Deep-diffeomorphic networks for conditional brain templates

Luke Whitbread[1,2,3*], Stephan Lau[1,2,3], Lyle J Palmer[1,4] and Mark Jenkinson[1,2,3]
for the Alzheimer's Disease Neuroimaging Initiative[†]
{luke.whitbread; stephan.lau; lyle.palmer; mark.jenkinson}@adelaide.edu.au

[1]Australian Institute for Machine Learning (AIML), The University of Adelaide
[2]South Australian Health and Medical Research Institute (SAHMRI)
[3]School of Computer and Mathematical Sciences, The University of Adelaide
[4]School of Public Health, The University of Adelaide

Abstract

Deformable brain templates are an important tool in many neuroimaging analyses. Conditional templates have advantages over single population templates by enabling improved registration accuracy and capturing common processes in brain development and degeneration. Conventional methods require large, evenly-spread cohorts to develop conditional templates, limiting their ability to create templates that could reflect richer combinations of clinical and demographic variables. More recent deep-learning methods, which can infer relationships in very high dimensional spaces, open up the possibility of producing conditional templates that are jointly optimised for these richer sets of conditioning parameters. We have built on recent deep-learning template generation approaches using a diffeomorphic framework to create a purely geometric method of conditional template construction that learns diffeomorphisms between: (i) a global or group template and conditional templates, and (ii) conditional templates and individual brain scans. We evaluated our method, as well as other recent deep-learning approaches, on a dataset of cognitively normal participants from the Alzheimer's Disease Neuroimaging Initiative (ADNI), using age as the conditioning parameter of interest. We assessed the effectiveness of these networks at capturing age-dependent anatomical differences. Our results demonstrate that while the assessed deep-learning methods have a number of strengths, they require further refinement to capture morphological changes in ageing brains with an acceptable degree of accuracy. The volumetric output of our method, and other recent deep-learning approaches, across four brain structures (grey matter, white matter, the lateral ventricles and the hippocampus), was measured and showed that although each of the methods captured some changes well, each method was unable to accurately track changes in all of the volumes. However, our method was able to produce T1-weighted conditional templates with high spatial fidelity and with consistent topology as age varies, making these conditional templates advantageous for spatial registrations. The use of diffeomorphisms in these deep-learning methods represents an important strength of these approaches, as it produces conditional templates that can be explicitly linked across age as well as to fixed, unconditional templates or brain atlases. The use of deep-learning in conditional template generation provides a framework for creating templates for more complex conditioning parameters, such as pathologies and demographic variables, in order to facilitate a broader application of conditional brain templates in neuroimaging studies.

*Corresponding author: luke.whitbread@adelaide.edu.au

[†]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

1 Introduction

Magnetic Resonance Imaging (MRI) has evolved into an indispensable tool for the study of the human brain, providing insights into its structural and functional architecture without the need for invasive procedures [1]. For instance, as the brain ages, it undergoes various structural changes, such as reductions in grey-matter volume [2]. MRI's ability to provide clear, detailed and differentiated images of soft tissues makes it well suited for obtaining insights into ageing processes in the brain and the progression of neurodegenerative diseases.

Deformable templates are reference images that can undergo geometric deformations to match participant scans contained in a dataset, and hence allow analyses of geometric variability [3]. Templates are utilised in a number of fields, including: (i) computer vision [4, 5, 6], (ii) graphics [7, 8], and medical image analysis [9, 10, 11, 12]. Brain templates and atlases are important tools for neuroimaging analyses, with applications including the characterisation of structural changes across the lifespan and the identification of pathological changes in the brain [13]. Conditional brain templates are tailored to reflect a particular set of conditioning parameters and have an important role in characterising changes in the brain that are associated with various factors such as age, environmental exposures and a range of other demographic, epidemiological, and clinical variables [14].

Conditional templates can aid researchers and clinicians in a variety of neuroimaging analyses, such as understanding how brain structure changes over time, under various interventions, or as disease progresses [15, 16]. By examining various structural measures of interest, including how volumes of specific brain regions (e.g., hippocampus, ventricles) vary as clinical and demographic variables change, researchers and clinicians can gain insight into the patterns of brain development and degeneration [17]. In research settings, it is also useful to have unbiased templates for studying older populations. These templates should be geometrically linked to a common reference template, to which standard atlases are also related. Such linkage is an essential facilitator of anatomical mapping across different templates, especially since it is not always feasible to create study-specific atlases.

Template construction methods have been studied for decades, undergoing many technical developments [18, 3]. These methods include both conventional (e.g., non-linear registration template construction approaches) and, more recently, deep-learning techniques, with deep-learning methods being initially established to focus on conditional factors, such as age; although, other joint factors have also been contemplated, such as gender and disease state [14]. Conventional methods produce templates using iterative processes [3], with templates typically constructed at a population level or for population-matched groups that reflect particular conditions, such as age. Using these methods, an individual template is constructed for each subgroup of interest (e.g., 80–90 year old cognitively normal participants). As such, these methods require huge amounts of data to produce templates that reflect a wide range or combination of clinical and demographic variables. Accordingly, reliance on pre-prepared publicly available population-based templates is widespread. However, it has become evident that population-based templates such as the Montreal Neuroimaging Institute (**MNI-152**) template do not accurately reflect global and regional brain structures among various subpopulations due to sampling biases inherent in the reference data [19]. Prior efforts using conventional methods to produce templates that reflect conditions such as age or disease state — for instance, the Mayo Clinic Adult Lifespan Template and Atlases [20] — have still suffered from significant limitations as they are often only produced for large age groupings (e.g., every 10 years) or for a particular status (e.g., patients with Alzheimer's disease). In contrast, deep-learning techniques for template construction offer a viable pathway to produce population level and conditional templates in a flexible manner. This could allow for the creation of templates that are conditioned on various combinations of parameters, such as age together with other demographic and clinical variables (e.g., biomarkers of disease progression), that are able to be jointly learned using a single model.

This work evaluates key deep-learning methods that have recently emerged for conditional template construction [3, 14], along with a method that builds on these key works, to establish how well they are able to represent underlying structural changes in the brain when conditioned on age as a demographic variable of interest. More complex sets of conditioning parameters (e.g., age, years of education and disease state) are of interest to neuroimaging analyses, although in the current study we focused on age to provide an evaluation of fundamental deep-learning conditional template construction methods.

1.1 Existing deep-learning methods

Dalca and colleagues have leveraged recent advances in deep-learning registration techniques [21] to produce a template construction method that induces templates by estimating network parameters that optimise a registration between templates and participant scans [3]. Their method is a two-stage network with: (i) the first-stage estimating a conditioned voxel-wise intensity difference map that is used to adjust an average T1-weighted (**T1w**) intensity scan to produce conditional templates; and (ii) the second stage registering conditional templates to T1w participant scans to optimise network parameters. This enables the method to directly learn a function from the data that can jointly optimise templates tailored to specific conditions, such as age.

The method of Dalca *et al.* employs a diffeormorphic deformation framework in the second network stage to produce deformation fields that are smooth and invertible, thus creating registrations that promote topological consistency and preserve biological plausibility. To optimise the network, a global reconstruction loss term is utilised that calculates the mean squared error (**MSE**) between predicted and ground truth participant scans. In addition, various regularisation techniques are applied to the network that encourage deformations to be smooth, minimal, and centred, in order to promote consistency between conditional templates and participant scans. Firstly, a penalty term is applied to the average total magnitude of deformations to encourage deformations to be minimal. Secondly, a penalty term is applied to the first-order spatial derivatives of deformations to support overall spatial smoothness. Finally, a penalty term is applied to voxel-wise mean deformations to promote minimal average deformations across the dataset, which encourages the production of unbiased conditional templates that are centred in the distribution of deformations from templates to participant scans. This mean deformation constraint is designed to act globally across the dataset, where all training scans contribute to a single mean deformation field that is tracked and used to penalise the network.

Yu and colleagues developed a method to produce conditional probabilistic templates based on a single stage deformation network that deforms a group probabilistic template to condition specific templates [14]. This network uses ground truth segmentations of individual participant scans that have been modified to create one-hot encoded channel-wise segmentations. The ground truth channel-wise segmentations are then used to calculate similarities between individual participant scan segmentations and conditional probabilistic templates. Specifically, the loss function was formulated as the Soft-Dice coefficient [22], calculated between condition-specific probabilistic templates and one-hot encoded segmentations of individual participant scans. Similar to the method of Dalca *et al.*, the method utilised a diffeomorphic deformation framework that estimates deformation parameters by scaling and squaring estimated stationary velocity fields. However, the method of Yu *et al.* does not utilise similar regularisation constraints to the method of Dalca *et al.* in order to encourage smooth or minimal deformations, with the loss function being comprised solely of the Soft-Dice coefficient. While this method has various applications, e.g., in conditional shape analyses of various brain regions, additional work is required in order to enable this method to effectively produce conditional structural templates (i.e., T1w conditional template images).

The aim of the current study was to extend the methods of Dalca *et al.* and Yu *et al.* to produce conditional deformable structural T1w templates that are geometrically linked to both a group (unconditional) template and individual participant scans.

