Unsupervised Learning of Probabilistic Diffeomorphic Registration for Images and Surfaces

Adrian V. Dalca
MIT and MGH
adalca@mit.edu

Guha Balakrishnan
MIT
balakg@mit.edu

John Guttag
MIT
guttag@mit.edu

Mert R. Sabuncu
Cornell University
msabuncu@cornell.edu

Abstract

Classical deformable registration techniques achieve impressive results and offer a rigorous theoretical treatment, but are computationally intensive since they solve an optimization problem for each image pair. Recently, learning-based methods have facilitated fast registration by learning spatial deformation functions. However, these approaches use restricted deformation models, require supervised labels, or do not guarantee a diffeomorphic (topology-preserving) registration. Furthermore, learning-based registration tools have not been derived from a probabilistic framework that can offer uncertainty estimates.

In this paper, we build a connection between classical and learning-based methods. We present a probabilistic generative model and derive an unsupervised learning-based inference algorithm that uses insights from classical registration methods and makes use of recent developments in convolutional neural networks (CNNs). We demonstrate our method on a 3D brain registration task for both images and anatomical surfaces, and provide extensive empirical analyses. Our principled approach results in state of the art accuracy and very fast runtimes, while providing diffeomorphic guarantees. Our implementation is available at `http://voxelmorph.csail.mit.edu`.

Keywords medical image registration · diffeomorphic registration · invertible registration · probabilistic modeling · convolutional neural networks · variational inference · machine learning

1 Introduction

Deformable registration computes a dense correspondence between two images, and is fundamental to many medical image analysis tasks. Classical registration techniques have been rigorously developed and studied, but require computationally intensive optimization for each image pair, often requiring tens of minutes to hours of compute time on a CPU. Recent, learning-based registration methods achieve fast runtimes by building on machine learning developments, but largely omit rigorous theoretical treatment of deformations and topology-preserving guarantees. In this work, we present an approach that builds on the strengths of both paradigms, and overcomes these shortcomings. We provide a rigorous connection between probabilistic generative models for deformations and learning algorithms based on convolutional neural networks (CNNs). We also demonstrate that the learning can be done end-to-end in an unsupervised fashion for this model. The resulting framework provides registration for a new image pair in under a second on a GPU, while maintaining guarantees developed for classical methods.

Our formulation casts registration as variational inference on a probabilistic generative model. This framework naturally results in an algorithm that leverages a collection of images to learn a global convolutional neural network with an intuitive cost function. Importantly, we introduce *diffeomorphic integration* layers combined with a spatial transform layer to enable unsupervised end-to-end learning for diffeomorphic registration. We demonstrate that our algorithm achieves state-of-the-art registration accuracy while providing diffeomorphic deformations and fast runtime, and can estimate of registration uncertainty. In our experiments we focus on the example of registering 3D MR brain scans, using a multi-study dataset of over 3,500 scans. However, the method is broadly applicable to many registration tasks.

This paper extends a preliminary version of this work presented at the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2018 conference [18]. We build on that work by providing theoretical extensions, new results, analysis, and discussion. We first expand the model, including a natural extension to anatomical surfaces. In our

experiments, we add baselines, new experiments on registration of both images and surfaces, and provide an analysis of the effect of our diffeomorphic implementation on field regularity and runtime. We implement our method as part of the registration framework called VoxelMorph, which is available at `http://voxelmorph.csail.mit.edu`.

1.1 Related Works

1.1.1 Classical Registration Methods

Classical methods solve an optimization over the space of deformations [5, 7, 8, 11, 19, 28, 66, 69, 70]. Common representations are displacement vector fields, including elastic-type models [8, 21, 62], free-form deformations with b-splines [61], statistical parametric mapping [6], Demons [56, 66], and more recently discrete methods [19, 30, 28].

Constraining the allowable transformations to diffeomorphisms ensures certain desirable properties, such as preservation of topology. Diffeomorphic transforms have seen extensive methodological development, yielding state-of-the-art tools, such as Large Diffeomorphic Distance Metric Mapping (LDDMM) [11, 14, 15, 32, 37, 49, 55, 70], DARTEL [5], diffeomorphic Demons [67], and symmetric normalization (SyN) [7]. In general, these tools demand substantial time and computational resources for a given image pair.

Some recent GPU-based iterative algorithms use these frameworks to develop faster algorithms by requiring a GPU to be available for each registration [51, 50]. Recent learning-based registration methods have demonstrated that they can provide good initializations to iterative GPU methods [10] to further improve runtime.

Probabilitic image registration methods specify priors on the deformation between two images, and likelihood models that describe image intensities [63, 70, 31, 58, 3]. These formulations also lead to iterative optimization methods, but can yield distributions of deformation fields. In this paper, we build on these models by presenting a general variational inference strategy to optimize a global neural network that efficiently outputs distributions of deformations.

1.1.2 Learning-based Registration

Recent methods have proposed to train neural networks that map a pair of input images to an output deformation. Most earlier approaches demonstrated the feasibility of deep learning based registration, and required ground truth registration fields [13, 42, 59, 64, 68]. Such ground truth deformations are often derived via more conventional registration tools or simulations, sometimes limiting their applicability.

Building on the successful demonstration of these methods, several recent papers [9, 10, 23, 22, 44] explore unsupervised, or end-to-end, strategies. These methods employ a neural network that computes spatial transformation [36] to warp one image to another, enabling end-to-end training. A recent approach builds on these methods by learning a spatially-adaptive regularizer within a registration model [54] These approaches use machine learning techniques to achieve efficient training and fast runtimes, but build on classical registration development, such as probabilistic models and diffeomorphic theory. In our work, we bridge these two paradigms to offer classical guarantees within a machine learning approach. We note the contemporaneous development of a method that uses a conditional variational auto encoder (CVAE) to learn diffeomorphic representations [43, 41]. Similar to our method, this approach uses a variational strategy to learn a network to predict a stationary velocity field (SVF). However, the authors focus on representing the SVF through the manifold of the CVAE, and focus on the anatomical variation captured through this encoding

Recent methods proposed using segmentation-based cost functions, such as Dice [25], to replace the image-based similarity term when segmentations are available during training for multi-modal registration, such as T2w MRI and 3D ultrasound, within the same subject [35, 34]. We extend this line of work by showing that our generative probabilistic model naturally describes the deformation of surfaces, therefore enabling the use of segmentations during training within a single cohesive framework. The extended model results in a combination of segmentation (surface) and image-based training losses.

1.1.3 Surface-based Registration

In this paper, we also present an extension to our main contribution which enables alignment of surfaces. In medical image registration, surface matching methods often use surface coordinates or geometric features extracted from anatomical structures [2, 26, 47, 57]. Several methods treat surfaces as 3D point sets with shape descriptors, and often use Iterated Closest Point based optimization methods to find the shape correspondences [12]. Currents, defined as unconnected oriented points, have been used to register surfaces, for example using Matching Pursuit algorithms [26]. Some methods combine volume and surface registration, often using surface registrations to initialize dense volumetric transforms [57]. Similar to volume registration, these classical surface matching methods use iterative optimization strategies, requiring significant computational resources. Building on these methods, we use a 3D point representation jointly with images to achieve fast registration with neural networks, enabled by a differentiable surface distance function.

1.2 Background: Diffeomorphic Registration

Although the method presented in this paper applies to a multitude of deformable representations, we choose to work with diffeomorphisms, and in particular with a stationary velocity field representation [5]. Diffeomorphic deformations are differentiable and invertible, and thus preserve topology. Let $\phi : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ represent the defor-

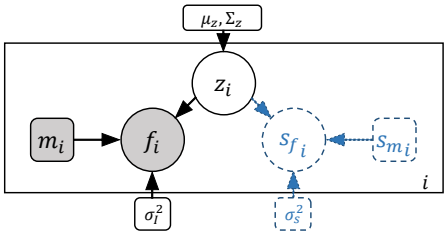

Figure 1: A graphical representation of our generative model. Circles indicate random variables and rounded squares represent parameters. Shaded quantities are observed *at test time*, and the plates indicate replication. $\boldsymbol{f}$ and $\boldsymbol{m}$ are the input images. The image intensities $\boldsymbol{f}$ are generated from a normal distribution centered at $\boldsymbol{m} \circ \boldsymbol{\phi}_z$. The registration prior is defined by normal parameters $\boldsymbol{\mu}_z$, and $\boldsymbol{\Sigma}_z$. In blue, the *optional* similar model structure is included for an anatomical surface, used purely for learning an improved posterior of registration.

mation that maps the coordinates from one image to coordinates in another image. In our implementation, the deformation field is defined through the following ordinary differential equation (ODE):

$$\frac{\partial \boldsymbol{\phi}^{(t)}}{\partial t} = \boldsymbol{v}(\boldsymbol{\phi}^{(t)}) \tag{1}$$

where $\boldsymbol{\phi}^{(0)} = Id$ is the identity transformation and t is time. We integrate the stationary velocity field $\boldsymbol{v}$ over $t = [0, 1]$ to obtain the final registration field $\boldsymbol{\phi}^{(1)}$ [53].

While we implement and evaluate several numerical integration techniques, we find *scaling and squaring* to be most efficient, and we briefly review the technique here [4]. The integration of a stationary ODE represents a one-parameter subgroup of diffeomorphisms. In group theory, $\boldsymbol{v}$ is a member of the Lie algebra and is exponentiated to produce $\boldsymbol{\phi}^{(1)}$, which is a member of the Lie group: $\boldsymbol{\phi}^{(1)} = \exp(\boldsymbol{v})$. From the properties of one-parameter subgroups, for any scalars t and t', $\exp((t + t')\boldsymbol{v}) = \exp(t\boldsymbol{v}) \circ \exp(t'\boldsymbol{v})$, where $\circ$ is a composition map associated with the Lie group. Starting from $\boldsymbol{\phi}^{(1/2^T)} = \boldsymbol{p} + \boldsymbol{v}(\boldsymbol{p})/2^T$ where $\boldsymbol{p}$ is a map of spatial locations, we use the recurrence $\boldsymbol{\phi}^{(1/2^{t-1})} = \boldsymbol{\phi}^{(1/2^t)} \circ \boldsymbol{\phi}^{(1/2^t)}$ to obtain $\boldsymbol{\phi}^{(1)} = \boldsymbol{\phi}^{(1/2)} \circ \boldsymbol{\phi}^{(1/2)}$. T is chosen so that $\boldsymbol{v}/2^T$ is very small.

2 Methods

We let $\boldsymbol{f}$ and $\boldsymbol{m}$ be 3D images, such as MRI volumes, and let $\boldsymbol{z}$ be a latent variable that parametrizes a transformation function $\boldsymbol{\phi}_{\boldsymbol{z}} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$. We propose a generative model that describes the formation of $\boldsymbol{f}$ by warping $\boldsymbol{m}$ via $\boldsymbol{m} \circ \boldsymbol{\phi}_{\boldsymbol{z}}$. We propose a variational inference approach that leverages a convolutional neural network with diffeomorphic integration and spatial transform layers. We learn network parameters in an unsupervised fashion, without access to ground truth registrations. We describe how the network yields fast diffeomorphic registration of a new image pair $(\boldsymbol{f}, \boldsymbol{m})$, in a probabilistic framework. We expand this treatment by including anatomical surface alignment, which enables training the network given (optional) anatomical segmentations.

