Computers in Biology and Medicine 179 (2024) 108811

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Improving brain atrophy quantification with deep learning from automated labels using tissue similarity priors

Albert Clèrigues [a,*], Sergi Valverde [b], Arnau Oliver [a], Xavier Lladó [a], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Institute of Computer Vision and Robotics, University of Girona, Spain
[b] Tensor Medical, Girona, Spain

ARTICLE INFO

Keywords:
Magnetic resonance imaging
Brain tissue segmentation
Deep learning
Brain atrophy quantification

ABSTRACT

Brain atrophy measurements derived from magnetic resonance imaging (MRI) are a promising marker for the diagnosis and prognosis of neurodegenerative pathologies such as Alzheimer's disease or multiple sclerosis. However, its use in individualized assessments is currently discouraged due to a series of technical and biological issues. In this work, we present a deep learning pipeline for segmentation-based brain atrophy quantification that improves upon the automated labels of the reference method from which it learns. This goal is achieved through tissue similarity regularization that exploits the a priori knowledge that scans from the same subject made within a short interval must have similar tissue volumes. To train the presented pipeline, we use unlabeled pairs of T1-weighted MRI scans having a tissue similarity prior, and generate the target brain tissue segmentations in a fully automated manner using the fsl_anat pipeline implemented in the FMRIB Software Library (FSL). Tissue similarity regularization is enforced during training through a weighted loss term that penalizes tissue volume differences between short-interval scan pairs from the same subject. In inference, the pipeline performs end-to-end skull stripping and brain tissue segmentation from a single T1-weighted MRI scan in its native space, i.e., without performing image interpolation. For longitudinal evaluation, each image is independently segmented first, and then measures of change are computed. We evaluate the presented pipeline in two different MRI datasets, MIRIAD and ADNI1, which have longitudinal and short-interval imaging from healthy controls (HC) and Alzheimer's disease (AD) subjects. In short-interval scan pairs, tissue similarity regularization reduces the quantification error and improves the consistency of measured tissue volumes. In the longitudinal case, the proposed pipeline shows reduced variability of atrophy measures and higher effect sizes of differences in annualized rates between HC and AD subjects. Our pipeline obtains a Cohen's d effect size of $d = 2.07$ on the MIRIAD dataset, an increase from the reference pipeline used to train it ($d = 1.01$), and higher than that of SIENA ($d = 1.73$), a well-known state-of-the-art approach. In the ADNI1 dataset, the proposed pipeline improves its effect size ($d = 1.37$) with respect to the reference pipeline ($d = 0.80$) and surpasses SIENA ($d = 1.33$). The proposed data-driven deep learning regularization reduces the biases and systematic errors learned from the reference segmentation method, which is used to generate the training targets. Improving the accuracy and reliability of atrophy quantification methods is essential to unlock brain atrophy as a diagnostic and prognostic marker in neurodegenerative pathologies.

1. Introduction

Global and regional brain atrophy quantification has been shown to be a relevant marker for prognosis of neurodegenerative pathologies, such as Alzheimer's disease (AD) [1–3] and multiple sclerosis (MS) [4–7]. In AD, improved atrophy quantification accuracy would allow for earlier and more reliable diagnosis, even in pre-symptomatic stages. After AD diagnosis, measures of atrophy could be used to assess the rate of disease progression, enabling more timely intervention and potentially slowing disease progression. Similarly, in MS, precise brain atrophy measurements could detect increased disease activity as an accelerated

* Correspondence to: Ed. P-IV, Campus Montilivi, University of Girona, 17003 Girona, Spain.
E-mail address: albert.clerigues@udg.edu (A. Clèrigues).

[1] 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

https://doi.org/10.1016/j.compbiomed.2024.108811
Received 12 December 2023; Received in revised form 31 May 2024; Accepted 24 June 2024
Available online 10 July 2024

rate of brain atrophy, aid in distinguishing MS phenotypes, and serve as a prognostic marker to evaluate the response to disease-modifying treatments. Ultimately, precise brain atrophy monitoring could be used to guide treatment decisions and improve their efficacy, supporting personalized data-driven adjustment of therapeutic strategies.

Magnetic resonance imaging (MRI) allows for noninvasive quantitative measures of global and regional atrophy of the brain parenchyma. These measurements are typically obtained from longitudinal T1-weighted (T1-w) images, on which there is good contrast between the cerebrospinal fluid (CSF) and the distinct gray matter (GM) and white matter (WM) components that form the brain parenchyma. Methods for brain atrophy quantification are currently affected by a number of confounding factors related to image acquisition, technical issues and pathophysiological changes [8], reducing their reliability and applicability. Although MRI-derived measurements of atrophy have proven useful for clinical population studies analyzing disease progression or treatment effects, they are still not considered sufficiently accurate or reliable for their use in individualized assessments [9]. Inaccurate atrophy quantification methods in the clinical setting could lead to false diagnoses or inappropriate treatment adjustments. Thus, to ensure accurate and reliable individualized atrophy measures, it is critical to reduce the impact of confounding factors such as differences in acquisition parameters, image artifacts or patient variability.

In general, longitudinal brain atrophy quantification methods can be classified into either segmentation-based or registration-based techniques. In segmentation-based methods, a target set of structures or tissues is independently segmented in each of the longitudinal scans, and atrophy is quantified from differences in the measured volumes. In practice, segmentation-based techniques are influenced by the quality of T1-weighted images, as they perform an indirect measure of atrophy through independent segmentations of each timepoint. In contrast, registration-based techniques derive measures of atrophy from the observed spatial deformation of structures or tissues between two longitudinal scans. For this, a non-linear registration is performed between the two image intensities which provides a higher degree of sensitivity to changes over time, more robustness and less dependency on the quality of MRI acquisitions. For these reasons, segmentation-based methods are typically regarded as less accurate and more variable than their registration-based counterparts, and their use has been discouraged for longitudinal studies [10].

Although several segmentation-based methods for cross-sectional brain volumetry from T1-w MRI have been proposed in the recent literature, only SIENA-XL [11] has been purposefully built for longitudinal imaging. Registration-based methods are typically preferred for longitudinal change analysis since they have lower quantification error and better sensitivity to atrophy changes [10]. SIENA [12] is a well-known and widely used registration-based atrophy quantification method based on the boundary shift integral (BSI) [13]. Within SIENA, atrophy is measured between two linearly registered scans from the surface displacement of the interface between GM and WM, which was obtained from FAST [14] tissue segmentations of each scan. Measures of longitudinal change can also be derived from the deformation fields obtained from nonlinear registration between baseline and follow-up. The work of Holland and Dale [15] used the deformation field to approximate voxels as irregular hexahedrons and directly compute the fractional volume change of a certain region between timepoints. More recently, methods based on Jacobian integration of displacement fields have shown further improvements, such as larger effect sizes and lower quantification error [16,17]. These methods measure volume changes by integrating the determinant of the Jacobian of a nonlinear transformation between two longitudinal scans. The region for integration is typically obtained from a cross-sectional segmentation of tissues or structures in one of the scans. It is worth noting that even within registration-based methods, some form of cross-sectional segmentation of tissue or structures is still needed.

In recent years, deep learning techniques have achieved higher levels of accuracy and performance in brain MRI segmentation tasks for Alzheimer's disease [18]. Several deep learning approaches have been recently proposed for cross-sectional brain tissue segmentation using a mix of automated and manually annotated data. While advances in methods for cross-sectional tissue segmentation contribute to better atrophy quantification, these methods are not optimized for dealing with the specific challenges of longitudinal analysis. Existing work often focuses on segmentation accuracy in a single scan, rather than ensuring consistency and minimizing error when analyzing changes in tissue volume over time. EA-Net [19] is a deep learning method that is fully trained on manual brain tissue segmentations and incorporates edge and boundary features to improve segmentation of the partial volumes between tissues. QuickNAT [20] is first trained on automated segmentations made with FreeSurfer [21] and then fine-tuned on manual delineations of brain tissue. In contrast, FastSurfer [22] and NeuroNet [23] are both trained solely on automated brain tissue segmentations made with FreeSurfer [21] and FSL [24], respectively, two of the most frequently used automated tools for neuroanatomical analysis. These approaches achieve greater consistency, reliability and shorter execution time than the reference methods on which they were trained. Moreover, both QuickNAT and FastSurfer also demonstrate improvements with respect to longitudinal brain atrophy quantification, having lower short interval error and higher sensitivity to atrophy changes. For pathological cases with brain lesions, Dorent et al. [25] used several disjoint heterogeneous datasets with manual annotations to learn a joint brain tissue and lesion segmentation model. This approach is much more robust to the volumetric errors introduced by the presence of abnormal brain lesions and can also deal with a variable number of input modalities. While advances in methods for cross-sectional brain tissue segmentation can be used by atrophy quantification approaches to improve their longitudinal results, these have not been purposefully built for improving brain tissue segmentation in the longitudinal case.

In this work, we present a deep learning pipeline for segmentation-based brain atrophy quantification that uses tissue similarity regularization to improve upon the automated labels of a reference method used for training. The presented pipeline is specifically tailored to longitudinal brain atrophy analysis and aims at improving the consistency and reliability of brain tissue segmentations over time. The proposed regularization exploits a priori knowledge that pairs of scans from the same subject made within a short interval must have similar brain tissue volumes. The pipeline is trained using a set of short-interval scan pairs from which training targets are generated in a fully automated manner using the fsl_anat pipeline provided in FSL. The reference tissue segmentations are obtained from fsl_anat in a similar fashion to SIENA-XL [11] by merging the resulting brain tissue segmentation of FAST [14] and the deep gray matter structures of FIRST [26]. Tissue similarity regularization is enforced during training through a weighted loss term that penalizes volume differences between similar scan pairs. In inference, the pipeline acts on a single T1-w scan in its native space and performs end-to-end skull stripping and brain tissue segmentation. For longitudinal evaluation, each image is independently segmented, and then change measures are computed. We performed a quantitative and qualitative evaluation of the improvements in brain atrophy quantification using two publicly accessible longitudinal MRI datasets, MIRIAD and ADNI1. The presented pipeline improves upon the reference method used for training by having a lower quantification error, better intracranial cavity consistency and higher sensitivity to differences in brain atrophy rates between healthy controls and Alzheimer's disease (AD) patients.

2. Materials

The datasets employed in this study consist of two publicly available MRI datasets are used with different characteristics, MIRIAD [27]

and ADNI1, which have short-interval and longitudinal imaging for healthy controls (HCs) and AD subjects. Datasets from AD patients are a particularly good fit as an evaluation framework for brain atrophy quantification methods. Typically, longitudinal scans from AD patients have a high pronounced rate of atrophy compared to other neurodegenerative pathologies and have less of the confounding factors commonly seen in other pathologies, such as the presence brain lesions. Additionally, these datasets have both short-interval and longitudinal imaging, which are required by the presented pipeline for training and evaluation.

The MIRIAD dataset provides a small set of homogeneous MRI scans taken with the same scanner and acquisition protocol, while the ADNI1 dataset has a larger number of subjects with more heterogeneous imaging acquired on different scanners and with varied voxel spacings. By using both a single-center (MIRIAD) and multi-center dataset with different scanner models (ADNI1), we can study the influence of training data heterogeneity on the model's performance and generalization. In this way, a cross-dataset evaluation can be performed, training on one dataset and testing on the other, to study if training on a diverse dataset with many scanners leads to improved generalization when evaluating on external datasets.

The datasets employed in this study were obtained from two publicly available and established sources, both of which adhere to rigorous ethical standards regarding patient consent and data privacy.

2.1. MIRIAD dataset

The Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) dataset [27] is a publicly accessible series of longitudinal T1 MRI scans of 46 mild–moderate Alzheimer's subjects with an average age of 69.4 ± 7.1 years and 23 healthy controls with an average age of 69.7 ± 7.2 years. The dataset consists of longitudinal scans taken at intervals of 2, 6, 14, 26, 38 and 52 weeks and 18 and 24 months from baseline, as well as rescan images at three of the timepoints, for both AD and controls. The rescan images were taken during three of the scanning sessions (0, 6 and 38 weeks) without repositioning of the subject. All scans were taken by the same radiographer on the same 1.5 T Signa MRI scanner (GE Medical systems, Milwaukee, WI, USA) with a voxel size of $0.9375 \times 1.5 \times 0.9375$ and total image dimensions of $256 \times 124 \times 256$. In our study, we consider both the rescan image pairs and the baseline to 2-weeks image pairs to have a tissue similarity prior that can be used for regularization. From the original dataset, some images were discarded due to poor scan quality or movement artifacts; details on which image pairs were used for training and evaluation can be found in the supplementary material.