2 Methods

2.1 Data

A key consideration when preparing conditional templates using deep-learning methods is the study population used as it directly impacts the generalisability of methods and their effectiveness at accurately representing a target demographic. We focused our study on the production of conditional templates for an ageing cohort, given the applicability of this cohort to the analysis and prognosis of neurodegenerative conditions. We have utilised a cognitively normal subset of the Alzheimer's Disease Neuroimaging Initiative (**ADNI**) study [23, 24] comprising 1534 T1-weighted MRI scans across 515 participants (55.02% female, 44.98% male; 50.5–95.7 years of age at scan date). The training dataset is comprised of 1230 scans, with the validation and testing datasets comprising 152 and 152 scans, respectively. In order to prevent cross-contamination among datasets, each of the datasets was constructed by first partitioning participants into each of the datasets such that any single participant was only represented in one of the training, validation or test datasets.

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimag-

ing Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

2.1.1 Dataset preprocessing

Our preprocessing pipeline included the following steps to address common MRI artefacts and enhance the comparability of scans across individuals:

1. ***Field of view adjustment (FOV)*** to remove lower head and neck voxels.
2. ***Skullstripping*** to remove non-brain voxels.
3. ***Bias-field adjustments*** to remove any bias-field artefacts remaining from acquisition.
4. ***Affine registration*** to a common space.

To focus analysis on the brain, we removed irrelevant sections of MRI scans, including the neck and lower portions of the head, using the *robustfov* tool from FSL [25]. We have then performed skullstripping with Freesurfer's *SynthStrip* tool [26] to further remove non-brain voxels from scans.

The common MRI artefact of bias-field is addressed using FSL's *FAST* tool [25], which is designed to efficiently segment tissue types at the same time as removing within-tissue low-spatial frequency intensity modulations. This has been well validated for the removal of bias-field artefacts.

Lastly, linearly aligning all MRI scans to a common reference space standardised the global orientation and size, which is helpful for many neuroimaging analyses. We have utilised FSL's *FLIRT* tool with 12 DOF for this task [27, 25].

2.2 Conditional template construction

We developed a deep-learning approach to conditional template construction by relying on a two-stage deformation process, combining elements of the approaches presented in [3] and [14], whereby deformations are used to: (i) deform a group T1w template to create conditional template T1w volumes, and (ii) deform conditional templates to participant scans using a deep-learning registration approach. Figure 1 provides a schematic description of our template construction method. A key advantage of our approach is the use of deformations in both network stages, so that conditional templates aim to be diffeomorphic to both the group T1w template and participant scans. Given that diffeomorphisms are invertible and differentiable, a network designed in this way should preserve the structural topology of the group template as it is deformed initially into conditional templates, and then subsequently to the participant scans.

2.2.1 Formulation of deformations

Our method builds upon the work of [3] and [14], but differed in that we utilised a single group T1w template, $T \in \mathcal{R}^{DxHxW}$, a discrete image with dimensions $DxHxW$, and applied deformations at both network stages to produce: (i) conditional templates (in stage 1), $T_{\boldsymbol{\alpha}} \in \mathcal{R}^{DxHxW}$, for conditioning parameters $\alpha_j \in \boldsymbol{\alpha}$ (e.g., age) that are associated with the jth participant $p_j \in \mathcal{P}$, where $\mathcal{P}$ is the set of all participants; and (ii) a deformed version of the conditional template to match participant T1w scans (in stage 2), for each scan, $X_{p_j} \in \mathcal{R}^{DxHxW}$. Deformation fields, $\phi_{\boldsymbol{\alpha}} \in \mathcal{R}^{3xDxHxW}$ and $\phi_{\mathcal{P}|\boldsymbol{\alpha}} \in \mathcal{R}^{3xDxHxW}$, represent the mapping between: (i) a group T1w template and conditional templates, and (ii) conditional templates and participant scans, respectively.

This mapping is defined as follows:

$$T_{\alpha_j}(\mathbf{x}) = T(\mathbf{x}') = T(\phi_{\alpha_j}(\mathbf{x})) \tag{1}$$

$$\hat{X}_{p_j}(\mathbf{x}) = T_{\alpha_j}(\phi_{p_j|\alpha_j}(\mathbf{x})) = T(\phi_{\alpha_j}(\phi_{p_j|\alpha_j}(\mathbf{x}))) \tag{2}$$

where interpolation is used to estimate values for non-integer coordinates.

We followed [28, 29, 30] to construct deformations from stationary velocity fields, $v_{\boldsymbol{\alpha}}$ and $v_{\mathcal{P}|\boldsymbol{\alpha}}$. Here, the deformation field is considered to be the solution of the differential equation:

$$\frac{\partial \phi_v^{(t)}(\mathbf{x})}{\partial t} = v(\phi_v^{(t)}(\mathbf{x})) \tag{3}$$

where t is time and $\phi_v^{(0)}$ represents the identity transformation. Deformation $\phi_v^{(1)}$ is calculated, in practice, by approximating the integration over the unit time interval with a set of scaling and squaring operations using the stationary velocity field vectors [31, 32, 28].

2.2.2 Model estimation

In order to estimate the deformation fields, $\phi_{\boldsymbol{\alpha}}$ and $\phi_{\mathcal{P}|\boldsymbol{\alpha}}$, we followed a similar approach to [3], using a maximum likelihood estimation of stationary velocity field parameters, $v_{\boldsymbol{\alpha}}$ and $v_{\mathcal{P}|\boldsymbol{\alpha}}$, as follows:

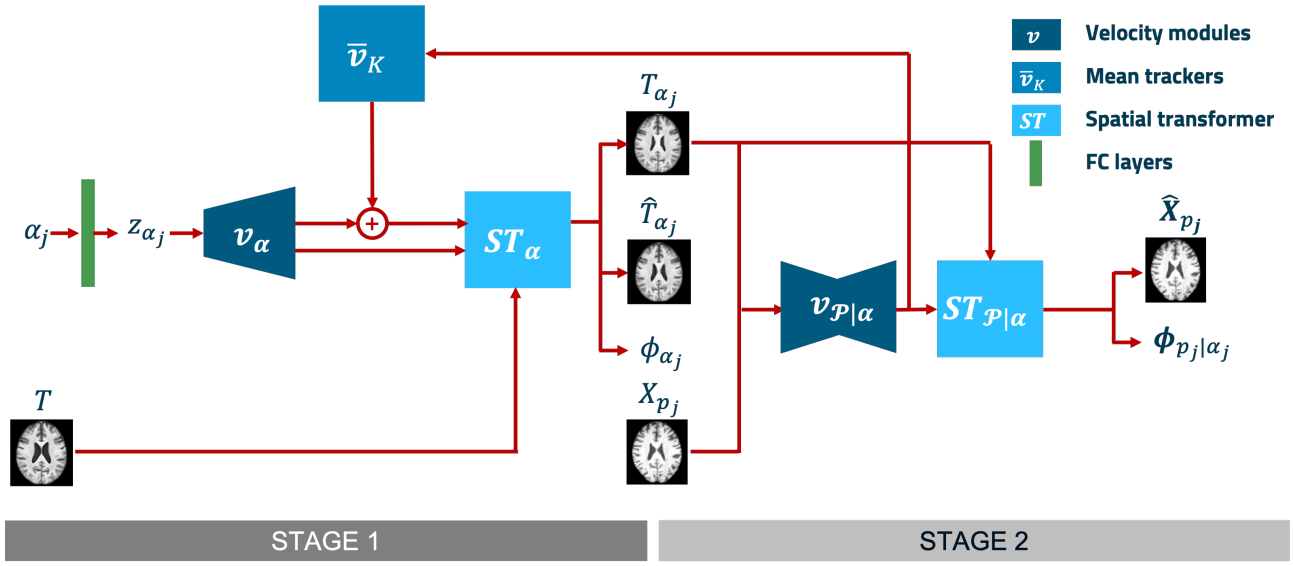

Figure 1: *Network schematic for our two stage conditional template construction network.* ***Stage 1****: conditional parameters are passed to fully-connected (FC) layers to parameterise a latent distribution (z_{α_j}), which is used to predict a conditional stationary velocity field, v_{α_j}. A spatial transformer layer integrates the conditional velocity field to produce a conditional deformation field, ϕ_{α_j} and deform the group template, T, to produce a conditional template, T_{α_j}.* ***Stage 2****: stationary velocity fields are predicted using a U-net architecture that ingests an individual participant scan, X_{p_j}, and conditional template, T_{α_j}, to produce a second stationary velocity field, $v_{p_j|\alpha_j}$. A second spatial transformer layer is then used to integrate $v_{p_j|\alpha_j}$ to produce a deformation field, $\phi_{p_j|\alpha_j}$, and deform the conditional template, T_{α_j}, to predict a participant scan, $\hat{X}_{p_j}$.* ***Mean-tracking****: in addition to the above, velocity fields from stage 2 of the network are tracked by mean-tracking modules ($\bar{v}_K$ in above schematic) in accordance with equations 6–7. Tracked mean velocity fields, $\bar{v}_k$, are combined with predicted stage 1 velocity fields in accordance with equation 8, which are used to predict a mean-adjusted conditional template, $\hat{T}_{\alpha_j}$.*

$$\begin{aligned}\tilde{v}_{\boldsymbol{\alpha}}, \tilde{v}_{\mathcal{P}|\boldsymbol{\alpha}} &= \underset{v_{\boldsymbol{\alpha}}, v_{\mathcal{P}|\boldsymbol{\alpha}}}{\operatorname{argmax}} \log p(v_{\boldsymbol{\alpha}}, v_{\mathcal{P}|\boldsymbol{\alpha}} | \mathcal{D}, \boldsymbol{\alpha}) \\ &= \underset{v_{\boldsymbol{\alpha}}, v_{\mathcal{P}|\boldsymbol{\alpha}}}{\operatorname{argmax}} \log p(\mathcal{D} | v_{\boldsymbol{\alpha}}, v_{\mathcal{P}|\boldsymbol{\alpha}}; \boldsymbol{\alpha}) \\ &\quad + \log p(v_{\boldsymbol{\alpha}}, v_{\mathcal{P}|\boldsymbol{\alpha}}; \boldsymbol{\alpha})\end{aligned} \tag{4}$$

where the first term is the likelihood for the data, where it is assumed that differences between the optimally warped conditional templates and participant scans have a Gaussian distribution; and the second term represents a prior (typically relating to smoothness) on the velocity fields, and $\mathcal{D}$ is the dataset.

