_pair}}$ for each model is calculated at a sparsity of the labeled δ (*e.g.*, with $\delta = 4$, $\phi_{\mathsf{vxm\_pair}}$ is a 28x28 vector field, given an image size of 112x112). Here, models with a lower δ value have a lower loss but, given a fixed δ, the b-spline and linear model versions have a similar loss.

Figure 5.9: Training loss plots for the Longitudinal VoxelMorph and baseline VoxelMorph models trained with different hyperparameter settings on the echocardiogram experiment. The solid line shows the average training loss over a window of 10 epochs, while the surrounding shaded area of the same color shows the minimum and maximum epoch training loss over the same period.

a **3.84% higher Dice score** and **3.50% lower MSD** than the pairwise models on average. In the reference frame of the fixed image (either the end-systolic or end-diastolic phase frame, whichever came first in that particular video), the Longitudinal VoxelMorph models achieved **1.55% higher Dice score** and **3.94% higher MSD** than the pairwise VoxelMorph models on average across our test set.

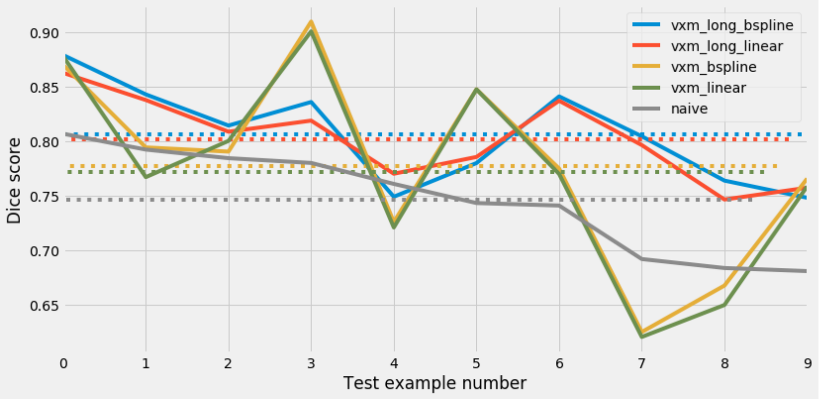
We found that there were negligible differences between the b-spline- and linear-versions of each model. The b-spline Longitudinal VoxelMorph model outperformed the other models in spatial consistency metrics, although all perform well, and the longitudinal models perform better on the temporal consistency evaluation than the pairwise models.

All models' evaluation on spatial- and temporal-consistency metrics is given in Table 5.4. We found that all models are highly spatially consistent, and the Longitudinal VoxelMorph models are much more temporally consistent than the baseline pairwise VoxelMorph models.

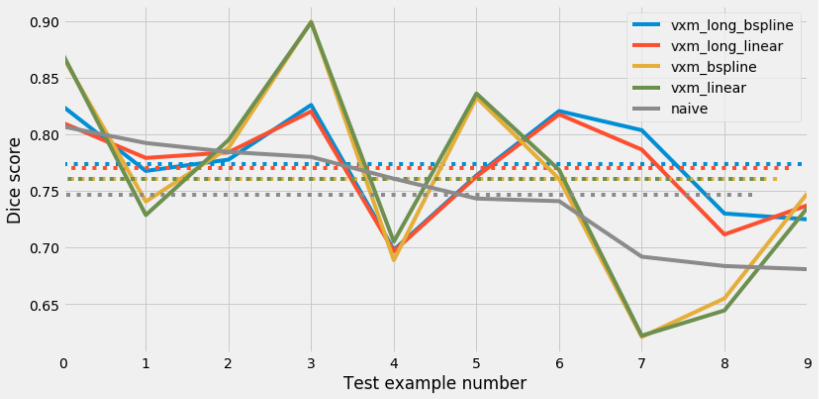
Figure 5.10 shows the pairwise Dice scores [30] for each of the models, and Figure 5.11 shows the pairwise MSD. The Longitudinal VoxelMorph models outperform the baseline pairwise VoxelMorph models in terms of Dice score. For MSD, the Longitudinal VoxelMorph models slightly outperform the pairwise VoxelMorph models when segmentations are compared in the reference frame of the 0th frame. When compared in the reference frame of the end-systolic or end-diastolic phase (whichever came first in the echocardiogram), the pairwise baseline VoxelMorph models perform slightly better. There is not a clear difference in either Dice score or MSD between b-spline and linear interpolation for these values of δ and δ_T. All of the learned models outperform the naive baseline model in all of the anatomical accuracy metrics. An explicit comparison of the models' average performance for Dice score and MSD can be found in Table 5.5.