2.2. ADNI1 data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging 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).

In our work, we consider a subset of subjects originally included in the "ADNI1: Complete 1Yr 1.5T" standardized data collection and use similarly preprocessed scans with corrected gradient nonlinearity and B1 and N3 nonuniformity correction. We consider 251 pairs of baseline and 1 year of follow-up scans from 105 AD patients with an average age of 75.9 ± 7.3 years and 146 healthy control subjects with an average age of 76.3 ± 5.4 years. We also consider 541 scan-rescan image pairs taken at each of the two timepoints, including some subjects who did not have both longitudinal scans. Within this cohort, scans were taken with varied voxel sizes ranging from $0.94 \times 0.94 \times 1.2$ to $1.3 \times 1.3 \times 1.2$, having the same voxel size between the scan-rescan and longitudinal image pairs of the same subject. In total, 8 different scanners from 2 manufacturers (GE and Siemens) were used for image acquisition. In 45 of the 250 subjects, a different scanner model from the same manufacturer was reported for the 1-year follow-up scan. Details of the image pair IDs used for training and inference can be found in the supplementary material.

3. Methods

We present a deep learning pipeline for segmentation-based brain volumetry that learns from automated tissue segmentations derived from fsl_anat while enforcing a tissue similarity regularization that improves longitudinal brain atrophy quantification. The proposed regularization exploits the assumption that two scans from the same subject taken within a short time interval should have similar brain tissue volumes. For training, we use a set of coregistered T1-w scan pairs having a tissue similarity prior and generate the segmentation targets in an automated manner using fsl_anat. In inference, the pipeline performs end-to-end skull stripping and brain tissue segmentation from a single image in its native space, i.e., without image interpolation. For longitudinal evaluation, each image is independently segmented first, and then measures of change are computed. In the following sections, we describe in detail how to prepare the training data and present the deep learning framework architecture, along with the procedures for network training and image inference.

3.1. Training data preparation

The pipeline is trained from a set of T1-w scan pairs belonging to the same subject and acquired within a short interval, thus having a tissue similarity prior, from which we generate the reference brain tissue segmentations in a fully automated manner, as shown in Fig. 1. For this purpose, the fully automated fsl_anat anatomical image processing pipeline implemented in FSL is applied to each T1-w scan to perform skull stripping, as well as segmentation of brain tissue using FAST [14] and deep gray matter structures using FIRST [26]. This tissue segmentation procedure is very similar to that done by SIENA-XL [11], which also used the fsl_anat pipeline to generate the tissue and subcortical structure segmentation. More specifically, fsl_anat performs skull-stripping through a nonlinear registration to the MNI standard space, which is used to transform a dilated MNI brain mask back into the native space of the T1-w image. From this skull-stripped image, brain tissue probabilities are obtained using FAST [14], which is run with the `--weakbias` option. Additionally, the deep gray matter of subcortical structures is segmented with the registration-based FIRST method [26]. Similar to SIENA-XL [11], we merge the FIRST subcortical structure segmentation into the FAST tissue probabilities by setting them as pure gray matter, obtaining the final reference FAST + FIRST segmentation in the native space of each T1-w scan.

As part of the fsl_anat pipeline, we also obtain a transform to an MNI T1-w structural template with 2 mm resolution. To avoid an additional registration step, we use the inverse of this transform to bring the 2 mm resolution MNI brain mask through nearest neighbor interpolation into the native T1-w scan space. This coarse brain mask is later used as a normalization mask to constrain the computation of image statistics to the brain tissues for performing the input normalization of our deep learning system.

Finally, each pair of similar T1-w scans is spatially aligned to be able to exploit their tissue similarity prior during training. This goal is achieved by performing a linear registration to a halfway space between them using the mri_robust_register method [28] implemented in the FreeSurfer image analysis suite with default parameters, using cubic interpolation to transform the images. Then, the reference FAST +

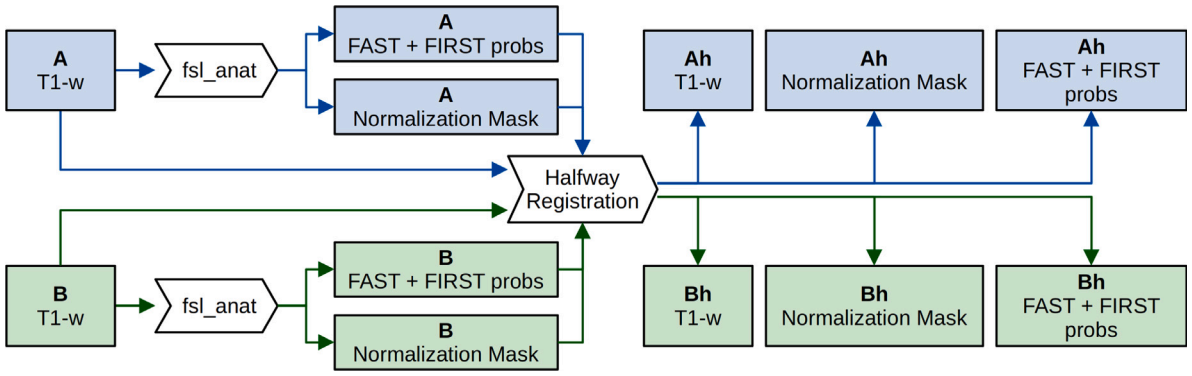


Fig. 1. Training data preparation diagram for each pair of short-interval T1-w scans, A and B, which have tissue similarity prior. The T1-w scans are first processed using the fsl_anat anatomical image processing pipeline to obtain the reference FAST + FIRST brain tissue probabilities and input normalization mask for each of them. Then, the T1-w scans and their segmentations are spatially aligned by linear registration to a halfway space, Ah and Bh, between them.

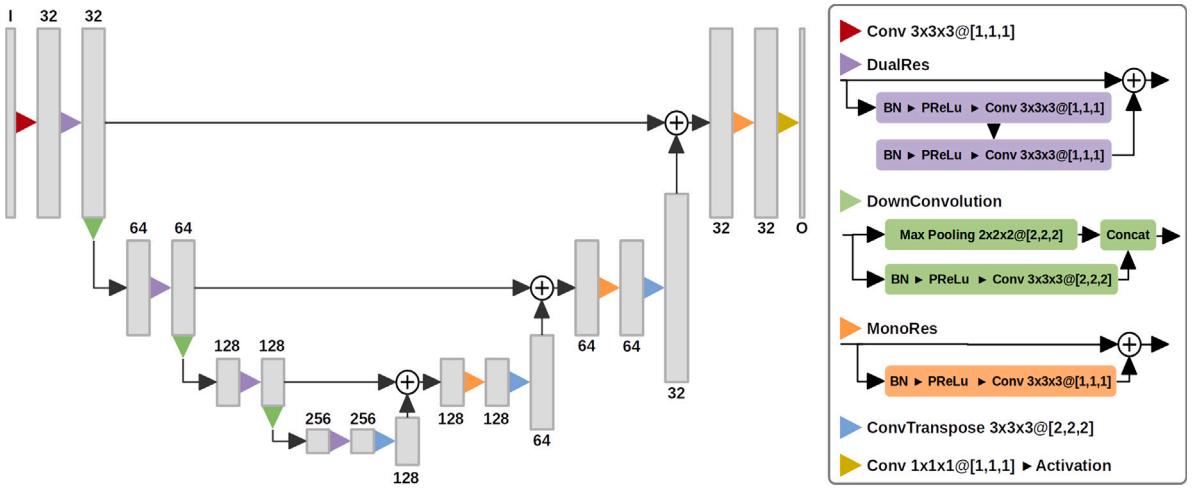


Fig. 2. Diagram of the 3D U-Net based model. The parameter distribution is asymmetrical, with the residual blocks of the encoder using two convolutional blocks, while a single one is used in the decoder. In the convolutional layers (Conv), $K_x \times K_y \times K_z@[S_x,S_y,S_z]$ indicates the kernel and stride dimensions on each axis. The gray boxes represent the feature maps with the number of channels indicated above or below it. The numbers of input and output feature maps are denoted by I and O, respectively.

FIRST tissue probabilities are also transformed through linear interpolation into the halfway space, along with the normalization mask, which is transformed using nearest neighbor interpolation. Note that the halfway registered T1-w scans and tissue probabilities are exclusively used during training, while image inference for evaluation is performed in the native space of each scan without any type of interpolation.

3.2. Deep learning pipeline

In this section, we describe in detail the network architecture and input normalization of the presented deep learning pipeline. We utilize a patch-based deep learning pipeline using a residual 3D architecture based on the U-Net [29], which performs both skull stripping and brain tissue segmentation from a single T1-w scan. As input, the network receives a single 3D patch with spatial dimensions of $32 \times 32 \times 32$ and outputs a brain tissue probability distribution among four classes (background, CSF, GM and WM) for each input voxel. The patch size of $32 \times 32 \times 32$ was selected through empirical tests to provide sufficient context for accurate segmentation while balancing class representation and improving training stability through the use of a larger batch size. The network architecture, depicted in Fig. 2, consists of a 3D U-Net model that uses residual convolution blocks and skip connections. All the convolutional layers use $3 \times 3 \times 3$ kernels and are always preceded, except for the input and output nodes, by a batch normalization (BN) layer [30] and a parametric rectified linear unit (PReLu) activation [31]. The parameter distribution is asymmetrical, with the residual blocks of the encoder part using two convolutional layers while a single one is used in the decoder. The network uses four different resolution levels, where the feature maps are downsampled by $2 \times 2 \times 2$ in each level of the encoder and upsampled back by the same factor in the decoder. Downsampling is performed by concatenating the result of a max pooling operation and strided convolution as proposed by Szegedy et al. [32], while upsampling is performed using a transposed convolution that learns the upsampling operator for each feature map. The last layer outputs a four-channel patch with the same $32 \times 32 \times 32$ spatial size as the input and is activated with a softmax to obtain a probability distribution among the background and three considered tissue classes.

Before extracting patches for either training or inference, we normalize the T1-w image intensities to standardize the input range and reduce the influence of outliers. More specifically, the intensity range is winsorized within the 0.05% and 99.95% percentiles and then the minimum and maximum intensities are mapped to the $[-1, 1]$ interval. To avoid influence from intensities not belonging to the brain tissues, image statistics are computed exclusively within a normalization mask, which is the coarse brain mask obtained through linear registration from a 2 mm resolution T1-w MNI template.

3.2.1. Training procedure

In this section, we provide a detailed overview of the procedure for training the presented pipeline, which includes the values for each hyperparameter and the network optimization strategy. The data used to

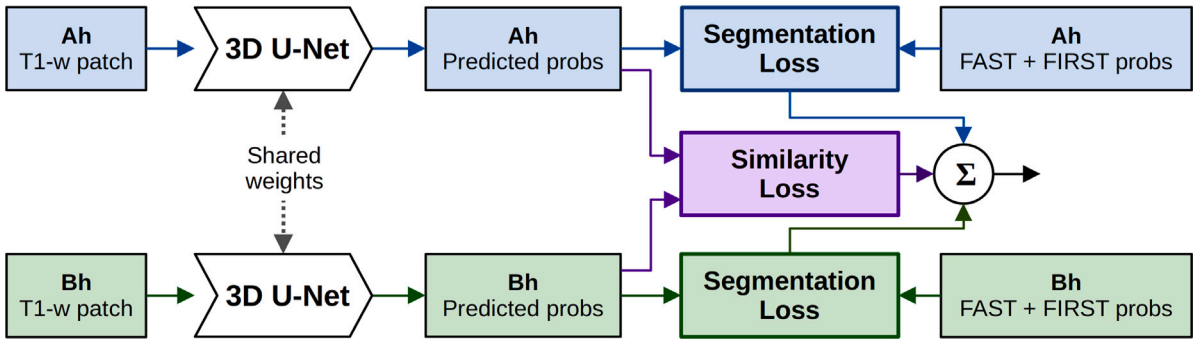


Fig. 3. Training iteration diagram of the proposed pipeline. The input comprises two patches extracted from a pair of halfway registered T1-w scans having a tissue similarity prior. As a segmentation target, we use the FAST + FIRST brain tissue probabilities computed in the native space of each T1-w scan and transformed to the halfway space. The two T1-w patches are predicted in two independent forward passes through the network, and two separate patch predictions are obtained. Then, a single backward pass is performed that updates the network weights to minimize the two segmentation loss terms, as well as the shared similarity loss term, which enforces the tissue similarity regularization.

train the proposed pipeline consists of the prepared halfway registered T1-w scans and their corresponding FAST + FIRST brain tissue probabilities derived from fsl_anat as the segmentation target. From these halfway registered scans, a patch set is generated with 100,000 pairs of samples, 85,000 for training and 15,000 for validation, extracted from the same spatial location of each halfway registered pair. The same number of patches is extracted from each of the available pairs with a deliberate sampling strategy to balance the representation of segmentation classes. For this purpose, we use the FAST + FIRST segmentations derived from fsl_anat as a guide to extract 25% of patches centered on CSF, 25% on GM, 25% on WM, 20% on the rest of the head and 5% from the background. To obtain a rough approximation of the nonparenchyma voxels, we define the head class as any nontissue voxel with a T1-w intensity greater than the mean of the image and the background class as any nontissue voxel with an intensity less than the mean. Additionally, a random 3D offset of up to half the patch size is applied to each sampled patch to increase the representation of class boundaries.

