Neuroinformatics
https://doi.org/10.1007/s12021-020-09499-z

ORIGINAL ARTICLE

Generating Longitudinal Atrophy Evaluation Datasets on Brain Magnetic Resonance Images Using Convolutional Neural Networks and Segmentation Priors

Jose Bernal[1] · Sergi Valverde[1] · Kaisar Kushibar[1] · Mariano Cabezas[1] · Arnau Oliver[1] · Xavier Lladó[1] · The Alzheimer's Disease Neuroimaging Initiative

Accepted: 6 November 2020

Abstract

Brain atrophy quantification plays a fundamental role in neuroinformatics since it permits studying brain development and neurological disorders. However, the lack of a ground truth prevents testing the accuracy of longitudinal atrophy quantification methods. We propose a deep learning framework to generate longitudinal datasets by deforming T1-w brain magnetic resonance imaging scans as requested through segmentation maps. Our proposal incorporates a cascaded multi-path U-Net optimised with a multi-objective loss which allows its paths to generate different brain regions accurately. We provided our model with baseline scans and real follow-up segmentation maps from two longitudinal datasets, ADNI and OASIS, and observed that our framework could produce synthetic follow-up scans that matched the real ones (Total scans=584; Median absolute error: 0.03 ± 0.02; Structural similarity index: 0.98 ± 0.02; Dice similarity coefficient: 0.95 ± 0.02; Percentage of brain volume change: 0.24 ± 0.16; Jacobian integration: 1.13 ± 0.05). Compared to two relevant works generating brain lesions using U-Nets and conditional generative adversarial networks (CGAN), our proposal outperformed them significantly in most cases ($p < 0.01$), except in the delineation of brain edges where the CGAN took the lead (Jacobian integration: Ours - 1.13 ± 0.05 vs CGAN - 1.00 ± 0.02; $p < 0.01$). We examined whether changes induced with our framework were detected by FAST, SPM, SIENA, SIENAX, and the Jacobian integration method. We observed that induced and detected changes were highly correlated (Adj. $R^2 > 0.86$). Our preliminary results on harmonised datasets showed the potential of our framework to be applied to various data collections without further adjustment.

Keywords Cerebral atrophy · Longitudinal atrophy synthesis · Image generation · Convolutional neural networks · Brain MRI

Introduction

Brain tissue segmentation and volume quantification are active research topics in medical image analysis as measurements in these regards are employed for diagnosing brain diseases and evaluating pathology progression and treatment effectiveness (Rovira et al. 2015; Steenwijk et al. 2016; Filippi et al. 2016; Storelli et al. 2018). Although slow shrinkage of the brain comes with ageing, changes in brain size and shape are also a consequence of disorders, such as schizophrenia, Alzheimer's disease (AD), and Multiple Sclerosis (MS) (Cover et al. 2011; Haijma et al. 2012; van Erp et al. 2016; Rocca et al. 2017). Thus, providing medical doctors with accurate and precise brain volume measurements is essential for understanding the nature of brain problems more rigorously.

Brain volume can be analysed at cross-sectional and longitudinal levels (Rocca et al. 2017). In the cross-sectional

JB and KK held FI-DGR2017 grants from the Catalan Government with reference numbers 2017FI B00476 and 2017FI B00372, respectively. MC holds a Juan de la Cierva - Incorporación grant from the Spanish Government with reference number IJCI-2016-29240. This work has been partially supported by Retos de Investigació TIN2015-73563-JIN and DPI2017-86696-R from the Ministerio de Ciencia y Tecnología. Data used in the preparation of this article were [in part] 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.

✉ Jose Bernal
jose.bernal@udg.edu

[1] Computer Vision and Robotics Institute, Universitat de Girona, Girona, Spain

Published online: 02 January 2021

studies, brain tissues and structures of scans acquired at a single time-point are examined. Such an analysis can be carried out using validated segmentation tools, such as FAST (Zhang et al. 2001), FIRST (Patenaude et al. 2011), SPM (Ashburner et al. 2012), and Freesurfer (Fischl et al. 2002), or whole-brain atrophy quantification algorithms, such as SIENAX (Smith et al. 2002). At a longitudinal level, atrophy quantification algorithms aim to find changes in brain volume between two scans taken at different time-points. Although SIENA (Smith et al. 2002) is a popular algorithm for this task, there exist more (Cover et al. 2011).

Longitudinal brain MRI datasets available to the public[1] are typically used for assessing repeatability and improvement on statistical power. To examine the former aspect, patients are scanned multiple times in different scanners in short periods of time, to ensure minimal brain changes, and brain volumetry methods are judged based on their precision. The latter aspect aims to determine whether these approximations can discern between patients undergoing different treatments/pathologies. Commonly, the exercise consists of running the algorithms over samples of two populations (e.g. dementia versus control), computing brain volume statistics (e.g. mean and standard deviation) for both groups, and calculating and comparing their sample sizes. The lower the sample size per arm, the better the algorithm. Nonetheless, this evaluation does not reflect the accuracy of the methods. In fact, accuracy is rarely assessed since manual segmentation is tedious, time-consuming, and error-prone, and conventional automatic segmentation tools exhibit inaccuracies (de Boer et al. 2010). Synthetic image generation could be used to address such a problem.

In medical image analysis, image generation approaches have been applied to assess registration, estimate and correct bias in longitudinal atrophy analyses, generate absent modalities and augment training sets (Karaçali and Davatzikos 2006; Ens et al. 2009; Roy et al. 2013; Sharma et al. 2013; Khanal et al. 2017; Wei et al. 2018; Chartsias et al. 2018; Shin et al. 2018; Frid-Adar et al. 2018; Costa et al. 2018; Salem et al. 2019). The techniques range from transformation models mimicking brain tissue loss to adversarial/generative networks with problem-specific loss functions.

Karaçali and Davatzikos (2006) devised a method for deforming magnetic resonance (MR) scans such that the atrophy extent corresponded to the requested one.[2] The downfall of such an approach is that resulting deformation patterns cannot be controlled locally and follow a topology-preserving strategy which might not permit mimicking multiple pathologies.

Roy et al. (2013) used patch-based dictionary learning to estimate a mapping function between two imaging sequences or image acquisition protocols,[3] e.g. making it appealing in retrospective harmonisation pipelines. However, its direct usage for atrophy generation might not be feasible since the technique does not deform the brain but finds matching intensity values between imaging modalities.

Chartsias et al. (2018) proposed a framework to synthesise MR modalities from others using encoder-decoder CNNs and modality-invariant latent spaces.[4] Apart from the modality synthesis, the authors showed the potential of the framework to in-paint white matter hyperintensities onto normal-appearing tissue and the usage of multiple losses to achieve realistic synthesis. Salem et al. (2019) devised a proposal, inspired by their work, to generate synthetic yet realistic MS lesions as an image augmentation strategy.[5] Evidently, a similar principle could be considered for generating atrophy.

Shin et al. (2018) developed a proposal in which realistic MR scans were generated from brain anatomy and tumour segmentation masks using conditional generative adversarial networks (CGAN). The authors showed that their approach could be used for dealing with the lack of diverse, sufficient, and correctly annotated data. Although their code is not available in principle, their proposal is inspired by the image-to-image translation with CGAN (Isola et al. 2017).[6] Up to our knowledge, these types of architectures have not been considered for longitudinal data generation, but they can be extended for this purpose by giving the network the baseline scan and the segmentation map of the follow-up acquisition.

In this work, we leverage deep learning to deform a given T1-w scan based on the information provided through tissue probability maps. This setting allows building longitudinal collections for assessing atrophy quantification methods as the tissue loss between original and generated scans is controlled, induced, and known beforehand. Our highlights are:

1. We present the first deep learning approach for generating synthetic atrophy change on brain MR using fully convolutional neural networks and tissue segmentation priors;
2. We use a multi-objective loss function to account for intensity similarity between the expected and generated scans at different brain regions;
3. We show qualitatively and quantitatively that our framework can generate a follow-up scan given a baseline

[1] See https://surfer.nmr.mgh.harvard.edu/fswiki/LongitudinalData

[2] Available at http://web.iyte.edu.tr/~bilgekaracali/VoxelVolumeMatching.tar.gz

[3] Available at https://www.nitrc.org/projects/image_synthesis/

[4] Available at https://github.com/agis85/multimodal_brain_synthesis.