2.1 Generative Model

We model the prior probability of the parametrization $\boldsymbol{z}$ as:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{\Sigma}_z), \tag{2}$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Our work applies to a wide range of representations $\boldsymbol{z}$. For example, $\boldsymbol{z}$ could be a dense displacement field, or a low-dimensional embedding of the displacement field. In this paper, we let $\boldsymbol{z}$ be a stationary velocity field that specifies a diffeomorphism through the ODE (1). We let $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ be the Laplacian of a neighborhood graph defined on the voxel grid, where $\boldsymbol{D}$ is the graph degree matrix, and $\boldsymbol{A}$ is a voxel neighbourhood adjacency matrix. We encourage *spatial smoothness* of the velocity field $\boldsymbol{z}$ by setting $\boldsymbol{\Sigma}_z^{-1} = \boldsymbol{\Lambda}_z = \lambda \boldsymbol{L}$, where $\boldsymbol{\Lambda}_z$ is a precision matrix and λ denotes a parameter controlling the scale of the velocity field $\boldsymbol{z}$.

We let $\boldsymbol{f}$ be a noisy observation of warped image $\boldsymbol{m}$:

$$p(\boldsymbol{f}|\boldsymbol{z}; \boldsymbol{m}) = \mathcal{N}(\boldsymbol{f}; \boldsymbol{m} \circ \boldsymbol{\phi}_{\boldsymbol{z}}, \sigma_I^2 \mathbb{I}), \tag{3}$$

where σ_I^2 captures the variance of additive image noise.

We aim to estimate the posterior registration probability $p(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})$. Using this, we can obtain the most likely registration field $\boldsymbol{\phi}_{\boldsymbol{z}}$ for a new image pair $(\boldsymbol{f}, \boldsymbol{m})$ via MAP estimation, along with an estimate of velocity field variance at each voxel. Figure 1 provides a graphical representation of our model.

maximum aposteriori

intractable := difficult

2.2 Learning

Given our assumptions, computing the posterior probability $p(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})$ is intractable. We use a variational approach, and introduce an approximate posterior probability $q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})$ parametrized by $\boldsymbol{\psi}$. We minimize the KL divergence

$$\begin{aligned}
&\min_{\boldsymbol{\psi}} \mathrm{KL}\left[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m}) || p(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})\right] \\
&= \min_{\boldsymbol{\psi}} \mathbf{E}_q\left[\log q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m}) - \log p(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})\right] \\
&= \min_{\boldsymbol{\psi}} \mathbf{E}_q\left[\log q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m}) - \log p(\boldsymbol{z}, \boldsymbol{f}; \boldsymbol{m})\right] + \log p(\boldsymbol{f}; \boldsymbol{m}) \\
&= \min_{\boldsymbol{\psi}} \mathrm{KL}\left[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m}) || p(\boldsymbol{z})\right] - \mathbf{E}_q\left[\log p(\boldsymbol{f}|\boldsymbol{z}; \boldsymbol{m})\right] + \text{const},
\end{aligned} \tag{4}$$

which yields the negative of the *variational lower bound* of the model evidence [39]. We model the approximate posterior $q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})$ as a multivariate normal:

$$q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{z|m,f}, \boldsymbol{\Sigma}_{z|m,f}), \tag{5}$$

where we let $\boldsymbol{\Sigma}_{z|m,f}$ be diagonal. To understand the effects of this assumption, we explore a non-diagonal covariance

if q an p are same then additional bits required are zero. By trying to minimize, we make both q and p to be same.

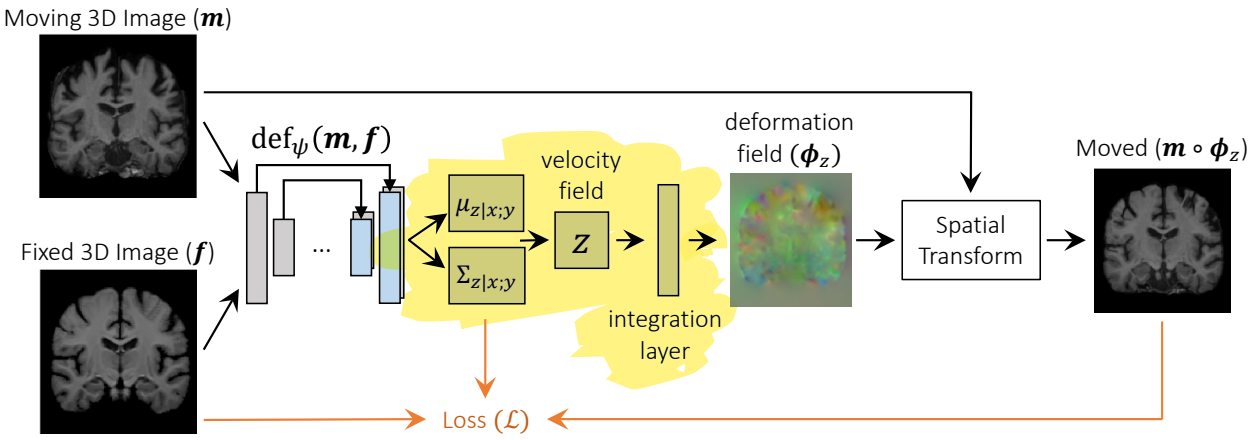

Figure 2: Overview of end-to-end unsupervised architecture. The first part of the network, $\text{def}_\psi(\boldsymbol{m}, \boldsymbol{f})$ takes the input images and outputs the approximate posterior probability parameters representing the velocity field mean, $\boldsymbol{\mu}_{z|m;f}$, and variance, $\boldsymbol{\Sigma}_{z|m;f}$. A velocity field $\boldsymbol{z}$ is sampled and transformed to a diffeomorphic deformation field $\boldsymbol{\phi}_z$ using novel differentiable *squaring and scaling* integration layers. Finally, a spatial transform warps $\boldsymbol{m}$ to obtain $\boldsymbol{m} \circ \boldsymbol{\phi}_z$. Figure 12 expands on this overview by including the optional surface-based loss.

in a later section. The statistics $\boldsymbol{\mu}_{z|m,f}$ and the diagonal of $\boldsymbol{\Sigma}_{z|m,f}$ can be interpreted as the voxel-wise mean and variance, respectively.

We estimate $\boldsymbol{\mu}_{z|m,f}$, and $\boldsymbol{\Sigma}_{z|m,f}$ using a convolutional neural network $\text{def}_\psi(\boldsymbol{f}, \boldsymbol{m})$ parameterized by ψ, as described in the next section. We learn parameters ψ by optimizing the variational lower bound (4) using stochastic gradient methods. Specifically, for each image pair $(\boldsymbol{f}, \boldsymbol{m})$ and sample $\boldsymbol{z}_k \sim q_\psi(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})$, we compute $\boldsymbol{m} \circ \boldsymbol{\phi}_{z_k}$, with the resulting loss (detailed derivation in supplementary material):

$$\begin{aligned}
&\mathcal{L}(\psi; \boldsymbol{f}, \boldsymbol{m}) = -\mathbf{E}_q \left[\log p(\boldsymbol{f}|\boldsymbol{z}; \boldsymbol{m})\right] \\
&+ \text{KL}\left[q_\psi(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})||p(\boldsymbol{z})\right] \\
&= \frac{1}{2\sigma^2 K} \sum_k ||\boldsymbol{f} - \boldsymbol{m} \circ \boldsymbol{\phi}_{z_k}||^2 \\
&+ \frac{1}{2}\left[\text{tr}(\lambda \boldsymbol{D}\boldsymbol{\Sigma}_{z|x;y} - \log \boldsymbol{\Sigma}_{z|x;y}) + \boldsymbol{\mu}_{z|m,f}^T \boldsymbol{\Lambda}_z \boldsymbol{\mu}_{z|m,f}\right] \\
&+ \text{const}, \qquad (6)
\end{aligned}$$

how are you getting this prior?

where K is the number of samples used to approximate the expectation. The first term encourages image $\boldsymbol{f}$ to be similar to the warped image $\boldsymbol{m} \circ \boldsymbol{\phi}_{z_k}$. The second term encourages the posterior to be close to the prior $p(\boldsymbol{z})$. Although the variational covariance $\boldsymbol{\Sigma}_{z|m,f}$ is diagonal, the last term spatially smoothes the mean, which can be seen by expanding $\boldsymbol{\mu}_{z|m,f}^T \boldsymbol{\Lambda}_z \boldsymbol{\mu}_{z|m,f} = \frac{\lambda}{2} \sum \sum_{j \in N(I)} (\boldsymbol{\mu}[i] - \boldsymbol{\mu}[j])^2$, where $N(i)$ are the neighbors of voxel i. We treat σ^2 and λ as fixed hyper-parameters that we investigate in our experiments, and use $K = 1$.

2.3 Neural Network Framework

We design the network $\text{def}_\psi(\boldsymbol{f}, \boldsymbol{m})$ that takes as input $\boldsymbol{f}$ and $\boldsymbol{m}$ and outputs $\boldsymbol{\mu}_{z|m,f}$ and $\boldsymbol{\Sigma}_{z|m,f}$, based on a 3D UNet-style architecture [60]. The network includes a convolutional layer with 32 filters, four downsampling layers with 64 convolutional filters and a stride of two, and three upsampling convolutional layers with 64 filters. We only upsample three times to predict the velocity field (and following integration steps) at every two voxels, to enable these operations to fit in current GPU card memory.

To enable unsupervised learning of parameters ψ using (6), we must form $\boldsymbol{m} \circ \boldsymbol{\phi}_z$ and compute the data term. We first implement a layer that samples a new $\boldsymbol{z}_k \sim \mathcal{N}(\boldsymbol{\mu}_{z|m,f}, \boldsymbol{\Sigma}_{z|m,f})$ using the "re-parameterization trick" [39]: $\boldsymbol{z}_k = \boldsymbol{\mu}_{z|m,f} + \sqrt{\boldsymbol{\Sigma}_{z|m,f}}\boldsymbol{r}$, where $\boldsymbol{r}$ is a sample from the standard normal: $\boldsymbol{r} \sim \mathcal{N}(0, \boldsymbol{I})$.

Given $\boldsymbol{z}_k$, we need to compute $\boldsymbol{\phi}_{z_k} = \exp(\boldsymbol{z}_k)$ as described in the introduction. We propose vector integration layers using *scaling and squaring* operations. Specifically, *scaling and squaring* operations involve compositions within the neural network architecture using a differentiable spatial transformation operation. Given two 3D vector fields $\boldsymbol{a}$ and $\boldsymbol{b}$, for each voxel p this operation computes $(\boldsymbol{a} \circ \boldsymbol{b})(p) = \boldsymbol{a}(\boldsymbol{b}(p))$, a non-integer voxel location $\boldsymbol{b}(\boldsymbol{p})$ in $\boldsymbol{a}$, using linear interpolation. Starting with $\boldsymbol{\phi}^{(1/2^T)} = \boldsymbol{p} + \boldsymbol{z}_k/2^T$, we compute $\boldsymbol{\phi}^{(1/2^{t-1})} = \boldsymbol{\phi}^{(1/2^t)} \circ \boldsymbol{\phi}^{(1/2^t)}$ recursively using these operations T times, leading to $\boldsymbol{\phi}^{(1)} \triangleq \boldsymbol{\phi}_{z_k} = \exp(\boldsymbol{z}_k)$. In our experiments, we extensively analyze the effect of the step size T on the runtime of the network, the accuracy of the registration, and the regularity of the deformation. We also implement vector integration layers using quadrature and ODE solvers, and in the ex-

periments show that these are significantly slower and can require significant memory.

