

Deep Group-wise Variational Diffeomorphic Image Registration

Tycho F.A. van der Ouderaa¹, Ivana Igum^{2,3},
Wouter B. Veldhuis⁴, and Bob D. de Vos¹

¹ Quantib-U, Utrecht, The Netherlands

Correspondence to: t.vanderouderaa@quantib.com

² Department of Biomedical Engineering and Physics, Amsterdam University Medical Center, University of Amsterdam,

³ Department of Radiology and Nuclear Medicine, Amsterdam University Medical Center, University of Amsterdam, Amsterdam,

⁴ Department of Radiology, University Medical Center Utrecht, Utrecht

Abstract. Deep neural networks are increasingly used for pair-wise image registration. We propose to extend current learning-based image registration to allow simultaneous registration of multiple images. To achieve this, we build upon the pair-wise variational and diffeomorphic VoxelMorph approach and present a general mathematical framework that enables both registration of multiple images to their geodesic average and registration in which any of the available images can be used as a fixed image. In addition, we provide a likelihood based on normalized mutual information, a well-known image similarity metric in registration, between multiple images, and a prior that allows for explicit control over the viscous fluid energy to effectively regularize deformations. We trained and evaluated our approach using intra-patient registration of breast MRI and Thoracic 4DCT exams acquired over multiple time points. Comparison with Elastix and VoxelMorph demonstrates competitive quantitative performance of the proposed method in terms of image similarity and reference landmark distances at significantly faster registration.

Keywords: deep learning · variational · diffeomorphic · group-wise · image registration

1 Introduction

Image registration, the process of aligning images, is a fundamental problem in medical image analysis [32]. For example, it can be used to align image volumes in longitudinal studies [25], or to align images acquired temporally during contrast enhancement [27].

A common way to register images is to apply a transformation, such as an affine transformation, B-splines [28], or a deformation (vector) field [29], and to optimize its parameters using some form of numerical optimization. The objective is to minimize misalignment defined by a loss, such as a squared error,

negative normalized cross correlation, or more advanced dissimilarity measures such as negative normalized mutual information [24]. Additionally, unrealistic deformations are mitigated by regularization or by using a constraint family of deformations.

Methods that constrain deformation fields to diffeomorphisms, such as LD-DMM [4], Diffeomorphic Demons [31], and SyN [1], guarantee preservation of topology, and have shown to be very effective.

Recently, it has been shown that deep learning models can be trained to perform image registration, leveraging on the increased availability of medical imaging data [34,2]. Once a model is trained, deformation fields can be obtained by inference, which is much faster than aligning images through an iterative optimization process. In [13] and [14], diffeomorphic neural network-based methods were proposed, further bridging the gap between learning-based and classical approaches. Although deep learning methods show great potential, they are restricted to image pairs or require a heuristically constructed template [8].

In this work, we extend learning-based methods to enable fast group-wise image registration. We build upon the variational framework proposed by [14] and stack multiple images along the channel axis of the input and output of the network. In addition, we propose a likelihood based on normalized mutual information and a prior which yields explicit control over the *viscous fluid energy*. We quantitatively evaluate our model and compare it with Elastix [18] and Voxelmorph [3].

2 Method

Let us consider random variables $\mathbf{x}$ and $\mathbf{y}$ each consisting of a collection of image volumes acquired at K different timepoints. We consider $\mathbf{y}$ to be unknown aligned ground-truth and let $\mathbf{z}$ be a latent variable that parameterizes transformations $\phi_{\mathbf{z}}^{-1}$ that describes the formation of $\mathbf{x}$ as a transformation of $\mathbf{y}$ through $\phi_{\mathbf{z}}^{-1} \circ \mathbf{y}$. All variables can be partitioned into different subsets for each timepoint $k \in \{1, \dots, K\}$, where $\mathbf{z}_k \in \mathbf{z}$ defines a transformation on 3D voxels $\phi_{\mathbf{z}_k}^{-1} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ that maps the volume $\mathbf{y}_k \in \mathbf{y}$ onto volume $\mathbf{x}_k \in \mathbf{x}$ via $\phi_{\mathbf{z}_k}^{-1} \circ \mathbf{y}_k$.

