A variational Bayesian method for similarity learning in non-rigid image registration

Daniel Grzech[1], Mohammad Farid Azampour[1,2,3],
Ben Glocker[1], Julia Schnabel[3,4], Nassir Navab[3,5], Bernhard Kainz[1,6], Loïc Le Folgoc[1]
[1] Imperial College London, [2] Sharif University of Technology,
[3] Technische Universität München, [4] King's College London, [5] Johns Hopkins University,
[6] Friedrich-Alexander-Universität Erlangen-Nürnberg

Abstract

We propose a novel variational Bayesian formulation for diffeomorphic non-rigid registration of medical images, which learns in an unsupervised way a data-specific similarity metric. The proposed framework is general and may be used together with many existing image registration models. We evaluate it on brain MRI scans from the UK Biobank and show that use of the learnt similarity metric, which is parametrised as a neural network, leads to more accurate results than use of traditional functions, e.g. SSD and LCC, to which we initialise the model, without a negative impact on image registration speed or transformation smoothness. In addition, the method estimates the uncertainty associated with the transformation. The code and the trained models are available in a public repository: https://github.com/dgrzech/learnsim.

1. Introduction

Image registration attempts to align images so that corresponding locations contain the same semantic information. It is a necessary pre-processing step for the statistical analysis of clinical imaging data, computer-aided diagnosis, and computer-assisted intervention. In order to calculate the transformation, traditional image registration methods minimise an energy function which consists of task-specific similarity and regularisation terms, e.g. [4, 27, 38]. The algorithm needs to be run independently for every pair of images to be aligned and optimisation of the energy function is performed in an iterative manner.

Traditional image registration methods minimise an energy function, which is similar to the training of neural networks by the minimisation of loss functions. However, using deep learning for medical image registration is difficult due to the lack of ground truth transformations. DLIR [15] and VoxelMorph (VXM) [5, 6, 13, 14] both use neural networks in order to learn in an unsupervised way a function that outputs a deformation field given a pair of input images, instead of optimising an energy function independently for each image pair. The calculation of transformations by evaluation of the neural network in a single forward pass speeds up the process by several orders of magnitude and maintains an accuracy comparable to traditional methods. The claim that deep learning models for image registration are limited to self- and unsupervised learning was recently countered by training a generative model exclusively on synthetic images and segmentations [25].

In this work, we present a new model for atlas-based diffeomorphic non-rigid image registration which, given a dataset of images, learns in an unsupervised way a suitable similarity metric for the task. The model implements the similarity metric as a neural network that takes as input two three-dimensional images and outputs the value of a function which needs to be minimised to align them. The few existing approaches for unsupervised similarity learning rely either on feature extraction used together with classical similarity metrics [12, 44, 45] or ad-hoc adversarial training [17, 18, 36]. In contrast to them, we refine the similarity metric itself, working within a rigorous Bayesian framework. The choice of a Bayesian model makes it possible to learn a data-specific similarity metric with relatively little data, improves robustness by proxy of the approximate variational posterior of the transformation parameters, and allows to quantify the uncertainty associated with the output. The following are the main contributions of our work:

1. We propose a novel variational Bayesian method for unsupervised similarity learning in atlas-based non-rigid medical image registration;

2. We show that the learnt metric outperforms traditional similarity metrics used in image registration on the examples of SSD and LCC, to which we initialise the model;

3. Furthermore, we show that the learnt metrics generalise well by comparing the accuracy of VXM trained using the baseline and learnt similarity metrics;

4. The proposed formulation also makes it possible to estimate the voxel-wise uncertainty associated with the diffeomorphic transformation.

Related work. State-of-the-art image registration models based on deep learning tend to rely on traditional similarity metrics, e.g. sum of squared differences (SSD) and local cross-correlation (LCC) in case of VXM [5, 6, 13, 14] or LCC and mutual information (MI) in case of DLIR [15, 16]. Deep learning has been used not only to learn a function that maps a pair of input images directly to a deformation field but also to improve image registration accuracy by learning image representations optimised for the task of image registration in a supervised [32] and weakly-supervised [8, 39] setting, a spatially-adaptive regulariser [34], and extracting features which were then used together with traditional similarity metrics [12, 44].

Traditionally, similarity metrics in medical image registration were designed manually rather than learnt, e.g. the modality-independent neighbourhood descriptor for multi-modal CT/MRI registration of the thorax [23]. Learnt metrics were used in *rigid* multi-modal registration for CT/MRI and PET/MRI [31], and for MRI/ultrasound [22]. Similarity learning in a *supervised* setting, which requires costly manual data annotation, was proposed for the registration of T1-T2 MRI brain scans [9], T1-T2 neonatal MRI brain scans [40], and CT/MRI head [11] and prostate scans [10].

The two existing methods for *unsupervised* similarity learning are closely related and used a generative adversarial network, with the discriminator network learning a similarity metric for the training of an image registration model [17, 18, 36]. In order to train the discriminator, they required pre-registered image patches that were generated from the dataset in an ad-hoc way, by defining patches of a weighted sum of the fixed and moving images and of the fixed image as positive samples, and patches of the warped moving image and of the fixed image as negative samples. These choices raise the question of what happens when the input images are similar prior to registration or accurately registered by the model. Moreover, only one of the models guaranteed diffeomorphic transformations [36].

Related non-rigid image registration models were previously adopted for the probabilistic inference of regularisation strength [41], uncertainty quantification [30], and learning a probabilistic model for diffeomorphic registration [28], but not for similarity learning.

2. Method

Background. We denote by $\mathcal{D} = \{(F, M_k) \mid k \in \{1, \ldots, K\}\}$ a dataset of image pairs, where $F \colon \Omega_F \to [0,1]$ and $M_k \colon \Omega_{M_k} \to [0,1]$ are a fixed and a moving image respectively. The aim of mono-modal image registration is to align the underlying domains Ω_F and Ω_{M_k} using a transformation $\varphi\left(w_k\right) : \Omega_F \to \Omega_{M_k}$, i.e. to find parameters w_k such that $F \simeq M_k(w_k) \coloneqq M_k \circ \varphi^{-1}\left(w_k\right)$. The transformation is often expected to possess some desirable properties, e.g. diffeomorphic transformations are smooth and invertible, with a smooth inverse.