Finally, we warp volume $\boldsymbol{m}$ according to the computed diffeomorphic field $\boldsymbol{\phi}_{z_k}$ using a spatial transform layer.

In summary, the network takes as input images $\boldsymbol{f}$ and $\boldsymbol{m}$, computes statistics $\boldsymbol{\mu}_{z|m,f}$ and $\boldsymbol{\Sigma}_{z|m,f}$, samples a new velocity field $\boldsymbol{z}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, computes a diffeomorphic $\boldsymbol{\phi}_{z_k}$ and warps $\boldsymbol{m}$. Since all the steps are designed to be differentiable, we learn the network parameters using stochastic gradient descent-based methods. This network results in three outputs, $\boldsymbol{\mu}_{z|m,f}$, $\boldsymbol{\Sigma}_{z|m,f}$ and $\boldsymbol{m} \circ \boldsymbol{\phi}_{z_k}$, which are used in the model loss (6). The framework is summarized in Figure 2.

2.4 Registration

Given learned parameters, we approximate registration of a new scan pair $(\boldsymbol{f}, \boldsymbol{m})$ using $\boldsymbol{\phi}_{\hat{z}_k}$. We first obtain the most likely velocity field $\hat{\boldsymbol{z}}_k$ using

$$\hat{\boldsymbol{z}}_k = \arg\max_{\boldsymbol{z}_k} p(\boldsymbol{z}_k|\boldsymbol{f}; \boldsymbol{m}) = \boldsymbol{\mu}_{z|m;f}, \tag{7}$$

by evaluating the neural network $\text{def}_\psi(\boldsymbol{f}, \boldsymbol{m})$. We then compute $\boldsymbol{\phi}_{\hat{z}_k}$ using the *scaling and squaring* based integration, altogether requiring less than a second on a GPU. We highlight that at test time, the diagonal covariance $\boldsymbol{\Sigma}_{z|m,f}$ is not used, however it enables an estimation of the deformation uncertainty. Analysis of uncertainty is beyond the scope of this paper, and is an interesting avenue for future study.

Using a stationary velocity field representation, computing the inverse deformation field $\boldsymbol{\phi}_{\boldsymbol{z}}^{-1}$ can be achieved by integrating the negative of the velocity field: $\boldsymbol{\phi}_{\boldsymbol{z}}^{-1} = \boldsymbol{\phi}_{-\boldsymbol{z}}$, since $\boldsymbol{\phi}_{\boldsymbol{z}} \circ \boldsymbol{\phi}_{-\boldsymbol{z}} = \exp(\boldsymbol{z}) \circ \exp(-\boldsymbol{z}) = \exp(\boldsymbol{z} - \boldsymbol{z}) = Id$ [5, 52]. This enables the computation of both fields inside one efficient network when desired.

2.5 Implementation

We implement our method as part of the VoxelMorph package [9], available online at `http://voxelmorph.csail.mit.edu`, using neuron [20] and Keras [16] with a Tensorflow [1] backend. We use a learning rate of $1e-4$ for the Adam optimizer [38], a batch size of 1 due to memory constraints, and Glorot uniform initialization for the convolution weights. We use a single sample ($K = 1$), which has been shown to lead to useful gradients for optimization while maintaining the memory footprint and implementation complexity low [39]. For large volumes, the number of samples is often constrained by the available GPU memory.

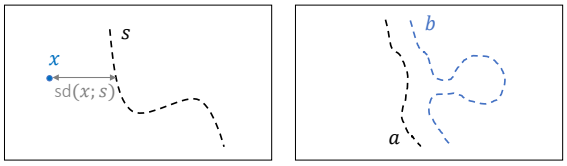

Figure 3: Left: an illustration of the surface distance function $\text{sd}(x; \boldsymbol{s})$. Right: asymmetric surface behavior requires that we compute the surface distance in both directions. For example, computing $\sum_v \text{sd}(\boldsymbol{a}[n], \boldsymbol{b})$ will be considerably smaller than $\sum_v sd(\boldsymbol{b}[n], \boldsymbol{a})$ due to surface points on the hairpin of $\boldsymbol{b}$ (recall that surface points are not directly corresponding.)

3 Method Extensions

3.1 Surface-based Registration

In various instances, anatomical segmentation maps for specific structures of interest may also be available with some of the training images. Recent papers have demonstrated that the use of segmentations can help in registration [10, 34]. Here, we show that our proposed model naturally extends to handle surfaces, enabling the use of segmentations during training within the same principled framework.

We focus on the case where one anatomical structure is segmented in the image. Given a segmentation map where each voxel is assigned the desired anatomical label or background, we extract the anatomical surface and let $\boldsymbol{s_f}$ represent the N surface *coordinates* of the anatomical structure for image $\boldsymbol{f}$, which can be stored as an $N \times 3$ matrix. Given the diffeomorphism $\boldsymbol{\phi}_z$ in the previous section, we model each surface location $\boldsymbol{s}_f[n]$, as formed by displacing a matching surface location $\boldsymbol{s}_m[n]$ according to $\boldsymbol{\phi}_z$, and adding (spatial) displacement noise:

$$p(\boldsymbol{s_f}|\boldsymbol{z}; \boldsymbol{s_m}) = \mathcal{N}(\boldsymbol{s_f}; \boldsymbol{s_m} \circ \boldsymbol{\phi_z}, \sigma_s^2 \mathbb{I}), \tag{8}$$

where the composition $\boldsymbol{s_m} \circ \boldsymbol{\phi_z}$ warps surface coordinates.

Given both images and segmentation maps during training, we extract surfaces of the desired structure and aim to estimate the posterior probability $p(\boldsymbol{z}|\boldsymbol{f}, \boldsymbol{s_f}; \boldsymbol{m}, \boldsymbol{s_m})$. As before, we use a variational approximation. Since segmentation maps, and hence surfaces, are usually derived from images, we assume that images are sufficient to approximate the posterior: $q(\boldsymbol{z}|\boldsymbol{f}, \boldsymbol{s_f}; \boldsymbol{m}, \boldsymbol{s_m}) = q_\psi(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})$. As before, we minimize the KL divergence between the true and approximate posterior (derived in supplementary material):

$$\begin{aligned} &\min_\psi \text{KL}\left[q_\psi(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})||p(\boldsymbol{z}|\boldsymbol{f}, \boldsymbol{s_f}; \boldsymbol{m}, \boldsymbol{s_m})\right] \\ =&\min_\psi \text{KL}\left[q_\psi(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})||p(\boldsymbol{z})\right] - \mathbf{E}_q\left[\log p(\boldsymbol{f}|\boldsymbol{z}; \boldsymbol{m})\right] \\ &- \mathbf{E}_q\left[\log p(\boldsymbol{s_f}|\boldsymbol{z}; \boldsymbol{s_m})\right], \end{aligned} \tag{9}$$

and arrive at the loss function:

$$\begin{aligned}
&\mathcal{L}(\psi; \boldsymbol{f}, \boldsymbol{s_f}, \boldsymbol{m}, \boldsymbol{s_m}) \\
&= \frac{1}{2}\left[\text{tr}(\lambda \boldsymbol{D}\boldsymbol{\Sigma}_{z|x;y} - \log \boldsymbol{\Sigma}_{z|x;y}) + \boldsymbol{\mu}_{z|m,f}^T \boldsymbol{\Lambda}_z \boldsymbol{\mu}_{z|m,f}\right] \\
&+ \frac{1}{2\sigma_I^2 K}\sum_k ||\boldsymbol{f} - \boldsymbol{m} \circ \boldsymbol{\phi}_{z_k}||^2 \\
&+ \frac{1}{2\sigma_s^2 K}\sum_k ||\boldsymbol{s_f} - \boldsymbol{s_m} \circ \boldsymbol{\phi}_{z_k}||^2. \qquad (10)
\end{aligned}$$

Compared to the original model loss (6), the additional third term encourages the deformation $\boldsymbol{\phi}_{z_k}$ to warp the moving surface close to the fixed surface $\boldsymbol{s}_f$. As described in the generative model (8), this requires *corresponding* surface points in $\boldsymbol{s_f}$ and $\boldsymbol{s_m}$. However, these correspondences are not available in practice, as segmentations are provided independently for each image. Therefore, the third term cannot be computed directly.

We propose an approximation of the surface term using surface *distance transforms*. Let $\text{sd}(x, \boldsymbol{s})$ be a *surface distance* function, which for location x returns the Euclidean distance to the closest surface point in $\boldsymbol{s}$ (Figure 3Left).[1] Noting that for two surfaces $\boldsymbol{a}$ and $\boldsymbol{b}$, $\sum_n \text{sd}(\boldsymbol{a}[n], \boldsymbol{b}) \neq \sum_n \text{sd}(\boldsymbol{b}[n], \boldsymbol{a})$ due to potential asymmetries in the surfaces (see Figure 3Right), we approximate the distance $\|\boldsymbol{s_f} - \boldsymbol{s_m} \circ \boldsymbol{\phi}_{z_k}\|^2$ by computing $\text{sd}(\cdot, \cdot)$ in both directions:

$$\begin{aligned}
&\|\boldsymbol{s_f} - \boldsymbol{s_m} \circ \boldsymbol{\phi}_{z_k}\|^2 \\
&\approx \frac{1}{2}\sum_n \text{sd}(\boldsymbol{s_f}[n] \circ \boldsymbol{\phi}_z^{-1}, \boldsymbol{s_m}) + \sum_n \text{sd}(\boldsymbol{s_m}[n] \circ \boldsymbol{\phi}_z, \boldsymbol{s_f}). \qquad (11)
\end{aligned}$$

We implement this function efficiently using distance transforms. Specifically, to compute $\text{sd}(\boldsymbol{s_m}[n] \circ \boldsymbol{\phi}_z, \boldsymbol{s_f})$, we first pre-compute distance transforms for the (fixed) given structure $\boldsymbol{s}_f$. We then sample $100,000$ points along $\boldsymbol{s}_m$, which we find to be sufficient to estimate accurate measures along the surface. We warp (move) them according to the deformation $\boldsymbol{\phi}_z$, and compute the distance transform of $\boldsymbol{s}_f$ at these locations. We take advantage of our diffeomorphic representation that enables computing the inverse $\boldsymbol{\phi}_z^{-1}$ efficiently within the network to similarly compute $\text{sd}(\boldsymbol{s_f}[n] \circ \boldsymbol{\phi}_z^{-1}, \boldsymbol{s_m})$.

In summary, since to compute the posterior approximation $q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})$ the neural network takes as input only the images $\boldsymbol{f}$ and $\boldsymbol{m}$, images alone are required at test time. Given a diffeomorphism $\boldsymbol{\phi}_{z_k}$, at training time the network uses both a warped image and a warped surface to evaluate the quality of the registration.

This model can also be used to register two surfaces when the images themselves are not available. The only modelling change required is removing the image likelihood terms and using the variational approximation $q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{S_f}; \boldsymbol{S_m})$, which uses the segmentation maps $\boldsymbol{S}_m$ and $\boldsymbol{S}_f$ as input. Surface-only registration is beyond the scope of this paper, and we leave it for future work. However, registration with images and surfaces is described here as an example of possible extensions of the model, and surface-*only* registration is beyond the scope of this paper.

The complete neural network framework, including the surface loss, is illustrated in supplemental Figure 12.