[5] Available at https://github.com/NIC-VICOROB/MS_Lesions_Generator

[6] Available at https://github.com/phillipi/pix2pix

volume and a follow-up tissue segmentation probability map accurately and better than state-of-the-art-inspired networks;
4. We show qualitatively and quantitatively that our framework allows training our network in one dataset and testing it in another one without affecting the overall performance;
5. We show quantitatively that our framework can generate varying extents of tissue loss which are detectable by established cross-sectional and longitudinal atrophy quantification tools.

Note that our aim is not to predict the atrophy that a patient will suffer in a certain amount of time, but a prediction of what would be the brain appearance given a tissue change (segmentation). The relevance of this work is two-fold. First, our proposal allows comparing atrophy quantification tools quantitatively (Evaluating Induced Changes with Brain Volumetry Methods). Second, it can serve as ground truth for training deep learning approaches for atrophy quantification.

The paper is organised as follows. Our proposal is described in "Methods". Datasets, performance metrics, experiments, and results are detailed in "Experiments and Results". Overall discussion and future work are presented in "Discussion".

Methods

Our proposed atrophy generation framework is depicted in Fig. 1. Given a baseline T1-w scan and its modified tissue probability maps, the goal of our framework is to alter the input such that brain tissues are altered as requested. In such a way the atrophy between the baseline and generated images is known in advance. We take a T1-w scan, segment its regions using conventional tissue segmentation tools, alter its segmentation probability maps manually or automatically, and plug both the baseline T1-w scan and the resulting probability maps into the generation network to create a synthetic volume.

Note that the way the framework has been structured is advantageous as a plethora of scans can be generated by modifying the input tissue segmentation maps (e.g. manually, using morphological operations, or pathology-related deformation fields (Krebs et al. 2019)). We apply real deformation fields to alter the original segmentation probability maps. Further details of the approach are discussed in the following sections.

Processing Pipeline

Our processing pipeline contemplates four essential components: pre-processing, data preparation, processing, and reconstruction.

Pre-processing consists of (i) skull stripping with ROBEX (Iglesias et al. 2011), (ii) histogram matching (Nyúl et al. 2000) to fix voxel values to a common range, and (iii) registration to the MNI space as harmonising step. The first step allows discarding non-relevant areas that may affect the generation process as they are commonly hyperintense in T1-w. We chose ROBEX since it is an unsupervised method that delivered consistent and robust results when compared to conventional methods. The second step allows mapping voxel intensities to a reference range. This procedure is essential to reduce issues regarding generalisability due to intensity shifts (Battaglini et al. 2018). The third step permits using the same network on various datasets as

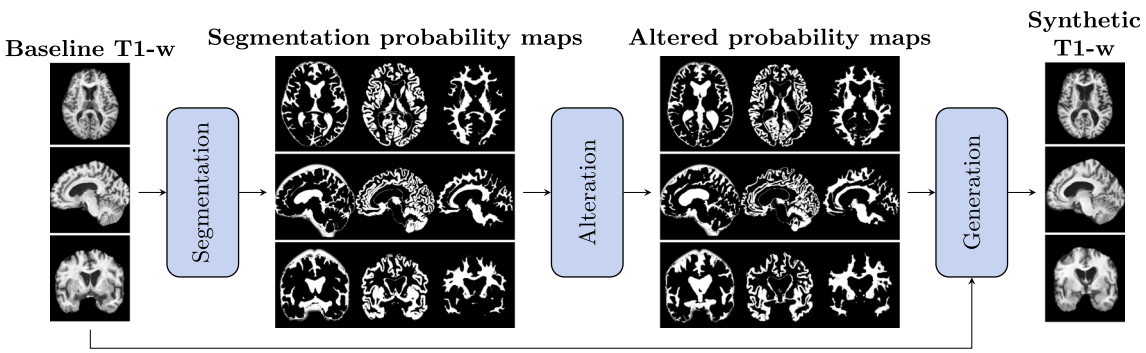


Fig. 1 Inducing controlled tissue variations. We take a baseline T1-w scan, segment it, alter its segmentation probability maps manually or automatically, and plug it into the generation network to create a synthetic volume. We apply conventional tools for tissue segmentation and real deformation fields to alter original segmentation probability maps. Given a baseline T1-w input image and modified tissue probability maps, the goal of our framework is to generate a T1-w scan in which the tissues are altered as requested. In this example, tissue changes were requested in both cortical and periventricular regions, e.g. lateral ventricles appear enlarged (all three views), Sylvian fissures have been altered (axial and coronal), and the third ventricle seems more atrophied (coronal)

reducing the heterogeneity of voxel spacing may enhance the overall performance (Bernal et al. 2019b).

Data preparation consists of splitting input volumes into patches. For both training and testing, we extract overlapping blocks to gather more samples, reduce block boundary artefacts, and enforce spatial consistency (Bernal et al. 2019b). Additionally, we discard empty or partially empty training patches to prevent building background-biased generators. Moreover, we opt to modify our loss function $\mathcal{L}$ to nullify the penalty coming from these misclassified background areas. We set the minimum content rate and overlap extent to 30% and 50%, respectively. Both values were favourable experimentally.

In the processing step, we pass each tuple of patches extracted from the baseline scan and modified probability maps through the network in batches of 32 elements at a time. We did not increase this parameter due to hardware constraints.

We overlay neighbouring predictions to reconstruct the synthetic volume and provide voxel-wise responses through averaging. We run histogram matching on the reconstructed volume to ensure intensity range similarity. No further post-processing is required.

Generation Architecture

Our proposed network follows a cascaded U-Net construction scheme, as illustrated in Fig. 2a. First, we input the baseline scan and its modified tissue probability maps into three networks arranged in parallel. Each one of these networks accounts separately for changes in cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM). Second, we append and pass the resulting individual latent representations to another u-shaped network which merges them effectively to produce the final output. In our implementation, the input and output patches have the same height, width and depth, 32 voxels in each dimension. The overall cascaded network is trained end-to-end, i.e. none of the sub-nets is trained independently.

Each U-Net module comprises a contracting path, performing consecutive convolution and down-sampling operations, and an expansive path, carrying successive up-sampling and convolutions. In this way, it is possible to output a patch with the same dimensions as the input while reducing response times. The architecture is illustrated in Fig. 2b. The network consists of $8 \times 2 + 1$ convolutional layers—eight pairs occur in *parallel*, as shown in the lower right corner of Fig. 2b—three down-sampling modules and three backward strided convolutions. The number of kernels doubles per contracting path layer from 2^4, in its shallowest, to 2^7, in its latent space, and afterwards halves per expansive path layer until the kernels are 2^4, in its deepest level. Strides for down-sampling and up-convolutions are set to $2 \times 2 \times 2$.

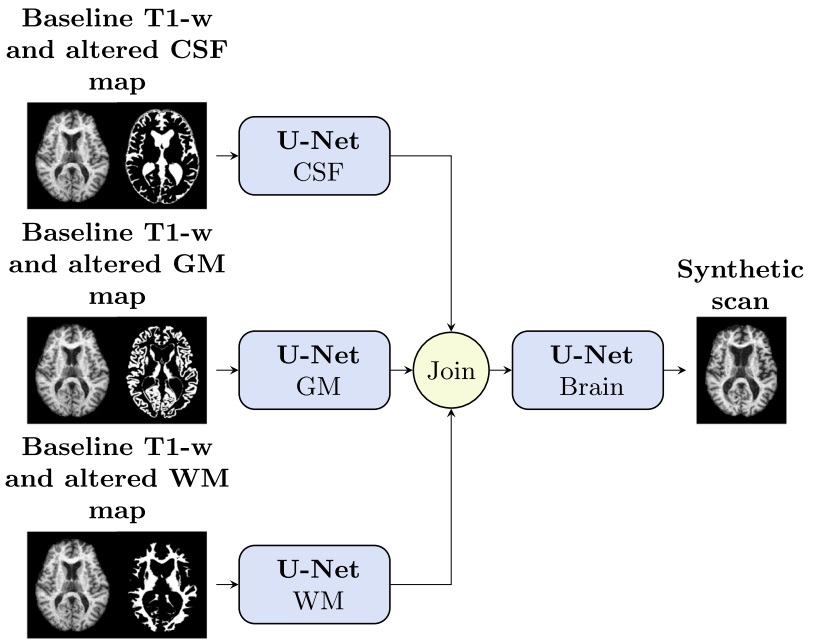
The U-Nets are equipped with filter banks of varied sizes in a Network-in-Network (NIN) resembling scheme (Lin et al. 2013; Szegedy et al. 2015). These modules, implemented as $1 \times 1 \times 1$-kernel layers, act similar to embedded multi-layer perceptrons which enhance the discriminant and representation power of the overall model. These processing components are referred to as *core elements* in Fig. 2b.