Model	% of $\lvert J(\phi_t)\rvert \leq 0$	$C_{\textsf{long}} = \lVert \dddot{\phi} \rVert$
vxm_long_bspline	0.0 (0.0)	9.901 (2.507)
vxm_long_linear	$3.943 \times 10^{-4} (6.711 \times 10^{-4})$	7.703 (2.047)
vxm_bspline	$7.972 \times 10^{-6} (1.305 \times 10^{-5})$	410.377 (74.576)
vxm_linear	$3.720 \times 10^{-5} (5.013 \times 10^{-5})$	396.553 (90.261)
naive_model	0.0 (0.0)	0.0 (0.0)

Table 5.4: Comparison of the models' spatial- and temporal-consistency evaluation results. The metrics are averages of the estimated deformation field for the same 10 test subjects as Fig. 5.10 and Fig. 5.11. We include the percentage of pixels for which the Jacobian of the deformation has a non-positive determinant (lower is better), meaning the deformation is not spatially diffeomorphic at that location. We also provide the L2 norm of the estimated acceleration (lower is better) implied by the deformation field. The standard deviations are given parenthetically.

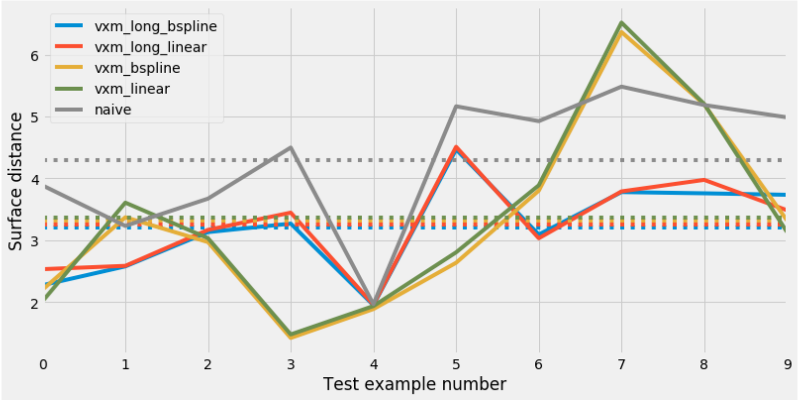


(a) The pairwise Dice scores across 10 testing examples, comparing all segmentations in the reference frame of the 0th frame in the sequence. The average pairwise Dice score (in the reference frame of the 0th frame) is shown as a dotted line for each model. Across these ten examples, the Longitudinal VoxelMorph models outperform the pairwise VoxelMorph models on average, and consistently outperform the naive baseline.

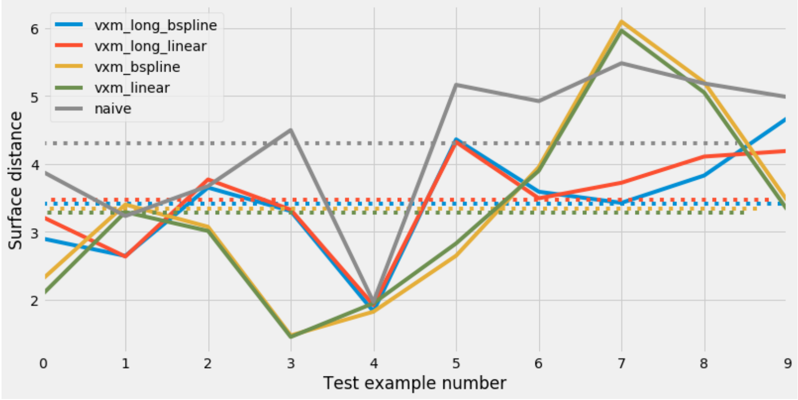


(b) The pairwise Dice scores across 10 testing examples, comparing all segmentations in the reference frame of either the end-systolic or end-diastolic phase, whichever came earlier in the echocardiogram. The models achieve this by first warping the moving segmentation to the reference frame of the 0th frame in the sequence, and then to the appropriate frame for the fixed segmentation. Again, the dotted line shows the average Dice score for each model. The Longitudinal VoxelMorph models continue to outperform the pairwise VoxelMorph models and the naive models on average. All of the VoxelMorph models have a slightly lower Dice score than when compared in the reference frame of the baseline image (as shown in Fig. 5.10a).

Figure 5.10: The pairwise Dice scores across 10 testing examples, compared in the reference frame of the 0th frame in the echocardiogram (Fig. 5.10a) and in the reference frame of the end-systolic or end-diastolic phase frame, whichever came first in the video (Fig. 5.10b). We randomly selected the testing examples from the subset of the test set that contained both end-systolic and end-diastolic phase frames before the 150th frame ($F = 150$), enabling the model to estimate the deformations. All testing examples were held out during the training and validation stages. The average Dice score of each model in each reference frame is also shown.