3.2 Non-diagonal Covariance

Approximating the velocity field covariance $\boldsymbol{\Sigma}_{z|m,f}$ using a diagonal matrix is a strong assumption that ignores spatial smoothness. As seen in (6), the spatially-smooth prior $p(\boldsymbol{z})$ encourages a smooth mean velocity field $\boldsymbol{\mu}_{z|m,f}$, but samples $\boldsymbol{z}_k \sim \mathcal{N}(\boldsymbol{\mu}_{z|m,f}, \boldsymbol{\Sigma}_{z|m,f})$ might still be noisy. In this section, we investigate the effects of this restriction, by providing a model expansion that computes a less restrictive covariance. In our experiments below, we analyze the effects of these different approximations.

To evaluate the effects of the diagonal covariance, we explore a second approximation $\boldsymbol{\Sigma}_{z|m,f} = \boldsymbol{C}_{\sigma_c}\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{C}_{\sigma_c}^T$ where $\boldsymbol{G}$ is a diagonal matrix returned by the neural network and $\boldsymbol{C}_{\sigma_c}$ is a fixed smoothing convolution matrix. Specifically, for each row of $\boldsymbol{C}_{\sigma_c}$ we create a flattened Gaussian smoothing kernel centered at a particular voxel, such that $\boldsymbol{C}_{\sigma_c}\boldsymbol{w}$ is equivalent to 3D convolution of image $\boldsymbol{w}$ by a gaussian filter with variance σ_c^2. We choose σ_c such that the smoothing operation matches the scale of the prior $p(\boldsymbol{z})$ determined by λ: $\frac{1}{\sqrt{2\pi\sigma_c^{3/2}}} = (\lambda * 6)^{-1}$.

During training, sampling from the posterior is achieved using the *reparametrization trick*: $\boldsymbol{z}_k = \boldsymbol{\mu}_{z|m,f} + \boldsymbol{C}_{\sigma_c}\boldsymbol{D}\boldsymbol{r}$, where $\boldsymbol{r}$ is a sample from the standard normal. Intuitively, compared to the diagonal $\boldsymbol{\Sigma}_{z|m,f}$ approximation, this sampling procedure smoothes the term $\boldsymbol{D}\boldsymbol{r}$ before adding the mean $\boldsymbol{\mu}_{z|m,f}$.

In our experiments, we show that this approximation yields smoother velocity fields during training, and the effect diminishes with higher λ values. However, the resulting deformation fields are diffeomorphic and accurate for *both* approximations, demonstrating that the diagonal covariance approximation is sufficient when working with diffeomorphisms.

4 Experiments

We perform a series of experiments demonstrating that the proposed probabilistic image registration framework achieves accuracy and runtime comparable to state-of-the-art methods while enabling diffeomorphic deformations. We also show the improvements enabled by the extended surface model, and analyze the effect of the various integration layers during test time.

[1] Function $\text{sd}(x, \boldsymbol{s})$ is a generating function for a distance transform image for the surface $\boldsymbol{s}$, by evaluating it at every grid point x

Method	Avg. Dice	GPU sec	CPU sec	mean $\lvert J_\Phi \rvert$	$\lvert J_\Phi \rvert \leq 0$
Affine only	0.584 (0.157)	0	0	1	0
ANTs (SyN)	0.749 (0.136)	-	9059 (2023)	1.001 (0.036)	7523 (4790)
NiftyReg (CC)	0.755 (0.143)	-	2347 (202)	1.072 (0.131)	33838 (8307)
VoxelMorph (CC)	0.753 (0.145)	0.45 (0.01)	57 (1.0)	1.032 (0.074)	19715 (3540)
Supervised-diff	0.730 (0.144)	0.35 (0.03)	82.6 (3.8)	1.088 (0.121)	0.05 (0.5)
`VoxelMorph-diff`	0.754 (0.139)	0.47 (0.01)	84.2 (0.1)	1.075 (0.124)	0.2 (1.0)

Table 1: Summary of results: mean Dice scores over all anatomical structures and subjects (higher is better), mean runtime; mean Jacobian determinant; and mean number of locations with a non-positive Jacobian determinants of each registration field (lower is better). All methods have comparable Dice scores, while our method and the other VoxelMorph variants are orders of magnitude faster than ANTs or NiftiReg. Only our presented method, VoxelMorph-diff, achieves both high accuracy and fast runtime while also having nearly zero non-negative Jacobian locations. All methods have mean Jacobian determinants close to 1, indicating smooth deformations. Each aspect of these results is studied in more details in the rest of the experiments and figures.

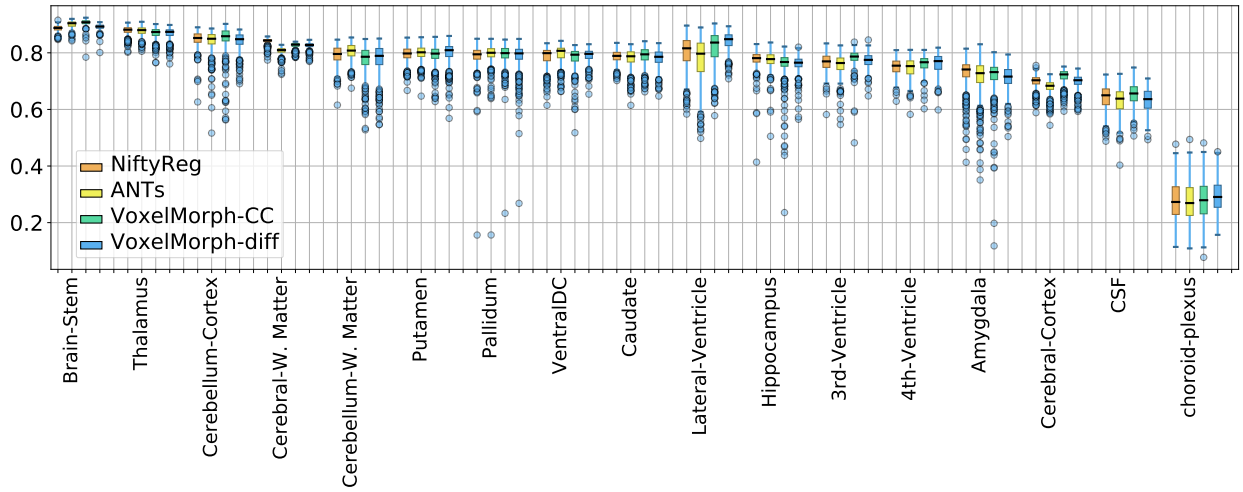

Figure 4: Boxplots indicating Dice scores for anatomical structures for baselines ANTs, NiftiReg, VoxelMorph (CC), and finally our algorithm VoxelMorph-diff. Left and right hemisphere structures are merged for visualization, and ordered by average ANTs Dice score. In general, all four algorithms demonstrate comparable results, each performing slightly better in some structures and slightly worse in others.

We focus on atlas-based registration, a common task in population analysis. Specifically, we register each scan to an atlas computed using external data [27, 65]. Because we implement our algorithm as part of the VoxelMorph framework, we will refer to it as `VoxelMorph-diff`.

4.1 Experiment setup

4.1.1 Data and Preprocessing

We use a large-scale, multi-site dataset of 3731 T1-weighted brain MRI scans from eight publicly available datasets: OASIS [45], ABIDE [24], ADHD200 [48], MCIC [29], PPMI [46], HABS [17], and Harvard GSP [33]. Acquisition details, subject age ranges and health conditions are different for each dataset. We performed standard pre-processing steps on all scans, including resampling to 1mm isotropic voxels, affine spatial normalization and brain extraction for each scan using FreeSurfer [27]. We crop the final images to $160 \times 192 \times 224$. Segmentation

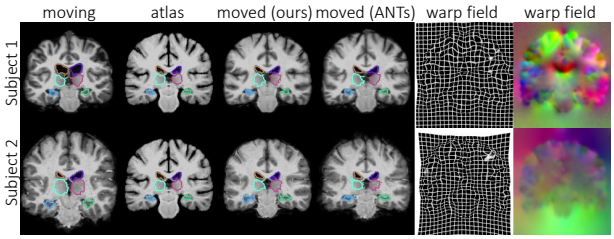

Figure 5: Example MR slices of input moving image, atlas, and resulting warped image for our method and ANTs, with overlaid boundaries of ventricles, thalami and hippocampi. Our resulting registration field is shown as a warped grid and RGB image, with the channels representing the x, y and z dimensions. We omit VoxelMorph (CC) and NiftyReg examples, which are visually similar to our results and ANTs. More examples are provided in the supplementary material Figure 13.

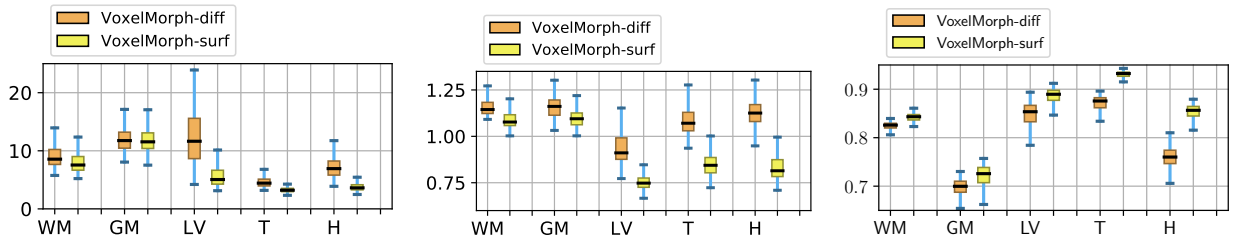

Figure 6: Surface results for the proposed VoxelMorph models. Left: maximum Euclidean surface distance (lower is better). Middle: median Euclidean surface distance (lower is better). Right: mean Dice (higher is better). VoxelMorph-surf trained with surfaces of the desired structures achieves significantly smaller surface distances and larger Dice scores on each structure. We use left hemisphere white matter (WM), gray matter (GM), lateral ventricle (LV), Thalamus (T), and hippocampus (H).

maps including 29 anatomical structures, obtained using FreeSurfer for each scan, are used in evaluating registration results. Each image contains roughly ~ 1.6 million brain voxels. We split the dataset into 3231, 250, and 250 volumes for train, validation, and test sets respectively, although we underscore that the training is unsupervised.

4.1.2 Evaluation Metrics

To evaluate a registration algorithm, we register each subject to an atlas, propagate the segmentation map using the resulting warp, and measure volume overlap using the Dice metric. For the surface experiments, we also employ the Euclidean surface distance, computed using the strategy described in (11).

We also evaluate the diffeomorphic property, a focus of our work. Specifically, the Jacobian matrix $J_\phi(p) = \nabla\boldsymbol{\phi}(p) \in \mathcal{R}^{3\times3}$ captures the local properties of $\boldsymbol{\phi}$ around voxel p. The local deformation is diffeomorphic, both invertible and orientation-preserving, only at locations for which $|J_\phi(p)| > 0$, where $|\cdot|$ is the determinant operator [5]. We count all other (folding) voxels, where $|J_\phi(p)| \leq 0$.

4.1.3 Baseline Methods

We compare our approach with the popular ANTs software package using Symmetric Normalization (SyN) [7], a top-performing algorithm [40]. We found that the default ANTs settings are sub-optimal for our task, and performed a wide parameter and similarity metric search across several datasets. We used the default geodesic implementation of SyN, which is most faithful to theoretical diffeomorphic development. Other versions, such as greedy SyN, would yield a slightly faster runtime, while giving less diffeomorphic deformations. We identified and use top performing parameter values for the Dice metric using: the cross-correlation (CC) loss function, SyN step size of 0.25, Gaussian smoothing of (9, 0.2) and three scales of 201 iterations. We also test the NiftyReg package, for which we use a multi-threaded CPU implementation as a GPU implementation is not currently available.[2] We experimented with different parameter settings for improved behavior, and used the following setting: CC cost function, grid spacing of 5, and 500 iterations.