We parametrise the transformation using stationary velocity fields (SVFs) [1, 2]. The ordinary differential equation that defines the transformation is given by:

$$\frac{\partial \varphi^{(t)}}{\partial t} = w_k\left(\varphi^{(t)}\right) \tag{1}$$

where $\varphi^{(0)}$ is the identity transformation and $t \in [0,1]$. Under the assumption of a spatially smooth velocity field w_k, the solution to Equation (1) is diffeomorphic [1]. Numerical integration is done by scaling and squaring, which uses the following recurrence relation with 2^T steps:

$$\varphi^{(1/2^{t-1})} = \varphi^{(1/2^t)} \circ \varphi^{(1/2^t)} \tag{2}$$

Mathematical foundation. Throughout registration, the model residuals will include voxel-wise error $e_k \coloneqq F - M_k(w_k)$ due to noise and the misalignment of the images[1]. Therefore, in order to find the registration parameters, in probabilistic image registration we maximise the log-likelihood of the fixed image $\log p\left(F \mid M_k, w_k\right)$ given the moving image and the transformation parameters. The expression for the log-likelihood depends on the assumptions about the distribution of the error, e.g. noise assumed to be independent and identically distributed across image voxels with the normal distribution corresponds to the SSD similarity metric [3]:

$$\log p\left(F \mid M_k, w_k\right) \propto -\frac{1}{2} e_k^\intercal \mathrm{K}^{-1} e_k \tag{3}$$

where K^{-1} is a precision matrix of the error in the image. To regularise the registration, a prior distribution on the transformation parameters w_k is used. The usual choice is a multivariate normal distribution [2, 19, 35]:

$$\log p\left(w_k\right) \propto -\frac{1}{2} \lambda_{\text{reg}} (\mathrm{L} w_k)^\intercal \mathrm{L} w_k \tag{4}$$

where λ_{reg} is the regularisation weight and L is the matrix of a differential operator. In what follows, we assume that L represents the gradient operator, which regularises the magnitude $\|\mathrm{L}w\|^2$ of the 1st derivative of the velocity field.

When using the maximum a posteriori method, a single value of parameters w_k is computed rather than the probability density function, aiming to find the most likely transformation parameters:

$$p\left(w_k \mid \mathcal{D}\right) = p\left(\mathcal{D} \mid w_k\right) \frac{p\left(w_k\right)}{p\left(\mathcal{D}\right)} \tag{5}$$

[1] To simplify notation, we omit the voxel index.

The objective function to be minimised is the negative logarithm, which corresponds to a sum of the log-likelihood and the prior, i.e. similarity and regularisation terms $\mathcal{E}_{\text{sim}}$ and $\mathcal{E}_{\text{reg}}$ respectively:

$$-\log p\left(w_k \mid \mathcal{D}\right) = \underbrace{-\log p\left(F \mid M, w_k\right)}_{\mathcal{E}_{\text{sim}}} \underbrace{-\log p\left(w_k\right)}_{\mathcal{E}_{\text{reg}}} \tag{6}$$

Model. We wish to find a similarity metric which maximises the likelihood of images in the dataset when registering them to the atlas. Let $g_\theta \colon \left(F, M_k\left(w_k\right)\right) \mapsto \mathbb{R}_+$ be a similarity metric, implemented as a neural network with parameters θ. We use a Boltzmann distribution as likelihood:

$$p\left(F \mid M_k, w_k, \theta\right) = \frac{1}{Z(\theta)} \cdot \exp\left(-g_\theta\left(F, M_k\left(w_k\right)\right)\right) \tag{7}$$

where $Z(\theta)$ is a normalisation constant. When using the Boltzmann distribution, maximising the log-likelihood in order to find the transformation parameters is equivalent to minimising the value of the similarity metric in traditional image registration.

In order to calculate the transformation parameters w and the neural network parameters θ, we use variational inference. The posterior distribution of the model parameters $p\left(w, \theta \mid \mathcal{D}\right)$ is approximated as a parametric probability distribution $q\left(w, \theta\right)$. We assume that w and θ are independent, and that w_k are mutually independent. Thus, the approximation of the posterior distribution factorises over the parameters:

$$q\left(w, \theta\right) = q\left(w\right) \cdot q(\theta) = \left\{\prod_{k=1}^{K} q_k\left(w_k\right)\right\} \cdot q\left(\theta\right) \tag{8}$$

We also assume that, for each image, the approximate posterior distribution of the transformation parameters follows a multivariate normal distribution $q_{wk} \sim \mathcal{N}\left(\mu_{wk}, \Sigma_{wk}\right)$, where $\mu_{wk} \in \mathbb{R}^{3N^3 \times 1}$, $\Sigma_{wk} \in \mathbb{R}^{3N^3 \times 3N^3}$ is a positive semi-definite covariance matrix, and N is the number of voxels along each dimension. Due to high dimensionality, the covariance matrix is approximated as a sum of diagonal and low-rank parts, i.e. $\Sigma_{wk} = \operatorname{diag}\left(\sigma_{wk}^2\right) + u_{wk} u_{wk}^\intercal$, with $\sigma_{wk} \in \mathbb{R}^{3N^3 \times 1}$ and $u_{wk} \in \mathbb{R}^{3N^3 \times R}$, where R is a hyperparameter that defines the rank of the parametrisation. This choice of an approximate posterior distribution is standard in image registration but in contrast to other recent models based on SVFs, e.g. [13,28], we use a diagonal + low-rank covariance matrix, rather than just diagonal.

To find the parameters μ_{wk}, Σ_{wk}, and θ, we maximise the evidence lower bound, which fits the model and penalises deviation of parameters from the priors [26]:

$$\begin{aligned}
\mathcal{L}\left(q\right) &= -\int_\theta \int_w q\left(w, \theta\right) \log \frac{q\left(w, \theta\right)}{p\left(\mathcal{D}, w, \theta\right)} \, \mathrm{d}w \, \mathrm{d}\theta \\
&= -\int_\theta \int_w q\left(w, \theta\right) \log \frac{q\left(w, \theta\right)}{p\left(F \mid M, w, \theta\right) p\left(w, \theta\right)} \, \mathrm{d}w \, \mathrm{d}\theta \\
&= \underbrace{\mathbb{E}_q\left[\log p\left(F \mid M, w, \theta\right)\right]}_{-\left\langle \mathcal{E}_{\text{sim}} \right\rangle_q} - \mathrm{D}_{\mathrm{KL}}(q \mid\mid p)
\end{aligned} \tag{9}$$

where $\mathrm{D}_{\mathrm{KL}}(q\left(w, \theta\right) \mid\mid p\left(w, \theta\right))$ is the Kullback-Leibler divergence between the approximate posterior $q\left(w, \theta\right)$ and the prior $p\left(w, \theta\right)$ and $\left\langle \cdot \right\rangle$ denotes the expected value. Similarly to maximum a posteriori, this corresponds to the sum of similarity and regularisation terms, with an additional entropy term $H\left(q\right)$:

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(q \mid\mid p) &= \underbrace{\int_\theta \int_w q\left(w, \theta\right) \log q\left(w, \theta\right) \mathrm{d}w \, \mathrm{d}\theta}_{-H(q)} \\
&- \underbrace{\int_w q\left(w\right) \log p\left(w\right) \mathrm{d}w}_{-\left\langle \mathcal{E}_{\text{reg}} \right\rangle_q} - \int_\theta q\left(\theta\right) \log p\left(\theta\right) \mathrm{d}\theta
\end{aligned} \tag{10}$$

We choose a flat prior on θ, so the gradient of the last term on the RHS in Equation (10) w.r.t. θ is zero. In order to reduce the computational overhead, we also assume that the approximate posterior $q(\theta)$ is the Dirac delta function.

We use contrastive divergence [24] to deal with the intractable normalisation constant $Z(\theta)$ in Equation (7), with $p\left(F \mid M_k, w_k, \theta\right)$ approximated by a multivariate normal distribution $q_F \sim \mathcal{N}\left(F, \Sigma_F\right)$. We have:

$$\begin{aligned}
\frac{\partial \mathcal{L}\left(q\right)}{\partial \theta} &= \frac{\partial Z(\theta)}{\partial \theta} - \left\langle \frac{\partial g_\theta\left(F, M_k\left(w_k\right)\right)}{\partial \theta} \right\rangle_q \\
&\approx \left\langle \frac{\partial g_\theta\left(F, M_k\left(w_k\right)\right)}{\partial \theta} \right\rangle_{q_F} - \left\langle \frac{\partial g_\theta\left(F, M_k\left(w_k\right)\right)}{\partial \theta} \right\rangle_q
\end{aligned} \tag{11}$$

We again assume that the covariance matrix $\Sigma_F = \operatorname{diag}\left(\sigma_F^2\right) + u_F u_F^\intercal$ is diagonal + low-rank, with $\sigma_F^2 \in \mathbb{R}^{N^3 \times 1}$, and $u_F \in \mathbb{R}^{N^3 \times R}$, which makes it easy to sample from the likelihood distribution when training the model.

Training. We optimise in turn parameters of the approximate variational posteriors and of the neural network, starting with the transformation parameters. We use the reparametrisation trick with two samples per update to backpropagate with respect to the parameters of the variational posteriors. For every moving image M_k, we sample $w_k \sim q_{w_k}$:

$$\begin{aligned}
w_k &= \mu_{wk} \pm \left(\operatorname{diag}\left(\sigma_{wk}\right) \cdot \epsilon + u_{wk} \cdot x\right) \\
&\epsilon \sim \mathcal{N}(0, I_{3N^3}),\ x \sim \mathcal{N}(0, I_R)
\end{aligned} \tag{12}$$

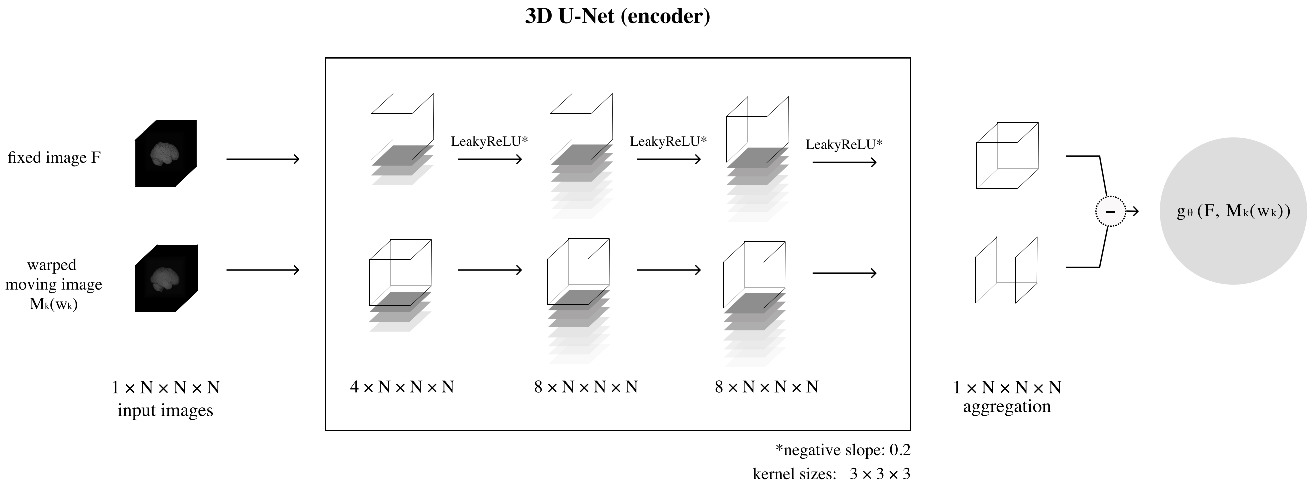


Figure 1. Neural network parametrising the similarity metric initialised to SSD. In case of LCC, we re-use the same architecture, with output of the aggregation layer convolved with a learnable $3 \times 3 \times 3$ kernel whose weights are initialised to one, in order to calculate the local intensity means and variances in the fixed and moving images.

In order to make optimisation less susceptible to local maxima of the loss function, we take advantage of Sobolev gradients [33]. Samples from q_w are convolved with a Sobolev kernel S, approximated for a given size s_{H^1} and value of λ_{H^1} by solving the linear system of equations:

$$(I_{s^3_{H^1}} - \lambda_{H^1}\Delta)S = v \tag{13}$$

where $v \in \mathbb{R}^{s^3_{H^1} \times 1}$ is a discretised Dirac impulse and Δ is the Laplacian matrix, discretised with a 7-point stencil [42]. To lower the computational overhead, we further approximate the three-dimensional kernel by three separable one-dimensional kernels by calculating the tensor higher-order singular value decomposition of S and retaining only the 1st singular vector from each resulting matrix, which is then normalised to unit sum [29, 42].