We modelled priors, $p(v_{\boldsymbol{\alpha}}, v_{\mathcal{P}|\boldsymbol{\alpha}})$, similarly to [3], under the assumption that deformations should be smooth, to maintain biological plausibility. In addition, the predicted conditional templates are likely to be geometrically biased towards the group template, which we want to avoid. Hence, we included terms to penalise: (i) deformation magnitude, (ii) deformation bending energy, and (iii) non-zero average deformations from conditional templates to participant scans that are clustered around a particular conditional parameter, such as age. Deformation magnitude and bending energy are calculated for both deformation fields, $\phi_{\boldsymbol{\alpha}}$ and $\phi_{\mathcal{P}|\boldsymbol{\alpha}}$, while non-zero average deformations (from conditional templates to participant scans) are corrected for by producing a second conditional template that corrects for the average component of locally tracked stage 2 deformations (i.e., tracked across instances of similar age). Differences between the two conditional template images are used to encourage stage 1 deformation fields to be centered with respect to local stage 2 deformations.

While we modelled deformation magnitude equivalently to [3], there was a difference in the penalty terms that we used to encourage smoothness and minimal average deformations from conditional templates to participant scans.

2.2.3 Model priors: smoothness

Smoothness was modelled by penalising bending energy as opposed to the first order spatial gradient. Following [28], bending-energy, $\mathcal{E}$, can be formulated as:

$$\mathcal{E} = \int_{\Omega} (\Delta\phi)^2 \, d\mathbf{x} \tag{5}$$

where Δ is the Laplacian operator and $\mathbf{x} \in \Omega$ represent spatial coordinates of a scan. This bending-energy formulation is based on the physics of thin elastic sheets or plates, where the optimisation of

this energy minimises deformation curvature rather than penalising first-order spatial gradients. As a consequence, this more directly encourages smoothness, without penalising global changes in the scale or size of brain regions that may be biologically valid representations for various conditional templates or participant scans, while still controlling local deformations in order to preserve anatomical fidelity.

2.2.4 Model priors: minimal average deformation from conditional templates to participant scans

The conditional brain templates should represent the average of the corresponding participant scans in terms of local shapes and locations, and therefore we want the deformations from conditional templates to these corresponding participant scans to have zero mean. In order to achieve this we extended the idea of using global average deformation penalties to local condition-specific penalties. Accordingly, it is necessary to track average deformations from conditional templates to participant scans, at all locations, for all participants in a neighbourhood of a conditional parameter setting, α_j. This poses certain challenges: in order to penalise the network using local average deformation parameters for a range of conditional parameter settings, we need to store and track multiple instances of average deformation fields. For example, if age is the conditioning parameter, we need to track a separate average deformation field for a number of age bins, say each 5 year interval between 55 to 85 years of age. While this is feasible for a simple low-dimensional conditioning parameter space (as with age), when the dimensionality grows, the number of average deformation fields that need to be tracked will grow combinatorially. Additionally, since batch sizes are typically too small to give a useful average when working with 3D images, we need to update average deformation parameters across batches for each of the tracked fields across batches.

Despite these difficulties, we implemented this approach here by creating a partition of the conditional parameter space, $\boldsymbol{\alpha}$, with one average deformation stored per partition. Updates to the averages were apportioned to the nearest two partitions (nearest neighbours) using a linear weighting scheme, similar to approaches used for kernel density estimation. Similarly, we linearly interpolate between stored average deformations when determining the deformation associated with an arbitrary conditioning parameter (e.g., an age that is between bin centres).

Specifically, for this work we partitioned a one-dimensional conditional parameter space (age), $\boldsymbol{\alpha}$, into a number of bins where the bin centres were used to create $k \in K$ stored average deformations, $\bar{v}_{\alpha_k}$ (with average deformations tracked in velocity space). Equations 6–8 describe this simple case.

For a particular α_j, the weighting to the two nearest α_k values ($\alpha_k \leq \alpha_j \leq \alpha_{k+1}$), is determined by:

$$\mu_{\alpha_j}^{(k)} = \frac{|\alpha_{k+1} - \alpha_k| - |\alpha_j - \alpha_k|}{|\alpha_{k+1} - \alpha_k|}, \quad \text{for} \quad \alpha_k \leq \alpha_j \leq \alpha_{k+1} \tag{6}$$

where $\mu_{\alpha}^{(k)}$ and $1 - \mu_{\alpha}^{(k)}$ is the membership weighting for $\bar{v}_{\alpha_k}$ and $\bar{v}_{\alpha_{k+1}}$, respectively, for each α_j parameter setting.

These stored average deformations were used to encourage zero-centred (i.e., unbiased) average deformations from conditional templates to participant scans. One option to do this would be to penalise the average deformation parameters directly, however, we found empirically that this technique is less effective than using average deformation parameters to directly construct a centred conditional template, $\hat{T}_{\alpha_j}$. We then penalised differences between predicted and centred conditional templates, T_{α_j} and $\hat{T}_{\alpha_j}$, respectively.

Average deformations were iteratively tracked during each forward pass of the network, updating stored average velocity fields, $\bar{v}_{\alpha_k}$ and $\bar{v}_{\alpha_{k+1}}$, using membership weighting, $\mu_{\alpha_j}^{(k)}$, as calculated in equation 6, as follows:

$$\begin{aligned} \bar{v}_{\alpha_k} &= (1 - \mu_{\alpha_j}^{(k)}(1-\lambda))\bar{v}_{\alpha_k} + \mu_{\alpha_j}^{(k)}(1-\lambda)v_{p_j|\alpha_j} \\ \bar{v}_{\alpha_{k+1}} &= (1 - (1 - \mu_{\alpha_j}^{(k)})(1-\lambda))\bar{v}_{\alpha_{k+1}} \\ &\quad + (1 - \mu_{\alpha_j}^{(k)})(1-\lambda)v_{p_j|\alpha_j} \end{aligned} \tag{7}$$

where λ is a hyperparameter, $0 < \lambda < 1$, that controls the update weighting (empirically set to 0.9). During each forward pass of the network, we construct centred conditional templates, $\hat{T}_{\alpha_j}$, by combining the regressed stage 1 velocity field, v_{α_j}, with the tracked average deformation velocity fields from stage 2 of our network as follows:

$$\hat{v}_{\alpha_j} = v_{\alpha_j} + \mu_{\alpha_j}^{(k)}\bar{v}_k + (1 - \mu_{\alpha_j}^{(k)})\bar{v}_{k+1} \tag{8}$$

Deformation fields, $\hat{\phi}_{\alpha_j}$, were derived (using equation 3), and applied to generate *mean-adjusted conditional templates*: $\hat{T}_{\alpha_j}(\mathbf{x}) = T(\mathbf{x}') = T(\hat{\phi}_{\alpha_j}(\mathbf{x}))$. Following this, a penalty term was applied to $|T_{\alpha_j} - \hat{T}_{\alpha_j}|$ to encourage the network to predict velocities, and hence deformations and conditional templates, that are centered for conditioning parameters α_j.

2.2.5 Conditional template network

Our method utilised a deep-learning probabilistic generative network in stage 1 to predict conditional stationary velocity fields, $v_{\boldsymbol{\alpha}}$. We used fully connected layers to parameterise latent distributions, $p(z|\boldsymbol{\alpha})$, and then sample instances of $z_{\boldsymbol{\alpha}}$. Convolutional transpose layers were then applied to produce predicted stationary velocity fields, $v_{\boldsymbol{\alpha}}$. Following equation 8, velocity fields, $v_{\boldsymbol{\alpha}}$, are then combined with tracked stage 2 velocity fields to produce $\hat{v}_{\boldsymbol{\alpha}}$. We used a spatial transformer layer [33] that: (i) integrates velocity fields, $v_{\boldsymbol{\alpha}}$ and $\hat{v}_{\boldsymbol{\alpha}}$, to produce deformation fields, $\phi_{\boldsymbol{\alpha}}$ and $\hat{\phi}_{\boldsymbol{\alpha}}$; and (ii) deforms our group template to produce conditional templates, $T_{\boldsymbol{\alpha}}$ and $\hat{T}_{\boldsymbol{\alpha}}$. Following [21], we utilised a U-Net [34] that ingests conditional templates, $T_{\boldsymbol{\alpha}}$, and participant scans, $X_{\mathcal{P}}$, to predict stationary velocity fields, $v_{\mathcal{P}|\boldsymbol{\alpha}}$. A second spatial transformer layer was then used to integrate velocity fields, $v_{\mathcal{P}|\boldsymbol{\alpha}}$, to produce deformation fields, $\phi_{\mathcal{P}|\boldsymbol{\alpha}}$, which were used to deform conditional templates, $T_{\boldsymbol{\alpha}}$, to predict participant scans, $\hat{X}_{\mathcal{P}}$.