To compare with recent learning-based registration approaches, we also test our recent CNN-based method, VoxelMorph, which produces state-of-the-art fast and accurate registration, but does not yield diffeomorphic results [9, 10]. We sweep the regularization parameter using our validation set, and use the optimal regularization parameter of 1 in our results.

We also compute a supervised baseline by training a VoxelMorph-diff network using ground truth deformations. We build a ground truth dataset by registering over 650 atlas-MRI subject training pairs using NiftyReg with the described settings. We then train a neural network to predict the resulting deformation fields using a mean squared error (MSE) loss. We explored several variants, and found that doubling the model capacity by doubling the number of features at each layer, as well as penalizing the deformations fields only within the proximity of the atlas brain, yielded optimal results. To enable direct comparison, we used the `VoxelMorph-diff` architecture, but without sampling of the velocity field.

4.2 Image Registration

Table 1 provides a summary of the results on the held-out test set. Figure 5 and supplementary material Figure 13 show representative results. Figure 4 illustrates Dice results on several anatomical structures. For better visualization, we combine the same structures from the two hemispheres, such as the left and right hippocampus. Our algorithm, `VoxelMorph-diff`, achieve state-of-the-art Dice results and runtimes, but produces diffeomorphic registration fields (nearly no folding voxels per scan) in a probabilistic framework.

All methods achieve comparable Dice results on each structure and overall, except the supervised method. Despite

[2] We compiled the latest source code from March, 2018 (tree [4e4525]).

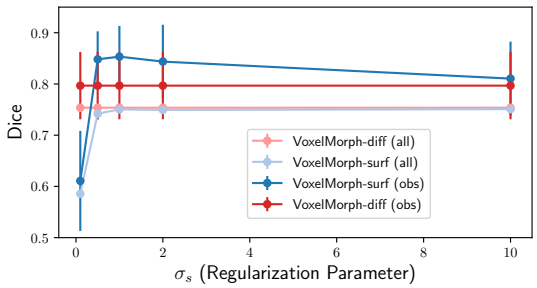

Figure 7: Average Dice score for VoxelMorph-surf models on the validation set. We test various values of the spatial noise parameter σ_s, for both the desired structures observed during training (obs) and all structures (all). For a range of values of $\sigma_s \in [0.5, 2.0]$, we find significant increases for observed surfaces when using the generative surface model.

training the latter on 650 subjects, we found that the supervised network leads to more diffeomorphic deformations than the training deformations, but results in a slight loss in Dice score. Learning-based methods require a fraction of the baseline runtimes to register two images: less than a second on a GPU, and less than a minute and a half on a CPU. Runtimes were computed for an NVIDIA TitanX GPU and a Intel Xeon (E5-2680) CPU, and exclude preprocessing common to all methods.

Our method outputs positive Jacobians at nearly all voxels, which we analyze in more detail in a later section.For `VoxelMorph-diff`, we find that for most scans, the deformation fields result in zero folding voxels. Very few volumes lead to a few or tens of grouped folding voxels, leading to a population average of less than a folding voxel per test scan. In contrast, the deformation fields resulting from the baseline methods contain a few thousand locations of non-positive Jacobians for each scan (Table 1), usually grouped in clusters. This may be alleviated with increased spatial regularization or more optimization iterations, but this in turn leads to a drop in performance on the Dice metric or even longer runtimes. The table also shows that, at the presented settings, all methods result in an average Jacobian determinant close to 1, with `VoxelMorph-diff` yielding smoothness statistics nearly identical to those given by NiftyReg, indicate smooth deformations.

4.3 Image and Surface Registration

In this section, we evaluate the generative surface model. We demonstrate the use of anatomical segmentation alongside images during training, and refer to this model as `VoxelMorph-Surf`. We focus on the setting where one structure of interest is available during training, and learn separate networks for the left white matter, gray matter, ventricle, thalamus and hippocampus. Our goal is to analyze how the additional surface model terms affect the accuracy and regularity of resulting deformations.

Method	$\lvert J\rvert \leq 0$	% of $\lvert J\rvert \leq 0$
ANTs SyN (CC)	9060 (4445)	0.545 (0.267)
NiftyReg (CC)	40425 (9901)	2.431 (0.595)
VoxelMorph (CC)	19077 (5928)	1.147 (0.360)
VoxelMorph-diff	0.1 (1.2)	6.1e-6 (7.6e-5)
VoxelMorph-surf (cer.w.m.)	3.0 (6.2)	1.8e-4 (3.8e-4)
VoxelMorph-surf (cer.cor.)	3.4 (6.4)	2.0e-4 (3.9e-4)
VoxelMorph-surf (lat.ven.)	4.0 (8.0)	2.3e-4 (4.8e-4)
VoxelMorph-surf (thalamus)	4.3 (8.2)	2.6e-4 (4.9e-4)
VoxelMorph-surf (hip.)	2.7 (5.8)	1.6e-4 (3.5e-4)

Table 2: Regularity measures for image and surface models on the test set. Leveraging diffeomorphic aspect of our joint image and surface model, VoxelMorph-surf preserved very low numbers of folding voxels even when training with example surfaces.

Figure 7 illustrates the behaviour of the model with respect to the hyper-parameter σ_s on the validation set. For a range of σ_s values, we find a significant improvement in terms of Dice for the desired structure. For very small values of σ_s, the training becomes unstable leading to poor generalization. A very large σ_s value leads to the model ignoring the surface term. Since the Dice scores are comparable in the range $\sigma_c \in [0.5, 2]$, for the rest of this section we use $\sigma_s = 2$, which exhibits slightly fewer folding voxels (≤ 5 compared to ~ 20 for $\sigma_c = 1$).

Figure 6 demonstrates the improvement on the test set in terms of Euclidean surface distance and Dice, compared to the image-only registration model `VoxelMorph-diff`. `VoxelMorph-surf` improves significantly in all measures for most desired structures. Additionally, Table 2 illustrates that with increased accuracy in both metrics, the number of folding voxels in the entire volume increases only very slightly (to an average 3.5 voxels per volume), which remains orders of magnitude fewer than the baseline methods (Table 1). Figure 14 in the supplementary material illustrates example results.

In summary, the principled joint diffeomorphic model enables the use of surfaces during training which dramatically improves registration near a given structure while preserving desired deformation properties. For example, given hippocampus surfaces at training, registration using `VoxelMorph-Surf` improves Dice by ~ 9 points over `VoxelMorph-diff`, improves maximum surface distance by more than three voxels, and preserving diffeomorphisms (less than three folding voxels per scan).

4.4 Analysis

4.4.1 Parameter Analysis

The two main hyper-parameters, smoothing precision λ and image noise σ_I^2, have physical meaning in our generative model. However, they share a single degree of freedom in the loss function. We set $\sigma_I^2 = 0.02$, and vary the precision scale λ between 0.5 and 100. Figure 9 shows average

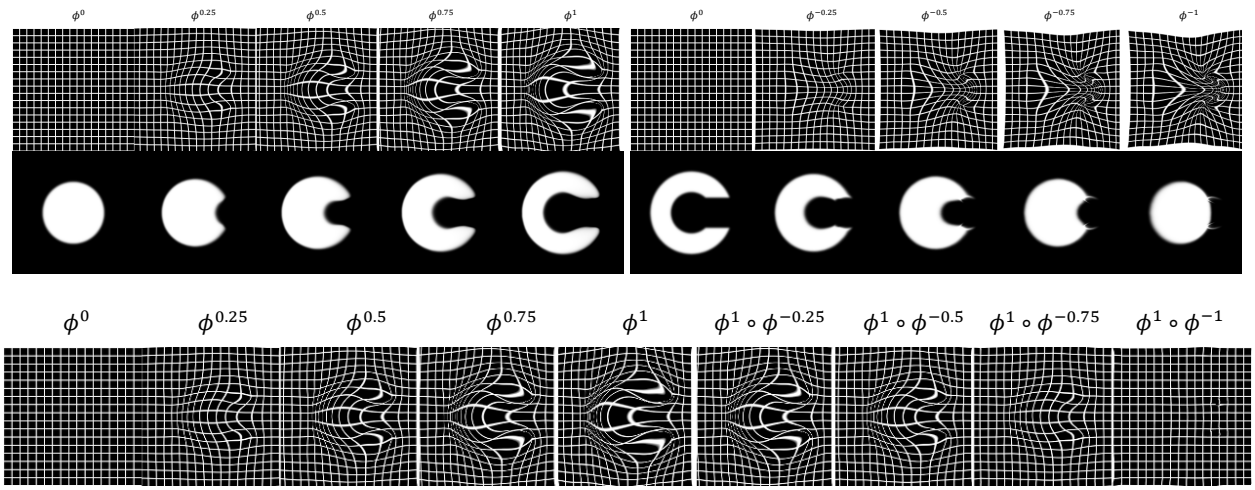

Figure 8: C-shape controlled experiments. We learn to warp disks to Cs of different radii, and illustrate the registration results for one example. The top row illustrates the integration of the velocity field at different time points, and the second row shows the resulting warp of the circle or C. Finally, on the bottom row, we illustrate deforming the grid with a composition of the forward warp and the inverse warp, demonstrating a return to identity.

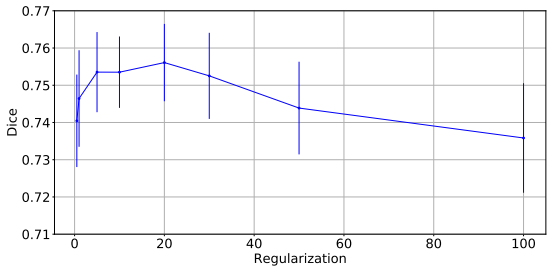

Figure 9: Dice score (computed using 50 validation scans) for VoxelMorph-diff with various values of the precision parameter λ.

Dice scores for 50 validation set scans for different parameter values, showing that the results vary smoothly over a large range, with reasonable behavior even near $\lambda \sim 0$. We use $\lambda = 20$ in our experiments above.

4.4.2 C-shape Registration

We also perform analysis on controlled experiments with C-shape synthetic images with intensities in $[0, 1]$. Specifically, we train a `VoxelMorph-diff` network to learn to register a disk with a radius ranging from one third to one fifth of the image, to a C-shape with variable radius and thickness. The outer radius of the C shape is sampled uniformly in the range $[1/3.5, 1/2.5]$ of the image size, whereas the inner radius is in the range $[1/6.5, 1/5.5]$. We increase hyper-parameter $\sigma_s = 0.06$ to account for the increase in maximum intensity. Figure 8 illustrates representative images and deformation fields. To obtain the fields at intermediate time points between 0 and 1, we employ Tensorflow ODE solver. We find that all the deformation fields lead to accurate registration between disks and C shapes, and have no folding voxels. We also find that the deformation fields are invertible, bringing the grid back to identity when the transforms are composed.

4.4.3 Integration Steps

During training, we hold the number of *scaling and squaring* steps fixed. However, this number can be varied at test time, affecting aspects of the resulting deformation field. In this section, we analyze the effects of the number of steps on accuracy, runtime, field regularity and invertability. We perform this experiment using 50 validation subjects and the image registration network VoxelMorph-diff trained with $T = 7$ integration steps and regularization parameter $\lambda = 20$. The velocity field is computed every two voxels, but all of the conclusions in this section are likely to apply to many reasonable field spacings.