(a) The pairwise MSD for 10 testing examples, comparing all segmentations in the reference frame of the 0th frame in the sequence. The average MSD is shown as a dotted line for each model. In this reference frame there is not a clear difference between the Longitudinal VoxelMorph and pairwise VoxelMorph models. All consistently outperform the naive model baseline.



(b) The pairwise MSD for 10 testing examples, comparing all segmentations in the reference frame of either the end-systolic or end-diastolic phase frame, whichever came earlier in the echocardiogram. The models achieve this by first warping the moving segmentation to the reference frame of the 0th frame in the sequence, and then to the appropriate frame for the fixed segmentation. Again, the dotted line shows the average MSD score for each model. Here, the pairwise VoxelMorph baselines appear to slightly outperform the Longitudinal models, and all still outperform the naive baseline.

Figure 5.11: The pairwise MSD for the same 10 testing examples as Fig. 5.10, compared in the reference frame of the 0th frame in the echocardiogram (Fig. 5.11a) and in the reference frame of the end-systolic or end-diastolic phase frame, whichever came first in the video (Fig. 5.11b). A lower MSD is better. The average MSD in each reference frame is also shown.

Model	Mean Dice Score (rf 0; rf Fixed)	Mean MSD (rf 0; rf Fixed)
vxm_long_bspline	0.791 (0.062) ; 0.774 (0.043)	3.204 (0.733) ; 3.421 (0.783)
vxm_long_linear	0.802 (0.036) ; 0.770 (0.041)	3.249 (0.720) ; 3.471 (0.707)
vxm_bspline	0.777 (0.083) ; 0.760 (0.085)	3.318 (1.433) ; 3.344 (1.374)
vxm_linear	0.771 (0.087) ; 0.760 (0.087)	3.363 (1.468) ; 3.286 (1.326)
naive_model	0.746 (0.044) ; 0.746 (0.044)	4.301 (1.048) ; 4.301 (1.048)

Table 5.5: Comparison of the models' segmentation-based anatomical-accuracy evaluation results. We include the average Dice score and MSD across the ten test subjects from Fig. 5.10 and 5.11. The metrics are shown first when compared in the reference frame of the 0th frame, and then when compared in the reference frame of the fixed frame (showing either the end-diastolic or end-systolic volume, whichever came first in the video). Higher values are better for Dice score; lower values are better for MSD. Standard deviations are given parenthetically.

Chapter 6

Conclusion and Discussion

In this work we presented Longitudinal VoxelMorph, a novel machine learning model for spatiotemporal medical image registration. It improves upon classical longitudinal registration methods by using a learning-based estimation network for spatiotemporal fields, increasing model efficiency, and enabling its practical use. Longitudinal VoxelMorph promises to estimate a more globally accurate deformation field than pairwise methods. It uses a sufficiently sparse representation to take advantage of all available input data in a time series, enabling it to outperform state-of-the-art pairwise models that can only utilize a small subset of the data at a time.

We evaluated Longitudinal VoxelMorph and several pairwise baseline models on real-world datasets. The results on cardiac cine-MRI data and echocardiogram data suggest that Longitudinal VoxelMorph and state-of-the-art pairwise models achieve similar anatomical accuracy in their estimated deformation, but that Longitudinal VoxelMorph is more temporally consistent. We did not find a significant improvement with b-spline interpolation as compared to linear interpolation, and hypothesize that this was because the estimated deformation, before it was brought to full resolution, was sufficiently dense for linear interpolation to perform well. The density of the presented deformation fields was possible with the data analyzed in this work, but is not scalable to datasets where the input images are larger.

Learning a sparser deformation field representation enables shorter training times, a smaller memory load, and a lower likelihood of overfitting to noise in the data than a denser deformation estimate. Although the experiments we presented in this work focused on 3-D data (2-D images plus time), in the future we will extend it to 4-D data (3-D images plus time). With 4-D data, the spatial- and temporal-sparsity factors will become even more important, since practical memory constraints will require a scalable estimation solution.

Such 4-D data also presents the need for Longitudinal VoxelMorph to estimate the longitudinal deformation in a pairwise manner. With larger, 4-D input data, model implementations may no longer be able to load all of the time series images in memory simultaneously. Instead, Longitudinal VoxelMorph could be updated to learn control-point placement in space and time while only ever loading pairs of images into memory.