Our network was constructed using the PyTorch deep-learning framework [35]. In order to optimise our network, we used the decoupled weight decay regularisation (AdamW) algorithm [36], with an initial learning rate of 10^{-4}, β_1 of 0.9, β_2 of 0.999, and weight decay of 10^{-2}. Our loss function consisted of 6 components: a reconstruction loss that calculates the reconstruction error (MSE) between predicted participant scans and ground truth participant scans ($\mathcal{L}_{\text{recon}}$), the bending energy of deformation fields predicted for stage 1 and 2 of our network ($\mathcal{L}^{(1)}_{\text{bend}}$ and $\mathcal{L}^{(2)}_{\text{bend}}$, respectively), the global mean deformation magnitude predicted for stage 1 and 2 of our network ($\mathcal{L}^{(1)}_{\text{disp}}$ and $\mathcal{L}^{(2)}_{\text{disp}}$, respectively), and reconstruction loss (MSE) between predicted conditional templates, T_{α_j} and mean-adjusted conditional templates, $\hat{T}_{\alpha_j}$ ($\mathcal{L}_{\text{mean}}$). We used the Kornia Filters package [37] to calculate spatial gradients for $\mathcal{L}^{(j)}_{\text{bend}}$. Deformation magnitude ($\mathcal{L}^{(j)}_{\text{disp}}$) was calculated as the mean of the squared voxel-wise deformation magnitude (as used in [3]). Our full loss function was formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \gamma_1\mathcal{L}_{\text{recon}} + \gamma_2\mathcal{L}^{(1)}_{\text{bend}} + \gamma_3\mathcal{L}^{(2)}_{\text{bend}} \\ +\gamma_4\mathcal{L}^{(1)}_{\text{disp}} + \gamma_5\mathcal{L}^{(2)}_{\text{disp}} + \gamma_6\mathcal{L}_{\text{mean}}\end{aligned} \tag{9}$$

where γ_j are weightings applied to each component of $\mathcal{L}_{\text{total}}$.

2.3 Evaluation

Important attributes of conditional templates include:

- topological consistency;
- structural fidelity; and
- accurate volumetric measurements for regions of interest.

We have evaluated these attributes, where reasonable to do so, for our method as well as for the methods of Dalca *et al.* and Yu *et al.*

2.3.1 Topological consistency

The preservation of topology promotes anatomical accuracy and biological plausibility, as adult brains that are free from obvious pathologies share the same essential topology, despite substantial geometric heterogeneity between individual brains. Conditional templates that maintain the brain's topology are more likely to represent plausible anatomical variations, as this should be preserved as conditioning parameters such as age vary.

To measure the topological consistency between different brain representations, the determinant of voxel-wise Jacobian matrices, $J_\phi(j) = \nabla_\phi(j) \in \mathcal{R}^{3\text{x}3}$, was calculated for deformation fields between images (e.g., unconditional and conditional templates). The determinant, $|J_\phi(j)|$, captures the local change in volume, such that deformations that do not preserve local topology have $|J_\phi(j)| \leq 0$. In addition, we show the distribution of the Jacobian determinants across the images to assess the scale of volume changes that each method was able to produce on a voxelwise basis.

2.3.2 Structural similarity

Well-constructed conditional templates should have good structural fidelity and be similar to individual brains (e.g., the shape and intensity of corresponding brain structures should be similar). Conditional templates with similar (adjacent) conditional parameters should exhibit greater structural similarity than those where the conditional parameters differ more.

The structural similarity of adjacent conditional templates can be measured using the Structural Similarity Index Measure (**SSIM**) [38], which extracts and compares three key images features: (i) luminance, $l(\cdot)$; (ii) contrast, $c(\cdot)$; and (iii) structure, $s(\cdot)$. It then combines these to produce a single overall measure: $f(l(\mathbf{x},\mathbf{y}), c(\mathbf{x},\mathbf{y}), s(\mathbf{x},\mathbf{y}))$ for images $\mathbf{x}$ and $\mathbf{y}$ (more details can be found in supplementary materials section S1).

While there should be substantial structural similarity between adjacent templates, these should not

be perfectly similar, as some differences are to be expected (e.g., due to age). There should exist a balance between structural similarity being consistently high between adjacent templates and changes of interest being apparent. We report SSIM before and after non-linear registration is applied, as this can provide an assessment that is less influenced by well-defined geometric changes in brain structure that can be captured and removed by the registration. The diffeomorphic symmetric normalisation algorithm (**SyN**) [39] was used to perform the registrations, with default hyperparameters as provided by the Advanced Normalization Tools (**ANTs**) [40] implementation.

The root mean squared difference (**RMSD**) between voxel intensities is also reported, as it is a straightforward measure (calculated equivalently to RMSE) that is a purely local, intensity-driven similarity measure. RMSD is an effective tool for detecting even small differences in the anatomy or appearance between conditional templates.

2.3.3 Volumetric measurements for regions of interest

Volumes of various regions of interest (**ROIs**) are of considerable interest in understanding developmental and degenerative processes. Therefore, the accuracy and reliability of these volume measures demonstrates whether conditional templates are practically effective at capturing anatomical change as a function of the conditioning parameter(s), such as age. Such validation is important for the adoption of methods in both research and clinical settings.

Volumes for the conditional template ROIs were obtained by applying a segmentation method independently to each template. We also normalised all volumes of a particular ROI by the mean volume of that ROI across all conditional templates, independently for each method. This provides a simpler comparison across methods and different ROIs (i.e., to show proportional changes in average volume), without undue influence from the values at the ends of the conditional parameter range, which are often under-represented and hence poorly estimated.

More specifically, the normalised volume is formulated as follows:

$$V_{\alpha_n}^{\mathrm{NORM},k} = \frac{V_{\alpha_n}^k}{\frac{1}{|N|}\sum_m V_{\alpha_m}^k} \tag{10}$$

where k is an index representing the ROI, and $n, m \in N$ iare indices representing conditional parameters (e.g., age).

Ground truth volumes were also obtained by segmenting the individual training scans. These were grouped into bins, based on conditional parameters values (age), and the average values from each bin were then normalised using equation 10. More specifically, the average normalised volume for training scans in each bin is:

$$V_b^{\mathrm{NORM},k} = \frac{V_b^k}{\frac{1}{|B|}\sum_c V_c^k} \tag{11}$$

where $b, c \in B$ represents the set of bins partitioning the training scans and V_b^k is the average volume for ROI_k in bin b.

These training scans volumes act as a simple unbiased estimate of what the volume should be at each age, and are plotted along with the corresponding standard deviation. This provides a standard for assessing the volumes obtained from each of the methods, to see if they are able to accurately capture morphological variations in brain anatomy.

3 Results

Evaluation criteria, as described in section 2.2, were assessed for our method as well as for the methods of Dalca *et al.* [3] and Yu *et al.* [14]. For the method of Dalca *et al.*, we implemented their conditional template network using the PyTorch deep-learning framework, following [3] and code available from their website[1]. For the method of Yu *et al.*, we utilised the code available from github[2], with adjustments made to their codebase to accommodate our dataset (see section 2.1).

Figure 2 provides an overview of conditional templates produced by our network with conditioning parameters of 55–95 years of age, where changes in the size of the lateral ventricles is clear.

3.1 Topological consistency

Jacobian determinants of deformation fields from unconditional to conditional templates may only be computed for our method and the method of Yu *et al.* The method of Dalca *et al.* does not use a deformation field to produce conditional templates, as it only uses deformation fields to deform conditional templates to participant scans. However, we have also computed Jacobian determinants of deformation fields for conditional templates to predicted participant scans for our method and the method of Dalca

[1] voxelmorph.csail.mit.edu (downloaded 23 Jul 2023)
[2] github.com/evanmy/conditional_deformation (downloaded 1 Nov 2023)