Initialisation of the similarity metric. The similarity metrics that are commonly used in non-rigid image registration include SSD, LCC, and MI. The function is chosen based on the dataset—SSD is the similarity metric of choice in case of mono-modal images with comparable intensity distributions, LCC is robust to linear intensity scaling and suitable when data was acquired with use of different imaging protocols, and MI is favoured in multi-modal image registration tasks.

Training a similarity metric from scratch would be difficult because a quantitative measure of whether a pair of images is aligned is required to register images to begin with. To solve this problem in a more rigorous way than previous methods for unsupervised similarity learning, we put the focus on functions that are useful in the context of inter-subject mono-modal registration, i.e. SSD and LCC, and initialise the neural networks such that:

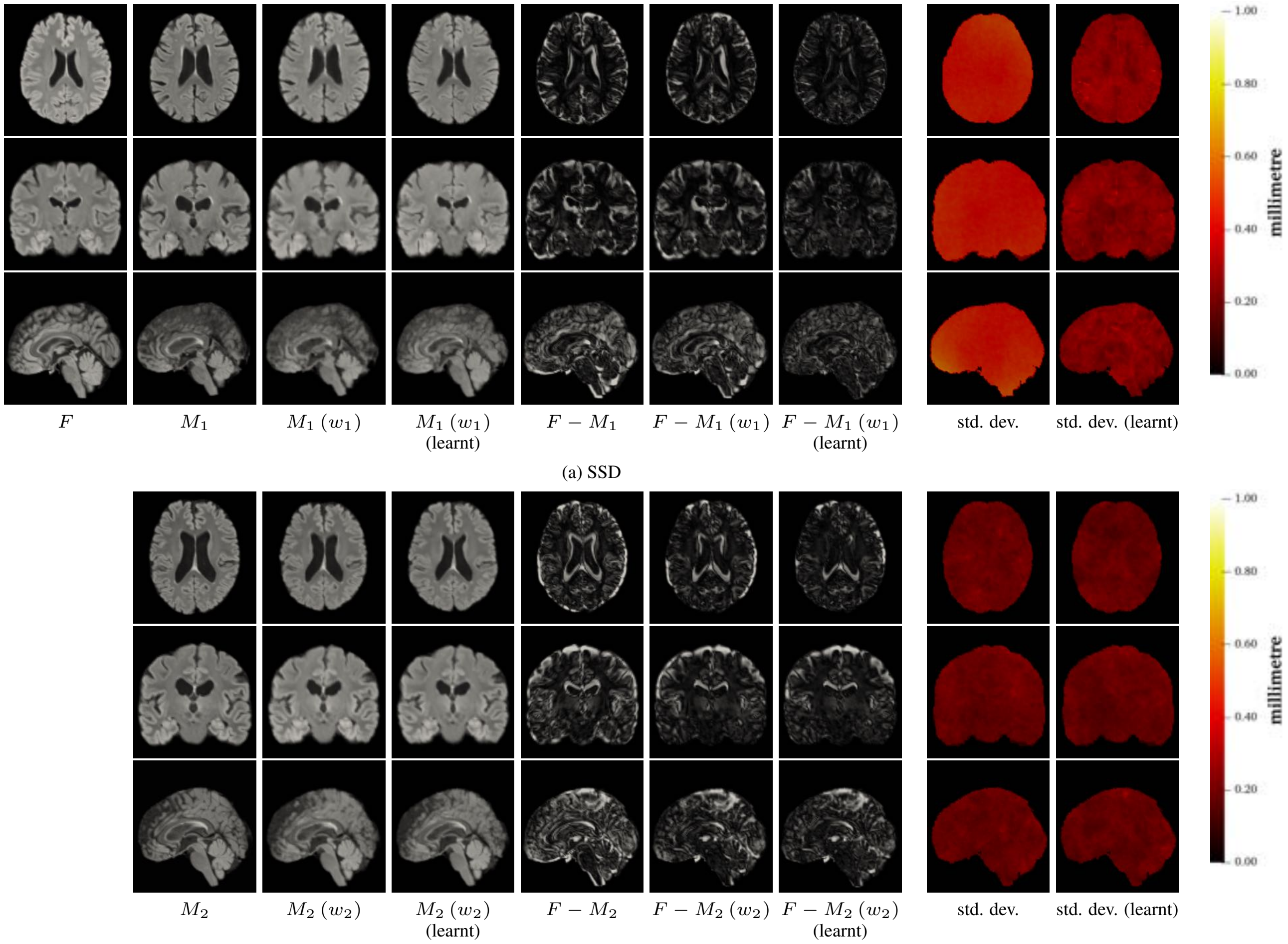
$$g_{\theta_{\text{SSD}}}(F, M) = \frac{1}{2}\|F - M(w)\|^2 \tag{14}$$

$$g_{\theta_{\text{LCC}}}(F, M) = -\frac{1}{2}\left\langle \frac{\widehat{F}}{\|\widehat{F}\|}, \frac{\widehat{M}}{\|\widehat{M}\|} \right\rangle^2 \tag{15}$$

where $\widehat{F}\colon x \mapsto \sum_{x' \in N(x)} {}^{F(x')}/_{n^3}$ is the local intensity mean of an image, $N(x)$ denotes the local neighbourhood of a voxel, and $n = |N(x)|$ is the count of voxels along each dimension inside the local neighbourhood.

For each similarity metric, the initialisation may proceed in two ways—by training in a supervised way a neural network that approximates the value of the chosen similarity metric for a given image pair or, more elegantly, by initialising a neural network in such a way that its output is approximately equal to the similarity metric. In Figure 1, we show the architecture for SSD used in the experiments, which consists of a 3D U-Net encoder [37] initialised to the Dirac delta function and followed by a 1D convolutional layer. Feature maps output by the 3D U-Net are used to compute a weighted sum returned by the aggregation layer. In case of LCC, we re-use this architecture, with its output convolved with a learnable $3 \times 3 \times 3$ kernel ($n^3 = 27$) whose weights are initialised to one, in order to calculate the local intensity means and variances in the fixed image and the moving image.

Neural networks can also be trained to approximate MI [7], so the proposed method is not limited to mono-modal image registration but potentially applicable to multi-modal registration problems as well.