In future work, we will evaluate more existing datasets. Both the cine-MRI and echocardiogram data capture cardiac motion over time. There is far more motion contained in a video of the heart, such as the expansion and contraction of ventricles and the opening and closing of valves, than in repeated brain MRIs or dental x-rays. In these other kinds of medical images where we would expect less change over time, noise may contribute more to the differences between images than in the cardiac case. We therefore believe that evaluating Longitudinal VoxelMorph on these fundamentally different kinds of medical image data would speak to the generalizability, strengths, and weaknesses of the model.

To transition to these new kinds of data, we also aim to apply Longitudinal VoxelMorph to a variable number of input images. For most kinds of repeated medical imaging there can be a variable number of follow-up scans, and we do not want the model to be limited to only considering a subset of the available image data for a patient.

Longitudinal VoxelMorph can be applied to other longitudinal tasks in the future, such as image and diagnostic prediction. We believe that such a prediction model, which takes advantage of all available temporal data, could have a significant clinical impact.

Bibliography

[1] Defibrillators to reduce risk by magnetic resonance imaging evaluation. URL https://clinicaltrials.gov/ct2/show/NCT00487279.

[2] Magnetic resonance imaging (mri). URL https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri.

[3] *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*, volume 1. Springer. URL https://www.springer.com/gp/book/9781441981943#aboutBook.

[4] Correction for holland et al., subregional neuroanatomical change as a biomarker for alzheimer's disease. *Proceedings of the National Academy of Sciences*, 107(14): 6551–6551, 2010. ISSN 0027-8424. doi: 10.1073/pnas.1001505107. URL https://www.pnas.org/content/107/14/6551.1.

[5] Cine imaging, 2013. URL https://www.med-ed.virginia.edu/courses/rad/cardiacmr/Techniques/Cine.html.

[6] Ultrasound, Jul 2016. URL https://www.nibib.nih.gov/science-education/science-topics/ultrasound.

[7] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

[8] J. Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38 (1):95–113, 2007. doi: http://dx.doi.org/10.1016/j.neuroimage.2007.07.007.

[9] J. Ashburner and K.J. Friston. Voxel-based morphometry – the methods. *NeuroImage*, 11:805–821, 2000.

[10] D J Atkinson and R R Edelman. Cineangiography of the heart in a single breath hold with a segmented turboflash sequence. *Radiology*, 178(2):357–360, 1991. doi: 10.1148/radiology.178.2.1987592. URL https://pubs.rsna.org/doi/10.1148/radiology.178.2.1987592.

[11] Brian B. Avants, Charles L. Epstein, Murray Grossman, and James C. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Anal.*, 12(1): 26–41, 2008. URL http://dblp.uni-trier.de/db/journals/mia/mia12.html#AvantsEGG08.

[12] R. Bajcsy and S. Kovacic. Multiresolution elastic matching. *Computer Vision, Graphics and Image Processing*, 46:1–21, 1989.

[13] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Trans Med Imaging*, Feb 2019.

[14] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John V. Guttag, and Adrian V. Dalca. An unsupervised learning model for deformable medical image registration. *CoRR*, abs/1802.02604, 2018. URL http://arxiv.org/abs/1802.02604.

[15] Isaac N. Bankman. *Handbook of medical image processing and analysis*. Elsevier - Academic Press, 2009. URL https://books.google.com/books?id=AnRPBKb7qHUC.

[16] Faisal Mirza Beg, I. Michael Miller, Alain Trouve, and Laurent Younes. Computing large deformation metric mappings via geodesic flows. *International Journal of Computer Vision*, 2004.

[17] Nikhil Bhagwat, Joseph D. Viviano, Aristotle N. Voineskos, M. Mallar Chakravarty, and Alzheimer's Disease Neuroimaging Initiative. Modeling and prediction of clinical symptom trajectories in alzheimer's disease using longitudinal data. *PLOS Computational Biology*, 14(9):1–25, 09 2018. doi: 10.1371/journal.pcbi.1006376. URL https://doi.org/10.1371/journal.pcbi.1006376.

[18] Alexandre Bône, Maxime Louis, Olivier Colliot, and Stanley Durrleman. Learning low-dimensional representations of shape data sets with diffeomorphic autoencoders. In Albert C. S. Chung, James C. Gee, Paul A. Yushkevich, and Siqi Bao, editors, *Information Processing in Medical Imaging*, pages 195–207, Cham, 2019. Springer International Publishing. ISBN 978-3-030-20351-1.

[19] Hamid Reza Boveiri, Raouf Khayami, Reza Javidan, and Ali Reza MehdiZadeh. Medical image registration using deep neural networks: A comprehensive review, 2020.