Once the training patch set is built, the randomly initialized network weights are iteratively trained following the procedure depicted in Fig. 3. Each training iteration consists of two separate forward passes through the network, obtaining a dense prediction for each halfway registered T1-w patch and a single backward pass that is used to update the network weights to minimize the loss function. In practice, each iteration is performed on a batch of 16 patch pairs, so that we first forward pass each of the 16 patch pairs and then perform a single backward pass from the average of their loss values. The network weights are updated through the Adadelta optimizer [33] with a learning rate of 0.05. To prevent overfitting, early stopping is performed when the loss on the validation set does not improve for 8 consecutive epochs.

As shown in Fig. 3, the training loss function comprises the sum of three terms: two of them come from segmentation loss terms, one for each T1-w patch, and the third is a shared similarity loss term that enforces the tissue similarity regularization during training. The probabilistic version of the cross-entropy loss (PCE) is used as the segmentation loss, targeting the partial volume probabilities of the FAST + FIRST segmentation derived from fsl_anat. Using probabilities as targets, instead of categorical labels, we encourage approximating the partial volume probabilities instead of attempting to maximize the probability of the most likely tissue class. More specifically, given a predicted probability distribution of a patch P over C classes with dimensions $C \times X \times Y \times Z$ and a target probability distribution T of the same dimensions, the probabilistic cross-entropy segmentation loss term is defined as:

$$\mathcal{L}_{\text{seg}}(P,T) = \frac{1}{XYZ} \sum_{x,y,z} \sum_{c_i=0}^{C-1} T(c_i,x,y,z) \cdot -\log \frac{\exp(P(c_i,x,y,z))}{\sum_{c_j=0}^{C-1} \exp(P(c_j,x,y,z))} \quad (1)$$

The tissue similarity regularization is implemented through the similarity loss term, which is taken as the sum of the L1 norm between the CSF, GM and WM percentages of the two predicted patches, ignoring the background class. More specifically, given two patches with output probability distributions, P_a and P_b, over C classes with dimensions $C \times X \times Y \times Z$, the similarity loss term is defined as:

$$\mathcal{L}_{\text{sim}}(P_a,P_b) = \sum_{c=1}^{C-1} \frac{100}{XYZ} \left| \sum_{x,y,z} P_a(c,x,y,z) - \sum_{x,y,z} P_b(c,x,y,z) \right| \quad (2)$$

Note that the two patches P_a and P_b used in the similarity loss term are forward passed separately so that the model cannot extract joint features between the short-interval scans to reduce the volume differences. In this way, the model is constrained to the use of cross-sectional features acting on a single patch to achieve this reduction. As a result, we obtain a model that performs inference on a single image at a time but does so with reduced quantification error thanks to the training regularization.

In summary, given two predicted tissue probability distributions P_a and P_b and their corresponding target probability distributions T_a and T_b, respectively, the loss function is defined as:

$$\mathcal{L}(P_a,P_b,T_a,T_b) = \mathcal{L}_{\text{seg}}(P_a,T_a) + \mathcal{L}_{\text{seg}}(P_b,T_b) + w_{\text{sim}} \cdot \mathcal{L}_{\text{sim}}(P_a,P_b) \quad (3)$$

where w_{sim} is a term that modulates the degree to which the model will be allowed to deviate from approximating the reference segmentation probabilities and instead focus on reducing tissue volume differences between similar patches. Setting $w_{\text{sim}} = 0.0$ would set the optimization target purely on approximating the target tissue probabilities as faithfully as possible, and any deviation from the target would be penalized by the segmentation loss terms. However, to avoid learning the biases and errors of the reference method, a level of disagreement is needed with respect to the target segmentations to allow room for improvement. By increasing the value of w_{sim}, we progressively shift the optimization target away from approximating the target probabilities and toward reducing the segmentation differences between short-interval scans. However, if w_{sim} is set too high, the learned segmentation model would be allowed to excessively ignore the FAST + FIRST segmentations to the point at which it might produce anatomically unfeasible results. For this reason, the preferred value for w_{sim} is the smallest one that provides sufficient improvement in brain atrophy quantification. The effect on segmentation accuracy and atrophy quantification of the proposed regularization is analyzed later in Section 5.1.

3.2.2. Image inference

Within the proposed pipeline, image inference performing end-to-end skull stripping and brain tissue segmentation is performed on a single T1-w scan in its native space, i.e., without image interpolation. First, input normalization is performed on the T1-w image as

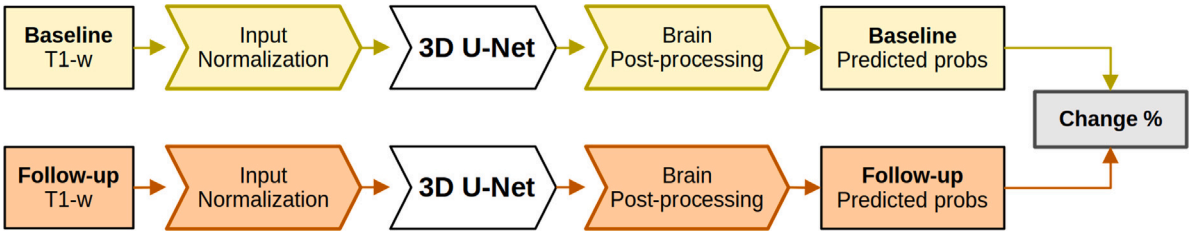


Fig. 4. Longitudinal inference procedure. The baseline and follow-up images are independently segmented in their native space and change measures are computed from the predicted tissue probability distributions.

previously described within a normalization mask obtained by linear transformation of a brain mask from a 2 mm resolution MNI template. Then, highly overlapping patches of size $32 \times 32 \times 32$ are extracted for inference at regular spatial steps of $10 \times 10 \times 10$. This level of overlap helps to reduce block boundary artifacts and improve spatial coherence. Before patch extraction, the T1-w image is edge padded on all sides by 16 voxels, which is half the patch size, to ensure that every voxel in the image is predicted with a similar degree of overlap. The extracted patches are then forward passed through the trained segmentation model, obtaining dense tissue probability distributions for each patch. The use of overlapping patches results in several brain tissue probability distributions for each voxel of the input image. To achieve the final whole image segmentation, the overlapping predictions are averaged and normalized to produce a single brain tissue probability distribution for each input image voxel.

Additionally, the brain tissue probabilities are postprocessed to improve the accuracy in intracranial cavity segmentation. Since the proposed pipeline performs end-to-end skull stripping and brain tissue segmentation, there is no assumption made regarding which voxels should be pure tissue or pure background, leading to small background probabilities appearing inside the intracranial cavity and small probabilities of tissue appearing outside of the brain. To reduce these small errors from compounding onto large volume measurement errors, postprocessing is performed based on the assumption that the intracranial cavity will be the largest connected component in the output segmentation. In practice, we first define a *pure tissue* mask as $p(\text{CSF}) + p(\text{GM}) + p(\text{WM}) > 0.99$, which is processed using morphological operators by filling holes and then keeping only the largest connected component. Within the *pure tissue* mask, the background probability is set to zero, and the remaining tissue probabilities are normalized to ensure that they total one. Outside of the *pure tissue* mask, the background probability is set to one, and all tissue probabilities are set to zero. From these probabilistic segmentations, measures of volume for each tissue are obtained by taking the brain tissue probability distribution of each voxel as an estimation of its partial volume mixture. In this way, the volume of each tissue class is calculated by totaling its voxelwise probability across the whole image and then normalizing by the voxel size to obtain the volume in mm^3.

For longitudinal evaluation, inference is performed independently for the baseline and follow-up images, in their native space, and then measures of change are computed from the predicted tissue probability distributions, as shown in Fig. 4.

3.3. Implementation details

The proposed method was implemented with Python using the Torch scientific computing framework [34]. All experiments were run on a GNU/Linux machine running the Ubuntu operating system, version 18.04, with 128 GB of RAM memory and an Intel®Core™ i7-7800X CPU. The versions of the software packages used were 6.0.4 for FSL and 6.0.0 for FreeSurfer. The network training and inference were performed with an NVIDIA 1080 Ti GPU (NVIDIA Corp., United States) with 12 GB G5X memory. The proposed network architecture has 7 million trainable parameters and takes 3.6 GB of GPU memory during training and 1.5 GB for inference. The time to perform inference for a whole image using the proposed pipeline within our system is between 2 and 3 min, depending on the image dimensions. The linear registration to obtain the normalization mask takes approximately 1 min, while inference of an image using the GPU takes between 1 and 2 min.

4. Evaluation

The evaluation is performed on two publicly available MRI datasets, MIRIAD and ADNI1, which have short-interval and longitudinal imaging for healthy controls (HCs) and AD subjects. The proposed pipeline is evaluated with a subject-wise cross-validation strategy in 3 folds, allocating in each fold a different two thirds of the subjects for training and validation and the remaining third for testing and evaluation, as detailed in Table 1. A 3-fold cross-validation was selected to achieve a balance between robust evaluation and computational efficiency. While a cross-validation using more folds could offer slightly more granular performance estimates, it would have substantially increased computational demands. Moreover, the use of fewer folds provides a more restrictive evaluation since in each iteration the model is trained on less data samples and evaluated on a larger amount of test images.

Within each fold, the coregistered short-interval scan pairs from the training and validation subjects are used to extract a total of 100,000 pairs of patches, 85,000 for training and 15,000 for validation, with size $32 \times 32 \times 32$ and used to train the pipeline as described in Section 3.2.1. In the MIRIAD dataset, we consider for training all the 182 scan-rescan pairs as well as 125 scan pairs made within 2 weeks, for a total of 307 scan pairs. For the ADNI1 cohort, we use all the 541 available scan-rescan pairs for training. Next, inference is performed for the subjects in the testing set, segmenting the maximum interval scan pairs, i.e., the first and last available timepoints, and also the short-interval scan pairs for evaluation of quantification error. While short-interval scans are registered to a halfway space for training and validation, inference on short-interval scan pairs is performed in their native space without image interpolation. After completing the three folds, segmentations for all the scans in the dataset are obtained and the evaluation metrics are calculated for the whole dataset.

We first studied the effect of tissue similarity regularization on brain volumetry with an ablation study performing several subject-wise cross-validations with increasing amounts of tissue regularization controlled by the w_{sim} parameter. More specifically, we performed seven cross-validation evaluations for each dataset considering values for w_{sim} from 0.0 to 0.6. The proposed pipeline was independently evaluated on each dataset using the available short-interval scan pairs having a tissue similarity prior and preparing the data for training as described in Section 3.1. From the results of this experiment, we set an optimal default value for w_{sim} and then performed a detailed quantitative analysis of the pipeline with the selected optimal weight. Finally, we also performed a cross-dataset evaluation to study domain shift performance, using the models trained on the MIRIAD dataset to test on ADNI1 and vice-versa.

We compared our results to the FAST + FIRST brain tissue probabilities derived from fsl_anat used as training targets and with SIENA, a

Table 1
Number of subjects and scan pairs for each set of the cross-validation folds. Training and validation is performed using the co-registered short-interval scan pairs while testing is done on the maximum interval and short-interval scan pairs in their native space.