Each sub-module uses residual connections to merge feature maps from higher-resolution layers with deconvolved maps to preserve localisation details and improve back-propagation (He et al. 2016). Moreover, each sub-module combines feature maps by adding them and not concatenating them as widespread (Brosch et al. 2016; Çiçek et al. 2016). This option is preferred to reduce the cardinality of the trainable parameter set. Note the different channels are processed in an early fusion fashion (Ghafoorian et al. 2017).

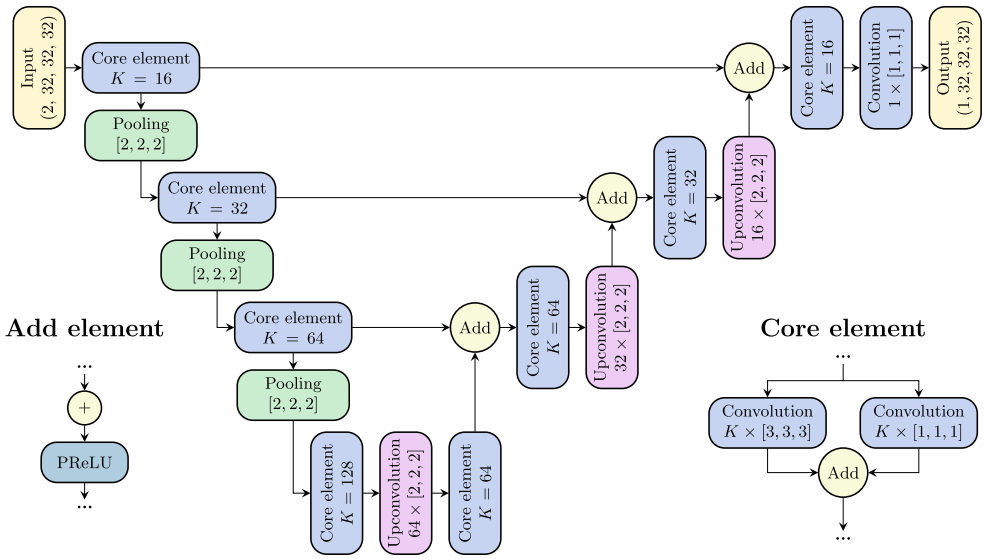
The design of the sub-modules is inspired by the work of Guerrero et al. (2018). The main differences are the dimensionality of the network, the downsampling approach, and the type and location of non-linear activation layers. First, the network is extended to process 3D data directly. This strategy is considered instead of a slice-by-slice approach to exploit the nature of MRI, incorporate contextual information from the three orthogonal planes, and produce more consistent results. Second, strided convolutions are used instead of max-pooling layers (Springenberg et al. 2015) to achieve improved performance. Third, the Rectified Linear Unit (ReLU) layers used in the original work are exchanged for Parametric ReLU (PReLU) (Trottier et al. 2017). This asset helps the model to cope with issues regarding the gradient update and empirical performance (He et al. 2015; Szegedy et al. 2017; Bernal et al. 2019a). Fourth, these rectifier layers are used after every addition of feature maps. This choice promotes sparsity within the network, i.e. a more resilient representation (Glorot et al. 2011).

Region-Wise Loss Function

Atrophy quantification algorithms perform tissue segmentation and/or linear and non-linear registration. These widespread practices impose three constraints on the generation: (i) tissue contrast should be sufficiently high to be segmentation-feasible, (ii) synthesised volumes should appear visually similar to the actual scans at intensity level, and (iii) brain boundaries should be well-defined. We propose a four-objective loss function to fulfil these needs and train the whole model properly. Each objective evaluates the similarity between the expected and synthesised volume in the CSF, GW, WM, and whole intracranial volume. Given a real scan, y, its corresponding tissue probability maps, s_{CSF}, s_{GM}, ands_{WM}, and an approximation obtained with



(a) Proposed generation network



(b) Specific implementation of each U-Net

our model, $\tilde{y}$, the region-wise mean square error (RWMSE) loss function is defined as follows

$$\mathcal{L}(y,\tilde{y}) = \underbrace{L(y,\tilde{y};\sum_{\text{ROI}} s_{\text{ROI}})}_{Combined} + \underbrace{\sum_{\text{ROI}} L(y,\tilde{y};s_{\text{ROI}})}_{Individual}, \quad (1)$$

$$L(y,\tilde{y};s) = \frac{1}{M\cdot N\cdot P}\sum_{v=1}^{M\cdot N\cdot P} H(s(v))\cdot ||y_v-\tilde{y}_v||_1, \quad (2)$$

where $H(a)$ is the discrete Heaviside step function. While the loss for overall reconstruction is back-propagated from the last layer of the network, the others affect the parallel U-Nets disjointly—i.e. one loss per path. Hence, the parallel sub-modules are in charge of generating tissue changes and the merging network of combining them smoothly.

This loss function requires segmentation priors of the follow-up volume, s_i in Eq. 1. This information is passed to the network to provide notions of the CSF, GM, and WM regions and specialise each path of the network towards generating realistic T1-w scans. This input can be obtained using a ground truth—if available—or validated

segmentation tools, such as FAST or SPM. In our case, we use FAST to obtain tissue probability maps.

Generating Controlled Evaluation Environments

Once the network is trained using real baseline and follow-up acquisitions, we use it to generate controlled atrophy change evaluation environments, as illustrated in Fig. 1. The process consists of gradually increasing the overall tissue loss to establish whether our tool can generate various extents of deformation accurately. Segmentation maps can be altered in various ways. For instance, they could be dilated or eroded using morphological operations. However, this will not mimic pathological processes altering brain tissue as atrophy changes are not necessarily even in all brain regions. Alternatively, real atrophy deformation fields could be used to modify the segmentation maps. We compute real deformation fields, using FNIRT (Andersson et al. 2007), from patients exhibiting the largest tissue loss and use them to alter baseline tissue segmentation maps. We multiply the resulting deformation vectors by scalars to obtain intermediate stages.

Implementation Details

Network Training

The steps to train our model on a given dataset are as follows. First, we split the training set into training and validation at random—70% and 30% of the volumes, respectively. Second, we train the network in batches of 32 (default parameter value) for a maximum of 100 epochs. At the end of each epoch, we compute the performance on the validation set. The training phase stops after 10 consecutive epochs without improvement. We retain the model leading to the lowest loss function value. We optimise the models using the Adam (Kingma and Ba 2014) optimisation method with an initial learning rate of 1×10^{-3}, a decay of 0, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ (i.e. default parameter values, as suggested in the original paper).

Network Testing

The steps to test a trained model on a given input MR volume are as follows. First, we divide the baseline input volume and the modified probability segmentation maps into patches. We extract these patches from the entire input and not from specific regions. Second, we input the patches to the network to obtain synthetic blocks. Third, as there is overlap between output blocks, we provide the final segmentation through means of averaging. We rearrange all synthetic patches to reconstruct the corresponding synthetic volume.

Software and Hardware

We implement all the architectures from scratch in Python, using the Keras library. We run all the experiments on a GNU/Linux machine box running Ubuntu 16.04, with 128GB RAM. We train and test our models using a single GeForce GTX 1080-TI GPU (NVIDIA Corp., United States) with 11GB RAM. The developed framework is available to download from our GitHub repository (See information sharing statement).

Experiments and Results

In this section, we describe the considered datasets, performance evaluation measurements, implementation details, and experiments evaluating our proposed model and corresponding results. The experiments assess loss function and architecture selection, image generation quality, and whether induced changes are detectable by conventional brain volumetry methods. Further details of each experiment and the outcomes are described in the following sections.