Depending on the task, we define images that require registration (i.e. moving images) as observations of $\mathbf{x}$ and the target images (i.e. fixed images) as observations of $\mathbf{y}$. We are not restricted to a predefined fixed timepoint and can describe all of our data as observations of $\mathbf{x}$, effectively letting $\mathbf{y}$ be an unobserved latent variable. We will refer to this case as *all-moving* group-wise image registration, and the case where one timepoint is chosen as fixed image as *all-to-one* group-wise image registration. Observe that we can obtain the former from the latter by constraining the deformation $\phi_{\mathbf{z}_p}^{-1} = \text{Id}$ at a fixed timepoint p to be the identity. Moreover, we can view pair-wise registration as a particular $K = 2$ instance in our group-wise framework, where one timepoint is kept fixed and the other one can move.

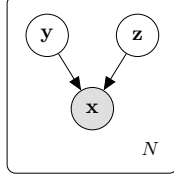


Fig. 1: Graphical model describing misaligned observations $\mathbf{x}$ as noisy observations of unknown aligned images $\mathbf{y}$ that have been transformed via $\phi_{\mathbf{z}}^{-1} \circ \mathbf{y}$.

2.1 Generative Model

We propose a generative model and aim to find the posterior probability of our transformation parameters $p(\mathbf{z}|\mathbf{x}, \mathbf{y})$ that govern the deformations $\phi_{\mathbf{z}}$. Because the posterior is intractable we approximate it using *variational inference*.

Prior We define a multivariate normal prior $\mathbf{z}$ with zero mean and precision matrix $\mathbf{\Lambda}$ as the inverse of the covariance:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid 0, \mathbf{\Lambda}^{-1}), \quad (1)$$

where we choose $\mathbf{\Lambda} = \lambda_v \mathbf{I} + \lambda_u \mathbf{L}$ with identity matrix $\mathbf{I}$ weighted by λ_v and the Laplacian neighbourhood graph $\mathbf{L} = \mathbf{D} - \mathbf{A}$ with degree matrix $\mathbf{D}$ and adjacency matrix $\mathbf{A}$, weighted by λ_u . We treat λ_v and λ_u as hyper-parameters that give control over the norm on the vectors in the field parameterized by $\mathbf{z}$ and the norm on its displacement, respectively. The prior is inspired on the spatial smoothing prior proposed in [13], which can be seen as a particular $\lambda_v = 0$ instance of our prior. If $\mathbf{z}$ describes the deformation field as is done in [3], it penalizes the *linear elastic energy* penalty [22] and in case ϕ is diffeomorphic and $\mathbf{z}$ describes the velocity of the deformation field, the prior is more akin to a fluid-based model and penalizes the *viscous fluid energy* [10,5,30].

Likelihood It is well known that (normalized) mutual information information is an effective similarity metric for image registration which is robust to varying intensity inhomogeneities [33,21]. Therefore, we model the probability of $\mathbf{x}$ as an exponential distribution of the average normalized mutual information (NMI) across all timepoint permutations:

$$p(\mathbf{x}|\mathbf{y}, \mathbf{z}) \propto \exp \left\{ \frac{1}{K(K-1)} \sum_{i \neq j} \text{NMI}(\phi_{\mathbf{z}_i} \circ \mathbf{x}_i, \phi_{\mathbf{z}_j} \circ \mathbf{x}_j) \right\} \quad (2)$$

where we define $\phi_{\mathbf{z}_p} = \text{Id}$ and $\mathbf{x}_p = \mathbf{y}_p$ in case of a fixed timepoint p . Various approaches exist to estimate the probability density functions (both marginal and joint) of the intensity values needed to compute the entropies for the mutual information [24]. In this work, we obtain differentiable estimates of the intensity values using a Gaussian Parzen window estimation [11,21] of the voxel intensities in each mini batch.