Figure 10 summarizes the analysis results. The runtime increases modestly with the number of steps, and is overall significantly smaller than the cost of the rest of the network (i.e. the deformation network computation of the velocity field, and the spatial transform of the full moving image). After four scaling and squaring steps, the method achieved maximum Dice score. We observe a steep decline in the number of folding voxels (note the log-scale vertical axis), reaching less than five voxels after five scaling and squaring steps, compared to classical methods which can include thousands of such voxels (Table 1). Finally, we measure the average displacement error after inverting the deformation fields: $\Delta \boldsymbol{u} = |Id - \boldsymbol{\phi} \circ \boldsymbol{\phi}^{-1}|$. We find that after five scaling and squaring steps, even the *worst* error is under a half voxel, indicating that five steps are sufficient to ensure invertible deformations.

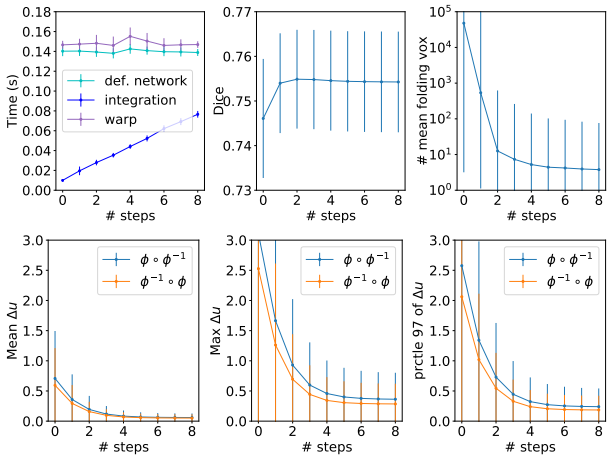

Figure 10: The effect of different number of scaling and squaring steps on the registration accuracy, runtime, deformation regularity and invertability. We find that after five scaling and squaring steps, our model, VoxelMorph-diff, is able to produce state-of-the-art accuracy while having essentially no folding (note the log-scale vertical axis in the top-right graph). Similarly, it is able to produce invertible deformations, as seen by the measure of the displacement error $\Delta \boldsymbol{u} = |I - \phi \circ \phi^{-1}|$. The total runtime cost of the scaling and squaring operations is below the runtime of the rest of the networks, indicating that increasing the number of steps improves deformation properties for trivial runtime cost.

In addition, we implemented the integration of the velocity field using Tensorflow ODE solvers and using standard quadrature, but found that these required significantly more runtime compared to the scaling and squaring strategy, consistent with literature findings [4, 5, 50]. Specifically, while five scaling and squaring operations required 0.06 ± 0.01 seconds, equivalent quadrature integration required 64 operations (occupying prohibitive amounts of memory) and 0.53 ± 0.01 seconds, and ODE-solver based integration with default parameter required a single layer and 2.9 ± 0.1 seconds. At comparable integration settings such as these, all three methods achieve similar Dice scores of 0.75 ± 0.01. While these alternative methods require significant resources, all three implementations are available in our source code for experimentation.

This analysis indicates that the proposed scaling and squaring network integration layer is efficient and accurate. Increasing the number of scaling and squaring layers incurs a negligible runtime cost while improving deformation field properties. We use $T = 7$ squaring steps in the test experiments above.

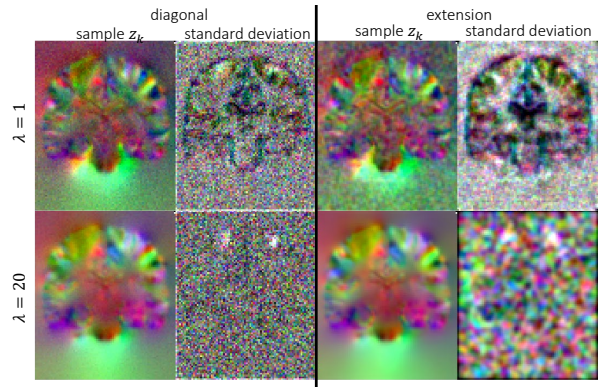

Figure 11: Illustration of voxel independence assumption in variational approximations for two prior parameters $\lambda = 1$ (top) and $\lambda = 20$ (bottom). Each row contains an example sample velocity field $\boldsymbol{z}_k$, and the voxel-wise standard deviation over 500 samples for that subject.

	Dice		$\lvert J\rvert \leq 0$	
λ	diagonal	extension	diagonal	extension
1	0.74 (0.01)	0.74 (0.01)	2934 (2007)	2720 (1593)
20	0.75 (0.01)	0.75 (0.01)	0.32 (0.96)	0.16 (0.57)

Table 3: Accuracy and deformation regularity for the different variational approximations and two dramatically different values for smoothing parameter λ. We find that for a given parameter value, the approximations lead to comparable accuracy and number of folding voxels.

4.4.4 Velocity Sampling and Uncertainty

We also evaluate the modeling assumptions of the variational covariance $\boldsymbol{\Sigma}_{z|m,f}$. Figure 11 illustrates example samples of the velocity field $\boldsymbol{z}_k$ and voxel-wise empirical variance for the two $\boldsymbol{\Sigma}_{z|m,f}$ approximations: diagonal covariance and the extended approximation in Section 3.2 that smooths samples $\boldsymbol{z}_k$. For under-regularized networks (very low values for hyper-parameter λ), the latter approximation yields smoother velocity fields. However, given a higher hyper-parameter λ value, such as the one used in our experiments, the network learns smaller values for the diagonal $\boldsymbol{\Sigma}_{z|m,f}$ approximation, and yields smooth samples $\boldsymbol{z}_k$ with either method. Futhermore, despite the difference in smoothness of the velocity field samples $\boldsymbol{z}_k$, the integration operation leads to equally regular and accurate deformations $\boldsymbol{\phi}_{z_k}$ for a given λ (Table 3).

Therefore, although the diagonal covariance has the potential to add noise to velocity field samples, the loss function coupled with the integration operation lead to smooth and accurate deformation fields $\boldsymbol{\phi}_z$ at reasonable λ values. Therefore, in the current setting, the diagonal and non-diagonal covariances give similar results. Nonetheless, in other applications the non-diagonal covariance might be important. For example, diagonal covariances would likely have negative effects in a different deformation model, for instance if z was modelled as the displacement field itself.

5 Discussion and Conclusion

In this work, we build a principled connection between classical registration methods and recent learning-based approaches. We propose a probabilistic model for diffeomorphic image registration and derive a learning algorithm that leverages a convolutional neural network and unsupervised, end-to-end learning for fast runtime. To achieve diffeomorphic transforms, we integrate stationary velocity fields through novel *scaling and squaring* differentiable network operations, and provide implementation and analysis for other integration layers.

Although the simplifying diagonal approximation to the velocity covariance $\boldsymbol{\Sigma}_{z|m,f}$ adds voxel-independent noise to every velocity field sample $\boldsymbol{z}_k$, the resulting deformation fields are well behaved because of our smoothing prior and diffeomorphic representation.

We also provide an anatomical surface deformation model. If image segmentations are available for a particular anatomical structure, the generative model incorporates them naturally in the same joint framework during training, while not requiring the surfaces at test time.

Our algorithm can infer the registration of new image pairs in under a second. Compared to traditional methods, our approach is significantly faster, and compared to recent learning based methods, our method offers diffeomorphic guarantees. We demonstrate that the surface extension to our model can help improve registration while preserving properties such as low runtime and diffeomorphisms.

Furthermore, several conclusions shown in recent papers apply to our method. For example, when only given very limited training data, deformation from VoxelMorph can still be used as initialization to a classical method, enabling faster convergence (Balakrishnan et al, 2019).

Our focus in this framework has been to present the technical connection between classical and learning paradigms, and show that diffeomorphisms are attainable in a very low runtime. Immediate extensions can enable other models and applications. For example, our derivation is generalizable to other formulations: $\boldsymbol{z}$ can be a low dimensional embedding representation of a deformation field, or the displacement field itself. Similarly, the variational covariance $\boldsymbol{\Sigma}_{z|m,f}$ enables an estimation of the uncertainty of the deformation field at each voxel, which can be informative in downstream tasks such as biomedical segmentation or population analysis. The model is also widely applicable to other applications, such as subject-to-subject registration, segmentation-only registration, or using multiple surfaces to improve image-based registration.

6 Acknowledgments

This research was funded by NIH grants R01LM012719, R01AG053949, and 1R21AG050122, NSF CAREER 1748377, NSF NeuroNex Grant 1707312, and Wistron Corporation. We thank John Ashburner for sharing code to simulate images for the C-shape controlled experiments.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Dror Aiger, Niloy J Mitra, and Daniel Cohen-Or. 4-points congruent sets for robust pairwise surface registration. In *ACM transactions on graphics (TOG)*, volume 27, page 85. Acm, 2008.

[3] A Amir-Khalili, G Hamarneh, R Zakariaee, I Spadinger, and R Abugharbieh. Propagation of registration uncertainty during multi-fraction cervical cancer brachytherapy. *Physics in Medicine & Biology*, 62(20):8116, 2017.

[4] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer, 2006.

[5] J. Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007.

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

[7] 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 analysis*, 12(1):26–41, 2008.

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

[9] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9252–9260, 2018.

[10] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: A learning framework for deformable medical image registration. *arXiv preprint arXiv:1809.05231*, 2019.

[11] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int. J. Comput. Vision*, 61:139–157, 2005.

[12] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures*, volume 1611, pages 586–607. International Society for Optics and Photonics, 1992.

[13] Xiaohuan Cao, Jianhua Yang, Jun Zhang, Dong Nie, Minjeong Kim, Qian Wang, and Dinggang Shen. Deformable image registration based on similarity-steered cnn regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 300–308. Springer, 2017.

[14] Yan Cao, Michael I Miller, Raimond L Winslow, and Laurent Younes. Large deformation diffeomorphic metric mapping of vector fields. *IEEE transactions on medical imaging*, 24(9):1216–1230, 2005.

[15] Can Ceritoglu, Kenichi Oishi, Xin Li, Ming-Chung Chou, Laurent Younes, Marilyn Albert, Constantine Lyketsos, Peter CM van Zijl, Michael I Miller, and Susumu Mori. Multi-contrast large deformation diffeomorphic metric mapping for diffusion tensor imaging. *Neuroimage*, 47(2):618–627, 2009.

[16] François Chollet. Keras. `https://github.com/fchollet/keras`, 2015.

[17] Alexander Dagley, Molly LaPoint, Willem Huijbers, Trey Hedden, Donald G McLaren, Jasmeer P Chatwal, Kathryn V Papp, Rebecca E Amariglio, Deborah Blacker, Dorene M Rentz, et al. Harvard aging brain study: dataset and accessibility. *NeuroImage*, 2015.

[18] Adrian V. Dalca, Guha Balakrishnan, John Guttag, and Mert R. Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 729–738, Cham, 2018. Springer.

[19] Adrian V Dalca, Andreea Bobu, Natalia S Rost, and Polina Golland. Patch-based discrete registration of clinical brain images. In *MICCAI-PATCHMI Patch-based Techniques in Medical Imaging*. Springer, 2016.

[20] Adrian V Dalca, John Guttag, and Mert R Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9290–9299, 2018.