		MIRIAD			ADNI1		
		Training	Validation	Testing	Training	Validation	Testing
Fold 1	Subjects	40	6	23	156	27	87
	Scan pairs	175	20	135	302	57	272
Fold 2	Subjects	40	6	23	156	27	89
	Scan pairs	185	20	125	299	57	277
Fold 3	Subjects	40	6	23	157	27	75
	Scan pairs	185	20	116	321	52	246

well-known and widely used state-of-the-art brain atrophy quantification method also implemented in FSL. In practice, SIENA is run with the `-R` option in MIRIAD, which iterates the skull stripping several times to robustly estimate the brain center, and with the `-B` option in ADNI1, which removes the neck present on the images.

Measures of whole-brain atrophy in segmentation-based methods are typically based on the volume change of brain parenchyma between the baseline and follow-up segmentations, which can be computed either from raw or from normalized volumes. Additionally, since our pipeline not only segments the parenchyma but also its distinct gray and white matter components, we also provide individualized measures of change for these tissues. To account for different time intervals between longitudinal scans of different subjects, all reported measures of change are annualized. Relative change measures are not computed relative to the baseline or follow-up volumes, but instead we do so with respect to their average as follows:

$$\text{Change } \% = 100 \cdot \frac{2(V_{\text{follow-up}} - V_{\text{baseline}})}{V_{\text{follow-up}} + V_{\text{baseline}}} \tag{4}$$

where V can be any measure of volume derived from the probabilistic brain tissue segmentations. The percentage of brain volume change (PBVC) is obtained when V is set as the raw volume of the brain parenchyma. Similarly, setting V as the raw volume of the GM or WM provides the percentage of GM volume change (PGMVC) or percentage of WM volume change (PWMVC), respectively. However, measures of change based on raw unscaled volumes are affected by a number of technical and physiological confounding factors [10]. A more robust measure of change can be obtained using tissue fractions, which are computed by normalizing the raw tissue volumes with respect to the intracranial volume (ICV), computed as the sum of all tissue volumes (CSF + GM + WM). In this way, the brain parenchymal fraction (BPF), gray matter fraction (GMF) and white matter fraction (WMF) are obtained by normalizing their respective raw volume measurements by the intracranial volume. Additionally, to study the longitudinal skull stripping consistency of the proposed pipeline, the ICV change is also measured by setting V as the raw intracranial volume. Although the ICV has been shown to decrease as a result of aging beyond adulthood [35], this change is expected to be close to zero within the time intervals between scans of the considered datasets.

These measures of volume change are also computed for the short-interval scan pairs to evaluate the quantification error. Between these images, an ideal atrophy quantification method should measure zero change between them; therefore, we consider any deviation from zero as quantification error.

In the absence of an atrophy ground truth, the measures of change by themselves are not indicative of the accuracy or quality of the brain tissue segmentations. However, the sensitivity of an atrophy quantification method to longitudinal changes can be assessed by quantifying the differences between two subject populations known to have different rates of change. In this way, atrophy quantification methods can be compared based on the assumption that better methods would detect larger and more pronounced differences between these two populations. In our case, we quantified differences between HC and AD subjects based on their annualized change measures. As in the work of Smith et al. [36], a measure of discriminative power can be obtained from the t statistic of Welch's unequal variances test, which quantifies confidence in the existence of differences between both groups. Welch's unequal variances test was selected due to its higher reliability within two unpaired populations which have unequal variances and possibly unequal sample sizes [37]. In this test, a large t provides a high level of evidence that the observed differences between the two populations are statistically significant — in other words, a low probability that the observed differences could be due to chance. However, t does not reflect the strength or size of these differences; for instance, a large t could be obtained for a very small difference in the magnitude of annualized change, which would not necessarily be of any practical significance or clinical importance. For this purpose, measures of effect size are typically used to quantify the magnitude or strength of observed differences. More specifically, we use the Cohen's d [38] to measure the effect size, which is calculated as:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s} \tag{5}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the set of annualized change measures of the HC and AD groups, respectively, and s, the pooled standard deviation, is defined for two independent populations as:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{6}$$

where n_1 and n_2 are the number of samples in each population, and s_1^2 and s_2^2 are the variances of each group, computed as:

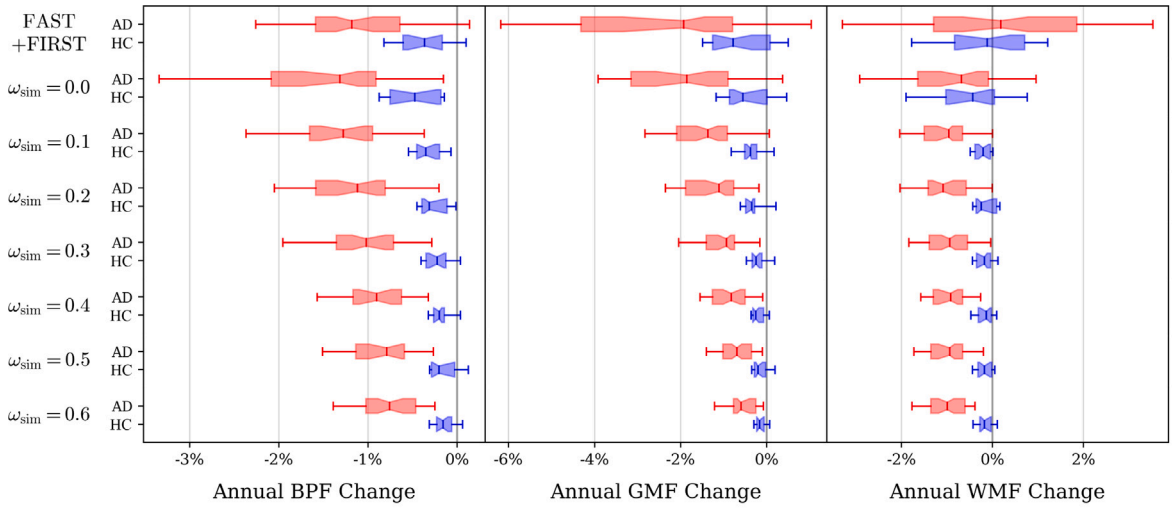
$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1,i} - \bar{x}_1)^2$$
$$s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{2,i} - \bar{x}_2)^2 \tag{7}$$

We also calculate the Dice similarity coefficient (DSC) between the segmentations of the proposed pipeline with respect to the FAST + FIRST reference derived from fsl_anat to quantify the extent to which the tissue similarity regularization shifts the segmentation away from the target. In practice, we calculate the DSC by first taking the argmax of the brain tissue probabilities to obtain a categorical multiclass segmentation.

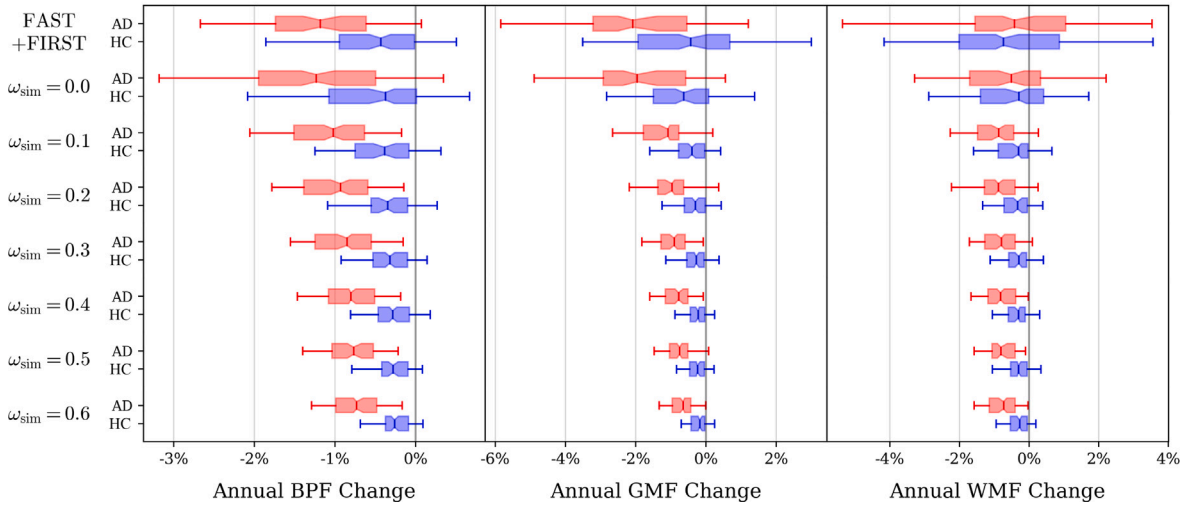
5. Results and discussion

5.1. Similarity weight analysis

In this experiment, we study the effect of the tissue similarity regularization on the proposed pipeline by performing seven cross-validations with increasing values for w_{sim}, from 0.0 to 0.6. To analyze the effect on brain atrophy quantification properties, we studied the response of annualized change measures as well as any improvement in the sensitivity to differences between healthy and AD subjects of these change measures. We also study how regularization affects the tissue segmentation model by calculating short-interval error measures, as well as the DSC with respect to the reference FAST + FIRST segmentations.



(a) MIRIAD dataset



(b) ADNI1 dataset

Fig. 5. Annualized change measures between maximum interval pairs of healthy controls (HC) in blue and Alzheimer's disease (AD) patients in red for the FAST + FIRST reference segmentations and the proposed pipeline with increasing similarity regularization weight. The boxes representing the interquartile range are notched within the confidence interval around the median, with the left and right whiskers set to the 5th and 95th percentiles, respectively.

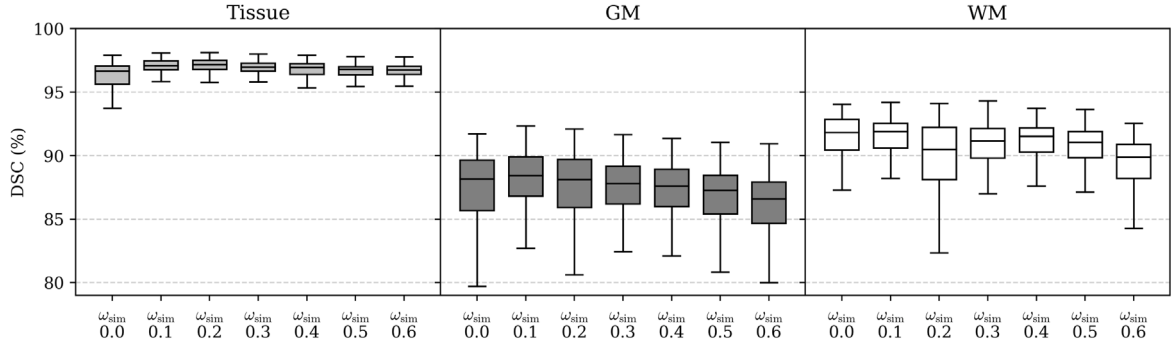
Figs. 5(a) and 5(b) show boxplots of annualized change measures of BPF, GMF and WMF between the maximum interval pairs of the MIRIAD and ADNI1 datasets, respectively. Overall, increasing values of w_{sim} reduce the standard deviation of all considered change measures in both datasets. The reduction in variability is more pronounced in the MIRIAD measures, especially for the healthy subjects, most likely due to the high similarity between images acquired with the same scanner and imaging protocol. In both datasets, higher values of w_{sim} reduce the median BPF and GMF change, while the median WMF change is negatively increased.

The effect of increasing regularization on the discriminative power and effect size of change measures between healthy and AD subjects is summarized in Table 2. The results show that tissue similarity regularization improves the discrimination and effect size between groups in all change measures in both the MIRIAD and ADNI1 datasets. In the MIRIAD dataset, the proposed pipeline with $w_{sim} = 0.0$ already improves the sensitivity of BPF, GMF and WMF change compared to the reference FAST + FIRST segmentations. When the regularization is enforced, increasing the similarity weight value further improves the results until $w_{sim} = 0.4$, where the improvement reaches its peak, and beyond this point, higher values actually worsen the differences between groups. In the ADNI1 dataset, the proposed pipeline without regularization ($w_{sim} = 0.0$) improves both the GMF and WMF change sensitivity while having a worse effect on BPF change compared with the reference FAST + FIRST segmentations. When the regularization is enforced, the sensitivity in all three measures steadily improves for higher values of w_{sim}. In contrast to the MIRIAD results, the sensitivity of ADNI1 measures does not peak at $w_{sim} = 0.4$, but beyond this point, improvement gains decrease rapidly.