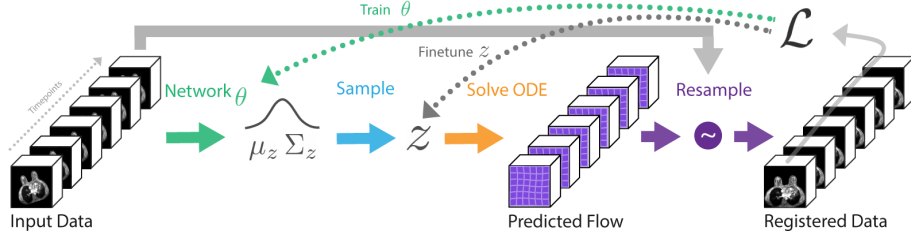


Fig. 2: GroupMorph: A neural network parameterized by θ is trained to output estimates over parameters of velocity fields $\mathbf{z}$ that can be integrated to obtain diffeomorphic flows that align the input images according to some loss $\mathcal{L}$.

2.2 Variational Lower Bound

Since the posterior probability $p(\mathbf{z}|\mathbf{x}, \mathbf{y})$ is intractable, we cannot directly obtain MAP estimates of $\mathbf{z}$. Instead, we estimate the true posterior with a variational approximating posterior by minimizing the negative evidence lower bound $\mathcal{L}(\theta|\mathbf{x}, \mathbf{y})$ which can be derived using the fact that the KL-divergence is non-negative [17]:

$$\log p_{\theta}(\mathbf{x}, \mathbf{y}) = \text{KL}(q_{\theta}(\mathbf{z}|\mathbf{x}, \mathbf{y}) \parallel p(\mathbf{z}|\mathbf{x}, \mathbf{y})) - \mathcal{L}(\theta|\mathbf{x}, \mathbf{y}) \quad (3)$$

$$\geq -\mathcal{L}(\theta|\mathbf{x}, \mathbf{y}) = -\mathbb{E}_q[-\log q_{\theta}(\mathbf{z}|\mathbf{x}, \mathbf{y}) + \log p(\mathbf{x}, \mathbf{y}, \mathbf{z})] \quad (4)$$

$$= -\mathbb{E}_q[\log p(\mathbf{x}|\mathbf{y}, \mathbf{z})] + \text{KL}(q_{\theta}(\mathbf{z}|\mathbf{x}, \mathbf{y}) \parallel p(\mathbf{z})) + \text{const.} \quad (5)$$

We let our variational approximate posterior [17] be a multivariate Gaussian:

$$\log q_{\theta}(\mathbf{z}|\mathbf{x}, \mathbf{y}) = \log \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}}), \quad (6)$$

where mean $\boldsymbol{\mu}_{\mathbf{z}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{z}}$ of the approximate posterior $q_{\theta}(\mathbf{z}|\mathbf{x}, \mathbf{y})$ are outputted by a neural network with parameters θ . To avoid differentiating through a random variable, we obtain the sample $\mathbf{z}$ using the *re-parameterization trick* [17] $\mathbf{z} = \boldsymbol{\mu}_{\mathbf{z}} + \boldsymbol{\Sigma}_{\mathbf{z}}\boldsymbol{\Sigma}_{\mathbf{z}}^T e$ where $e \sim \mathcal{N}(0, \mathbf{I})$ is normally distributed.

2.3 Resulting Loss

We find the resulting loss for a single Monte Carlo sample from q :