[21] Christos Davatzikos. Spatial transformation and registration of brain images using elastically deformable models. *Computer Vision and Image Understanding*, 66(2):207–222, 1997.

[22] Bob D de Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.

[23] Bob D de Vos, Floris F Berendsen, Max A Viergever, Marius Staring, and Ivana Išgum. End-to-end unsupervised deformable image registration with a convolutional neural network. In *DLMIA*, pages 204–212. Springer, 2017.

[24] A. Di Martino et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

[25] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[26] Stanley Durrleman. *Statistical models of currents for measuring the variability of anatomical curves, surfaces and their evolution*. PhD thesis, Université Nice Sophia Antipolis, 2010.

[27] B. Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[28] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios. Dense image registration through mrfs and efficient linear programming. *Medical image analysis*, 12(6):731–741, 2008.

[29] Randy L Gollub, Jody M Shoemaker, Margaret D King, Tonya White, Stefan Ehrlich, Scott R Sponheim, Vincent P Clark, Jessica A Turner, Bryon A Mueller, Vince Magnotta, et al. The MCIC collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics*, 11(3):367–388, 2013.

[30] Mattias P Heinrich, Mark Jenkinson, Michael Brady, and Julia A Schnabel. Mrf-based deformable registration and ventilation estimation of lung ct. *IEEE transactions on medical imaging*, 32(7):1239–1248, 2013.

[31] Mattias P Heinrich, Ivor JA Simpson, BartŁomiej W Papież, Michael Brady, and Julia A Schnabel. Deformable image registration by combining uncertainty estimates from supervoxel belief propagation. *Medical image analysis*, 27:57–71, 2016.

[32] 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.

[33] Avram J Holmes, Marisa O Hollinshead, Timothy M O'Keefe, Victor I Petrov, Gabriele R Fariello, Lawrence L Wald, Bruce Fischl, Bruce R Rosen, Ross W Mair, Joshua L Roffman, et al. Brain genomics superstruct project initial data release with structural, functional, and behavioral measures. *Scientific data*, 2, 2015.

[34] Yipeng Hu, Marc Modat, Eli Gibson, Nooshin Ghavami, Ester Bonmati, Caroline M Moore, Mark Emberton, J Alison Noble, Dean C Barratt, and Tom Vercauteren. Label-driven weakly-supervised learning for multimodal deformable image registration. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 1070–1074. IEEE, 2018.

[35] Yipeng Hu, Marc Modat, Eli Gibson, Wenqi Li, Nooshin Ghavami, Ester Bonmati, Guotai Wang, Steven Bandula, Caroline M Moore, Mark Emberton,

et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Medical image analysis*, 49:1–13, 2018.

[36] Max Jaderberg, Karen Simonyan, and Andrew Zisserman. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

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

[38] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[39] D.P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.

[40] Arno Klein, Jesper Andersson, Babak A Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E Christensen, D Louis Collins, James Gee, Pierre Hellier, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. *Neuroimage*, 46(3):786–802, 2009.

[41] Julian Krebs, Hervé e Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE transactions on medical imaging*, 2019.

[42] Julian Krebs et al. Robust non-rigid registration through agent-based action learning. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 344–352, Cham, 2017. Springer International Publishing.

[43] Julian Krebs, Tommaso Mansi, Boris Mailhé, Nicholas Ayache, and Hervé Delingette. Unsupervised probabilistic deformation modeling for robust diffeomorphic registration. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 101–109. Springer, 2018.

[44] H. Li and Y. Fan. Non-rigid image registration using fully convolutional networks with deep self-supervision. *arXiv preprint arXiv:1709.00799*, 2017.

[45] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

[46] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011.

[47] Michael I Miga, Tuhin K Sinha, David M Cash, Robert L Galloway, and Robert J Weil. Cortical surface registration for image-guided neurosurgery using laser-range scanning. *IEEE Transactions on medical Imaging*, 22(8):973–985, 2003.

[48] Michael P Milham, Damien Fair, Maarten Mennes, Stewart HMD Mostofsky, et al. The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience*, 6:62, 2012.

[49] 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.

[50] 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):024003, 2014.

[51] 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.

[52] Marc Modat, Ivor JA Simpson, Manual Jorge Cardoso, David M Cash, Nicolas Toussaint, Nick C Fox, and Sébastien Ourselin. Simulating neurodegeneration through longitudinal population analysis of structural and diffusion weighted mri data. *Medical Image Computing and Computer-Assisted Intervention*, LNCS 8675:57–64, 2014.

[53] Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, 45(1):3–49, 2003.

[54] Marc Niethammer, Roland Kwitt, and Francois-Xavier Vialard. Metric learning for image registration. *arXiv preprint arXiv:1904.09524*, 2019.

[55] Kenichi Oishi, Andreia Faria, Hangyi Jiang, Xin Li, Kazi Akhter, Jiangyang Zhang, John T Hsu, Michael I Miller, Peter CM van Zijl, Marilyn Albert, et al. Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and alzheimer's disease participants. *Neuroimage*, 46(2):486–499, 2009.

[56] Xavier Pennec, Pascal Cachier, and Nicholas Ayache. Understanding the "demon's algorithm": 3d non-rigid registration by gradient descent. pages 597–605, 1999.

[57] Gheorghe Postelnicu, Lilla Zollei, and Bruce Fischl. Combined volumetric and surface registration. *IEEE transactions on medical imaging*, 28(4):508–522, 2008.

[58] Petter Risholm, Firdaus Janoos, Isaiah Norton, Alex J Golby, and William M Wells III. Bayesian characterization of uncertainty in intra-subject non-rigid registration. *Medical image analysis*, 17(5):538–555, 2013.

[59] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. SVF-Net: Learning deformable image registration using shape matching. In *MICCAI*, pages 266–274. Springer, 2017.

[60] O. Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[61] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformation: Application to breast mr images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

[62] D. Shen and C. Davatzikos. Hammer: Hierarchical attribute matching mechanism for elastic registration. *IEEE Trans. Med. Imag.*, 21(11):1421–1439, 2002.

[63] Ivor JA Simpson, Mark W Woolrich, Jesper LR Andersson, Adrian R Groves, and Julia A Schnabel. A probabilistic non-rigid registration framework using local noise estimates. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 688–691. IEEE, 2012.

[64] Hessam Sokooti, Bob de Vos, Floris Berendsen, Boudewijn PF Lelieveldt, Ivana Išgum, and Marius Staring. Nonrigid image registration using multi-scale 3d convolutional neural networks. In *MICCAI*, pages 232–239, Cham, 2017. Springer.

[65] R. Sridharan, A.V. Dalca, K.M. Fitzpatrick, L. Cloonan, A. Kanakis, O. Wu, K.L. Furie, J. Rosand, N.S. Rost, and P. Golland. Quantification and analysis of large multimodal clinical image studies: Application to stroke. In *International Workshop on Multimodal Brain Image Analysis*, pages 18–30. Springer International Publishing, 2013.

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

[67] Tom Vercauteren et al. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009.

[68] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage*, 158:378–396, 2017.

[69] BT 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 transactions on medical imaging*, 29(7):1424–1441, 2010.

[70] 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 *International Conference on Information Processing in Medical Imaging*, pages 559–570. Springer, 2017.

Supplementary Material

Derivation of Main Loss

$$
\begin{aligned}
\mathcal{L}(\psi; \boldsymbol{m}, \boldsymbol{f}) &= -\mathbf{E}_q \left[\log p(\boldsymbol{m}|\boldsymbol{z}; \boldsymbol{f})\right] + \mathrm{KL}\left[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{m}; \boldsymbol{f})||p(\boldsymbol{z})\right] \\
&= -\mathbf{E}_q \left[\log \mathcal{N}(\boldsymbol{m}; \boldsymbol{z} \circ \boldsymbol{f}; \sigma^2 \mathbb{I})\right] + \mathrm{KL}\left[\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{z|m,f}, \boldsymbol{\Sigma}_{z|m,f})||\mathcal{N}(\boldsymbol{z}; \underline{0}, \boldsymbol{\Lambda}_z)\right] \\
&= \frac{1}{2}\mathbf{E}_q \left[\log 2\pi\sigma^{2d} + \frac{1}{\sigma^2}\|\boldsymbol{m} - \boldsymbol{f} \circ \boldsymbol{\phi}_z\|^2\right] + \mathrm{KL}\left[\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{z|m,f}, \boldsymbol{\Sigma}_{z|m,f})||\mathcal{N}(\boldsymbol{z}; \underline{0}, \boldsymbol{\Lambda}_z)\right] \\
&= \frac{1}{2\sigma^2}\mathbf{E}_q \left[\|\boldsymbol{m} - \boldsymbol{f} \circ \boldsymbol{\phi}_{z_k}\|^2\right] + \frac{1}{2}\left[\log \frac{|\boldsymbol{\Lambda}_z^{-1}|}{|\boldsymbol{\Sigma}_{z|m,f}|} - 3d + \mathrm{tr}(\boldsymbol{\Lambda}_z \boldsymbol{\Sigma}_{z|m,f}) + \boldsymbol{\mu}_{z|m,f}^T \boldsymbol{\Lambda}_z \boldsymbol{\mu}_{z|m,f}\right] + \text{const.}
\end{aligned}
$$

Using the facts that $\log|\boldsymbol{\Lambda}_z|$ is constant, $\log|\boldsymbol{\Sigma}_{z|m,f}| = \mathrm{tr}\log\boldsymbol{\Sigma}_{z|m,f}$, and $\mathrm{tr}(\boldsymbol{\Lambda}_z\boldsymbol{\Sigma}_{z|m,f}) = \mathrm{tr}((\lambda\boldsymbol{D} - \boldsymbol{A})\boldsymbol{\Sigma}_{z|m,f}) = \mathrm{tr}(\lambda\boldsymbol{D}\boldsymbol{\Sigma}_{z|m,f})$, and approximating the expectation with K samples $z_k \sim q_z$, we obtain

$$
\mathcal{L}(\psi; \boldsymbol{m}, \boldsymbol{f}) = \frac{1}{2\sigma^2 K}\sum_k \|\boldsymbol{m} - \boldsymbol{f} \circ \boldsymbol{\phi}_{z_k}\|^2 + \frac{1}{2}\left[\mathrm{tr}(\lambda\boldsymbol{D}\boldsymbol{\Sigma}_{z|x;y} - \log\boldsymbol{\Sigma}_{z|x;y}) + \boldsymbol{\mu}_{z|m,f}^T\boldsymbol{\Lambda}_z\boldsymbol{\mu}_{z|m,f}\right] + \text{const.} \quad (12)
$$