(a) SSD

(b) LCC

Figure 2. The output on two sample images in the test split when using the baseline and the learnt similarity metrics. In case of SSD, the average improvement in DSC over the baseline on the image above is approximately 27.2 percentage points and in case of LCC, it is approximately 6.5 percentage points. The uncertainty estimates are visualised as the standard deviation of the displacement field, based on 50 samples. Use of the learnt similarity metric which was initialised to SSD results in better calibration of uncertainty estimates than in case of the baseline, e.g. higher uncertainty within regions with homogeneous voxel intensities.

3. Evaluation

The model is implemented in PyTorch 1.7.1. We manually select an atlas image without white matter hyperintensities and use a random sample of 1,500 moving images from the 13,401 three-dimensional T2-FLAIR MRI brain scans in the UK Biobank dataset [43]. 80% of the images are used for training and 20% for testing. The input is pre-registered with the affine component of *drop2* [20] and then resampled to $N = 128$ isotropic voxels of length $1.82\,\text{mm}$ along each dimension.

In order to show that the learnt metrics generalise well, we also train VXM using the baseline and learnt similarity metrics[2]. VXM is trained on a random 80/20 split of the whole UK Biobank dataset, using the same atlas image as our models. To make the comparison fair, we set the hyperparameters of VXM trained with the baseline similarity metrics to ensure diffeomorphic transformations, and then use the same hyperparameter values, including the regularisation weight that determines the transformation smoothness, for training VXM with the learnt similarity metrics.

Implementation details. We determined $\lambda_{\text{reg}} = 1.8$ for SSD and $\lambda_{\text{reg}} = 2.8$ for LCC to be the minimum values of regularisation weight which guaranteed diffeomorphic transformations. The integration of SVFs is done in 2^{12} steps. In order to start training in a stable way with small displacements, we set the rank hyperparameter to $R = 1$ and initialise μ_w to zero, σ_w to half a voxel in every direction, and u_w to a tenth of a voxel in every direction. We ob-

[2] The official VXM implementation is available on GitHub: https://github.com/voxelmorph/voxelmorph.

served that a variational posterior of transformation parameters with only a diagonal covariance matrix is too restrictive for accurate image registration. Moreover, use of the rank parameter set to $R \geq 2$ in the diagonal + low-rank approximation is not possible due to constraints on GPU memory. The parameters of q_F are initialised to $\sigma_F = u_F = 0.1$. We set the Sobolev kernel width to $s_{H^1} = 7$ and the smoothing parameter to $\lambda_{H^1} = 0.5$. We use the Adam optimiser with a step size of 1×10^{-1} and 2×10^{-2} for the variational posterior q_w respectively in case of SSD and LCC, 1×10^{-3} for q_F, and 1×10^{-5} for θ. Between every update to q_F and θ, we run 1,024 and 1,344 updates to the variational parameters of q_w respectively in case of SSD and LCC, which is sufficient for convergence.

We run training of each similarity metric for 5 epochs. It takes approximately 6 days on a system with an Intel Core i9-1090X CPU, 128 GB RAM, and two GeForce RTX 3090 GPUs, and requires 4 GB of memory per image in a mini-batch. The registration of one image takes approximately 1 to 3 min, depending on the similarity metric (cf. Table 2).

Results. First, we show that, *ceteris paribus*, our trained models outperform the baseline SSD and LCC similarity metrics. To register images in the test split, we calculate the variational posteriors of transformation parameters using the same number of iterations and the same initialisation as during training. The neural network parameters are held constant. In case of our model, we sample five transformations for every image in the test split, which gives a total of 1,500 samples. VXM is deterministic, so the results related to it are based on a single transformation per image, i.e. a total of 2,679 samples. In Figures 2 and 3, we show the result on four MRI brain scans in the test split for models initialised to SSD and LCC, as well as for VXM trained with the baseline and the learnt similarity metrics. The improvement over image registration with the baseline similarity metrics is clearly visible.

In Figure 4, we report the average surface distances (ASDs) and Dice scores (DSCs) of subcortical structure segmentations for the two baseline and learnt similarity metrics, and for VXM trained with the baseline and learnt similarity metrics. For the majority of subcortical structures, image registration with the learnt similarity metrics yields consistently better ASDs and DSCs. We observe an average increase in DSC of 4.1 percentage points per structure in case of SSD, 0.6 percentage points in case of LCC, 2.0 percentage points in case of VXM + SSD, and 1.5 percentage points in case of VXM + LCC. There is a corresponding average decrease in ASD of 0.1 mm per structure in case of SSD, 0.01 mm in case of LCC, 0.06 mm in case of VXM + SSD, and 0.06 mm in case of VXM + LCC. We performed one-tailed Welch's t-tests at the 0.05 significance level to determine if the improvement in accuracy over the baseline models is statistically significant. We found that this was the case in terms of ASD for $12/15$ structures for SSD, $10/15$ for LCC, $10/15$ for VXM + SSD, and $14/15$ for VXM + LCC, and in terms of DSC for $13/15$ structures for SSD, $11/15$ for LCC, $10/15$ for VXM + SSD, and $15/15$ for VXM + LCC.