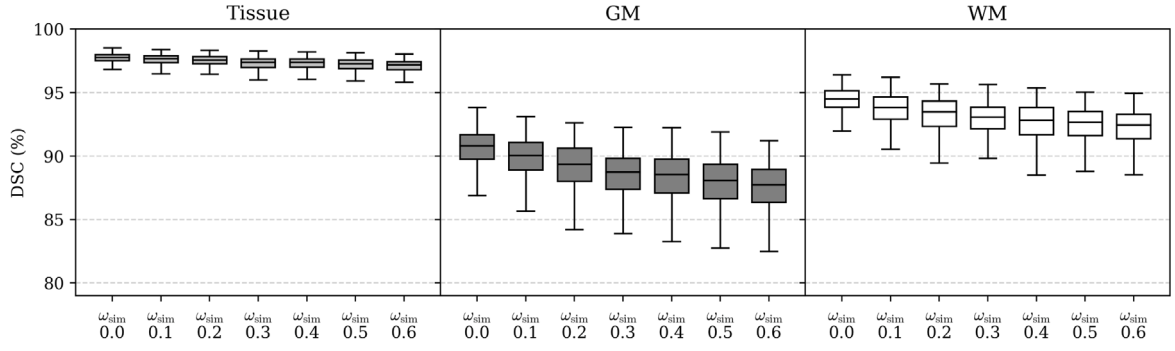
Fig. 6 shows DSC measures between the reference FAST + FIRST segmentations and the proposed pipeline with increasing w_{sim} values. The reported DSC results are calculated from the argmax classification of the probabilistic brain tissue segmentations and are given separately for parenchyma (GM+WM) as well as for its GM and WM components. As expected, higher amounts of regularization decrease the similarity with respect to the reference FAST + FIRST brain tissue segmentations. Moreover, it can also be observed that the DSC of GM and WM components that form the parenchyma decrease much more rapidly with increasing regularization than those of the parenchyma itself. This outcome suggests that the dissimilarity is due to a redistribution of probabilities between GM and WM classes. By increasing w_{sim}, the learning focus is progressively shifted away from approximating

Table 2
Discrimination and effect size of annualized change measures between healthy controls (HC) and Alzheimer's disease (AD) subjects for the maximum interval scan pairs. The discriminative power is measured with the t statistic from Welch's unequal variances test, while the effect size is measured using Cohen's d.

Method	MIRIAD						ADNI1					
	ΔBPF		ΔGMF		ΔWMF		ΔBPF		ΔGMF		ΔWMF	
	t	d	t	d	t	d	t	d	t	d	t	d
FAST + FIRST	4.66	1.01	3.35	0.71	−0.49	0.10	6.19	0.80	5.29	0.69	−0.56	0.07
$w_{sim} = 0.0$	5.74	1.21	4.91	1.05	1.24	0.27	5.24	0.70	6.18	0.80	0.67	0.09
$w_{sim} = 0.1$	9.02	1.86	6.01	1.21	6.61	1.26	8.30	1.11	7.08	0.92	5.08	0.66
$w_{sim} = 0.2$	9.02	1.84	6.47	1.30	6.62	1.29	9.14	1.23	7.73	1.02	5.87	0.77
$w_{sim} = 0.3$	9.30	1.86	7.42	1.49	7.37	1.45	9.69	1.31	8.49	1.12	7.01	0.93
$w_{sim} = 0.4$	10.61	2.07	7.77	1.59	10.05	2.05	10.07	1.37	8.68	1.14	7.16	0.95
$w_{sim} = 0.5$	9.44	1.90	7.03	1.47	9.36	1.87	10.29	1.40	8.78	1.18	7.52	1.00
$w_{sim} = 0.6$	9.53	1.90	6.13	1.28	9.60	2.03	10.28	1.40	8.84	1.18	7.48	0.99



(a) MIRIAD dataset



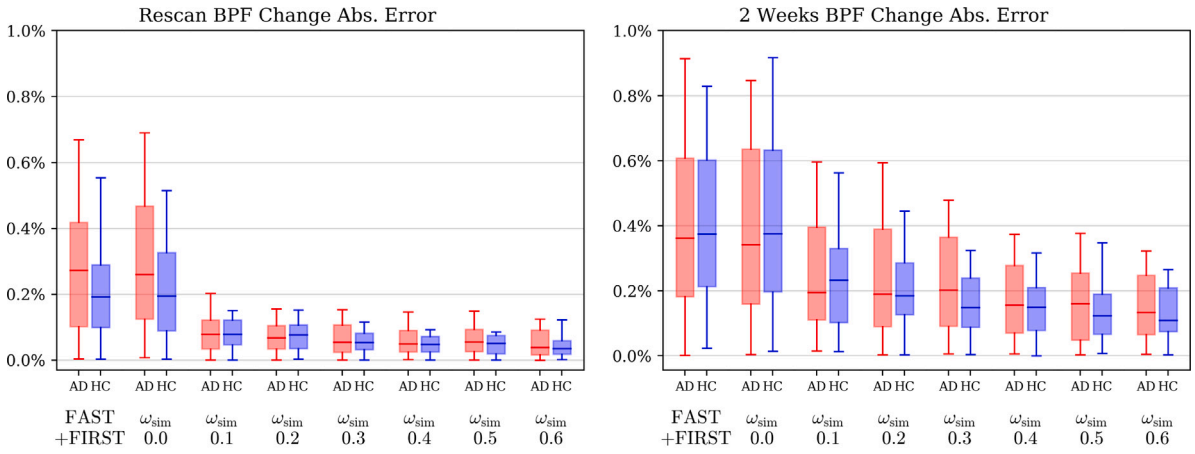
(b) ADNI1 dataset

Fig. 6. DSC of the maximum interval pair segmentations of the proposed pipeline with respect to reference FAST + FIRST segmentations used for training with increasing tissue similarity regularization weights.

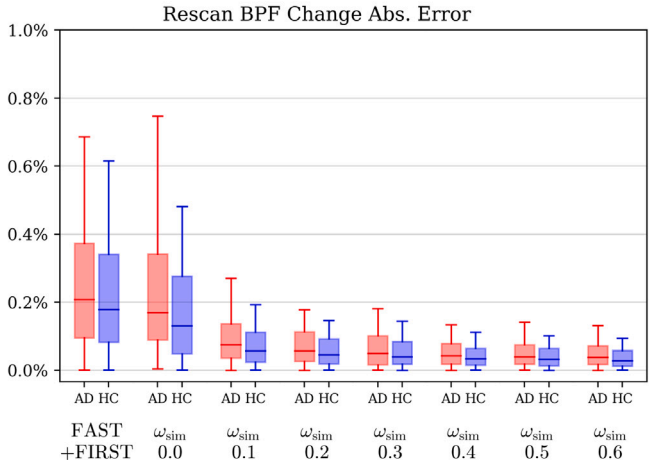
the FAST + FIRST probabilities, and a greater degree of deviation is allowed to reduce the segmentation differences between short-interval scans.

Figs. 7(a) and 7(b) show the absolute BPF change error between short-interval scans of the two considered datasets. Without regularization ($\omega_{sim} = 0.0$), the proposed pipeline exhibits levels of error similar to those of the reference FAST + FIRST segmentations. Even when the smallest amount of regularization is enforced ($\omega_{sim} = 0.1$), the error is greatly reduced, with higher weights providing smaller improvements thereafter. Fig. 7(a) also shows that, despite both the rescan and 2-week pairs of the MIRIAD dataset being used for training, the rescan pairs without repositioning show a much greater reduction in error than the 2-week pairs. This outcome suggests that the scan differences due to repositioning and/or small time intervals are larger than those of the rescan images. We expected that the use of the 2-week pairs with repositioning for training in the MIRIAD dataset would also help in improving the scan-rescan error. However, the scan-rescan error obtained for MIRIAD is only slightly lower than the one obtained in ADNI1, which only used rescan images without repositioning. These results suggest that the proposed regularization does not especially benefit from images with repositioning to reduce the quantification error.

The experiments performed have shown that the proposed regularization can improve the sensitivity of atrophy measures to differences between healthy and AD subjects. However, these improvements are obtained at the cost of decreasing the segmentation similarity with respect to the reference FAST + FIRST segmentations. In our experiments, the sensitivity to differences between groups reached its maximum at $w_{sim} = 0.4$ for the MIRIAD dataset while reaching a point of diminishing returns at $w_{sim} = 0.4$ for the ADNI1 dataset. Thus, we decided on $w_{sim} = 0.4$ as an optimal default value for the proposed pipeline, providing the most improvement for the least deviation from the reference segmentations.



(a) MIRIAD dataset



(b) ADNI1 dataset

Fig. 7. Short interval error of BPF for the MIRIAD and ADNI1 datasets with increasing tissue similarity regularization.

Table 3
Annualized measures of ICV change (mean ± std. dev.) for the proposed and fsl_anat pipelines.

		MIRIAD		ADNI1	
		fsl_anat	Proposed ($w_{sim} = 0.4$)	fsl_anat	Proposed ($w_{sim} = 0.4$)
|ΔICV|	HC	0.38 ± 0.61%	0.27 ± 0.35%	0.80 ± 1.49%	0.43 ± 0.69%
	AD	0.67 ± 0.91%	0.62 ± 0.61%	0.83 ± 1.42%	0.42 ± 0.30%
ΔICV	HC	−0.23 ± 0.69%	−0.17 ± 0.41%	0.20 ± 1.68%	−0.10 ± 0.81%
	AD	−0.25 ± 1.11%	−0.38 ± 0.78%	0.30 ± 1.62%	−0.26 ± 0.44%

5.2. Longitudinal atrophy quantification analysis

In the previous section, we studied the effect of varying degrees of tissue similarity regularization on the presented deep learning pipeline for brain tissue segmentation in both single-site and multisite datasets. In this section, we now perform a detailed quantitative and qualitative evaluation of the presented pipeline trained with $w_{sim} = 0.4$, the empirically selected optimal regularization weight.

5.2.1. Intracranial volume change

The results for ICV measurements of the selected model can be found in Table 3. On average, the absolute ICV change of the proposed pipeline is significantly lower in ADNI1 ($p < 10^{-4}$) and marginally lower in MIRIAD than for the fsl_anat reference, suggesting a more consistent intracranial volume between longitudinal scans. In terms of ICV change, the brain masks from fsl_anat show a similarly negative rate in the MIRIAD dataset for both HC and AD subjects, while the ADNI1 dataset shows positive ICV changes for both groups, with a slightly higher average rate for the AD subjects. The proposed pipeline obtains a much more consistent ICV change between datasets and subject groups, having a small negative ICV change for the healthy subjects and a larger negative change for the AD subjects. The results suggest that the learned skull stripping of our pipeline is somehow affected by global atrophy since the longitudinal ICV change is negative and more pronounced for the AD group. This outcome would be caused by the way in which skull stripping is performed by fsl_anat, which nonlinearly registers a dilated brain mask to segment the brain parenchyma. In this way, instead of attempting to segment the entire intracranial cavity, fsl_anat essentially sets a fixed band around the parenchyma that does not encompass the entire intracranial cavity. Thus, in cases with greater amounts of atrophy, the fixed band around a more shrunken parenchyma means that there will be a larger amount of the intracranial cavity which will not be segmented by fsl_anat. Within the presented pipeline, tissue similarity regularization cannot reduce

Table 4
Scan-rescan absolute error (mean ± std. dev. (median)) for raw and normalized change measures. In the case of SIENA, which uses the BSI method, PBVC is the only provided measure of atrophy.

	MIRIAD			ADNI1		
	FAST + FIRST	Proposed ($w_{sim} = 0.4$)	SIENA	FAST + FIRST	Proposed ($w_{sim} = 0.4$)	SIENA
PBVC	0.36 ± 0.51% (0.23%)	0.17 ± 0.20% (0.11%)	0.31 ± 0.35% (0.20%)	0.37 ± 0.62% (0.20%)	0.14 ± 0.69% (0.06%)	0.33 ± 0.43% (0.18%)
ΔBPF	0.31 ± 0.35% (0.23%)	0.06 ± 0.06% (0.05%)	N/A	0.31 ± 0.45% (0.19%)	0.07 ± 0.13% (0.04%)	N/A
ΔGMF	1.13 ± 1.13% (0.76%)	0.09 ± 0.08% (0.07%)	N/A	1.00 ± 1.09% (0.62%)	0.09 ± 0.21% (0.05%)	N/A
ΔWMF	1.27 ± 1.46% (0.74%)	0.10 ± 0.12% (0.07%)	N/A	1.18 ± 1.24% (0.76%)	0.11 ± 0.32% (0.07%)	N/A

Table 5
Absolute error between 2 week interval pairs of the MIRIAD dataset (mean ± std. dev. (median)) for raw and normalized change measures. In the case of SIENA, which uses the BSI method, PBVC is the only provided measure of atrophy.