Derivation of Surface VLB and Loss

We derive the variational lower bound and loss for the generative surface model. We start by minimizing the KL divergence between the true and approximate posteriors:

$$
\begin{aligned}
&\min_{\psi} \mathrm{KL}\left[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})||p(\boldsymbol{z}|\boldsymbol{f}, \boldsymbol{s_f}; \boldsymbol{m}, \boldsymbol{s_m})\right] \\
&= \min_{\psi} \mathbf{E}_q \left[\log q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m}) - \log p(\boldsymbol{z}|\boldsymbol{f}, \boldsymbol{s_f}; \boldsymbol{m}, \boldsymbol{s_m})\right] \\
&= \min_{\psi} \mathbf{E}_q \left[\log q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m}) - \log p(\boldsymbol{z}, \boldsymbol{f}, \boldsymbol{s_f}; \boldsymbol{m}, \boldsymbol{s_m})\right] + \log p(\boldsymbol{f}, \boldsymbol{s_f}; \boldsymbol{m}, \boldsymbol{s_m}) \\
&= \min_{\psi} \mathbf{E}_q \left[\log q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m}) - \log p(\boldsymbol{z}) - \log p(\boldsymbol{f}, \boldsymbol{s_f}|\boldsymbol{z}; \boldsymbol{m}, \boldsymbol{s_m})\right] + \text{const} \\
&\overset{\star}{=} \min_{\psi} \mathbf{E}_q \left[\log q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m}) - \log p(\boldsymbol{z}) - \log p(\boldsymbol{f}|\boldsymbol{z}; \boldsymbol{m}) - \log p(\boldsymbol{s_f}|\boldsymbol{z}; \boldsymbol{s_m})\right] + \text{const} \\
&= \min_{\psi} \mathrm{KL}\left[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{f}; \boldsymbol{m})||p(\boldsymbol{z})\right] - \mathbf{E}_q \left[\log p(\boldsymbol{f}|\boldsymbol{z}; \boldsymbol{m})\right] - \mathbf{E}_q \left[\log p(\boldsymbol{s_f}|\boldsymbol{z}; \boldsymbol{s_m})\right] + \text{const}, \quad (13)
\end{aligned}
$$

where in $\star$ we used the assumptions that the fixed image is independent of anatomical surfaces given the moving image and the deformation, and the fixed surface is independent of either image given the moving surface and the deformation. Following this variational lower bound, the loss follows the previous section closely (see (13)), with the additional term:

$$
\begin{aligned}
\mathbf{E}_q \left[\log p(\boldsymbol{s_f}|\boldsymbol{z}; \boldsymbol{s_m})\right] &= \mathbf{E}_q \left[\log \mathcal{N}(\boldsymbol{m}; \boldsymbol{z} \circ \boldsymbol{f}; \sigma^2 \mathbb{I})\right] \\
&= \frac{1}{2}\mathbf{E}_q \left[\log 2\pi\sigma_s^{2d} + \frac{1}{\sigma_s^2}\|\boldsymbol{s_f} - \boldsymbol{s_m} \circ \boldsymbol{\phi}_{z_k}\|^2\right] \\
&= \frac{1}{2}\mathbf{E}_q \left[\frac{1}{\sigma_s^2}\|\boldsymbol{s_f} - \boldsymbol{s_m} \circ \boldsymbol{\phi}_{z_k}\|^2\right] + \text{const.} \quad (14)
\end{aligned}
$$

Combining this term with (12), and approximating expectations with k samples leads to the final loss (10):

$$
\begin{aligned}
\mathcal{L}(\psi; \boldsymbol{f}, \boldsymbol{s_f}, \boldsymbol{m}, \boldsymbol{s_m}) = &\frac{1}{2\sigma^2 K}\sum_k ||\boldsymbol{f} - \boldsymbol{m} \circ \boldsymbol{\phi}_{z_k}||^2 + \frac{1}{2\sigma_s^2 K}\sum_k ||\boldsymbol{s_f} - \boldsymbol{s_m} \circ \boldsymbol{\phi}_{z_k}||^2 \\
&+ \frac{1}{2}\left[\mathrm{tr}(\lambda\boldsymbol{D}\boldsymbol{\Sigma}_{z|x;y} - \log\boldsymbol{\Sigma}_{z|x;y}) + \boldsymbol{\mu}_{z|m,f}^T\boldsymbol{\Lambda}_z\boldsymbol{\mu}_{z|m,f}\right] + \text{const.} \quad (15)
\end{aligned}
$$

Overview figure with surface loss

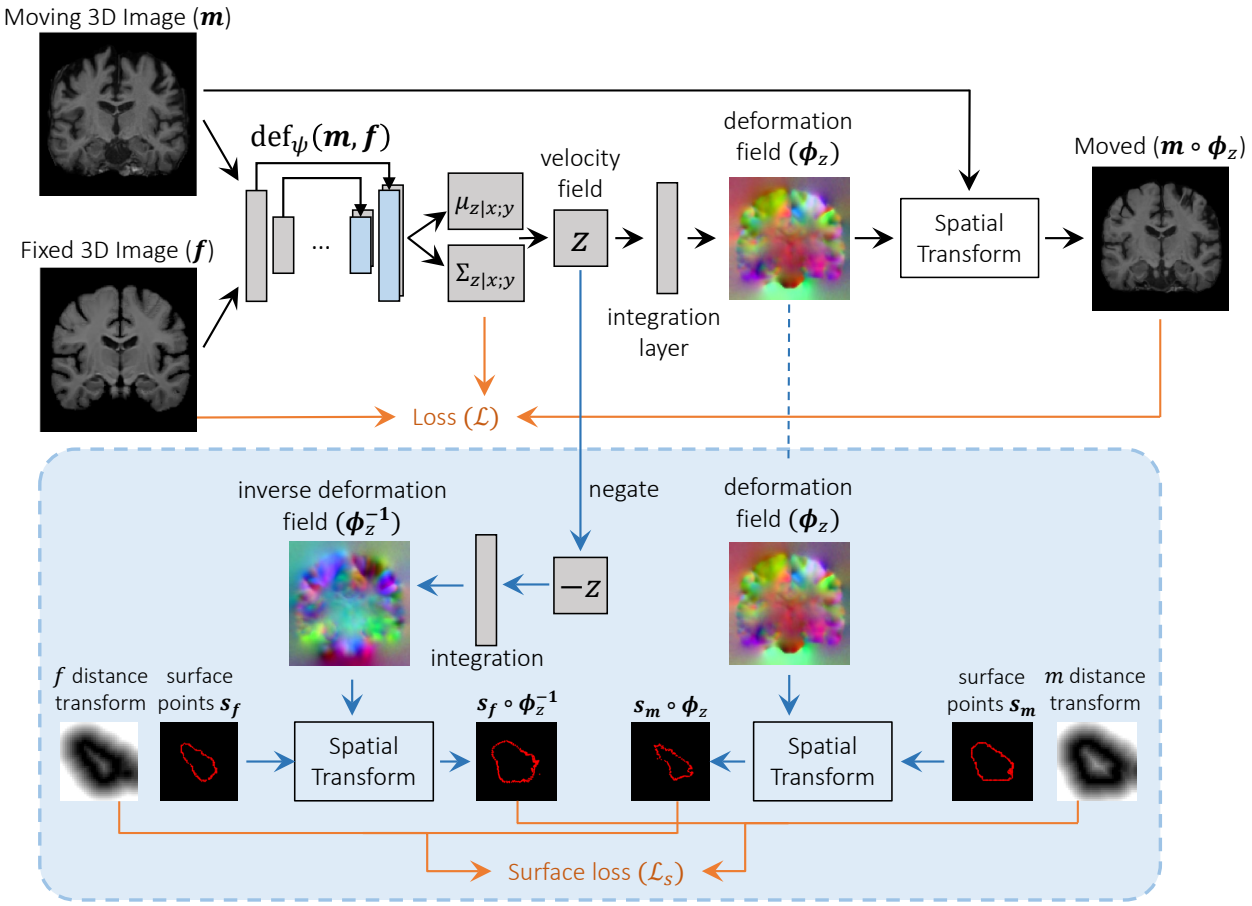

Figure 12: Overview of end-to-end unsupervised architecture building on Figure 2. The first part of the network, $\text{def}_\psi(\boldsymbol{m}, \boldsymbol{f})$ takes the input images and outputs the approximate posterior probability parameters representing the velocity field mean, $\boldsymbol{\mu}_{z|m;f}$, and variance, $\boldsymbol{\Sigma}_{z|m;f}$. A velocity field $\boldsymbol{z}$ is sampled and transformed to a diffeomorphic deformation field ϕ_z using novel differentiable *squaring and scaling* layers. Finally, a spatial transform warps $\boldsymbol{m}$ to obtain $\boldsymbol{m} \circ \phi_z$. The blue window illustrated the computation of *optional* surface registration loss. The surface points and distance transform are computed for the both the moving and fixed surfaces. The surface points are warped by the resulting deformation, and a distance is computed using distance transforms.

Additional Figures

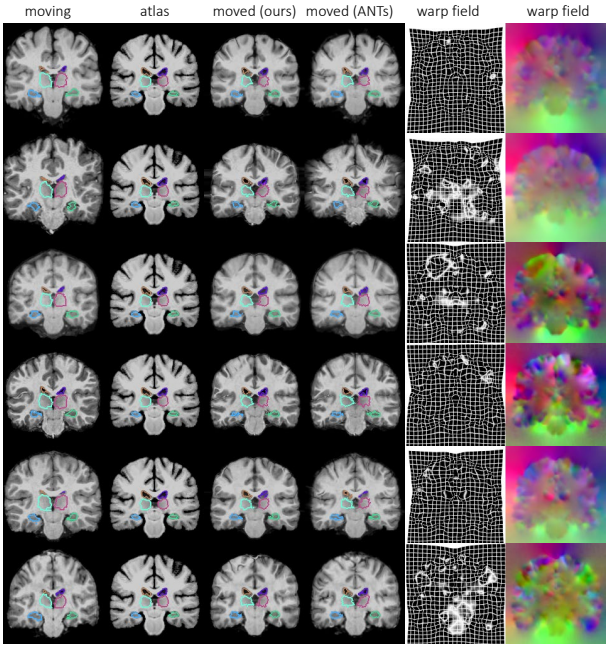

Figure 13: Additional example MR slices of input moving image, atlas, and resulting warped image for our method (VoxelMorph-diff) and ANTs, with overlaid boundaries of ventricles, thalami and hippocampi. Each row is a different scan. Our resulting registration field is shown as a warped grid and RGB image, with each channel representing a dimension. We omit VoxelMorph and NiftyReg examples, which are visually similar to our results and ANTs.

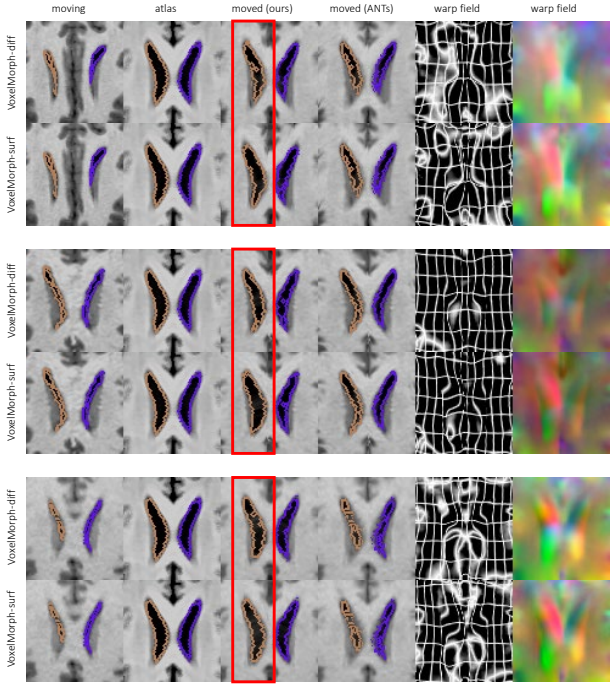

Figure 14: Example surface-driven results. For three subjects, we show cropped MR slices of input moving image, atlas, and resulting warped image for our method and ANTs, with overlaid boundaries of ventricles for VoxelMorph-diff (top) and VoxelMorph-surf (bottom). For each set of two rows, we highlight in the red box an improvement in the segmentation of the ventricle from the top row to the bottom row.