$$\begin{aligned} \mathcal{L}(\theta|\mathbf{x}, \mathbf{y}) &= \underbrace{-\mathbb{E}_q[\log p(\mathbf{x}|\mathbf{y}, \mathbf{z})]}_{\text{Reconstruction Error}} + \underbrace{\text{KL}[q_{\theta}(\mathbf{z}|\mathbf{x}, \mathbf{y}) \parallel p(\mathbf{z})]}_{\text{Regularization}} \\ &\simeq \frac{1}{K(K-1)} \sum_{i \neq j} -\text{NMI}(\phi_{\mathbf{z}_i} \circ \mathbf{x}_i, \phi_{\mathbf{z}_j} \circ \mathbf{x}_j) + \frac{1}{2} \left[\text{tr}((\lambda_u \mathbf{D} + \lambda_v) \boldsymbol{\Sigma}_{\mathbf{z}} - \log \boldsymbol{\Sigma}_{\mathbf{z}}) + \boldsymbol{\mu}_{\mathbf{z}}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\mu}_{\mathbf{z}} \right] + \text{const.} \end{aligned} \quad (7)$$

where we define $\phi_{\mathbf{z}_p} = \text{Id}$ and $\mathbf{x}_p = \mathbf{y}_p$ in case of a fixed timepoint p and where we can expand the last term $\boldsymbol{\mu}_{\mathbf{z}}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\mu}_{\mathbf{z}} = \lambda_u \frac{1}{2} \sum_i \sum_{j \in N(i)} (\boldsymbol{\mu}[i] - \boldsymbol{\mu}[j])^2 + \lambda_v \sum_i \boldsymbol{\mu}[i]^2$ where $N(i)$ are the neighbors of voxel i .

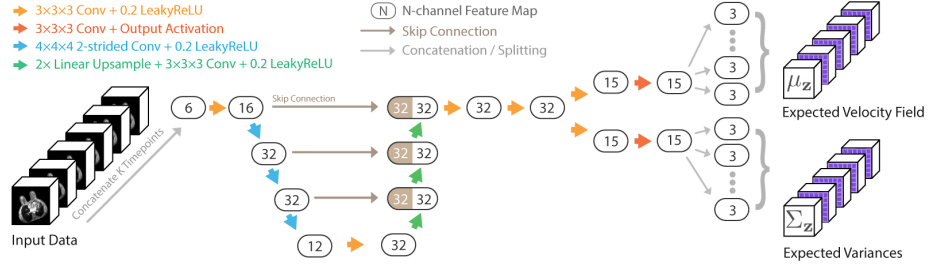


Fig. 3: Illustration of the 3D U-Net architecture used to output mean $\mu_{\mathbf{z}}$ and variance $\Sigma_{\mathbf{z}}$ of distribution over velocity fields $q_{\theta}(\mathbf{z}|\mathbf{x}, \mathbf{y})$ used for diffeomorphic image registration.

2.4 Diffeomorphic Transformations

Diffeomorphic deformation fields are invertible (ϕ^{-1} exists) and have found to be very valuable in image registration as they preserve topology, by design. For diffeomorphic GroupMorph, we follow [3] and describe transformations $\phi_{\mathbf{z}_k}$ for each timepoints k as an ordinary differential equation (ODE):

$$\frac{\partial \phi_{\mathbf{z}_k}^{(t)}}{\partial t} = v_{\mathbf{z}_k}(\phi_{\mathbf{z}_k}^{(t)}) \quad (8)$$

where $\phi_{\mathbf{z}_k}^{(0)} = \text{Id}$ is the identity transformation and t is time. For every timepoint k , we let $\mathbf{z}_k$ parameterize the vectors of the velocity field $v_{\mathbf{z}_k}$ and solve deformation field $\phi_{\mathbf{z}_k}^{(1)}$ using the *scaling and squaring* solver, proposed by [13] in the context of neural-based diffeomorphic image registration. For non-diffeomorphic models, we let $\mathbf{z}_k$ parameterize the vectors of the deformation field $\phi_{\mathbf{z}_k}$ directly.

3 Implementation

We let a 3D U-net [26,2,3], output the parameters $\mu_{\mathbf{z}}$ and $\Sigma_{\mathbf{z}}$ of the approximate posterior $q_{\theta}(\mathbf{z}|\mathbf{x}, \mathbf{y})$, as illustrated in Figure 2. Fixed and moving image volumes are concatenated along the channel dimension and form the input of the network. Figure 3 gives a more detailed description of the network architecture, including filter sizes and channel counts. All layers are followed by a Leaky ReLU [20] with 0.2 slope, except for the $\mu_{\mathbf{z}}$ and $\Sigma_{\mathbf{z}}$ heads which are respectively followed by a hyperbolic tangent scaled with $\alpha = 15$ that limits the output to $[-\alpha, \alpha]$ voxels and a Softplus⁵ activation to ensure positive variances.