Considered Datasets

We considered two publicly available longitudinal MRI repositories: the Open Access Series of Imaging Studies (OASIS) (Marcus et al. 2010) and the Alzheimer's Disease Neuroimaging Initiative (ADNI).[7] Relevant information of each dataset is presented in Table 1. The OASIS2 dataset was split, for easing downloading, into two sets. We refer to those as $O1$ and $O2$ from hereon. The former set contains 169 pairs of baseline follow-up cases and the second one 126. The ADNI collection contains a plethora of longitudinal cases and, hence, we opted to filter some cases. We used only cases of ADNI2 subjects with Alzheimer's disease which scans were bias field corrected and coregistered correctly using FLIRT (Jenkinson and Smith 2001; Jenkinson et al. 2002). Unlike in the OASIS2 case, the database was not divided in principle. Thus, we split it into two sets, $A1$ and $A2$, with 153 and 136 pairs of cases, respectively. For the sake of reproducibility, we attach the list of selected cases as Supplementary material.

The distribution of relative CSF change between baseline and follow-up scans for OASIS and ADNI2 is illustrated

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

Table 1 Relevant information from the two considered datasets

Item	OASIS2	ADNI
No. of pairs	295	289
No. of time-points (max.)	5	5
Voxel spacing	$1.0 \times 1.0 \times 1.3$	$1.2 \times 1.0 \times 1.0$
Reconstruction matrix	$256 \times 256 \times 128$	$196 \times 256 \times 256$
Bias-field corrected	No	Yes
Intensity standardised	No	No
Skull stripped	No	No
Sets and no. of pairs	$O_1 : 169$, $O_2 : 126$	$A_1 : 153$, $A_2 : 136$

The items to describe each dataset are listed in the first column. Although the average reconstruction matrix of the ADNI dataset is the one indicated below, the actual dimensions vary. Pairs refer to tuples of baseline and follow-up acquisitions

in Fig. 3. The majority of cases were concentrated within [0.45, 0.55] for the OASIS2 dataset and [0.30, 0.50] for the ADNI2 dataset, but ADNI contained more cases with values above 1.00.

Evaluation Metrics

Our generation framework should produce synthetic scans of such a quality that they resemble real ones. In this work, we scrutinised generation quality by comparing real and synthetic scans in terms of their perceptual properties and their tissue segmentation and cerebral atrophy quantification results.

Image Quality

We assessed the quality of our generations with respect to that of real scans locally and globally. Locally, we measured voxel-wise intensity differences between a real scan, y, and its approximation, $\tilde{y}$, using the following expression

$$MAE(y, \tilde{y}) = \text{median } |y - \tilde{y}|. \tag{3}$$

The MAE approaches zero as voxel-wise differences between y and $\tilde{y}$ decrease. Globally, we quantified similarity between images through the structural similarity index (SSIM) (Wang et al. 2004) as it has been found correlated with the quality of perception of the human visual system (Hore and Ziou 2010) and accounts jointly for variations in luminance, contrast, and structure (correlation):

$$SSIM(y, \tilde{y}) = \underbrace{\frac{2\mu_y\mu_{\tilde{y}} + c_1}{\mu_{\tilde{y}}^2\mu_{\tilde{y}}^2 + c_1}}_{\text{Luminance}} \cdot \underbrace{\frac{2\sigma_y\sigma_{\tilde{y}} + c_2}{\sigma_y^2\sigma_{\tilde{y}}^2 + c_2}}_{\text{Contrast}} \cdot \underbrace{\frac{\text{cov}(y, \tilde{y}) + c_3}{\sigma_y\sigma_{\tilde{y}} + c_3}}_{\text{Structure}}, \tag{4}$$

where μ and σ denote the mean and standard deviation values of the luminance of the images, $\text{cov}(y, \tilde{y})$ the covariance between y and $\tilde{y}$, and c_i constants to avoid a null denominator (Hore and Ziou 2010). The SSIM values range within zero and one, where the former indicates null similarity white the latter implies that y and $\tilde{y}$ are equal. We expected our framework to produce synthetic scans of such perceptual quality that MAE and SSIM values tended to zero and one, respectively.

Segmentation Agreement

Segmentation-based atrophy quantification algorithms segment brain tissues and measure volumetric differences (Rudick et al. 1999; Jia et al. 2016) or brain boundary shifts (Smith et al. 2002; Nakamura et al. 2014; Freeborough and Fox 1997; Fox et al. 2000). This situation

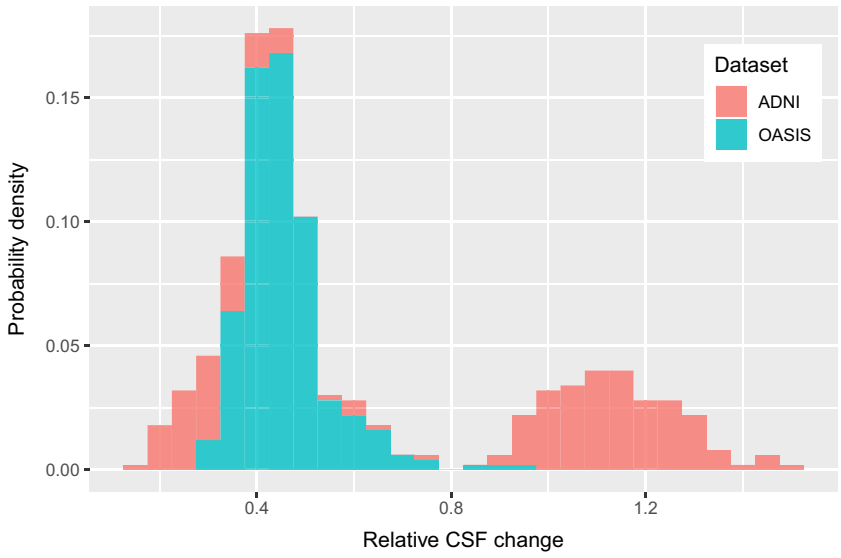


Fig. 3 Distribution of relative cerebrospinal fluid enlargement among pairs of baseline and follow-up volumes on the OASIS and ADNI datasets. Of note, these values may be affected by skull stripping results. CSF: cerebrospinal fluid

required our framework to produce synthetic brain scans in which tissue contrast is good enough for algorithms to detect grey matter, white matter, and cerebrospinal fluid. To evaluate that, we segmented brain tissues in both real and generated scans using FAST (Zhang et al. 2001). With the Dice similarity coefficient (DSC) (Dice 1945; Crum et al. 2006). With the DSC, we determined the extent of overlap between the segmentation masks obtained from synthetic scans and the ground truth used to generate them in the first place. Given tissue probability maps for a real scan, s_{CSF}, s_{GM}, and s_{WM}, and those for its corresponding approximation, $\tilde{s}_{CSF}$, $\tilde{s}_{GM}$, and $\tilde{s}_{WM}$, the DSC is mathematically expressed as

$$DSC\,(s, \tilde{s}) = 2\,\frac{\sum H(s) \cdot H(\tilde{s})}{\sum H(s) \cdot \sum H(\tilde{s})}, \tag{5}$$

where $H(a)$ is the discrete Heaviside step function. The values for DSC range from zero to one, where zero indicates null similarity between segmentation masks and one exact agreement. We expected our framework to produce synthetic scans such that their segmentations are comparable to those used for generating them, i.e. DSC values close to one.

Cerebral Atrophy

As the ultimate goal of our generation framework was to predict the appearance of a baseline T1-w scan after being altered as requested, we studied whether induced variations matched the request. We considered two atrophy quantification methods for assessing this aspect: SIENA (Smith et al. 2002) and the Jacobian determinant integration method (Nakamura et al. 2014)—segmentation-based and registration-based methods, respectively. Once our model deformed the baseline scan according to the input probability maps, we used these two tools to quantify potential atrophy variations between the generated and real scans. Ideally, the percentage of whole-brain volume change (PBVC) yielded by SIENA and the integral of Jacobian determinants yielded by the Jacobian method should be close to zero and one, respectively. Since these two methods address atrophy quantification from two different perspectives, they allowed us to verify whether tissue variations were induced effectively and whether brain boundaries were well-defined.