	MIRIAD		
	FAST + FIRST	Proposed ($w_{sim} = 0.4$)	SIENA
PBVC	0.65 ± 0.83% (0.39%)	0.50 ± 0.76% (0.24%)	0.44 ± 0.50% (0.33%)
ΔBPF	0.49 ± 0.47% (0.38%)	0.18 ± 0.14% (0.15%)	N/A
ΔGMF	1.40 ± 1.32% (1.05%)	0.20 ± 0.17% (0.16%)	N/A
ΔWMF	1.52 ± 1.64% (0.94%)	0.24 ± 0.18% (0.20%)	N/A

the learning of this fixed band bias, and the measured ICV is affected by the brain shrinkage observed on follow-up scans, which is higher for AD subjects than for healthy controls.

5.2.2. Short interval error

Table 4 shows the percentage of absolute volume error between scan-rescan pairs for raw and normalized measures of change in the ADNI1 and MIRIAD dataset. Compared with the reference FAST + FIRST results and those of SIENA, the short interval error of our pipeline is significantly lower in all measures of both datasets ($p < 10^{-6}$). Moreover, while the reference FAST + FIRST segmentations have much greater error for individual GM and WM tissue than for the parenchyma, in the proposed pipeline, the error is much more similar between the parenchyma and its GM and WM components. These results show that tissue similarity regularization not only reduces the quantification error of the pipeline but also increases the consistency of measured GM and WM volumes between short-interval scans.

Table 5 shows the quantification error between pairs of scans in the MIRIAD dataset taken within a 2 week interval, which should ideally be close to zero. As expected, the quantification error is larger than the one measured for the scan-rescan pairs. Compared with FAST + FIRST, the proposed pipeline obtains a significantly lower error in all measures of both datasets ($p < 10^{-4}$). Although the median error of our pipeline is lower than SIENA, our method has a marginally larger mean, which suggests a lower error in the majority of cases and larger errors in the more confounding ones. Similarly to the scan-rescan results, the error for ΔGMF and ΔWMF has a greater reduction than the one obtained for the ΔBPF, showing that the volumes measured by our pipeline are much more consistent between 2 week interval pairs than the reference FAST + FIRST segmentations.

5.2.3. Annualized atrophy rates

Table 6 shows the annualized rates of ΔBPF, ΔGMF, ΔWMF, PBVC, PGMVC and PWMVC of all maximum interval pairs for the reference FAST + FIRST segmentations and the proposed pipeline. In general, our pipeline shows much less variability in all measures of change than the FAST + FIRST reference segmentations. Compared with the results of FAST + FIRST, the annualized PBVC in the MIRIAD dataset is slightly reduced, especially for the HC subjects, while it is slightly increased in the ADNI1 dataset for both subject groups. In terms of BPF changes, the average rate of the proposed pipeline is reduced in both datasets compared to the reference FAST + FIRST segmentations. This finding would be mostly explained by the generally smaller ICV obtained by our pipeline for the follow-up scans, especially for the AD subjects, which slightly biases the follow-up tissue fractions toward larger values and reduces the apparent atrophy rate.

It can also be observed that the WMF change, as measured from the FAST + FIRST segmentations suggests that healthy controls have greater WM atrophy than AD subjects. In contrast, our pipeline shows greater WM atrophy for the AD subjects than for the healthy controls, which makes more intuitive sense in the context of a generalized brain atrophy process. As seen in Fig. 5, the amount of regularization is directly related to the lowering of the median WMF change, suggesting that the segmentation of WM is directly improved by tissue similarity regularization.

For comparison, we also calculated the annualized atrophy rates with SIENA [12]. Our PBVC results in the MIRIAD dataset (HC: −0.26 ± 0.43%; AD: −1.31 ± 0.86%) shows reduced average rates for both groups, with a pronounced reduction of variability for the AD group, compared to those of SIENA (HC: −0.53 ± 0.45%; AD: −2.43 ± 1.34%). In the ADNI1 dataset, the annualized PBVC of our pipeline (HC: −0.41 ± 0.92%; AD: −1.14 ± 0.76%) also shows lower average rates when compared to SIENA (HC: −0.61 ± 0.75%; AD: −1.85 ± 1.13%). The higher rates measured by SIENA might be caused by the way in which the BSI method extrapolates the PBVC for the whole image based on the displacement of the boundary surface between gray matter (GM) and white matter (WM) between timepoints. By deriving the measure of longitudinal change from the analysis of a comparatively small region within the image, it makes SIENA more sensitive to changes in this particular area. In contrast, our pipeline independently segments the brain tissue at each timepoint and aggregates the partial volumes across the entire brain to quantify longitudinal volume change. As a result, a relatively small change in the GM/WM boundary region yields a more pronounced percentage change in SIENA than for our method.

5.2.4. Sensitivity to differences between groups

Table 7 shows the sensitivity to differences between groups of normalized and unnormalized measures of change for the MIRIAD and ADNI1 dataset. For both of the segmentation-based methods, FAST + FIRST and proposed pipeline, the normalized measures of change provide much larger sensitivity to differences than the PBVC. In both the MIRIAD and ADNI datasets, the ΔBPF is improved by tissue similarity regularization results with respect to the FAST + FIRST reference, and is also better than the results achieved by SIENA. The sensitivity of the ΔGMF and ΔWMF measures is also improved with respect to the

Table 6
Annualized measures of atrophy (mean ± std. dev.) from maximum interval scan pairs.

		MIRIAD		ADNI1	
		FAST + FIRST	Proposed ($w_{sim} = 0.4$)	FAST + FIRST	Proposed ($w_{sim} = 0.4$)
ΔBPF	HC	−0.35 ± 0.51%	−0.18 ± 0.15%	−0.52 ± 0.84%	−0.29 ± 0.30%
	AD	−1.13 ± 0.87%	−0.91 ± 0.42%	−1.23 ± 0.92%	−0.80 ± 0.45%
ΔGMF	HC	−0.46 ± 1.50%	−0.19 ± 0.23%	−0.51 ± 2.12%	−0.25 ± 0.44%
	AD	−2.21 ± 2.83%	−0.87 ± 0.50%	−2.07 ± 2.41%	−0.79 ± 0.52%
ΔWMF	HC	−0.21 ± 1.25%	−0.17 ± 0.21%	−0.54 ± 2.41%	−0.34 ± 0.44%
	AD	0.02 ± 2.61%	−0.97 ± 0.45%	−0.36 ± 2.53%	−0.81 ± 0.55%
PBVC	HC	−0.58 ± 0.74%	−0.35 ± 0.42%	−0.32 ± 1.46%	−0.39 ± 0.91%
	AD	−1.38 ± 1.35%	−1.30 ± 0.94%	−0.93 ± 1.45%	−1.06 ± 0.71%
PGMVC	HC	−0.69 ± 1.39%	−0.36 ± 0.47%	−0.31 ± 2.27%	−0.35 ± 0.86%
	AD	−2.47 ± 3.15%	−1.25 ± 0.91%	−1.77 ± 2.50%	−1.05 ± 0.73%
PWMVC	HC	−0.44 ± 1.59%	−0.33 ± 0.44%	−0.34 ± 2.83%	−0.44 ± 1.07%
	AD	−0.24 ± 2.64%	−1.35 ± 1.03%	−0.06 ± 2.95%	−1.07 ± 0.80%

Table 7
Discrimination and effect size of annualized change measures between healthy controls (HC) and Alzheimer's disease (AD) subjects for the maximum interval scan pairs. The discriminative power is measured with the t statistic from Welch's unequal variances test, while the effect size is measured using Cohen's d. In the case of SIENA, which uses the BSI method, PBVC is the only provided measure of atrophy.

Method	MIRIAD				ADNI1			
	ΔBPF		PBVC		ΔBPF		PBVC	
	t	d	t	d	t	d	t	d
FAST + FIRST	4.66	1.01	3.21	0.68	6.19	0.80	3.27	0.42
Proposed ($w_{sim} = 0.4$)	10.61	2.07	5.76	1.17	10.07	1.37	6.48	0.80
SIENA	–	–	8.99	1.73	–	–	9.78	1.33

Table 8
Results of leave-one-out cross-validation on the MRBrains18 challenge dataset calculated using the evaluation scripts provided by the challenge organizers.

	DSC (%)	HD95 (mm)	Volume similarity	IoU
CSF	83.13 ± 1.63	2.40 ± 0.37	0.97 ± 0.02	0.712 ± 0.02
GM	84.84 ± 2.08	1.79 ± 0.66	0.95 ± 0.03	0.737 ± 0.03
WM	87.34 ± 2.80	2.26 ± 0.64	0.94 ± 0.03	0.776 ± 0.04

reference FAST + FIRST segmentations. Overall, the effect size of all measures is larger for the MIRIAD dataset than for ADNI1, most likely due to the more consistent imaging parameters that introduce a lower level of confounding factors.

5.2.5. Tissue segmentation

The cross-sectional brain tissue segmentation performance of the presented pipeline is evaluated both quantitatively and qualitatively. Quantitative evaluation is performed using the international MRBrains18 challenge dataset [39], which provides 8 cases with semi-automated gold standard tissue segmentations revised by human raters. In the qualitative evaluation, the effect of regularized deep learning is assessed by comparing overall segmentation differences between the output of the presented pipeline and the FAST + FIRST segmentations.

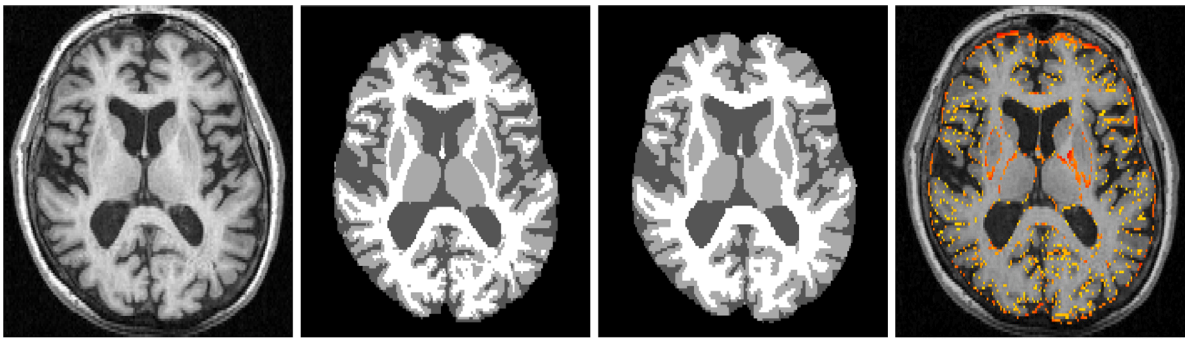
Quantitative results on the MRBrains18 dataset were obtained with a leave-one-out cross-validation strategy, and calculating the challenge metrics using the evaluation code provided by the organizers. To train our longitudinal pipeline on the cross-sectional MRBrains18 dataset, the same image is taken as both scan and rescan and the value of the similarity loss weight is set to zero ($\omega_{sim} = 0.0$). Additionally, out of the three MRI modalities included in the challenge dataset (T1w, T1w-IR and T2w-FLAIR), only the T1w is used for training and inference with the presented pipeline. Table 8 shows the results obtained from the leave-one-out cross-validation, which are in line with state-of-the-art results also using only the T1w modality for training and inference [40].

To study the effect of regularized deep learning on the resulting brain tissue segmentations, we perform a qualitative evaluation comparing the reference FAST + FIRST segmentations to those of the presented pipeline. Fig. 8 shows the tissue segmentation results of FAST + FIRST and our pipeline for two representative cases of MIRIAD and ADNI1. In both datasets, the segmentation of our pipeline presents some differences with respect to the reference, having generally smoother edges between tissues and less noise. The largest segmentation differences are located in the outer brain interface, where our pipeline tends to segment a larger area as brain, and in the borders of subcortical structures, which depending on the case are either enlarged or shrunken. Smaller segmentation differences are also observed in the interfaces between tissues throughout the cortex, where our pipeline tends to segment less WM and more GM than the reference FAST + FIRST segmentation.