We train for 100k iterations using Adam [16] with a batch size of 1 and a learning rate of $1e^{-5}$ ($\beta_1 = 0.9, \beta_2 = 0.999$) decayed with cosine annealing [19]. Inputs are normalized using 1-99 percentile normalization [23] and training samples consist of $96 \times 96 \times 96$ patches for all timepoints at a randomly sampled location. For the hyperparameters we choose $\lambda_u = 1$ and $\lambda_v = 0.01$.

⁵ Softplus(x) = $\log(\exp(x) + 1)$

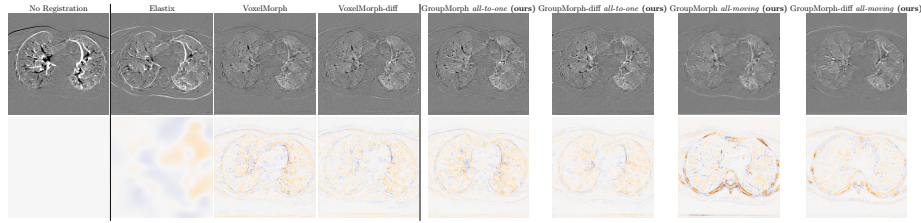


Fig. 4: Qualitative evaluation of first patient of *DIR-Lab 4DCT* dataset after applying different registration models. TOP: Axial slice of average timepoint after subtracting the first timepoint. BOTTOM: Axial slice of the Jacobian determinant field averaged over all timepoints.

4 Datasets

Breast MR Dataset The *Breast MR* dataset comprises 270 training and 68 evaluation dynamic contrast enhancement series of subjects with extremely dense breast tissue (Volpara Density Grade 4). Each series contains DCE-MRI images ($384 \times 384 \times 60$ voxels with spacing $0.966 \times 0.966 \times 3.000$ mm resampled to 1mm^3) acquired on a 3.0T Achieva or Ingenia Philips system at $K = 6$ timepoints in the axial plane with bilateral anatomic coverage. The interval between timepoints between the first two timepoints is 90 seconds in which a 0.1 mL dose of gadobutrol per kilogram of body weight is injected at a rate of 1 mL/sec and 1 minute for the other timepoints.

DIR-Lab 4DCT This set contains 4D chest CT scans of 10 different patients that have been acquired as part of the radiotherapy planning process for the treatment of thoracic malignancies. For each patient 3D CT volumes have been acquired at $K = 6$ timepoints during the expiratory phase with an average voxel spacing of $1.04 \times 1.04 \times 2.5$. A detailed description of the data acquisition can be found in [7] and [6]. The dataset is a subset of the publicly available *DIR-Lab 4DCT* dataset [7] and contains 75 landmark reference points for all timepoints. The volumes have been cropped in-plane to 256×256 voxels in such a way that all reference points are present. We report performance on the first scan (‘Patient 1’) and used the others for training.

5 Experiments and Results

Using the proposed GroupMorph model, we have performed intra-patient registration of multiple 3D scans acquired at different timepoints in two different ways. First, we performed *all-to-one* registration where the scan acquired at the first time point was used as fixed image and all other scans of the same patient as moving images. Thereafter, we performed *all-moving* registration, where we compute deformation fields for all images and the register to a geodesic average.

To compare performance of the proposed GroupMorph model with state-of-the-art conventional and neural-based methods, we performed registration using Elastix [18] and VoxelMorph [3,14]. For registration with Elastix we use configuration described for registration of DCE-MRI described in [15]. For fair comparison with VoxelMorph, we use the same loss, network architecture and hyper-parameters in VoxelMorph and GroupMorph. We only vary the channels in the first and last layers to match required dimensions and set $\lambda_v = 0$ for VoxelMorph.