Statistical Differences

We used the Wilcoxon signed-rank test to assess statistical significance of differences among methods. We considered p-values below 0.01 to be statistically significant.

Architectural Directives and Loss Functions

The first experiment compared the generation quality of four strategies: two of them inspired by relevant data generation strategies and two of our networks optimised with two different loss functions. Some details as follows:

(A) **3D CGAN - MSE**: A network inspired by the work of Shin et al. (2018), consisting of a U-Net generating three brain regions and a discriminator determining whether the generated scan is realistic enough or not. We optimised such networks using the mean square error (generator) and the categorical cross-entropy (discriminator).

(B) **Baseline U-Net - MSE**: A network inspired by the work of Chartsias et al. (2018) and Salem et al. (2019), consisting of three parallel U-Nets generating three brain regions separately and a final addition module to merge them into a single T1-w scan. Each u-shaped subnetwork resembles the design illustrated in Fig. 3 in (Salem et al. 2019). We optimised such a network using a mean square error loss as in the original papers.

(C) **Cascaded U-Nets - MSE**: Our proposed network, as depicted in Fig. 2, consisting of three parallel U-Nets generating three brain regions separately and a final U-Net merging them into a single T1-w scan. We optimised this network using a mean square error loss.

(D) **Cascaded U-Nets - RWMSE**: Our proposed network, as depicted in Fig. 2, consisting of three parallel U-Nets generating three brain regions separately and a final U-Net merging them into a single T1-w scan. We optimised this network using our proposed region-wise mean square error, described in Eq. 1.

We implemented the aforementioned strategies and compared their generation quality. We provided the networks with baseline volumes and actual follow-up tissue segmentation probability maps and evaluated the similarity between the actual follow-up and the approximated one. We trained all networks using the same scheme, i.e. same optimiser, training data, training stopping policy, and machine. Data were taken from the O_2 collection and tested on the O_1 set and vice versa. The results of this experiment are presented in Table 2.

The cascaded U-Net trained with the mean square error loss performed significantly better than its baseline in most cases ($n = 295$; Cascaded-MSE vs Baseline-MSE, p-value; MAE: 0.08 ± 0.04 vs 0.12 ± 0.06, $p < 0.01$; SSIM: 0.95 ± 0.02 vs 0.88 ± 0.07, $p < 0.01$; DSC-CSF: 0.94 ± 0.02 vs 0.92 ± 0.03, $p < 0.01$; DSC-GM: 0.89 ± 0.04 vs 0.88 ± 0.03, $p < 0.01$; PBVC: 0.26 ± 0.21 vs 2.68 ± 0.74, $p < 0.01$), except in terms of the segmentation of white matter, where they both obtained similar Dice scores

Table 2 Generation quality scores obtained with four different strategies

Train → Test	n	Approach	Intensity		Segmentation			Atrophy	
			MAE	SSIM	DSC - CSF	DSC - GM	DSC - WM	PBVC	Jacobian Int
$O2 \rightarrow O1$	126	3D CGAN - MSE	0.03 ± 0.01	0.95 ± 0.02	0.83 ± 0.16	0.69 ± 0.21	0.78 ± 0.20	2.19 ± 5.70	$\mathbf{0.99 \pm 0.01}$
		Baseline U-Net - MSE	0.08 ± 0.04	0.90 ± 0.04	0.92 ± 0.02	0.87 ± 0.03	0.90 ± 0.02	2.60 ± 1.01	1.13 ± 0.06
		Cascaded - MSE	0.05 ± 0.03	0.96 ± 0.01	0.93 ± 0.02	0.87 ± 0.05	0.91 ± 0.04	0.33 ± 0.25	1.16 ± 0.06
		Cascaded - RWMSE	$\mathbf{0.02 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.94 \pm 0.03}$	$\mathbf{0.95 \pm 0.02}$	0.27 ± 0.16	1.14 ± 0.05
$O1 \rightarrow O2$	169	3D CGAN - MSE	0.03 ± 0.01	0.95 ± 0.01	0.80 ± 0.15	0.71 ± 0.20	0.79 ± 0.21	2.19 ± 5.70	$\mathbf{1.00 \pm 0.02}$
		Baseline U-Net - MSE	0.15 ± 0.05	0.87 ± 0.08	0.92 ± 0.03	0.88 ± 0.02	0.91 ± 0.01	2.73 ± 0.43	1.12 ± 0.05
		Cascaded - MSE	0.11 ± 0.03	0.95 ± 0.02	0.94 ± 0.02	0.90 ± 0.02	0.91 ± 0.03	0.21 ± 0.16	1.14 ± 0.05
		Cascaded - RWMSE	$\mathbf{0.01 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.96 \pm 0.02}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.95 \pm 0.01}$	0.19 ± 0.14	1.14 ± 0.05

The results were obtained from training on O_2 and testing on O_1 and vice versa. The variable n represents the number of test cases. The values in bold are significantly higher (p-value < 0.01) than the ones yielded by the other three approaches. MAE: median absolute error. SSIM: structural similarity. DSC: Dice similarity coefficient. CSF: cerebrospinal fluid. GM: grey matter. WM: white matter. PBVC: percentage of brain volume change. CGAN: conditional generative adversarial network. MSE: mean square error. RWMSE: region-wise mean square error

(DSC-WM: 0.91 ± 0.04 vs 0.91 ± 0.02, $p = 1.00$), and Jacobian integral, where the latter outperformed the former (Jacobian Int: 1.15 ± 0.06 vs 1.12 ± 0.6, $p < 0.01$). This outcome suggested that scans generated with the cascaded U-Nets trained with the mean square error appear more similar to real follow-up acquisitions and exhibited better tissue contrast than those generated with the baseline U-Net, but brain edges were more blurred.

The use of the region-wise mean square error resulted in significantly improved performance compared to that obtained using the original mean square error ($n = 295$; Cascaded-RWMSE vs Cascaded-MSE, p-value; MAE: 0.01 ± 0.01 vs 0.08 ± 0.04, $p < 0.01$; SSIM: 0.99 ± 0.01 vs 0.95 ± 0.02, $p < 0.01$; DSC-CSF: 0.96 ± 0.02 vs 0.94 ± 0.02, $p < 0.01$; DSC-GM: 0.94 ± 0.03 vs 0.89 ± 0.04, $p < 0.01$; DSC-WM: 0.95 ± 0.01 vs 0.91 ± 0.04, $p < 0.01$; PBVC: 0.22 ± 0.15 vs 0.26 ± 0.21, $p < 0.01$), except for the Jacobian integral, where the difference between their scores was not significant (Jacobian Int: 1.14 ± 0.05 vs 1.15 ± 0.06, $p > 0.01$). These results suggest that the proposed loss function allows the network to generate more faithful reconstructions versus the accustomed loss. However, the proposed loss did not seem to help to sharpen brain edges.

Notably, the image-to-image translation conditional adversarial network inspired by the work of Isola et al. (2017) and Shin et al. (2018) obtained Jacobian integration values close to one (1.00 ± 0.02), i.e. brain edges were delineated almost perfectly according to this metric. In this regard, this network outperformed all other approaches significantly ($p < 0.01$). Nevertheless, its performance according to the rest of the metrics was significantly lower than our cascaded U-Net trained with the region-wise mean square loss function ($n = 295$; Cascaded-RWMSE vs 3D CGAN-MSE, p-value; MAE: 0.01 ± 0.01 vs 0.03 ± 0.01, $p < 0.01$; SSIM: 0.99 ± 0.01 vs 0.95 ± 0.01, $p < 0.01$; DSC-CSF: 0.96 ± 0.02 vs 0.81 ± 0.15, $p < 0.01$; DSC-GM: 0.94 ± 0.03 vs 0.70 ± 0.20, $p < 0.01$; DSC-WM: 0.95 ± 0.01 vs 0.79 ± 0.21, $p < 0.01$; PBVC: 0.22 ± 0.15 vs 2.19 ± 5.70, $p < 0.01$). Compared to the rest of the models, scans generated using the adversarial model presented lower tissue contrast that prevented them from being segmented properly.