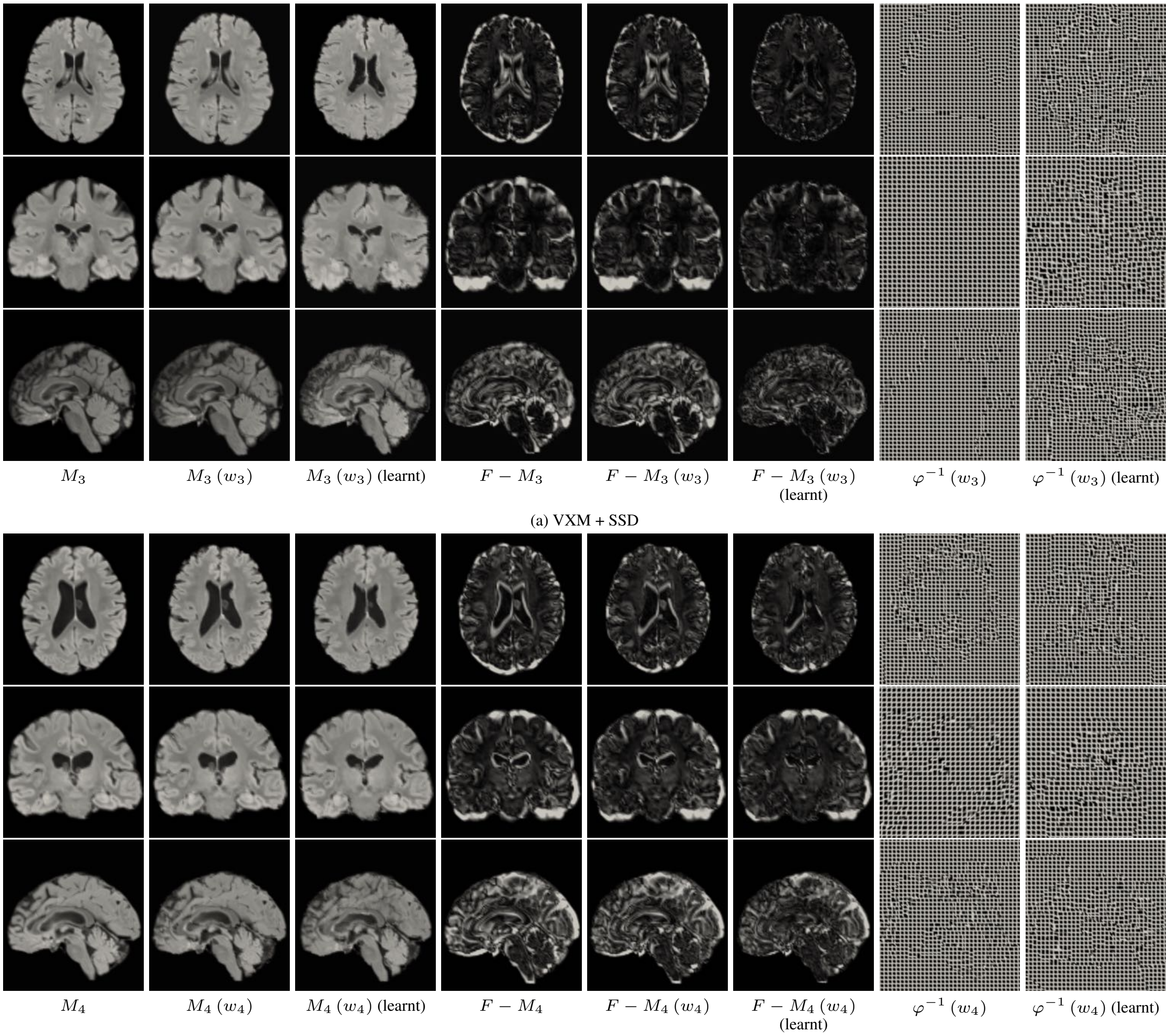
method	$\lvert \det J_{\varphi^{-1}} \rvert \leq 0$	% ($\times 10^{-5}$)
baseline (SSD)	0.00 (0.00)	0.00 (0.00)
learnt	0.10 (0.39)	0.00 (0.00)
baseline (LCC)	0.00 (0.04)	0.00 (0.00)
learnt	0.00 (0.00)	0.00 (0.00)
VXM + SSD	0.00 (0.00)	0.00 (0.00)
VXM + learnt	0.03 (0.38)	0.00 (0.00)
VXM + LCC	0.00 (0.00)	0.00 (0.00)
VXM + learnt	0.00 (0.00)	0.00 (0.00)

Table 1. Mean and standard deviation of the number and percentage of voxels where the sampled transformation is not diffeomorphic on the test data. The methods produce only a small number of voxels where the sampled transformations are not diffeomorphic. The use of learnt similarity metrics does not have a negative impact on the transformation smoothness.

Because it is trivial to improve accuracy at the expense of smoothness, e.g. by lowering the regularisation weight, it is also necessary to show that the proposed method does not have a negative impact on transformation smoothness. To do this, we count the number of voxels where the sampled transformations are not diffeomorphic, i.e. where the Jacobian determinant of the sampled transformation is non-positive, denoted by $\lvert \det J_{\varphi^{-1}} \rvert \leq 0$. In Table 1, we report the mean and the standard deviation of the values. The statistics are nearly zero for both the baseline and the learnt similarity metrics. The learnt similarity metric which was initialised to LCC not only improves accuracy but also reduces the count of voxels with a non-positive determinant of the transformation Jacobian. There is no evidence that the learnt similarity metrics have a negative impact on smoothness of the transformations.

4. Discussion

It is difficult to interpret the data-specific similarity metrics but the fact that VXM trained with the learnt similarity metrics is more accurate than VXM trained with the baseline functions indicates that the model does learn meaningful features. There are other methods to improve probabilistic image registration that cause a lower computational overhead than the proposed method, e.g. virtual decimation [21, 41]. However, unlike virtual decimation, the proposed method is not specific to SSD and could be easily integrated with many existing deterministic and probabilistic image registration algorithms. Moreover, the ideas for



(a) VXM + SSD

(b) VXM + LCC

Figure 3. The output on two sample images in the test split when using VXM with the baseline and the learnt similarity metrics. In order to make the comparison fair, we use the exact same hyperparameter values for VXM trained with the baseline and the learnt similarity metrics. We also use the same atlas image as in Figure 2. In case of VXM + SSD, the average improvement in DSC over the baseline on the image above is approximately 25.3 percentage points and in case of LCC, it is approximately 11.8 percentage points.

unsupervised similarity learning that we present are not limited to medical image registration.

Limitations. Training of the similarity metric initialised to SSD does not converge despite the promising results, due to the negative samples used in Equation (11) which make the optimisation numerically unstable. The method is slower than image registration models based on deep learning that use neural networks to map a pair of input images directly to a deformation field but it can be seamlessly integrated with them. Finally, use of a fixed regularisation strength leads to sub-optimal accuracy in case of large datasets, so the proposed method would also benefit from a reliable way to infer regularisation strength for a given pair of images.