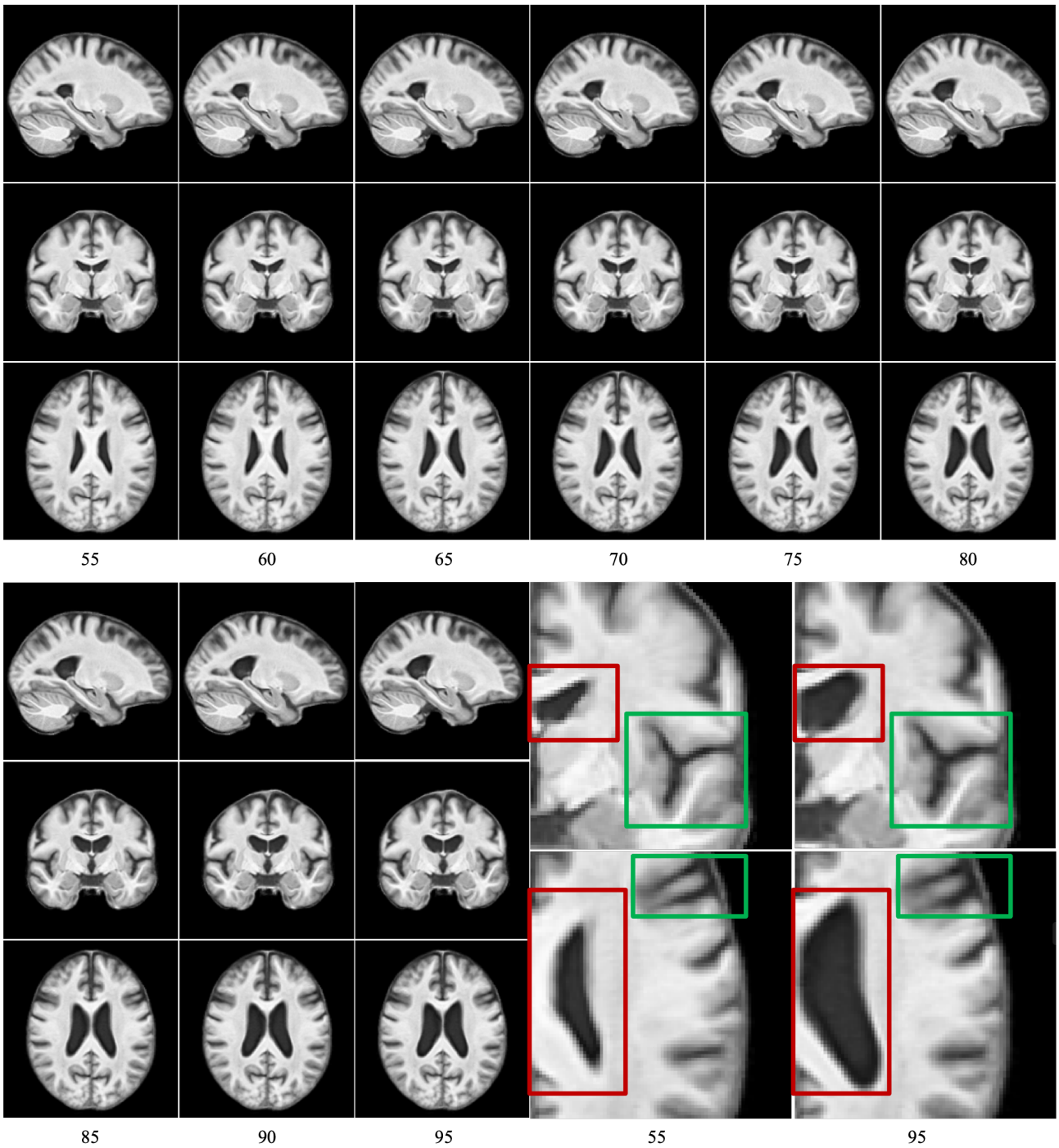

Figure 2: *Conditional templates produced by our network with conditioning parameters of 55–95 years of age (first row: sagittal, second row: coronal, third row: axial). Magnified versions of the coronal and axial slices are shown for the conditioning parameters of 55 and 95 years of age — red and green highlight boxes show differences in the size of the lateral ventricles and width of sulci, respectively.*

et al., given that the method of Yu *et al.* is a one-stage deformation network.

Figure 3 displays distributions of the Jacobian determinants for the stage 1 conditional deformation fields (ours and Yu *et al.*) as well as stage 2 deformation fields (ours and Dalca *et al.*). From these results, it is evident that the deformations produced by our method, in both cases, and for the method of Yu *et al.*, are well centred around 1.0. It is apparent that our method exhibits a greater range of contraction and expansion of voxels, which possibly indicates that our method is less regularised and able to model changes over a wider range more flexibly. Additionally, neither our method nor the method of Yu *et al.* exhibits negative Jacobian determinants, which indicates that these methods can effectively maintain topological consistency through deformations. The method of Dalca *et al.* exhibits a much wider range of Jacobian determinants, with voxel determinant values spanning -2.7 to 22.2. A number of voxels exhibited negative Jacobian determinants within these fields, with an average of 508 voxels for each of the stage 2 deformation fields in the test dataset. This indicates that topological inconsistencies are introduced between the conditional template and predicted participant scans, which could be a result of less regularised registrations or topological alterations in their conditional templates.

3.2 Structural similarity

We calculated SSIM for our method, as well as the method of Dalca *et al.* The method of Yu *et al.* is not amenable to these comparisons as the conditional templates that are produced are conditional probabilistic atlases.

Figure 4 displays SSIM results for the method of Dalca *et al.* and our method for adjacent conditional templates separated by age gaps of 5 years. While SSIM metrics are above 0.95 for each method, on both pre- and post-registration bases, SSIM performance for the method of Dalca *et al.* changes very little with registration and degrades at the tails of the age distribution. Given that the method of Dalca *et al.* generates conditional templates from a difference map, the network is less constrained compared to producing conditional templates through purely geometric deformations of a group template.

Figure 5 displays RMSD results. It should be noted that the RMSD results are reasonably close to zero for each of the methods, which indicates that there is considerable similarity between each of the pairs of conditional templates. However, a low RMSD value may result from negligible differences in every voxel, or more likely, a strong difference in a small number of voxels (e.g., near the edges of various ROIs). Of interest, our method produced higher (less similar) RMSD results when calculated on a pre-registration basis. Such a result may be due to greater differences exhibited by macro-brain-structures (e.g., the lateral ventricles) as the conditioning age parameter is adjusted; which is supported by the post-registration RMSD results, where the registration can remove such geometric changes, being comparable to the RMSD results for the method of Dalca *et al.*

3.3 Volumetric measurements for regions of interest

We focused our volumetric measurements on particular ROIs and tissues that are implicated in ageing and neurodegeneration, namely: (i) hippocampus, (ii) lateral ventricles, (iii) grey-matter, and (ii) white-matter. These measurements provided a benchmark to compare with those reported in the literature [3, 14]. To measure structural volumes for each of the deep-learning methods, we adopted two approaches. For the method of Dalca *et al.* and our method, we have used the FastSurfer segmentation tool [41] to segment and calculate volumes for each conditional template produced by these deep-learning models as trained on the ADNI dataset (described above in section 2.1). For the method of Yu *et al.*, we calculated volumes by thresholding the predicted conditional probabilistic atlases and summed voxels across each channel that reflected a particular ROI; we believe that this is the most direct way to calculate volumes for ROIs as the method outputs conditional probabilistic atlases that are deformed from an unconditional probabilistic atlas, and not intensity-based images suitable for input to FastSurfer. We also used FastSurfer to segment and calculate volumes directly from the training data. For each of the methods and training scans, we then normalised volumes by dividing each volume by the mean of all volumes, across all bins, independently for each ROI and method (see equations 10 and 11).

Figure 6 depicts normalised volumes, for each ROI, from each of the deep-learning methods being evaluated, as well as volumes calculated directly from the training dataset, using bins with a width of 2 years. These results demonstrate that each of the methods performs differently across the various ROIs considered. Both the method of Yu *et al.* and our method produce volumes for the lateral ventricles as a function of age that are more consistent with those obtained from the training data as compared to the method of Dalca *et al.* For white-matter volumes, the

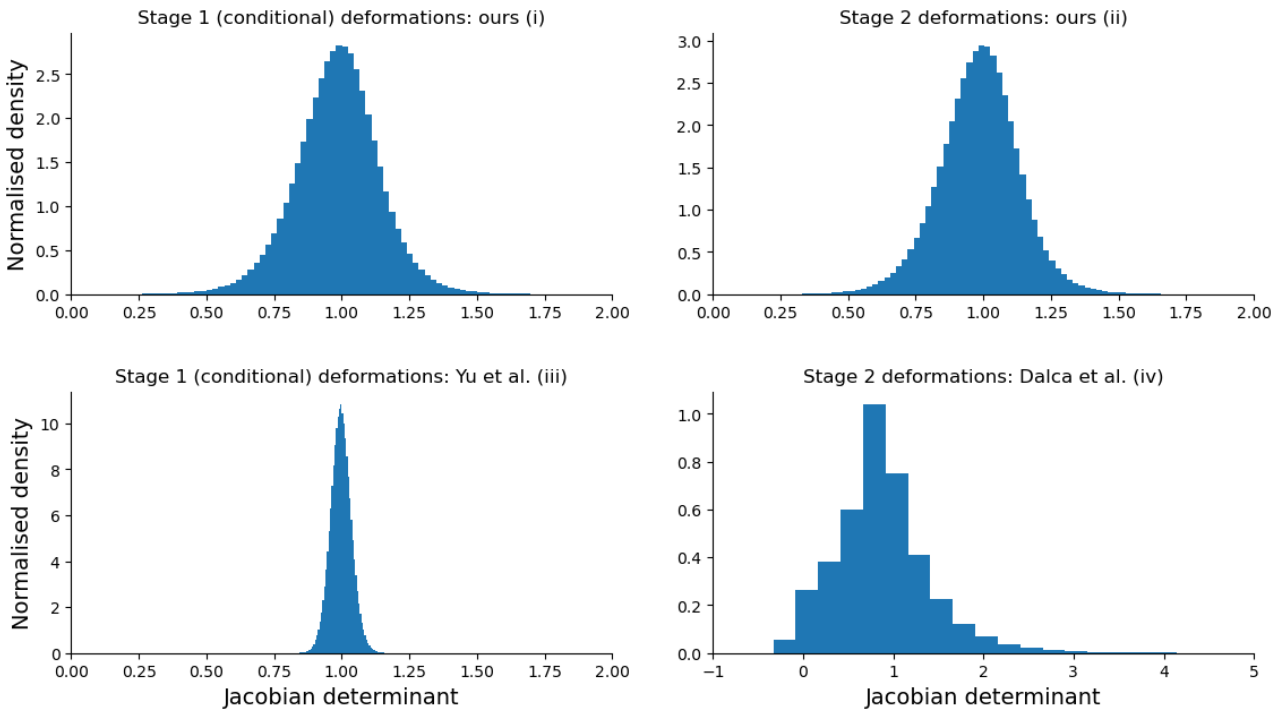

Figure 3: *Distributions of Jacobian determinants for: (i) our method for deformation fields from unconditional to conditional templates, and (ii) conditional templates to predicted participant scans; (iii) deformations from unconditional to conditional templates for the method of Yu et al.; and (iv) deformations from conditional templates to predicted participant scans for the method of Dalca et al.*