Experiments were performed using both breast MRI and chest CT scans. For breast MR scans no landmarks were available, hence evaluation was performed with root mean square error (RMSE) and NMI. For chest CT, a set of 75 landmarks was provided at all timepoints within DIR-Lab, hence performance was evaluated by measuring the average distance between the landmarks after registration and over all timepoint combinations. Moreover, for both sets we report percentages of negative Jacobian determinants ($\% |J_\phi| < 0$) that indicate undesirable folding. Additionally, we evaluate the speed of registrations. Results are listed in Table 1⁶. They show that the group-wise models perform better in terms of RMSE and NMI on both sets, especially in the *all-moving* case. We find that diffeomorphic registration models (indicated with ‘-diff’) lead to lower amount of negative Jacobian determinants, which means that less folding occurs and indicates that the fields are smoother than those of non-diffeomorphic counterpart. Average registration times per series (Table 1) show that registration times of the GroupMorph model are significantly lower than the registration times of the baselines, especially compared with Elastix.

Figure 4 illustrates the axial slices of the first volume in the *DIR-Lab* dataset. Subtraction images illustrate improved alignment for all registration models, especially at borders and in particular for *all-moving* group-wise models. From the Jacobian determinant fields, we can observe particularly strong local expansions and compression in the *all-moving* models and most notably in the non-diffeomorphic variant.

⁶ We measured time to calculate all required deformation fields for Elastix on two 10-core Intel Xeon Silver 4114 2.20GHz CPUs and of the neural-based methods on a NVIDIA GeForce RTX 2080 GPU.

Table 1: Quantitative performance using different registration methods. LEFT: Models and average registration time per serie. CENTER: RMSE and NMI on *Breast MR*. RIGHT: RMSE, NMI and average displacement on *DIR-Lab 4DCT* evaluation set.

Model	Time (s)	<i>Breast MR</i>				<i>DIR-Lab 4DCT</i>			
		RMSE	NMI	% $ J_\phi < 0$		RMSE	NMI	Avg Displacement (mm)	% $ J_\phi < 0$
No Registration (Identity)	-	24.15	0.24	-		63.39	0.47	1.83 ± 2.24	-
Elastix	652	23.42	0.25	0		36.84	0.49	1.09 ± 1.30	0
VoxelMorph	5	22.43	0.26	3.31e-2		28.42	0.50	1.18 ± 1.33	4.32e-6
VoxelMorph-diff	5	22.12	0.26	1.55e-4		25.37	0.51	1.12 ± 1.33	7.79e-7
<i>All-to-one</i> GroupMorph (ours)	2	21.97	0.28	4.20e-5		27.11	0.51	1.16 ± 1.36	3.05e-5
<i>All-to-one</i> GroupMorph-diff (ours)	2	22.00	0.27	1.60e-5		25.26	0.51	1.08 ± 1.35	2.60e-7
<i>All-moving</i> GroupMorph (ours)	2	21.47	0.32	2.00e-4		23.15	0.53	1.39 ± 1.32	4.98e-6
<i>All-moving</i> GroupMorph-diff (ours)	2	21.57	0.32	1.90e-4		22.09	0.54	1.52 ± 1.34	9.74e-7

6 Discussion and Conclusion

In this study, we proposed a variational and diffeomorphic learning-based registration method for application in group-wise image registration. We evaluated our approach using intra-patient registration with sets of multiple 3D scans of two different modalities (MR and CT) showing different anatomies (breast and chest). In addition, we provided a likelihood based on normalized mutual information, a well-performing image similarity metric in registration, between multiple images and a prior that allows for explicit control over the viscous fluid energy to effectively control regularization of deformation fields.