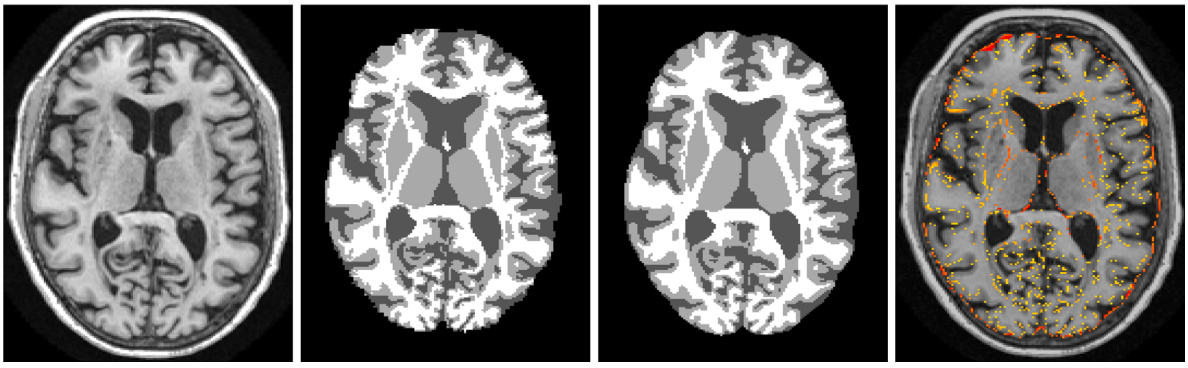
Fig. 9 shows the median differences between the probabilistic segmentations of FAST + FIRST and the proposed pipeline across all of the cases from each dataset. In practice, we subtract the FAST + FIRST probabilities of each tissue from those of our pipeline, which are then transformed to the MNI space, where we obtain the voxelwise median across all available cases for each dataset. The differences for both datasets show a very similar behavior, with MIRIAD displaying stronger differences, most likely due to its more homogeneous single-center images. In terms of CSF, the blue color around the outer brain border indicates a tendency for our pipeline to segment more CSF in that region compared to the reference method. Conversely, the red color in the midline region, ventricle borders and temporal lobes suggests that our pipeline segments less CSF in these regions when compared to FAST + FIRST. Median GM differences display a generalized blue color throughout the cortex, while the WM differences take on a red color, indicating that the presented pipeline segments less WM and more GM in those regions than FAST + FIRST. Another area showing large differences consists of the subcortical structures; the red color in their inner borders suggests that our pipeline tends to reduce their size compared to FAST + FIRST. However, this behavior is reversed when examining the outer borders, where more GM is segmented in favor of reducing the WM.

5.3. Cross-dataset evaluation

We study the domain shift sensitivity of the presented pipeline by doing a cross-dataset evaluation, where we train and validate our model using one of the two datasets and subsequently test it on the other one. In practice, we inference each scan once with each of the three models trained during the subject-wise cross-validation and then perform a voxel-wise average to obtain a single final tissue probability map. From these segmentations, we then compute the relevant evaluation metrics for each dataset. To study the effect of tissue similarity regularization on domain shift performance, we show the results for the case without regularization ($w_{sim} = 0.0$) as well as for the empirically selected optimal weight ($w_{sim} = 0.4$).



(a) MIRIAD dataset



(b) ADNI1

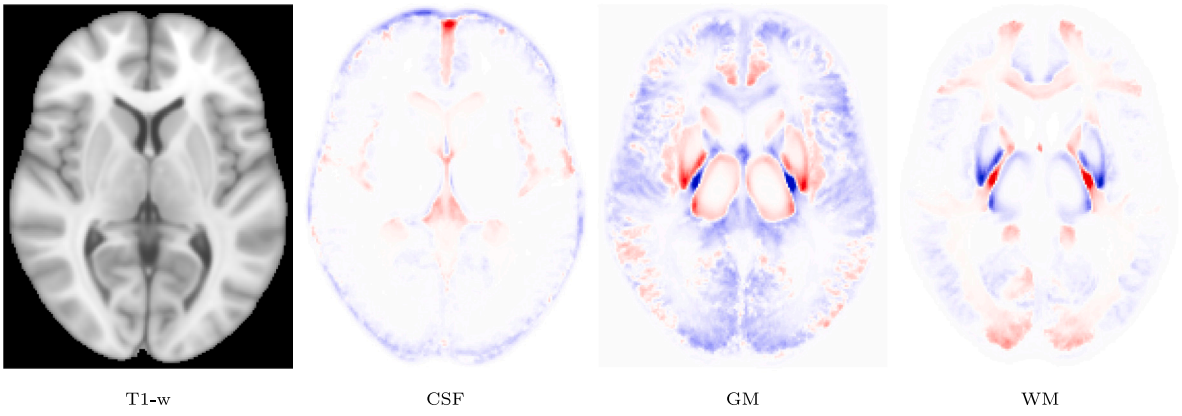
Fig. 8. Comparison of argmax segmentation results between FAST + FIRST and the proposed pipeline for a representative case of each dataset. The last column shows the absolute probability differences of voxels changing their most likely tissue class overlaid with a yellow to red colormap, where yellow corresponds to a difference greater than 0.0 and red to a difference of 1.0 in the voxelwise sum of absolute probability differences. Differences for both datasets are mainly located in the cortex, in the interfaces between subcortical structures and in the outer brain border.

Table 9
Baseline results of the ADNI1 dataset and cross-dataset results of the ADNI1 dataset using the proposed pipeline trained on the MIRIAD dataset.

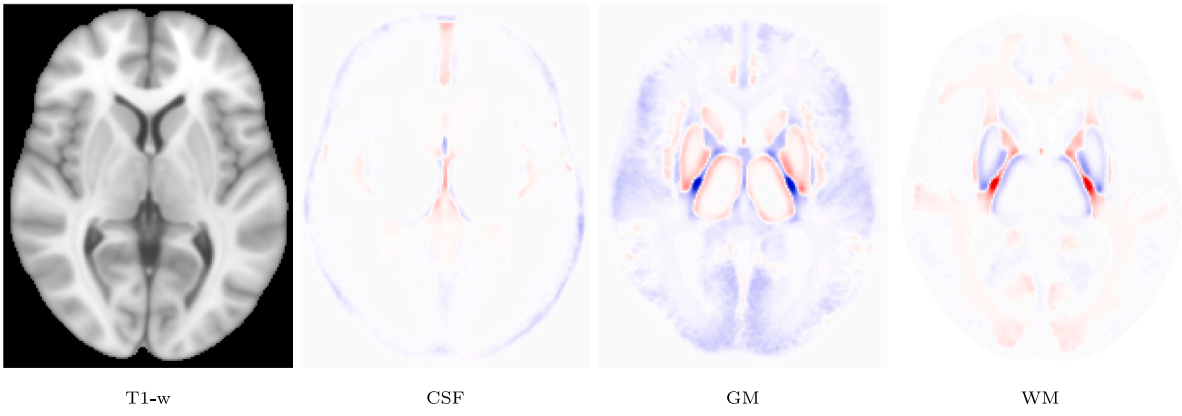
		ADNI1 baseline		Trained on MIRIAD	
		$w_{sim} = 0.0$	$w_{sim} = 0.4$	$w_{sim} = 0.0$	$w_{sim} = 0.4$
ΔBPF	HC	−0.52 ± 0.94%	−0.29 ± 0.30%	−0.39 ± 1.18%	−0.25 ± 0.39%
	AD	−1.27 ± 1.23%	−0.80 ± 0.45%	−1.15 ± 1.58%	−0.74 ± 0.53%
Differences between groups (ΔBPF)	t	5.24	10.07	4.18	8.02
	d	0.70	1.37	0.56	1.08
Rescan error (ΔBPF)		0.31 ± 0.59% (0.15%)	0.07 ± 0.13% (0.04%)	0.60 ± 1.05% (0.31%)	0.15 ± 0.20% (0.08%)

Table 9 shows the baseline evaluation results on the ADNI1 dataset as well as the cross-dataset results on ADNI1 with the proposed pipeline trained on the MIRIAD dataset. Compared with the baseline models, the results are worse in the cross-dataset evaluation for both unregularized and regularized models. Particularly, the rescan error shows a twofold increase in models trained on the MIRIAD dataset when compared to the baseline models trained on ADNI1. However, in terms of sensitivity to group differences, the model trained on MIRIAD with $w_{sim} = 0.4$ continues to outperform the baseline model trained without regularization. In this case, the MIRIAD dataset utilized for training purposes consists of images from a single scanner and a specific voxel size, which does not seem to prepare the model to deal with the variations in scanners and voxel sizes encountered in the ADNI1 dataset and yields poor results.

Table 10 shows the baseline evaluation results on the ADNI1 dataset as well as the cross-dataset results of ADNI1 with the proposed pipeline trained on the MIRIAD dataset. In this case, the observed performance degradation due to the domain shift is considerably reduced. With respect to short-interval error, while the rescan error of the cross-dataset models are significantly worse than the baseline ones trained with the same weight ($p < 0.05$), the 2-week error is only marginally higher. In terms of sensitivity to group differences, the model trained on ADNI1 with $w_{sim} = 0.0$ demonstrates superior performance compared to the MIRIAD baseline model trained with the same parameter. In the case of $w_{sim} = 0.4$, the cross-dataset model trained on ADNI1 achieves a high Cohen's d effect size of $d = 1.79$, which does not reach the same effect size as the MIRIAD baseline model ($d = 2.07$), but surpasses the $d = 1.73$ obtained by SIENA. The results suggest that training a model with a varied set of images acquired with different scanners and voxel sizes, such as the ones in ADNI1, make the pipeline more robust to the domain shift caused by the use of different scanners and voxel sizes.



(a) MIRIAD dataset



(b) ADNI1

Fig. 9. Median probability differences between the probabilistic segmentations of the proposed ($w_{sim} = 0.4$) pipeline with respect to the reference FAST + FIRST segmentations. For this purpose, the tissue probability maps of FAST + FIRST and the presented pipeline from each case are subtracted and then transformed to the MNI space for joint analysis across the whole dataset. The differences are displayed per tissue in a red to white to blue colormap, where red corresponds to a median difference of −0.25 or less, white to 0.0 and blue to an increase in median probability of 0.25 or higher.

Table 10
Baseline results of the MIRIAD dataset and cross-dataset results of the MIRIAD dataset using the proposed pipeline trained on the ADNI1 dataset.

		MIRIAD baseline		Trained on ADNI1	
		$w_{sim} = 0.0$	$w_{sim} = 0.4$	$w_{sim} = 0.0$	$w_{sim} = 0.4$
ΔBPF	HC	−0.45 ± 0.52%	−0.18 ± 0.15%	−0.41 ± 0.29%	−0.25 ± 0.17%
	AD	−1.51 ± 1.00%	−0.91 ± 0.42%	−1.38 ± 0.93%	−0.89 ± 0.42%
Differences between groups (ΔBPF)	t	5.74	10.61	6.48	9.03
	d	1.21	2.07	1.25	1.79
Rescan error (ΔBPF)		0.31 ± 0.29% (0.23%)	0.06 ± 0.06% (0.05%)	0.36 ± 0.41% (0.23%)	0.09 ± 0.09% (0.06%)
2 week error (ΔBPF)		0.46 ± 0.42% (0.36%)	0.18 ± 0.14% (0.15%)	0.53 ± 0.62% (0.37%)	0.21 ± 0.18% (0.17%)

5.4. Limitations

This study has some limitations related to the evaluation of atrophy measures and the clinical applicability of the presented pipeline. Within this work, we have not been able to evaluate the quality or accuracy of either the brain tissue segmentation model learned from fsl_anat outputs or the measured atrophy rates. For this purpose, we would need a dataset similar to those considered in this work having both short interval and longitudinal scan pairs from healthy and AD subjects with sufficiently accurate manual delineations of brain tissue. Despite this limitation, we have evaluated our pipeline on several metrics typically used in the literature to assess longitudinal atrophy quantification and have shown that it improves over extensively validated state-of-the-art methods. In this sense, the comparison with fsl_anat is nuanced since our data-driven pipeline has been previously trained and optimized for the evaluation domain, whereas fsl_anat was not. However, the main goal of these comparisons is only to quantify the relative improvement of our pipeline, which is trained from these fsl_anat outputs. Despite tuning the SIENA execution parameters to obtain the best performance on each dataset, the comparison is also nuanced in the same way since it was not trained or optimized beforehand for the evaluation domain.