[20] Can Ceritoglu, Kenichi Oishi, Xin Li, Ming-Chung Chou, Laurent Younes, Marilyn Albert, Constantine Lyketsos, Peter C.M. van Zijl, Michael I. Miller, Susumu Mori, and et al. Multi-contrast large deformation diffeomorphic metric mapping for diffusion tensor imaging, May 2009. URL https://www.sciencedirect.com/science/article/pii/S1053811909004194.

[21] François Chollet et al. Keras. https://keras.io, 2015.

[22] Claire Cury, Stanley Durrleman, David M. Cash, Marco Lorenzi, Jennifer M. Nicholas, Martina Bocchetta, John C. van Swieten, Barbara Borroni, Daniela Galimberti, Mario Masellis, and et al. Spatiotemporal analysis for detection of pre-symptomatic shape changes in neurodegenerative diseases: Initial application to the genfi cohort, Dec 2018. URL https://www.sciencedirect.com/science/article/pii/S105381191832144X.

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

[24] Adrian V Dalca, Andreea Bobu, Natalia S Rost, and Polina Golland. Patch-based discrete registration of clinical brain images. *Patch Based Tech Med Imaging*, 9993: 60–67, 2016 Oct 2016.

[25] Adriyana Danudibroto, Jørn Bersvendsen, Olivier Gérard, Oana Mirea, Jan D'hooge, and Eigil Samset. Spatiotemporal registration of multiple three-dimensional echocardiographic recordings for enhanced field of view imaging. *Journal of Medical Imaging*, 3(3):1 – 10, 2016. doi: 10.1117/1.JMI.3.3.037001. URL https://doi.org/10.1117/1.JMI.3.3.037001.

[26] Christos Davatzikos. Spatial transformation and registration of brain images using elastically deformable models. *Comput. Vis. Image Underst.*, 66(2):207–222, 1997. doi: 10.1006/cviu.1997.0605. URL https://doi.org/10.1006/cviu.1997.0605.

[27] C De Boor. *A Practical Guide to Splines.* Springer, 1978.

[28] Emily L Dennis, Xue Hua, Julio Villalon-Reina, Lisa M Moran, Claudia Kernan, Talin Babikian, Richard Mink, Christopher Babbitt, Jeffrey Johnson, Christopher C Giza, and et al. Tensor-based morphometry reveals volumetric deficits in moderate=severe pediatric traumatic brain injury, May 2016. URL https://www.ncbi.nlm.nih.gov/pubmed/26393494.

[29] John A Detre. *Magnetic Resonance Imaging*, page 793–800. Academic Press, 2007.

[30] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. ISSN 00129658, 19399170. URL http://www.jstor.org/stable/1932409.

[31] S. Durrleman, X. Pennec, A. Trouve, J. Braga, G. Gerig, and N. Ayache. Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. *Int J Comput Vis*, 103(1):22–59, May 2013.

[32] Carissa Fonseca, Michael Backhaus, Jae Chung, Wenchao Tao, Pau Medrano-Gracia, Brett Cowan, Peter Hunter, J. Finn, Kalyanam Shivkumar, Abel Lima, David Bluemke, Alan Kadish, Daniel Lee, and Alistair Young. The cardiac atlas project: Rationale, design and procedures. volume 6364, pages 36–45, 09 2010. doi: 10.1007/978-3-642-15835-3_4.

[33] E C Ford, G S Mageras, E Yorke, and C C Ling. Respiration-correlated spiral ct: a method of measuring respiratory-induced anatomic motion for radiation treatment planning, Jan 2003. URL https://www.ncbi.nlm.nih.gov/pubmed/12557983.

[34] G. Gerig, J. Fishbaugh, and N. Sadeghi. Longitudinal modeling of appearance and shape and its potential for clinical use. *Med Image Anal*, 33:114–121, 10 2016.

[35] Ben Glocker, Nikos Komodakis, Georgios Tziritas, Nassir Navab, and Nikos Paragios. Dense image registration through mrfs and efficient linear programming. *Medical Image Anal.*, 12(6):731–741, 2008. URL http://dblp.uni-trier.de/db/journals/mia/mia12.html#GlockerKTNP08.

[36] Suicheng Gu, Xin Meng, C Frank Sciurba, Hongxia Ma, Joseph Leader, Naftali Kaminski, David Gur, and Jiantao Pu. Bidirectional elastic image registration using b-spline affine transformation. *Comput. Medical Imaging Graph.*, pages 306–314, 2014.

[37] Said Hajouli and Dipesh Ludhwani. Heart failure and ejection fraction, Jan 2020. URL https://www.ncbi.nlm.nih.gov/books/NBK553115/.

[38] Nazanin Sadat Hashemi, Roya Babaie Aghdam, Atieh Sadat Bayat Ghiasi, and Parastoo Fatemi. Template matching advances and applications in image analysis. *CoRR*, abs/1610.07231, 2016. URL http://arxiv.org/abs/1610.07231.