Moral, political, and societal issues. Everyday use of non-rigid image registration in medical image analysis remains in the distant future but accurate image registration models could be employed both to help and to further disad-

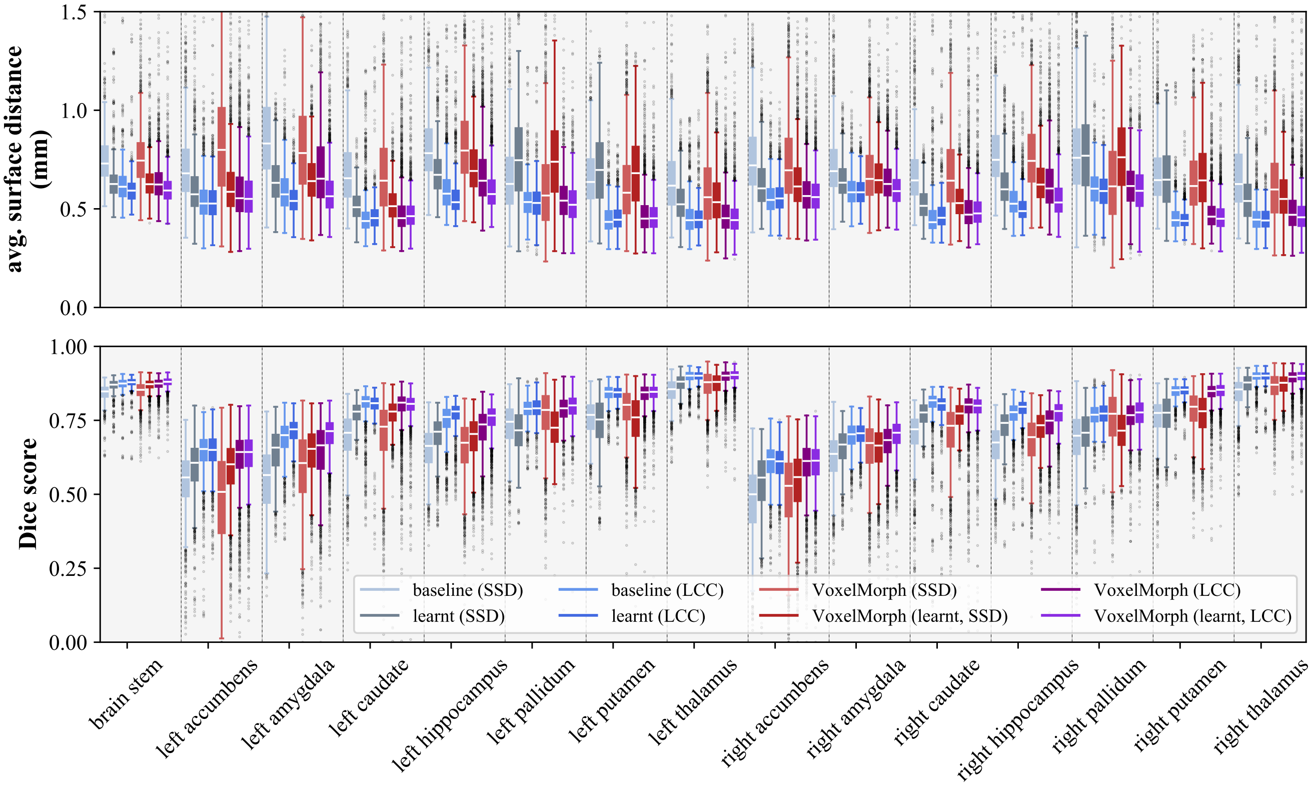


Figure 4. ASDs and DSCs calculated on the subcortical structure segmentation when aligning images in the test split using the baseline and learnt similarity metrics. For the probabilistic methods, we use five samples per image, which results in a total of 1,500 samples. The learnt models show clear improvement over the baselines. On average, when comparing different methods, DSC increases in the range of 0.6 to 4.1 percentage points and ASD decreases in the range of 0.01 to 0.1 mm. We provide details on the statistical significance of the improvement in the main text.

method	training time	registration time
baseline (SSD)	—	1 min
learnt (SSD)	6 d	1 min
baseline (LCC)	—	2 min
learnt (LCC)	6.5 d	3 min
VXM + SSD	1.5 d	<1 sec
VXM + learnt	1.5 d	<1 sec
VXM + LCC	1.5 d	<1 sec
VXM + learnt	1.5 d	<1 sec

Table 2. Training and image registration time for a single image. Use of the learnt similarity metrics does not result in a large increase in the training time of the models or the inference time. The baseline methods do not require training.

vantage marginalised groups, e.g. by comparing individuals to healthy populations and to deny them access to healthcare services based on the result. For this reason, care needs to be taken to ensure that the benefits of fast and accurate image registration, such as democratisation of access to specialised healthcare and knowledge gained through medical imaging population studies, outweigh the political and societal costs.

5. Conclusion

In this paper we presented a new method for unsupervised similarity learning in medical image registration. We showed on the examples of SSD and LCC that it can significantly improve the result of image registration without a negative impact on the transformation smoothness. We also showed that the data-specific similarity metrics generalise well and may be used together with other existing image registration models to improve accuracy.

Acknowledgments. We would like to thank Rhea Jiang from the Harvard Graduate School of Design for the figures. This research used UK Biobank resources ID 12579 and was supported by EPSRC EP/L015226/1, EP/P023509/1, EP/S013687/1, and the UKRI London Medical Imaging & Artificial Intelligence Centre for Value-Based Healthcare.

References

[1] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *MICCAI*, 2006. 2

[2] John Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1), 2007. 2

[3] John Ashburner and Karl J. Friston. Nonlinear spatial normalization using basis functions. *Human Brain Mapping*, 7(4), 1999. 2

[4] Brian B. Avants, Nicholas J. Tustison, Michael Stauffer, Gang Song, Baohua Wu, and James C. Gee. The insight toolkit image registration framework. *Frontiers in Neuroinformatics*, 8, 2014. 1

[5] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. An unsupervised learning model for deformable medical image registration. In *CVPR*, 2018. 1, 2

[6] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE TMI*, 38(8), 2019. 1, 2

[7] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *ICML*, 2018. 4

[8] Max Blendowski, Lasse Hansen, and Mattias P. Heinrich. Weakly-supervised learning of multi-modal features for regularised iterative descent in 3d image registration. *MedIA*, 67, 2021. 2

[9] Michael M. Bronstein, Alexander M. Bronstein, Fabrice Michel, and Nikos Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, 2010. 2

[10] Xiaohuan Cao, Jianhuan Yang, Li Wang, Zhong Xue, Qian Wang, and Dinggang Shen. Deep learning based inter-modality image registration supervised by intra-modality similarity. In *MLMI MICCAI*, 2018. 2

[11] Xi Cheng, Li Zhang, and Yefeng Zheng. Deep similarity learning for multimodal medical images. In *MICCAI Workshop on Deep Learning in Medical Image Analysis*, 2018. 2