As shown by the cross-dataset experiment, another limitation is that the performance of deep learning methods is degraded when applied to images that differ in excess from those seen during training, i.e., from

a different image domain, such as one acquired with a different MRI scanner, acquisition protocol or voxel spacing. Our results suggest that training the pipeline with a diverse set of scanners, voxel sizes and acquisition parameters reduces domain shift sensitivity and improves generalization performance. Furthermore, a pretrained model could be fine-tuned for the specific domain using transfer learning or one-shot domain adaptation techniques for deep learning methods [41]. In this case, training the proposed pipeline would only require a set of unlabeled short-interval scan pairs from the target domain.

6. Conclusion

In this work, we have presented a novel deep learning pipeline for segmentation-based brain atrophy quantification that uses tissue similarity regularization to improve upon the reference automated segmentation method from which it is trained. We have analyzed the tissue similarity regularization effect and empirically selected $w_{\text{sim}} = 0.4$ as an optimal default value for the similarity weight loss term, which performs well across single-site and multisite datasets.

In general, the presented pipeline improves upon atrophy evaluation metrics and produces smoother and less noisy segmentations than the reference method used for training. The regularization introduces differences in the segmentation of GM/WM in the cortex, the outer brain interface and borders of subcortical structures compared with the reference method. Our evaluation results on short-interval scan pairs show that the proposed regularization lowers the quantification error and improves the overall tissue segmentation consistency, especially for the gray and white matter components. In this sense, our pipeline shows lower and more similar levels of error between the parenchyma and its distinct GM and WM components, whereas the reference method had much larger errors for GM and WM than for the parenchyma. In the longitudinal case, we observed lower variability in atrophy rates and greater sensitivity to differences between healthy controls and AD subjects. Furthermore, while the reference method measured higher levels of WM atrophy for healthy controls than for the AD group, which does not make intuitive sense within a generalized atrophy process, the proposed regularization in our pipeline reverses this tendency and shows more coherent WM atrophy rates between the HC and AD groups.

The presented pipeline is based on the idea that regularized deep learning can exploit data priors to reduce the biases and systematic errors learned from a reference segmentation method. We have shown that the proposed regularization, which aims at reducing short-interval scan differences, can directly improve brain atrophy quantification in the longitudinal case. Data-driven approaches have the potential to surpass their classical counterparts and unlock brain atrophy as a useful diagnostic and prognostic marker for neurodegenerative pathologies.

CRediT authorship contribution statement

Albert Clèrigues: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Sergi Valverde:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Arnau Oliver:** Writing – review & editing, Supervision, Funding acquisition. **Xavier Lladó:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Albert Clèrigues holds an FPI grant from the Ministerio de Ciencia e Innovación, Spain with reference number PRE2018-083507. This work has been supported by DPI2020-114769RB-I00 from the Ministerio de Ciencia, Innovación y Universidades, Spain. The authors gratefully acknowledge the support of the NVIDIA Corporation with their donation of the TITAN X GPU used in this research. This work has been also supported by ICREA Academia program.

Data used in the preparation of this article were obtained from the MIRIAD database. The MIRIAD investigators did not participate in analysis or writing of this report. The MIRIAD dataset is made available through the support of the UK Alzheimer's Society (Grant RF116). The original data collection was funded through an unrestricted educational grant from GlaxoSmithKline (Grant 6GKC).

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.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108811.

References

[1] M. Bobinski, M. De Leon, J. Wegiel, S. Desanti, A. Convit, L. Saint Louis, H. Rusinek, H. Wisniewski, The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in Alzheimer's disease, Neuroscience 95 (1999) 721–725.

[2] J.L. Whitwell, D.W. Dickson, M.E. Murray, S.D. Weigand, N. Tosakulwong, M.L. Senjem, D.S. Knopman, B.F. Boeve, J.E. Parisi, R.C. Petersen, et al., Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: A case-control study, Lancet Neurol. 11 (2012) 868–877.

[3] L. Pini, M. Pievani, M. Bocchetta, D. Altomare, P. Bosco, E. Cavedo, S. Galluzzi, M. Marizzoni, G.B. Frisoni, Brain atrophy in Alzheimer's disease and aging, Ageing Res. Rev. 30 (2016) 25–48.

[4] N.D. Stefano, P.M. Matthews, M. Filippi, F. Agosta, M.D. Luca, M.L. Bartolozzi, L. Guidi, A. Ghezzi, E. Montanari, A. Cifelli, A. Federico, S.M. Smith, Evidence of early cortical atrophy in MS, Neurology 60 (2003) 1157–1162.

[5] K. Morgen, G. Sammer, S.M. Courtney, T. Wolters, H. Melchior, C.R. Blecker, P. Oschmann, M. Kaps, D. Vaitl, Evidence for a direct association between cortical atrophy and cognitive impairment in relapsing–remitting MS, NeuroImage 30 (2006) 891–898.

[6] R.A. Rudick, J.C. Lee, K. Nakamura, E. Fisher, Gray matter atrophy correlates with ms disability progression measured with MSFC but not edss, J. Neurol. Sci. 282 (2009) 106–111.

[7] M.A. Rocca, M. Battaglini, R.H. Benedict, N.D. Stefano, J.J. Geurts, R.G. Henry, M.A. Horsfield, M. Jenkinson, E. Pagani, M. Filippi, Brain MRI atrophy quantification in MS, Neurology 88 (2017) 403–413.
[8] J. Sastre-Garriga, D. Pareto, M. Battaglini, M.A. Rocca, O. Ciccarelli, C. Enzinger, J. Wuerfel, M.P. Sormani, F. Barkhof, T.A. Yousry, N.D. Stefano, M. Tintoré, M. Filippi, C. Gasperini, L. Kappos, J. Río, J. Frederiksen, J. Palace, H. Vrenken, X. Montalban, Àlex. Rovira, Magnims consensus recommendations on the use of brain and spinal cord atrophy measures in clinical practice, Nat. Rev. Neurol. 2020 16:3 16 (2020) 171–182.
[9] A. Rovira, M.P. Wattjes, M. Tintoré, C. Tur, T.A. Yousry, M.P. Sormani, N.D. Stefano, M. Filippi, C. Auger, M.A. Rocca, F. Barkhof, F. Fazekas, L. Kappos, C. Polman, D. Miller, X. Montalban, Magnims consensus guidelines on the use of MRI in multiple sclerosis—clinical implementation in the diagnostic process, Nat. Rev. Neurol., 2015 11:8 11 (2015) 471–482.
[10] J. Sastre-Garriga, D. Pareto, Àlex. Rovira, Brain atrophy in multiple sclerosis: Clinical relevance and technical aspects, Neuroimaging Clin. 27 (2017) 289–300.
[11] M. Battaglini, M. Jenkinson, N.D. Stefano, Siena-XL for improving the assessment of gray and white matter volume changes on brain MRI, Hum. Brain Map. 39 (1063) (2018).
[12] S.M. Smith, Y. Zhang, M. Jenkinson, J. Chen, P.M. Matthews, A. Federico, N.D. Stefano, Accurate, robust, and automated longitudinal and cross-sectional brain change analysis, NeuroImage 17 (2002) 479–489.
[13] P.A. Freeborough, N.C. Fox, The boundary shift integral: An accurate and robust measure of cerebral volume changes from registered repeat MRI, IEEE Trans. Med. Imaging 16 (1997) 623–629.
[14] Y. Zhang, M. Brady, S. Smith, Segmentation of brain MR images through a hidden Markov random field model and the expectation–maximization algorithm, IEEE Trans. Med. Imaging 20 (2001) 45–57.
[15] D. Holland, A.M. Dale, Nonlinear registration of longitudinal images and measurement of change in regions of interest, Med. Image Anal. 15 (2011) 489–497.
[16] K. Nakamura, N. Guizard, V.S. Fonov, S. Narayanan, D.L. Collins, D.L. Arnold, Jacobian integration method increases the statistical power to measure gray matter atrophy in multiple sclerosis, NeuroImage: Clinical 4 (2014) 10–17.
[17] D. Smeets, A. Ribbens, D.M. Sima, M. Cambron, D. Horakova, S. Jain, A. Maertens, E.V. Vlierberghe, V. Terzopoulos, A.M.V. Binst, M. Vaneckova, J. Krasensky, T. Uher, Z. Seidl, J.D. Keyser, G. Nagels, J.D. Mey, E. Havrdova, W.V. Hecke, Reliable measurements of brain atrophy in individual patients with multiple sclerosis, Brain Behav. 6 (2016) e00518.
[18] N. Yamanakkanavar, J.Y. Choi, B. Lee, Mri segmentation and classification of human brain using deep learning for diagnosis of Alzheimer's disease: A survey, Sensors (Switzerland) 20 (2020) 1–31.
[19] Q. Hu, Y. Wei, X. Li, C. Wang, J. Li, Y. Wang, Ea-net: Edge-aware network for brain structure segmentation via decoupled high and low frequency features, Comput. Biol. Med. 150 (2022) 106139.
[20] A. Guha Roy, S. Conjeti, N. Navab, C. Wachinger, Quicknat: A fully convolutional network for quick and accurate segmentation of neuroanatomy, NeuroImage 186 (2019) 713–727.
[21] B. Fischl, D.H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, A.M. Dale, Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain, Neuron 33 (2002) 341–355.
[22] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, M. Reuter, Fastsurfer - A fast and accurate deep learning based neuroimaging pipeline, NeuroImage 219 (2020) 117012.
[23] M. Rajchl, N. Pawlowski, D. Rueckert, P.M. Matthews, B. Glocker, Neuronet: Fast and robust reproduction of multiple brain image segmentation pipelines, 2018, arXiv preprint arXiv:1806.04224.
[24] M. Jenkinson, C.F. Beckmann, T.E. Behrens, M.W. Woolrich, S.M. Smith, FSL, NeuroImage 62 (2012) 782–790.
[25] R. Dorent, T. Booth, W. Li, C.H. Sudre, S. Kafiabadi, J. Cardoso, S. Ourselin, T. Vercauteren, Learning joint segmentation of tissues and brain lesions from task-specific hetero-modal domain-shifted datasets, Med. Image Anal. 67 (2021) 101862.
[26] B. Patenaude, S.M. Smith, D.N. Kennedy, M. Jenkinson, A Bayesian model of shape and appearance for subcortical brain segmentation, NeuroImage 56 (2011) 907–922.
[27] I.B. Malone, D. Cash, G.R. Ridgway, D.G. MacManus, S. Ourselin, N.C. Fox, J.M. Schott, Miriad—public release of a multiple time point Alzheimer's MR imaging dataset, NeuroImage 70 (2013) 33–36.
[28] M. Reuter, B. Fischl, Avoiding asymmetry-induced bias in longitudinal image processing, NeuroImage 57 (2011) 19–21.
[29] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, vol. 9351, 2015, pp. 234–241.
[30] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, 2015, pp. 448–456.
[31] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: International Conference on Machine Learning, 2010, pp. 807–814.
[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[33] M.D. Zeiler, Adadelta: An adaptive learning rate method, 2012, arXiv preprint arXiv:1212.5701.
[34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch, Neural Inform. Process. Syst. (2017).
[35] N.A. Royle, M.C. Hernández, S.M. Maniega, B.S. Arabisala, M.E. Bastin, I.J. Deary, J.M. Wardlaw, Influence of thickening of the inner skull table on intracranial volume measurement in older people, Magnet. Reson. Imaging 31 (2013) 918–922.
[36] S.M. Smith, A. Rao, N.D. Stefano, M. Jenkinson, J.M. Schott, P.M. Matthews, N.C. Fox, Longitudinal and cross-sectional analysis of atrophy in Alzheimer's disease: Cross-validation of BSI, Siena and Sienax, NeuroImage 36 (2007) 1200–1206.
[37] B.L. Welch, The generalization of 'Student's' problem when several different population varlances are involved, Biometrika 34 (1947) 28–35.
[38] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Academic Press, New York, 1977.
[39] H.J. Kuijf, E. Bennink, K.L. Vincken, N. Weaver, G.J. Biessels, M.A. Viergever, MR brain segmentation challenge 2018 data, 2024.
[40] A. Anand, N. Anand, Fast brain volumetric segmentation from t1 MRI scans, in: K. Arai, S. Kapoor (Eds.), Advances in Computer Vision, Springer International Publishing, Cham, 2020, pp. 402–415.
[41] J.M. Valverde, V. Imani, A. Abdollahzadeh, R.D. Feo, M. Prakash, R. Ciszek, J. Tohka, Transfer learning in magnetic resonance brain imaging: A systematic review, J. Imaging 2021 7 (2021) 66, 7, 66.