An example of generated scans using the five strategies is presented in Fig. 4. We displayed the generation on the case 157 of the OASIS2 dataset as it exhibited the maximum relative CSF change in this dataset and, thus, generation issues were visually evident. Qualitatively speaking, literature inspired strategies did not lead to appealing results. The conditional adversarial network generated scans with sharp yet noisy edges and inaccuracies in the lateral ventricles that appear as if the model laid the ground truth over the baseline and failed at amalgamating intensities accurately. The baseline U-Net learnt identity mapping as the only visual differences are in terms of the noise, reduced in synthetic scans. Scans generated using our cascaded U-Nets trained with the mean square error exhibited artefacts; the reconstructions provided by each branch seem to be merged in an uncoordinated way as tissues seem superimposed. On the contrary, both axial slices generated using our proposed RWMSE loss function appear similar to the expected follow-up scan as tissues were altered as expected. Our proposal reduced speckle noise and delineated better some structures (e.g. sub-cortical structures) compared to the real follow-up scans, i.e. the contrast of the image was enhanced.

Taking the aforementioned information into account, our proposed cascaded U-Net model optimised with the region-wise loss function evidenced improved performance both

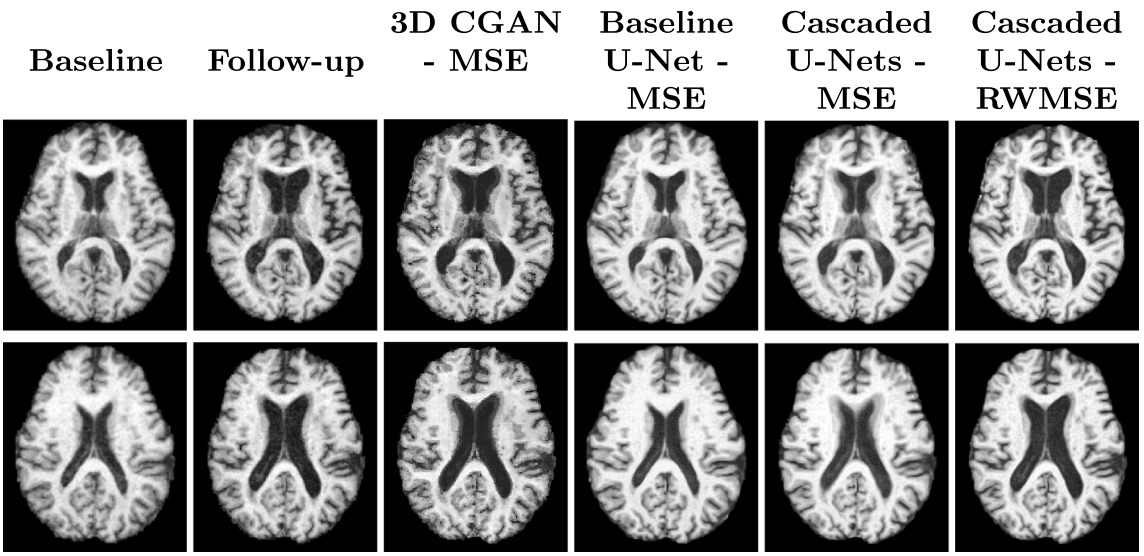


Fig. 4 Example of scans generated with different architectures and loss functions. The first and second column correspond to the real baseline and follow-up scans. From the third to the sixth column, scans generated with the conditional generative adversarial network trained using the mean square error loss, with the baseline U-Net trained using the mean square error loss, with our proposed design trained with a mean square error, and with our proposed architecture optimised with our region-wise mean square error. CGAN: conditional generative adversarial network. MSE: mean square error. RWMSE: region-wise mean square error

qualitatively and quantitatively. Henceforth, we computed our results using such a model.

Generation Quality (Same Dataset)

We ran a second experiment to evaluate the quality of the generation of our tool. We assessed generation when synthesising a scan from a baseline, i.e. we provide the network with a baseline T1-w volume and three tissue probability maps of the corresponding follow-up T1-w acquisition.

The results obtained by our proposal on the considered datasets are displayed in Table 3. Our model generated volumes that were quantitatively similar to the actual follow-up scans at intensity, segmentation and atrophy levels. Regarding intensity, our method yielded MAE values below 0.11 and SSIM values above 0.90. Concerning segmentation, our tool produced images with tissue masks comparable to the ones of the actual volumes as all DSC values are above 0.80. Nevertheless, the obtained segmentation errors were within the FAST accuracy and reproducibility ranges (de Boer et al. 2010). Regarding the volume change detected by atrophy quantification algorithms, our method reported low values overall and within reproducibility rates (Cover et al. 2011).

Our method yielded better results intensity-wise on the OASIS set than on the ADNI one. This might be a consequence of increased lousy skull stripping of ROBEX on the latter set in comparison to the former. If a synthetic scan is compared to a follow-up volume which skull has not been entirely removed, the scores for MAE and SSIM will

Table 3 Comparison between generated and actual volumes concerning intensity, segmentation, and atrophy dissimilarities

Train → Test	n	Intensity		Segmentation			Atrophy	
		MAE	SSIM	DSC - CSF	DSC - GM	DSC - WM	PBVC	Jacobian Int
$O2 \rightarrow O1$	169	0.02 ± 0.01	0.99 ± 0.01	0.96 ± 0.01	0.94 ± 0.03	0.95 ± 0.02	0.27 ± 0.16	1.14 ± 0.05
$O1 \rightarrow O2$	126	0.01 ± 0.01	0.99 ± 0.01	0.96 ± 0.02	0.94 ± 0.02	0.95 ± 0.01	0.19 ± 0.14	1.14 ± 0.05
$A2 \rightarrow A1$	153	0.03 ± 0.03	0.97 ± 0.02	0.95 ± 0.02	0.92 ± 0.02	0.96 ± 0.01	0.24 ± 0.18	1.11 ± 0.05
$A1 \rightarrow A2$	136	0.04 ± 0.03	0.98 ± 0.01	0.95 ± 0.02	0.93 ± 0.03	0.96 ± 0.01	0.23 ± 0.16	1.13 ± 0.03
$OASIS \rightarrow ADNI$	289	0.03 ± 0.01	0.97 ± 0.02	0.96 ± 0.01	0.94 ± 0.03	0.96 ± 0.01	0.15 ± 0.15	1.12 ± 0.04
$ADNI \rightarrow OASIS$	295	0.02 ± 0.01	0.97 ± 0.01	0.96 ± 0.01	0.94 ± 0.02	0.95 ± 0.02	0.22 ± 0.35	1.15 ± 0.06

The column n shows the cardinality of test set. The segmentation scores correspond to the DSC values computed using FAST masks. MAE: median absolute error. SSIM: structural similarity. DSC: Dice similarity coefficient. CSF: cerebrospinal fluid. GM: grey matter. WM: white matter. PBVC: percentage of brain volume change

be lower than when non-brain areas have been completely masked out.

Generation Quality (Cross-Dataset)

The third experiment consisted of evaluating the performance of the proposal when training on a certain dataset and testing on a different one (OASIS→ADNI and ADNI→OASIS). The results are shown in Table 3 and two cases depicted in Fig. 5. The generation per se did not seem significantly affected as none of the intensity, segmentation, or atrophy values differed significantly from the performance measurements obtained when training and testing on the same dataset. This outcome makes our proposal appealing as it shows that by pre-processing the incoming data (e.g. harmonisation by registering to a common space and matching intensity histograms), the network might be used in a different domain without requiring retraining.

Evaluating Induced Changes with Brain Volumetry Methods

The fourth experiment consisted of exploring whether induced tissue variations could be detected by atrophy quantification algorithms. We created the dataset as follows. Initially, we selected ADNI subjects which exhibited the maximum atrophy over time. The atrophy was measured as relative enlargement of the CSF region. We computed the deformation field between the baseline and latest follow-up scans. Then, we multiplied the resulting deformation vectors by scalars between zero and one to obtain intermediate scans. We considered five scalars: 0%, 25%, 50%, 75%, and 100%. Of note, this is an approximation to the pathological process as we would assume that atrophy change varies spatially at the same time in all directions. Afterwards, we ran FAST on the baseline scans to segment each tissue and altered the resulting tissue probability maps using the various deformation fields. Finally, we input each pair of baseline volume and modified tissue maps to generate a synthetic scan. In total, we generated 216 synthetic scans.