In spite of using diffeomorphic transformations, we observed some negative Jacobian determinants likely resulting from approximation errors. Hence, future work could investigate options for further improvements. It can be expected that training with more data, further tuning of the network architecture and hyperparameters will lead to higher performance. Other potential advancements include a 4D network [9] to more explicitly model the temporal component, fine tuning of $\mathbf{z}$ after inference (see dotted grey arrow in Figure 2), time-dependent velocity fields to allow more flexible diffeomorphic transformations, and a bi-directional cost [12] and non-diagonal covariance [14] to further improve smoothness of the deformation fields.

To conclude, we found that our GroupMorph model can simultaneously register multiple scans with performance similar to conventional registration with Elastix and learning-based VoxelMorph, enabling close to real-time group-wise registration without delaying clinical reading.

Acknowledgements

Authors would like to thank Jorrit Glastra, Tim Henke and Rob Hesselink for insightful discussions and support.

References

1. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis* **12**(1), 26–41 (2008)
2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: An unsupervised learning model for deformable medical image registration. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 9252–9260 (2018)
3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging* **38**(8), 1788–1800 (2019)
4. Beg, M.F., Miller, M.I., Trounev, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision* **61**(2), 139–157 (2005)
5. Bro-Nielsen, M., Gramkow, C.: Fast fluid registration of medical images. In: *International Conference on Visualization in Biomedical Computing*. pp. 265–276. Springer (1996)
6. Castillo, E., Castillo, R., Martinez, J., Shenoy, M., Guerrero, T.: Four-dimensional deformable image registration using trajectory modeling. *Physics in Medicine & Biology* **55**(1), 305 (2009)
7. Castillo, R., Castillo, E., Guerra, R., Johnson, V.E., McPhail, T., Garg, A.K., Guerrero, T.: A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. *Physics in Medicine & Biology* **54**(7), 1849 (2009)
8. Che, T., Zheng, Y., Cong, J., Jiang, Y., Niu, Y., Jiao, W., Zhao, B., Ding, Y.: Deep group-wise registration for multi-spectral images from fundus images. *IEEE Access* **7**, 27650–27661 (2019)
9. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3075–3084 (2019)
10. Christensen, G.E.: *Deformable shape models for anatomy* (1994)
11. Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., Marchal, G.: Automated multi-modality image registration based on information theory. In: *Information processing in medical imaging*. vol. 3, pp. 263–274. Citeseer (1995)
12. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Improved probabilistic diffeomorphic registration with cnns
13. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 729–738. Springer (2018)
14. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis* **57**, 226–236 (2019)
15. Gubern-Mérida, A., Martí, R., Melendez, J., Hauth, J.L., Mann, R.M., Karssemeijer, N., Platel, B.: Automated localization of breast cancer in dce-mri. *Medical image analysis* **20**(1), 265–274 (2015)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)

17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
18. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a tool-box for intensity-based medical image registration. *IEEE transactions on medical imaging* **29**(1), 196–205 (2009)
19. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
20. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: *Proc. icml*. vol. 30, p. 3 (2013)
21. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.: Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging* **16**(2), 187–198 (1997)
22. Miller, M.I., Christensen, G.E., Amit, Y., Grenander, U.: Mathematical textbook of deformable neuroanatomies. *Proceedings of the National Academy of Sciences* **90**(24), 11944–11948 (1993)
23. Patrice, D., et al.: *Image Processing and Analysis: A Primer*, vol. 3. World Scientific (2018)
24. Pluim, J.P., Maintz, J.A., Viergever, M.A.: Mutual-information-based registration of medical images: a survey. *IEEE transactions on medical imaging* **22**(8), 986–1004 (2003)
25. Reuter, M., Fischl, B.: Avoiding asymmetry-induced bias in longitudinal image processing. *Neuroimage* **57**(1), 19–21 (2011)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. pp. 234–241. Springer (2015)
27. Rueckert, D., Hayes, C., Studholme, C., Summers, P., Leach, M., Hawkes, D.J.: Non-rigid registration of breast mr images using mutual information. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 1144–1152. Springer (1998)
28. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Non-rigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging* **18**(8), 712–721 (1999)
29. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: A survey. *IEEE transactions on medical imaging* **32**(7), 1153–1190 (2013)
30. Tian, J.: *Molecular imaging: Fundamentals and applications*. Springer Science & Business Media (2013)
31. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage* **45**(1), S61–S72 (2009)
32. Viergever, M.A., Maintz, J.A., Klein, S., Murphy, K., Staring, M., Pluim, J.P.: A survey of medical image registration—under review (2016)
33. Viola, P., Wells III, W.M.: Alignment by maximization of mutual information. *International journal of computer vision* **24**(2), 137–154 (1997)
34. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I.: A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis* **52**, 128–143 (2019)