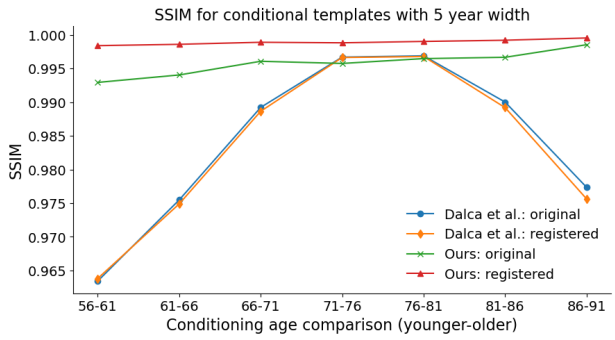

Figure 4: *Structural Similarity Index Measure (**SSIM**) for predicted conditional templates separated by 5 years, on both a pre- (original) and post-registration basis. For the registered images, each lower age conditional template has been registered to the next higher age conditional template, with the registered image and the higher age template used to calculate SSIM.*

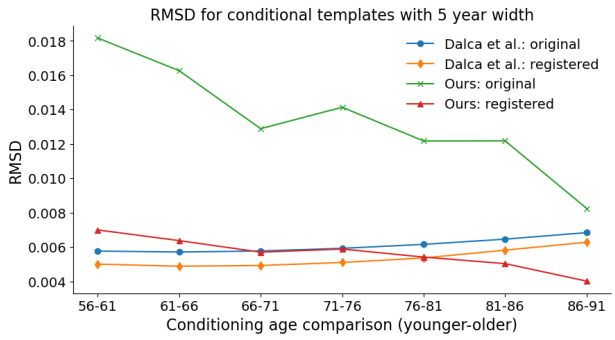

Figure 5: *Root mean squared distance (**RMSD**) for predicted conditional templates separated by 5 years, on both a pre- (original) and post-registration basis. For the registered images, each lower age conditional template has been registered to the next higher age conditional template, with the registered image and the higher age template used to calculate RMSD.*

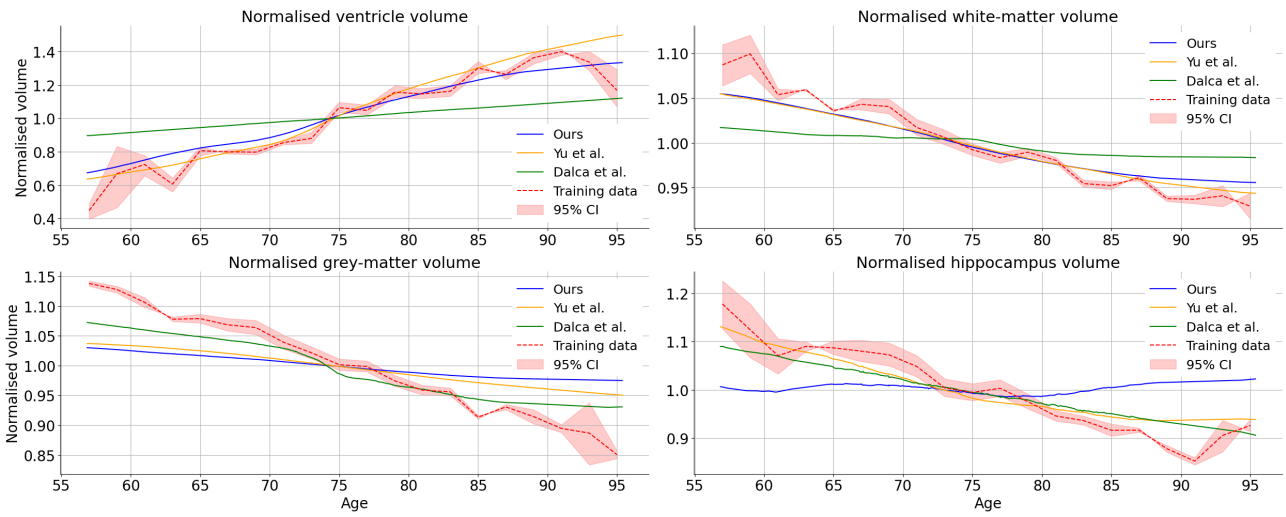

Figure 6: *Volumetric measurements for various ROIs: grey-matter, white-matter, lateral ventricles and hippocampus. Volumes are presented on a normalised basis, where volumes are normalised by the mean of all volumes across the range of conditional parameters, independently for each ROI and method. In addition, a 95% confidence interval is also presented for the empirical estimates made from training data within 2-year age bins.*

method of Dalca *et al.* produces volume differences that are less consistent with the changes exhibited in the training data when compared to the method of Yu *et al.* and our method, although none of the methods display changes of the same magnitude as those measured directly from the training dataset scans. The method of Dalca *et al.* showed better agreement with the training dataset results for grey-matter volume than both the method of Yu *et al.* and our method, although none of these methods show the same degree of change as those measured directly from the training dataset. For hippocampal volumes, our method produced very minimal changes, which are inconsistent with what we would expect to see as age increases. While both the methods of Yu *et al.* and Dalca *et al.* exhibited larger changes, they potentially underestimate the change shown by the volumes directly measured from the training images. However, with low numbers of participants in the higher and lower age bins, the values at the ends of the age range are not as trustworthy as those in the middle of the age range.

4 Discussion

The use of deep-learning methods for the construction of conditional brain templates has many advantages. In order for such methods to gain wider acceptance and provide meaningful utility to both researchers and clinicians, our analysis suggested that they require further refinement. For instance, while methods may be able to capture some morphological differences associated with the ageing adult brain (e.g., volumetric changes in the ventricles), it is also apparent that they do not accurately represent all changes across the brain (e.g., volumetric changes in grey-matter). Moreover, although our method produced visually high quality conditional T1w templates, these failed to capture important ageing characteristics of hippocampal degeneration. Conversely, while the methods of Dalca *et al.* and Yu *et al.* were able to more accurately capture the volumetric distribution of hippocampal volumes, these methods either do not produce conditional T1w templates (Yu *et al.*) or produce conditional T1w templates with lower image quality when compared to our method (as demonstrated by the results in figure 4).

When we consider topological consistency, our method outperformed the method of Dalca *et al.* for deformations from conditional templates to individual participant scans (stage 2 of the respective networks) — for these deformations, our network did not produce any negative Jacobian determinants while the method of Dalca *et al.* produced a number of very small and negative Jacobian determinants, which indicates that there were several instances of topological inconsistency between conditional templates and individual scans, with an average of 508 voxels that exhibited negative Jacobian determinants for each of the predicted deformation fields. Additionally, while the distribution of Jacobian determinants for stage 2 deformations was reasonably symmetric for our method, the method of Dalca *et al.* produced a right-skewed distribution with Jacobian determinant values spanning -2.7 to 22.2. For deformations from the group template to conditional

templates (stage 1 of our network), both our method and the method of Yu *et al.* did not produce any negative Jacobian determinants, which indicates that they were both able to maintain topological consistency between group and conditional templates. An interesting contrast between the methods exists in the noticeably different variances exhibited by the distribution of Jacobian determinants for the method of Yu *et al.* when compared to our method. Here, given the amount of volumetric change exhibited by certain structures, such as the ventricles, where there is up to 40% change in volume from the central age to the ends of the age distribution, it is reasonable to expect the distribution of Jacobian determinants to stretch accordingly for a non-negligible proportion of voxels, which is what we see in our Jacobian determinant distribution. Furthermore, for our method, the distribution of Jacobian determinants is similar and symmetrical across both network stages, which may imply that it is effective at factoring out changes that are common across the population and associated with age, from those that are associated with individual anatomical differences.

From a structural similarity perspective, our method can produce adjacent conditional templates that share substantial similarities as conditioning parameters are adjusted, with high SSIM metric values for adjacent conditional templates, prior to any additional registration, across the range of age comparisons considered. Additionally, after additional non-linear registration, our method produced SSIM metric values that are consistently improved compared to the pre-registration values, which we would see if there was useful and measurable geometric change being captured with varying age that registration could undo. Conversely, the method of Dalca *et al.* produced similar SSIM metric values both before and after registration. This implies that the age-related (conditional parameter) changes were not able to be easily explained in terms of geometric changes, such as the size and location of brain structures, by the registration process, even though we would expect to see such age-related changes in geometry. Additionally, SSIM results for the method of Dalca *et al.* exhibited decreasing similarity at both ends of the age distribution, which indicates that greater structural discrepancies were observed, between adjacent 5-year intervals, at the edges of the conditional parameter space.