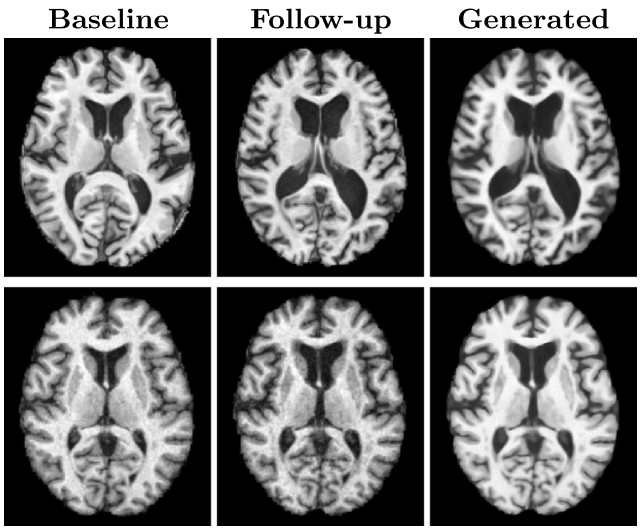


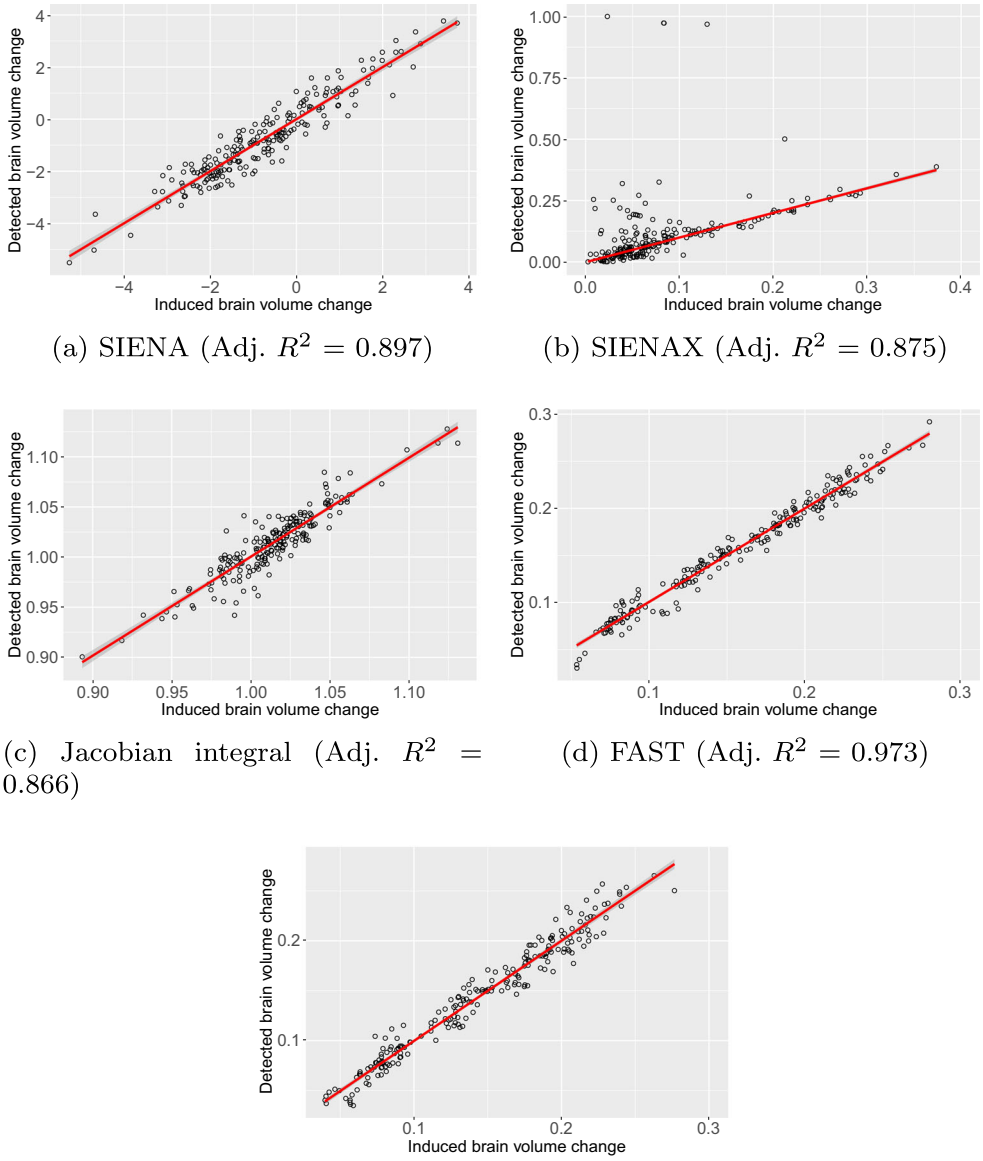
Fig. 5 Cross-dataset generation examples: an ADNI follow-up scan using a network trained on OASIS (top) and an OASIS follow-up scan using a network trained on ADNI (bottom)

We evaluated the capacity of our framework to generate detectable tissue variations using four methods: three atrophy quantification algorithms, SIENA, SIENAX, and the Jacobian determinant integration method, and two tissue segmentation algorithms, FAST and SPM. We computed a robust multiple linear regression model (Li 1985) using relative absolute volumetric differences, tissue-wise average symmetric surface changes (Heimann et al. 2009), and Dice coefficients (as surrogate measures for tissue displacement) as predictor variables and detected or observed brain volume change as a response variable. The results are shown in Fig. 6. Overall, our induced tissue variations correlated well with the detected volume change (adjusted correlation coefficient R^2 above 0.86). For SPM and FAST, the linear model was close to $x = y$ as $R^2 \approx 1$ and y-intercept ≈ 0. This outcome implies that the induced tissue variations were detected correctly by conventional cross-sectional and longitudinal atrophy quantification tools.

Discussion

In this paper, we proposed a CNN-based framework for creating longitudinal evaluation environments given a set of T1-w baseline scans and follow-up tissue probability maps. Our pipeline contemplates four stages: preprocessing, data preparation, generation and postprocessing. Initially, we skull-stripped, intensity corrected and registered all volumes to a common space. Then, we tiled up the baseline and altered tissue probability maps into overlapping blocks and passed them through our network, a cascaded u-shaped network. Finally, once all blocks had been processed, we reconstructed and intensity corrected the resulting synthetic volume.

The network consisted of four processing modules: one dedicated to generating changes on each class (namely, CSF, GM, and WM) and the last one in charge of fusing them. We optimised all components end-to-end using a region-aware multi-objective loss function. We followed state-of-the-art design patterns to devise our network. Overall, the devised framework produced synthetic scans accurately in terms of intensity, segmentation and tissue volume similarity. The proposal was assessed through four experiments exploring architecture directives and loss



(a) SIENA (Adj. $R^2 = 0.897$)

(b) SIENAX (Adj. $R^2 = 0.875$)

(c) Jacobian integral (Adj. $R^2 = 0.866$)

(d) FAST (Adj. $R^2 = 0.973$)

(e) SPM (Adj. $R^2 = 0.957$)

Fig. 6 Actual versus fitted values obtained using robust linear regression models for the five different methods. The models were built using the average symmetric surface distance, the relative absolute volumetric difference, and Dice coefficients between original and deformed tissue maps as predictor variables and detected volume change as response variable. Data points and regression lines are represented by empty circles and red lines, respectively

functions, generation quality when training the network on a particular dataset and testing on the same or different one, and the ability of our framework to generate acceptable and detectable changes.

The first experiment compared our proposal against two literature-inspired networks based on the work on latent space representations using U-Nets of Chartsias et al. (2018) and Salem et al. (2019) and the label-to-image translation conditional generative adversarial network described by Shin et al. (2018). The assessment consisted of generating follow-up scans using baseline data and real follow-up tissue segmentation maps and measuring the similarity between the generated and real scans in terms of their perceptual properties, their segmentation results, and their atrophy extents. Quantitatively, our cascaded U-Net optimised with our region-wise mean square error objective function outperformed both state-of-the-art approximations in most cases, except in delineating brain edges sharply where the conditional generative adversarial network took the lead. Qualitatively, scans generated with our proposal did not exhibit visual artefacts unlike those synthesised with other approximations, as illustrated in Fig. 4. Having the aforementioned aspects in mind, we chose our proposal over the considered literature-inspired models.