Supplementary Material

Derivation of Main Loss

By following [17] and the Supplementary Material of [14] adjusting for our proposed likelihood and prior we obtain

$$\begin{aligned}
\mathcal{L}(\theta; \mathbf{x}, \mathbf{y}) &= \\
&= -\mathbb{E}_q[\log p(\mathbf{x}|\mathbf{y}, \mathbf{z}) + \text{KL}[q_\theta(\mathbf{z}|\mathbf{x}, \mathbf{y})||p(\mathbf{z})]] \\
&= -\mathbb{E}_q \left[\log \exp \frac{1}{K(K-1)} \sum_{i \neq j} \overbrace{\text{NMI}(\phi_{\mathbf{z}_i} \circ \mathbf{x}_i, \phi_{\mathbf{z}_j} \circ \mathbf{x}_j)}^{p(\mathbf{x}|\mathbf{y}, \mathbf{z})} \right] + \text{KL}[\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}})||\mathcal{N}(\mathbf{z}; 0, \boldsymbol{\Lambda})] + \text{const.} \\
&= \frac{1}{K(K-1)} \sum_{i \neq j} \mathbb{E}_q [-\text{NMI}(\phi_{\mathbf{z}_i} \circ \mathbf{x}_i, \phi_{\mathbf{z}_j} \circ \mathbf{x}_j)] + \frac{1}{2} \left[\log \frac{|\boldsymbol{\Lambda}_{\mathbf{z}}^{-1}|}{|\boldsymbol{\Sigma}_{\mathbf{z}}|} - 3d + \text{tr}(\boldsymbol{\Lambda}_{\mathbf{z}} \boldsymbol{\Sigma}_{\mathbf{z}}) + \boldsymbol{\mu}_{\mathbf{z}}^T \boldsymbol{\Lambda}_{\mathbf{z}} \boldsymbol{\mu}_{\mathbf{z}} \right] + \text{const.}
\end{aligned}$$

We know that $\log |\boldsymbol{\Lambda}|$ is constant, $\log |\boldsymbol{\Sigma}_{\mathbf{z}}| = \text{tr} \log \boldsymbol{\Sigma}_{\mathbf{z}}$, and $\text{tr}(\boldsymbol{\Lambda}_{\mathbf{z}} \boldsymbol{\Sigma}_{\mathbf{z}}) = \text{tr}((\lambda_u \mathbf{D} - \mathbf{A} + \lambda_v \mathbf{I}) \boldsymbol{\Sigma}_{\mathbf{z}}) = \text{tr}((\lambda_u \mathbf{D} + \lambda_v \mathbf{I}) \boldsymbol{\Sigma}_{\mathbf{z}})$, and approximating the expectation with 1 sample $z_k \sim q_z$, to find

$$\mathcal{L}(\theta; \mathbf{x}, \mathbf{y}) = \frac{1}{K(K-1)} \sum_{i \neq j} -\text{NMI}(\phi_{\mathbf{z}_i} \circ \mathbf{x}_i, \phi_{\mathbf{z}_j} \circ \mathbf{x}_j) + \frac{1}{2} \left[\text{tr}((\lambda_u \mathbf{D} + \lambda_v) \boldsymbol{\Sigma}_{\mathbf{z}} - \log \boldsymbol{\Sigma}_{\mathbf{z}}) + \boldsymbol{\mu}_{\mathbf{z}}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\mu}_{\mathbf{z}} \right] + \text{const.}$$