The volumetric measurement results for several ROIs that are known to vary in the ageing brain are mixed, with each method having advantages and disadvantages over other methods across different structures. For instance, our method could track the distribution of volumes for structures like the ventricles with good accuracy, and to a lesser extent, for white-matter tissue. However, our method tracked the distribution of hippocampal volumes poorly, and was also deficient in tracking the distribution of grey-matter tissue volume. On the basis of volumetric measurements, the method of Yu *et al.* was the strongest approach considered for the various structures examined, with good accuracy, compared to the empirical ranges estimated from the training scans, for the volumes of the ventricles, white-matter tissue and hippocampus. However, while the method of Yu *et al.* performed slightly better than our method at tracking the distribution of grey-matter volume, it was still ineffective at tracking this. The method of Dalca *et al.* was able to track the hippocampal volumes with good accuracy, and although it was the best for grey-matter out of the methods compared here, it was still only capturing about half of the empirically observed change in volume. For the ventricles and white-matter, the method of Dalca *et al.* performed poorly.

Another output of interest from these methods are the template images themselves and how suitable they are for registration targets and how well they capture a qualitative representation of the ageing process. For the method of Yu *et al.*, conditional probabilistic templates were produced that have a number of practical uses, including the tracking of various morphological features, such as structural volumes; however, further work is required to enable such an approach to effectively produce T1w structural templates that appear realistic and would make good targets for registrations. In contrast, the method of Dalca *et al.* employed a relatively unconstrained approach in their first network stage, with the network able to make independent voxel-wise intensity changes to the conditional template images. As such, this kind of architecture can introduce intensity artefacts into the conditional templates and lacks the strong underpinning for finding anatomical correspondences across the age range of a method based on spatial deformations. The use of a diffeomorphic deformation framework in both stage 1 and 2 of our network means that conditional templates will typically have few artefacts introduced, with it focusing on geometric changes of interest.

4.1 Limitations

One limitation of our method is the inability of our network to make non-geometric changes to conditional templates as conditioning parameters, such as

age, are adjusted. For instance, it is known that during ageing the human brain goes through a process of demyelination [42, 43], leading to an increase in T1-relaxation time for the affected tissue, and hence decreased signal intensity in T1-weighted images. These kinds of changes cannot easily be captured by a purely geometric network architecture. We have investigated the inclusion of network modules that are able to make non-geometric changes to conditional templates and predicted participant scans, however, we found that they made the model unstable and less able to capture geometric changes of interest.

Another limitation of our approach, and its potential, lies in the ability of local mean constraints to be effectively scaled-up to use with higher-dimensional combinations of conditional parameters. Given that our method partitioned the conditional parameter space to create a discrete number of voxel-wise mean-tracking deformations, the combinatorial increase in the number of deformations required with higher dimensions may become difficult to manage computationally. For instance, if we were to jointly model 4 conditioning parameters, each partitioned such that we were tracking 10 discrete representations for each conditioning parameter of interest, we would need to manage 10^4 local mean deformations if they were all allowed to be independent. Additionally, the amount of data used to inform each mean deformation becomes smaller in higher dimensions due to data sparsity. The first issue may be ameliorated by managing relevant data-structures on disk or in CPU memory, as opposed to GPU memory. However, the sparsity of participant scans in the conditional space means that either different representations over conditional parameter range are needed, which might be coarser or better at modeling/interpolating, or that additional training data is required to reduce the sparsity. Relaxing the independence assumption and imposing a better model is something that deep learning may be able to provide, although the exact form of the model and the amount of bias that it induces would require careful management. While the method of Dalca *et al.* utilised a global mean-tracking constraint across the dataset, which has some benefits to encourage centrality, such a global constraint may suffer from bias towards particular parts of conditional parameter space. For instance, this kind of network constraint is likely to become highly optimised for the most common conditioning parameters.

While we have prepared our structural volumetric analysis for four structures that are strongly implicated in ageing and neurodegeneration, and also commonly reported on, there are many others that may be useful. Hence the set of structures considered should be enlarged in future studies. Additionally, the calculation of volumes used here was based on the output from an automated segmentation tool, albeit one with proven performance, but it is not known whether these segmentations are biased or creating errors that may impact these results.

Another consideration is the size of the dataset utilised for this study, namely 1534 scans across 515 individual participants. It is possible that some or all of the methods that we have evaluated perform significantly better on much larger datasets. This notwithstanding, the use of a more constrained dataset has certain advantages for our evaluation: (i) some populations, especially of rarer diseases that might be of interest, will necessitate working with smaller datasets; and (ii) when introducing multiple conditioning factors, the density of datapoints in the higher dimensional space will decrease — hence, methods will need to be able to cope in regimes where the density of datapoints is lower than the density of datapoints found in large single-factor datasets.

4.2 Future directions and conclusion

An interesting direction for future work could be to mix the different elements from the methods compared here. For instance, the method of Yu *et al.* used probabilistic atlases, which could potentially be leveraged in conjunction with intensity templates to provide additional information to a network when producing T1-weighted conditional templates. Here, we can envisage probabilistic atlases augmenting templates to provide richer information about key brain structures in the network. Additionally, in future work, we aim to develop methods that are able to account for multiple clinical and demographic variables jointly, given the potential utility of such conditional template construction methods.

In conclusion, while our approach, and those of Dalca *et al.* and Yu *et al.* offer a number of useful contributions to conditional template construction, current deep-learning methods require further refinement to achieve stronger calibration with morphological changes exhibited by ground truth data in ageing and degenerative brains. Furthermore, when considering building methods that can work with a number of conditional parameters jointly, conventional methods often have limitations in terms of the large amount of data required or the restricted modeling options. Hence, deep-learning approaches have some inherent advantages in terms of flexible data-driven modeling, without necessarily requiring as much data, and so future work in this area should be

pursued while actively working to improve the ability of these methods to accurately track volumetric changes.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] I -F Talos et al. "Magnetic resonance and the human brain: anatomy, function and metabolism". In: *Cellular and Molecular Life Sciences CMLS* 63 (2006), pp. 1106–1124.

[2] Antonio Giorgio et al. "Age-related changes in grey and white matter structure throughout adulthood". In: *Neuroimage* 51.3 (2010), pp. 943–951.

[3] Adrian V. Dalca et al. "Learning conditional deformable templates with convolutional networks". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[4] Pedro F. Felzenszwalb and Joshua D. Schwartz. "Hierarchical Matching of Deformable Shapes". In: *2007 IEEE Conference on Computer Vision and Pattern Recognition* (2007), pp. 1–8. URL: https://api.semanticscholar.org/CorpusID:6782979.

[5] Anil K. Jain, Yu Zhong, and Sridhar Lakshmanan. "Object Matching Using Deformable Templates". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 18.3 (Mar. 1996), pp. 267–278. ISSN: 0162-8828. DOI: 10.1109/34.485555. URL: https://doi.org/10.1109/34.485555.

[6] Jaechul Kim et al. "Deformable Spatial Pyramid Matching for Fast Dense Correspondences". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2013.

[7] Iasonas Kokkinos et al. "Intrinsic shape context descriptors for deformable shapes". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), pp. 159–166. URL: https://api.semanticscholar.org/CorpusID:13014994.

[8] Carsten Stoll et al. "Template Deformation for Point Cloud Fitting". In: Jan. 2006, pp. 27–35. DOI: 10.2312/SPBG/SPBG06/027-035.

[9] Stéphanie Allassonnière, Yali Amit, and Alain Trouvé. "Towards a Coherent Statistical Framework for Dense Deformable Template Estimation". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69 (Feb. 2007), pp. 3–29. DOI: 10.1111/j.1467-9868.2007.00574.x.

[10] B. Davis, P. Lorenzen, and S. Joshi. "Large deformation minimum mean squared error template estimation for computational anatomy". In: *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)* (2004), 173–176 Vol. 1. DOI: 10.1109/ISBI.2004.1398502.

[11] S. Joshi et al. "Unbiased diffeomorphic atlas construction for computational anatomy". In: *NeuroImage* 23 (2004). Mathematics in Brain Imaging, S151–S160. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2004.07.068. URL: https://www.sciencedirect.com/science/article/pii/S1053811904003842.

[12] Jun Ma et al. "Bayesian template estimation in computational anatomy". In: *NeuroImage* 42.1 (Aug. 2008), pp. 252–261. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2008.03.056. URL: https://www.sciencedirect.com/science/article/pii/S1053811908002991.

[13] David Alexander Dickie et al. "Whole brain magnetic resonance image atlases: a systematic review of existing atlases and caveats for use in population imaging". In: *Frontiers in neuroinformatics* 11 (2017), p. 1.

[14] Evan M. Yu, Adrian V. Dalca, and Mert R. Sabuncu. "Learning Conditional Deformable Shape Templates for Brain Anatomy". In: *Machine Learning in Medical Imaging*. Ed. by Mingxia Liu et al. Cham: Springer International Publishing, 2020, pp. 353–362. ISBN: 978-3-030-59861-7.