[39] Rolf A Heckemann, Joseph V Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. Automatic anatomical brain mri segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126, 2006.

[40] Mattias P. Heinrich. Closing the gap between deep and conventional image registration using probabilistic dense displacement networks, 2019.

[41] Monica Hernandez, Matias N Bossa, and Salvador Olmos. Registration of anatomical images using paths of diffeomorphisms parameterized with stationary vector field flows. *International Journal of Computer Vision*, 85(3):291–306, 2009.

[42] Sarang C Joshi and Michael I Miller. Landmark matching via large deformation diffeomorphisms. *IEEE transactions on image processing*, 9(8):1357–1370, 2000.

[43] Beant Kaur, Amandeep Kaur, and Gurmeet Kaur. Applications of image registration. 2016.

[44] Fahmi Khalifa, Garth Beache, Georgy Gimel'farb, Jasjit Suri, and Ayman El-Baz. *State-of-the-Art Medical Image Registration Methodologies: A Survey*, volume 1, pages 235–280. 05 2011. doi: 10.1007/978-1-4419-8195-0_9.

[45] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[46] Julian Krebs, Tommaso Mansi, Nicholas Ayache, and Hervé Delingette. Probabilistic motion modeling from medical image sequences: Application to cardiac cine-mri, 2019.

[47] Manuel Lang, Oliver Wang, Tunc Aydin, Aljosa Smolic, and Markus Gross. Practical temporal consistency for image-based graphics applications. *ACM Transactions on Graphics - TOG*, 31, 07 2012. doi: 10.1145/2185520.2185530.

[48] Daniel C Lee. Determine. URL http://www.cardiacatlas.org/studies/determine/.

[49] Ruizhi Liao, Esra A. Turk, Miaomiao Zhang, Jie Luo, Elfar Adalsteinsson, P. Ellen Grant, and Polina Golland. Temporal registration in application to in-utero mri time series, 2019.

[50] Jianwei Lin, Yuanjie Zheng, Wanzhen Jiao, Bojun Zhao, Shaoting Zhang, James Gee, and Rui Xiao. Groupwise registration of sequential images from multispectral imaging (msi) of the retina and choroid. *Opt. Express*, 24(22):25277–25290, Oct 2016. doi: 10.1364/OE.24.025277. URL http://www.opticsexpress.org/abstract.cfm?URI=oe-24-22-25277.

[51] Ma, Li, Jing, Bin, Liu, Li, Dan, Li, and Haiyun. Identify the atrophy of alzheimer's disease, mild cognitive impairment and normal aging using morphometric mri analysis, Oct 2016. URL https://www.frontiersin.org/articles/10.3389/fnagi.2016.00243/full.

[52] N. Emily Manning, K. Kelvin Leung, M. Jennifer Nicholas, B. Ian Malone, Jorge M. Cardoso, M. Jonathan Schott, C. Nick Fox, and Josephine Barnes. A comparison of accelerated and non-accelerated mri scans for brain volume and boundary shift

integral measures of volume change: Evidence from the adni dataset. *Neuroinformatics*, pages 215–226, 2017.

[53] Domnic Marera. Ethical consideration when using x-ray examination for none medical purposes. *International Journal of Science and Research*, 4 (5):1528–1531, May 2015. URL https://pdfs.semanticscholar.org/ff8d/f784c0a8c43d83e8cba5b9950dca73e4dca1.pdf.

[54] Michael I. Miller, M. Faisal Beg, Can Ceritoglu, and Craig Stark. Increasing the power of functional maps of the medial temporal lobe by using large deformation diffeomorphic metric mapping. *Proceedings of the National Academy of Sciences*, 102(27):9685–9690, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0503892102. URL https://www.pnas.org/content/102/27/9685.

[55] Pawel Mlynarski, Hervé Delingette, Hamza Alghamdi, Pierre-Yves Bondiau, and Nicholas Ayache. Anatomically consistent CNN-based segmentation of organs-at-risk in cranial radiotherapy. *Journal of Medical Imaging*, 7(1):1 – 21, 2020. doi: 10.1117/1.JMI.7.1.014502. URL https://doi.org/10.1117/1.JMI.7.1.014502.

[56] Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98(3):278–284, 2010.

[57] Marc Modat, David M. Cash, Pankaj Daga, Gavin P. Winston, John S. Duncan, and Sébastien Ourselin. Global image registration using a symmetric block-matching approach. *Journal of Medical Imaging*, 1(2):1 – 6, 2014. doi: 10.1117/1.JMI.1.2.024003. URL https://doi.org/10.1117/1.JMI.1.2.024003.