[12] Steffen Czolbe, Oswin Krause, and Aasa Feragen. DeepSim: Semantic similarity metrics for learned image registration. In *Medical Imaging Meets NeurIPS*, 2020. 1, 2

[13] Adrian V. Dalca, Guha Balakrishnan, John Guttag, and Mert R. Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *MICCAI*, 2018. 1, 2, 3

[14] Adrian V. Dalca, Guha Balakrishnan, John Guttag, and Mert R. Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *MedIA*, 57, 2019. 1, 2

[15] 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. *MedIA*, 52, 2019. 1, 2

[16] Bob D. de Vos, Bas H.M. van der Velden, Jörg Sander, Kenneth G.A. Gilhuijs, Marius Staring, and Ivana Išgum. Mutual information for unsupervised deep learning image registration. In *Medical Imaging 2020: Image Processing*. International Society for Optics and Photonics, 2020. 2

[17] Jingfan Fan, Xiaohuan Cao, Qian Wang, Pew-Thian Yap, and Dinggang Shen. Adversarial learning for mono-or multi-modal registration. *MedIA*, 58, 2019. 1, 2

[18] Jingfan Fan, Xiaohuan Cao, Zhong Xue, Pew-Thian Yap, and Dinggang Shen. Adversarial similarity network for evaluating image alignment in deep learning based registration. In *MICCAI*, 2018. 1, 2

[19] James C. Gee and Ruzena K. Bajcsy. Elastic matching: Continuum mechanical and probabilistic analysis. In *Brain Warping*, volume 2. Elsevier, 1998. 2

[20] Ben Glocker, Nikos Komodakis, Georgios Tziritas, Nassir Navab, and Nikos Paragios. Dense image registration through mrfs and efficient linear programming. *MedIA*, 12(6), 2008. 5

[21] Adrian R. Groves, Christian F. Beckmann, Steve M. Smith, and Mark W. Woolrich. Linked independent component analysis for multimodal data fusion. *NeuroImage*, 54(3), 2011. 6

[22] Grant Haskins, Jochen Kruecker, Uwe Kruger, Sheng Xu, Peter A. Pinto, Brad J. Wood, and Pingkun Yan. Learning deep similarity metric for 3d mr–trus image registration. *International Journal of Computer Assisted Radiology and Surgery*, 14(3), 2019. 2

[23] Mattias P. Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V. Gleeson, Michael Brady, and Julia A. Schnabel. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. *MedIA*, 16(7), 2012. 2

[24] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8), 2002. 3

[25] Malte Hoffmann, Benjamin Billot, Juan E. Iglesias, Bruce Fischl, and Adrian V. Dalca. Learning mri contrast-agnostic registration. In *ISBI*, 2021. 1

[26] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37(2), 1999. 3

[27] Stefan Klein, Marius Staring, Keelin Murphy, Max A. Viergever, and Josien P.W. Pluim. Elastix: a toolbox for intensity-based medical image registration. *IEEE TMI*, 29(1), 2009. 1

[28] Julian Krebs, Hervé Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE TMI*, 38(9), 2019. 2, 3

[29] Joseph B. Kruskal. Rank, decomposition, and uniqueness for 3-way and N-way arrays. *Multiway data analysis*, 1989. 4

[30] Loïc Le Folgoc, Herve Delingette, Antonio Criminisi, and Nicholas Ayache. Quantifying registration uncertainty with sparse Bayesian modelling. *IEEE TMI*, 36(2), 2016. 2

[31] Daewon Lee, Matthias Hofmann, Florian Steinke, Yasemin Altun, Nathan D. Cahill, and Bernhard Schölkopf. Learning similarity measure for multi-modal 3d image registration. In *CVPR*, 2009. 2

[32] Matthew C.H. Lee, Ozan Oktay, Andreas Schuh, Michiel Schaap, and Ben Glocker. Image-and-spatial transformer networks for structure-guided image registration. In *MICCAI*, 2019. 2

[33] John Neuberger. *Sobolev gradients and differential equations*. Springer, 1997. 4

[34] Marc Niethammer, Roland Kwitt, and François-Xavier. Metric learning for image registration. In *CVPR*, 2019. 2

[35] Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *IJCV*, 66(1), 2006. 2

[36] Chen Qin, Bibo Shi, Rui Liao, Tommaso Mansi, Daniel Rueckert, and Ali Kamen. Unsupervised deformable registration for multi-modal images via disentangled representations. In *IPMI*, 2019. 1, 2

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[38] Julia A. Schnabel, Daniel Rueckert, Marcel Quist, Jane M. Blackall, Andy D. Castellano-Smith, Thomas Hartkens, Graeme P. Penney, Walter A. Hall, Haiying Liu, Charles L. Truwit, F. Gerritsen, D. Hill, and D. Hawkes. A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In *MICCAI*, 2001. 1

[39] Alireza Sedghi, Lauren J. O'Donnell, Tina Kapur, Erik Learned-Miller, Parvin Mousavi, and William M. Wells III. Image registration: Maximum likelihood, minimum entropy and deep learning. *MedIA*, 69, 2021. 2

[40] Martin Simonovsky, Benjamín Gutiérrez-Becker, Diana Mateus, Nassir Navab, and Nikos Komodakis. A deep metric for multimodal registration. In *MICCAI*, 2016. 2

[41] Ivor J.A. Simpson, Julia A. Schnabel, Adrian R. Groves, Jesper L.R. Andersson, and Mark W. Woolrich. Probabilistic inference of regularisation in non-rigid registration. *NeuroImage*, 59(3), 2012. 2, 6

[42] Miroslava Slavcheva, Maximilian Baust, and Slobodan Ilic. SobolevFusion: 3D reconstruction of scenes undergoing free non-rigid motion. In *CVPR*, 2018. 4

[43] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, and Others. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3), 2015. 5

[44] Guorong Wu, Minjeong Kim, Qian Wang, Brent C. Munsell, and Dinggang Shen. Scalable high-performance image registration framework by unsupervised deep feature representations learning. *IEEE Transactions on Biomedical Engineering*, 63(7), 2015. 1, 2

[45] Junshen Xu, Eric Z. Chen, Xiao Chen, Terrence Chen, and Shanhui Sun. Multi-scale neural ODEs for 3D medical image registration. In *MICCAI*, 2021. 1