[15] Kathryn L. Mills and Christian K. Tamnes. "Methods and considerations for longitudinal structural brain imaging analysis across development". In: *Developmental Cognitive Neuroscience* 9 (2014), pp. 172–190. ISSN: 1878-9293. DOI: https://doi.org/10.1016/j.dcn.2014.04.004. URL: https://www.sciencedirect.com/science/article/pii/S1878929314000310.

[16] Emma Lawrence et al. "A Systematic Review of Longitudinal Studies Which Measure Alzheimer's Disease Biomarkers". In: *Journal of Alzheimer's Disease* 59 (July 2017), pp. 1–21. DOI: 10.3233/JAD-170261.

[17] Antonio Giorgio and Nicola De Stefano. "Clinical use of brain volumetry". In: *Journal of Magnetic Resonance Imaging* 37.1 (2013), pp. 1–14.

DOI: https://doi.org/10.1002/jmri.23671. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.23671. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.23671.

[18] Pravat K Mandal, Rashima Mahajan, and Ivo D Dinov. "Structural brain atlases: design, rationale, and applications in normal and pathological cohorts". In: *Journal of Alzheimer's Disease* 31.s3 (2012), S169–S188.

[19] Naren P. Rao et al. "Population differences in brain morphology: Need for population specific brain template". In: *Psychiatry Research: Neuroimaging* 265 (2017), pp. 1–8. ISSN: 0925-4927. DOI: https://doi.org/10.1016/j.pscychresns.2017.03.018. URL: https://www.sciencedirect.com/science/article/pii/S0925492716301147.

[20] Christopher G Schwarz et al. "[IC-P-122]: THE MAYO CLINIC ADULT LIFE SPAN TEMPLATE: BETTER QUANTIFICATION ACROSS THE LIFE SPAN". In: *Alzheimer's & Dementia* 13.7S_Part_2 (2017), P93–P94.

[21] Adrian V. Dalca et al. "Unsupervised Learning for Fast Probabilistic Diffeomorphic Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2018. URL: https://api.semanticscholar.org/CorpusID:43995396.

[22] Carole H Sudre et al. "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer. 2017, pp. 240–248.

[23] Susanne G. Mueller et al. "The Alzheimer's disease neuroimaging initiative". English (US). In: *Neuroimaging Clinics of North America* 15.4 (Nov. 2005), pp. 869–877. ISSN: 1052-5149. DOI: 10.1016/j.nic.2005.09.008.

[24] Michael W Weiner et al. "The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement". In: *Alzheimer's & Dementia* 13.5 (2017), pp. 561–571.

[25] Mark Jenkinson et al. "FSL". In: *NeuroImage* 62.2 (2012). 20 YEARS OF fMRI, pp. 782–790. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2011.09.015. URL: https://www.sciencedirect.com/science/article/pii/S1053811911010603.

[26] Andrew Hoopes et al. "SynthStrip: skull-stripping for any brain image". In: *NeuroImage* 260 (2022), p. 119474. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2022.119474. URL: https://www.sciencedirect.com/science/article/pii/S1053811922005900.

[27] Mark Jenkinson et al. "Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images". In: *NeuroImage* 17.2 (2002), pp. 825–841. ISSN: 1053-8119. DOI: https://doi.org/10.1006/nimg.2002.1132. URL: https://www.sciencedirect.com/science/article/pii/S1053811902911328.

[28] John Ashburner. "A fast diffeomorphic image registration algorithm". In: *NeuroImage* 38.1 (2007), pp. 95–113. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2007.07.007. URL: https://www.sciencedirect.com/science/article/pii/S1053811907005848.

[29] Julian Krebs et al. "Learning a Probabilistic Model for Diffeomorphic Registration". In: *IEEE Transactions on Medical Imaging* PP (Feb. 2019), pp. 1–1. DOI: 10.1109/TMI.2019.2897112.

[30] Marc Modat et al. "Parametric non-rigid registration using a stationary velocity field". In: *Proceedings of the Workshop on Mathematical Methods in Biomedical Image Analysis* (Jan. 2012). DOI: 10.1109/MMBIA.2012.6164745.

[31] Cleve Moler and Charles Van Loan. "Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later". In: *SIAM Review* 45.1 (2003), pp. 3–49. DOI: 10.1137/S00361445024180. eprint: https://doi.org/10.1137/S00361445024180. URL: https://doi.org/10.1137/S00361445024180.

[32] Vincent Arsigny et al. "A Log-Euclidean Framework for Statistics on Diffeomorphisms". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*. Ed. by Rasmus Larsen, Mads Nielsen, and Jon Sporring. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 924–931. ISBN: 978-3-540-44708-5.

[33] Max Jaderberg et al. *Spatial Transformer Networks*. 2016. arXiv: 1506.02025 [cs.CV].

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab et al. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4.

[35] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[36] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2019. arXiv: 1711.05101 [cs.LG].

[37] Edgar Riba et al. "Kornia: an Open Source Differentiable Computer Vision Library for PyTorch". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Mar. 2020.

[38] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (Apr. 2004), pp. 600–612. ISSN: 1941-0042. DOI: 10.1109/TIP.2003.819861.

[39] B.B. Avants et al. "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain". In: *Medical Image Analysis* 12.1 (2008). Special Issue on The Third International Workshop on Biomedical Image Registration – WBIR 2006, pp. 26–41. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2007.06.004. URL: https://www.sciencedirect.com/science/article/pii/S1361841507000606.

[40] Brian B Avants, Nick Tustison, Gang Song, et al. "Advanced normalization tools (ANTS)". In: *Insight j* 2.365 (2009), pp. 1–35.

[41] Leonie Henschel et al. "FastSurfer - A fast and accurate deep learning based neuroimaging pipeline". In: *NeuroImage* 219 (2020), p. 117012. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2020.117012. URL: https://www.sciencedirect.com/science/article/pii/S1053811920304985.

[42] Carlos López-Otín et al. "The hallmarks of aging". In: *Cell* 153.6 (2013), pp. 1194–1217.

[43] Robin JM Franklin, Chao Zhao, and Fraser J Sim. "Ageing and CNS remyelination". In: *Neuroreport* 13.7 (2002), pp. 923–928.

Supplementary materials

S1 Structural Similarity Index Measure (SSIM) derivation

SSIM extracts and calculates comparisons for three key images features, namely: (i) luminance, $l(\cdot)$; (ii) contrast, $c(\cdot)$; and (iii) structure, $s(\cdot)$; combining the components to produce an overall similarity measure for images $\mathbf{x}$ and $\mathbf{y}$: $S(\mathbf{x},\mathbf{y}) = f(l(\mathbf{x},\mathbf{y}), c(\mathbf{x},\mathbf{y}), s(\mathbf{x},\mathbf{y}))$.

The luminance comparison function, $l(\mathbf{x},\mathbf{y})$, for images $\mathbf{x}$ and $\mathbf{y}$, compares the mean pixel or voxel intensities, $\mu_{\mathbf{x}}$ and $\mu_{\mathbf{y}}$, for two images, as follows:

$$l(\mathbf{x},\mathbf{y}) = \frac{2\mu_{\mathbf{x}}\mu_{\mathbf{y}} + C1}{\mu_{\mathbf{x}}^2 + \mu_{\mathbf{y}}^2 + C1} \tag{12}$$

where $C1 = (K_1 L)^2$ is a constant term for numerical stability, L is the pixel or voxel dynamic range and $K_1 << 1$ is a small constant.

The contrast comparison function, $c(\mathbf{x},\mathbf{y})$, takes a similar form to $l(\cdot)$, comparing the standard deviation of the intensity values taken across all voxels within an image, $\sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{y}}$, for two images, as follows:

$$c(\mathbf{x},\mathbf{y}) = \frac{2\sigma_{\mathbf{x}}\sigma_{\mathbf{y}} + C2}{\sigma_{\mathbf{x}}^2 + \sigma_{\mathbf{y}}^2 + C2} \tag{13}$$

where $C2$ is similarly calculated to $C1$ for $K_2 << 1$.

To perform a structural comparison between images $\mathbf{x}$ and $\mathbf{y}$, luminance is subtracted from the images and they are normalised by variance (i.e., $(\mathbf{x} - \mu_{\mathbf{x}})/\sigma_{\mathbf{x}}$ and $(\mathbf{y} - \mu_{\mathbf{y}})/\sigma_{\mathbf{y}}$), with the correlation between these an effective measure to quantify structural similarity, as given by:

$$s(\mathbf{x},\mathbf{y}) = \frac{\sigma_{\mathbf{xy}} + C3}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}} + C3} \tag{14}$$

where $\sigma_{\mathbf{xy}}$ represents the covariance of the intensity values taken across all voxels in images $\mathbf{x}$ and $\mathbf{y}$. Each of the three components is then weighted to give an overall measure of SSIM:

$$S(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^{\alpha} \cdot [c(\mathbf{x},\mathbf{y})]^{\beta} \cdot [s(\mathbf{x},\mathbf{y})]^{\gamma} \tag{15}$$

where $\alpha, \beta, \gamma > 0$ adjust the importance of each component. Typically, each of these exponents is set equal to 1; which we have adopted for the purposes of our analysis.