[58] J. K. Molitoris, T. Diwanji, J. W. Snider, S. Mossahebi, S. Samanta, S. N. Badiyan, C. B. Simone, and P. Mohindra. Advances in the use of motion management and image guidance in radiation therapy treatment for lung cancer. *J Thorac Dis*, 10 (Suppl 21):S2437–S2450, Aug 2018.

[59] Pinar Muyan-Özçelik, D. John Owens, Junyi Xia, and S. Sanjiv Samant. Fast deformable registration on the gpu: A cuda implementation of demons. *ICCSA Workshops*, pages 223–233, 2008.

[60] Sayan Nag. Image registration techniques: A survey. Nov 2017. doi: 10.31224/osf.io/rv65c. URL http://dx.doi.org/10.31224/osf.io/rv65c.

[61] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020. doi: 10.1038/s41586-020-2145-8. URL https://www.nature.com/articles/s41586-020-2145-8.

[62] Nicholas M. Patrikalakis, Takashi M. Maekawa, and Wonjoon M. Cho. *Shape interrogation for computer aided design and manufacturing.* Springer, 2010. URL http://web.mit.edu/hyperbook/Patrikalakis-Maekawa-Cho/.

[63] Xavier Pennec, Pascal Cachier, and Nicholas Ayache. Understanding the "demon's algorithm": 3d non-rigid registration by gradient descent. In Chris Taylor and Alain Colchester, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI'99*, pages 597–605, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

[64] Eugenio Picano. Economic and biological costs of cardiac imaging. *Cardiovascular Ultrasound*, 3(1), 2005. doi: 10.1186/1476-7120-3-13.

[65] Josien P. W. Pluim, J. B. Antoine Maintz, and Max A. Viergever. Image registration by maximization of combined mutual information and gradient information. In Scott L. Delp, Anthony M. DiGoia, and Branislav Jaramaz, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2000*, pages 452–461, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[66] Marianne Rakic, John Guttag, and Adrian V. Dalca. Anatomical predictions using subject-specific medical data. In *Medical Imaging with Deep Learning*, 2020. URL https://openreview.net/forum?id=dPCZhAHmIl.

[67] Martin Reuter, Nicholas J. Schmansky, Herminia Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4):1402–1418, 2012. doi: 10.1016/j.neuroimage.2012.02.084. URL http://dx.doi.org/10.1016/j.neuroimage.2012.02.084.

[68] Gustavo K Rohde, Alan S Barnett, Peter J Basser, and Carlo Pierpaoli. Estimating intensity variance due to noise in registered images: Applications to diffusion tensor mri, Apr 2005. URL http://imagedatascience.com/rohde-neuroimage-05-final.pdf.

[69] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

[70] D. Rueckert, Luke Sonoda, C. Hayes, Derek Hill, M.O. Leach, and D.J. Hawkes. Nonrigid registration using free-form deformations: Application to breast mr images. *Medical Imaging, IEEE Transactions on*, 18:712 – 721, 09 1999. doi: 10.1109/42.796284.

[71] Daniel Rueckert, Paul Aljabar, Rolf A. Heckemann, Joseph V. Hajnal, and Alexander Hammers. Diffeomorphic registration using b-splines. In Rasmus Larsen,

Mads Nielsen, and Jon Sporring, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, pages 702–709, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-44728-3.

[72] Robin Sandkühler, Christoph Jud, Simon Andermatt, and Philippe C. Cattin. Airlab: Autograd image registration laboratory, 2018.

[73] I. J. Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions. *Quarterly of Applied Mathematics*, 4(1):45–99, Jan 1946. doi: 10.1090/qam/15914.

[74] Siyuan Shan, Xiaoqing Guo, Wen Yan, Eric Chang, Yubo Fan, and Yan Xu. Unsupervised end-to-end learning for deformable medical image registration. 11 2017.

[75] Dinggang Shen and Christos Davatzikos. HAMMER: heirarchical attribute matching mechanism for elastic registration. *IEEE Trans. Med. Imaging*, 21(8), 2002.

[76] Erik Smistad, Thomas Falch, Mohammadmehdi Bozorgi, Anne Elster, and Frank Lindseth. Medical image segmentation on gpus - a comprehensive review. *Medical Image Analysis*, 20:1–18, 02 2015. doi: 10.1016/j.media.2014.10.012.

[77] Rebecca Smith-Bindman, Diana L. Miglioretti, Eric Johnson, Choonsik Lee, Heather Spencer Feigelson, Michael Flynn, Robert T. Greenlee, Randell L. Kruger, Mark C. Hornbrook, Douglas Roblin, Leif I. Solberg, Nicholas Vanneman, Sheila Weinmann, and Andrew E. Williams. Use of Diagnostic Imaging Studies and Associated Radiation Exposure for Patients Enrolled in Large Integrated Health Care Systems, 1996-2010. *JAMA*, 307(22):2400–2409, 06 2012. ISSN 0098-7484. doi: 10.1001/jama.2012.5960. URL https://doi.org/10.1001/jama.2012.5960.

[78] A Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *Medical Imaging, IEEE Transactions on*, 32(7):1153–1190, July 2013. ISSN 0278-0062. doi: 10.1109/TMI.2013.2265603. URL http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6522524.

[79] Jean-Philippe Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical Image Anal.*, 2(3):243–260, 1998. URL http://dblp.uni-trier.de/db/journals/mia/mia2.html#Thirion98.

[80] Arthur Toga and Paul Thompson. The role of image registration in brain mapping. *Image and vision computing*, 19:3–24, 01 2001. doi: 10.1016/S0262-8856(00)00055-X.

[81] Raj Varadhan, Grigorios Karangelis, Karthik Krishnan, and Susanta Hui. A framework for deformable image registration validation in radiotherapy clinical applications. *Journal of applied clinical medical physics / American College of Medical Physics*, 14:4066, 01 2013. doi: 10.1120/jacmp.v14i1.4066.

[82] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1, Supp.1):S61–S72, March 2009. doi: 10.1016/j.neuroimage.2008.10.040. URL http://www.inria.fr/sophia/asclepios/Publications/Tom.Vercauteren/DiffeoDemons-NeuroImage08-Vercauteren.pdf.

[83] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: https://doi.org/10.1038/s41592-019-0686-2.

[84] Jianfeng Wu, Jie Zhang, Jie Shi, Kewei Chen, J. Richard Caselli, M. Eric Reiman, and Yalin Wang. Hippocampus morphometry study on pathology-confirmed alzheimer's disease patients with surface multivariate morphometry statistics. *ISBI*, pages 1555–1559, 2018.

[85] Zhen Xiong and Yun Zhang. A critical review of image registration methods. *International Journal of Image and Data Fusion*, 1(2):137–158, 2010. doi: 10.1080/19479831003802790. URL https://doi.org/10.1080/19479831003802790.

[86] H. Xu and X. Li. Consistent feature-aligned 4D image registration for respiratory motion modeling. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 584–587, 2013.

[87] Yan Cao, M. I. Miller, R. L. Winslow, and L. Younes. Large deformation diffeomorphic metric mapping of vector fields. *IEEE Transactions on Medical Imaging*, 24(9):1216–1230, 2005.

[88] B T Thomas Yeo, Mert R Sabuncu, Tom Vercauteren, Daphne J Holt, Katrin Amunts, Karl Zilles, Polina Golland, and Bruce Fischl. Learning task-optimal registration cost functions for localizing cytoarchitecture and function in the cerebral cortex. *IEEE Trans Med Imaging*, 29(7):1424–41, 2010 Jul 2010. ISSN 1558-254X. doi: 10.1109/TMI.2010.2049497.

[89] Miaomiao Zhang, Ruizhi Liao, Adrian V. Dalca, Esra A. Turk, Jie Luo, P. Ellen Grant, and Polina Golland. Frequency diffeomorphisms for efficient image registration. In Marc Niethammer, Martin Styner, Stephen R. Aylward, Hongtu Zhu, Ipek Oguz, Pew-Thian Yap, and Dinggang Shen, editors, *Information Processing in Medical Imaging - 25th International Conference, IPMI 2017, Boone, NC, USA,*

June 25-30, 2017, Proceedings, volume 10265 of *Lecture Notes in Computer Science*, pages 559–570. Springer, 2017. doi: 10.1007/978-3-319-59050-9_44. URL https://doi.org/10.1007/978-3-319-59050-9_44.

[90] Shengyu Zhao, Yue Dong, Eric Chang, and Yan Xu. Recursive cascaded networks for unsupervised medical image registration. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. doi: 10.1109/iccv.2019.01070. URL http://dx.doi.org/10.1109/ICCV.2019.01070.

[91] Shengyu Zhao, Tingfung Lau, Ji Luo, Eric I-Chao Chang, and Yan Xu. Unsupervised 3d end-to-end medical image registration with volume tweening network, 2019.

[92] Qiao Zheng, Hervé Delingette, Nicolas Duchateau, and Nicholas Ayache. 3d consistent & robust segmentation of cardiac images by deep learning with spatial propagation. *CoRR*, abs/1804.09400, 2018. URL http://arxiv.org/abs/1804.09400.