The second experiment gauged the capacity of our proposal to generate synthetic follow-up scans which were similar to the actual images when training and testing on the same domain. The similarity was evaluated regarding intensity using MAE and SSIM, tissue segmentation mask overlap using FAST and DSC, and atrophy change

using SIENA and the Jacobian integration method. The experiment considered four collections: two from the OASIS and two from the ADNI. In all of them, our proposal yielded high similarity scores. We observed that skull stripping errors resulted in increased dissimilarity scores, as indicated previously in the literature (Nakamura et al. 2018). Nonetheless, all the values were within the reproducibility ranges reported in the literature.

The third experiment explored whether the framework could be used in unseen and different data collections. We trained our network on a particular selection (e.g. OASIS2) and tested on another one (e.g. ADNI) and vice versa to determine how robust was the entire framework to these sort of variations. Our preliminary results showed that our framework may cope with this situation without affecting its performance considerably and without requiring additional adjustments, but further testing in this regard is needed. Evidently, this outcome is appealing as our ultimate goal is to apply our pipeline to datasets with possibly varying acquisition parameters.

The fourth experiment examined whether conventional tools detected synthetically induced changes. This is key in this research as our primary goal is to create high-quality synthetic scans for which tissue variations (loss) with respect to the baseline scans are known beforehand. We used real tissue displacement vectors to alter baseline segmentation masks, input them into our framework, and gauged changes using SIENA, SIENAX, the Jacobian integration method, SPM, and FAST. All changes detected by these five tools highly correlate with our induced changes (Adj. R^2 values above 0.86), showing common tissue segmentation and volumetry methods can detect brain alterations generated by our proposal. Note that even algorithms that were not used at any point within our framework correlated with the induced changes.

A direct and fair comparison with other works in the area is not straightforward as inputs and generation mechanisms vary. For example, in (Karaçali and Davatzikos 2006), the tool is provided with an MR scan and a number indicating the desired level of expected tissue loss and the tool outputs another scan in which the brain volume has been altered to match the requested value. The deformation of the volume follows the topology of the brain rather than a pathology-oriented pattern per se. Khanal et al. (2017) proposed a tool for prescribing local atrophy changes given segmentation and atrophy maps, in which the user indicates modifiable and not modifiable regions and the expected degree of atrophy, respectively. We did not compare to their work since we would need to build both maps appropriately and accurately. Thus, we compared our proposal against networks inspired by previous works on image and lesion synthesis (Chartsias et al. 2018; Salem et al. 2019) and data augmentation (Shin et al. 2018) since their code was either publicly available and/or used established and well-known strategies.

The motivation behind our proposal is two-fold. First, we aim to generate controlled environments to evaluate atrophy quantification strategies. Following the urgent challenges in GM atrophy measurement exposed by Amiri et al. (2018), pipelines could be compared under the same settings, and their pros and cons could be adequately analysed using our tool. This would be a way to extend the clinical validation of existing tools. Second, we target using the deep learning power to craft a more precise and accurate method for measuring tissue loss. As it is well-known in the literature, deep learning has outperformed traditional machine learning methods in scenarios where lots of data are available. Thus, we could train networks to achieve improved measurements using our tool.

Our proposal exhibits limitations regarding segmentation, model assumptions, domain dependence, and bias. First, it is well-known that the segmentation performance of FAST in basal ganglia is not accurate enough. Although we did not observe problems in this regard (see Figs. 4 and 5), better segmentation strategies need to be considered. Second, unlike model-based proposals (Karaçali and Davatzikos 2006), there are no assumptions on how tissues are altered to match the input segmentation maps. On the one hand, this favours the flexibility of the generation scheme. On the other hand, it does not follow a specific pathology-oriented deformation strategy. Third, the core network may produce undesired outcomes when the intensity range of an input scan differs considerably from the training intensity interval. Nonetheless, this issue was mitigated by performing intensity standardisation and registering input scans to the training space. Fourth, the current strategy for generating controlled environments requires image segmentation and registration, i.e. generation is biased towards them. Nonetheless, we observed that our method could generate tissue changes that were highly correlated with SPM, a method that was not considered in the training pipeline.

In the future, we plan to use deep learning to learn pathology specific tissue deformations using a conditional generative network and add this module to our framework. This will open the doors to developing novel tools that can be later used in investigating atrophy-related pathologies. Additionally, the image-to-image translation conditional generative adversarial network proposed by Isola et al. (2017) generates edges that are of better quality than those generated with our current proposal. We plan to implement and evaluate the effectiveness of other generative adversarial models to further decrease the error that our current proposal presents. Moreover, we aim to devise a fully deep learning-based framework to provide medical doctors with robust and reliable longitudinal atrophy

measurements. To encourage other researchers to use the implemented framework, we released a publicly accessible version of it in our GitHub repository (See information sharing statement).

Information Sharing Statement

We implemented our framework using the Python programming language and the Keras library. We made our implementation publicly available at our GitHub repository: github.com/NIC-VICOROB/atrophy-generation. We listed all installation requirements in the README.md file in the same repository. Most dependencies needed to run our framework can be installed using "pip", except for FSL (fsl.fmrib.ox.ac.uk/fsl/fslwiki/FslInstallation) and ROBEX (www.nitrc.org/projects/robex). The original OASIS and ADNI datasets are available to download from www.oasis-brains.org/ and http://adni.loni.usc.edu/, respectively. We attached the list of selected cases from the ADNI dataset as Supplementary material.

Acknowledgements Data used in the preparation of this article were [in part] obtained from the OASIS dataset: Longitudinal: Principal Investigators: D. Marcus, R, Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382. Data collection and sharing for the ADNI 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 organisation 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.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Supplementary Information The online version contains supplementary material available at (10.1007/s12021-020-09499-z)

References

Amiri, H., de Sitter, A., Bendfeldt, K., Battaglini, M., Wheeler-Kingshott, C.A.G., Calabrese, M., Geurts, J.J., Rocca, M.A., Sastre-Garriga, J., Enzinger, C., et al. (2018). Urgent challenges in quantification and interpretation of brain grey matter atrophy in individual MS patients using MRI. *NeuroImage: Clinical*, *19*, 466–475.

Andersson, J.L., Jenkinson, M., Smith, S., et al. (2007). Non-linear registration aka Spatial normalisation FMRIB Technial Report TR07JA2. FMRIB Analysis Group of the University of Oxford.

Ashburner, J., Barnes, G., Chen, C. (2012). SPM12 Manual. www.fil.ion.ucl.ac.uk (Online; Accessed 21 Jun 2018.

Battaglini, M., Jenkinson, M., De Stefano, N., Initiative, A.D.N. (2018). SIENA-XL for improving the assessment of gray and white matter volume changes on brain MRI. *Human Brain Mapping*, *39*(3), 1063–1077.

Bernal, J., Kushibar, K., Asfaw, D.S., Valverde, S., Oliver, A., Martí, R., Lladó, X. (2019a). Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial Intelligence in Medicine*, *95*, 64–81.

Bernal, J., Kushibar, K., Cabezas, M., Valverde, S., Oliver, A., Lladó, X. (2019b). Quantitative analysis of patch-based fully convolutional neural networks for tissue segmentation on brain magnetic resonance imaging. *IEEE Access*, *7*, 89986–90002.

Brosch, T., Tang, L.Y., Yoo, Y., Li, D.K., Traboulsee, A., Tam, R. (2016). Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Transactions on Medical Imaging*, *35*(5), 1229–1239.

Chartsias, A., Joyce, T., Giuffrida, M.V., Tsaftaris, S.A. (2018). Multimodal MR synthesis via modality-invariant latent representation. *IEEE Transactions on Medical Imaging*, *37*(3), 803–814.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention* (pp. 424–432): Springer.

Costa, P., Galdran, A., Meyer, M.I., Niemeijer, M., Abràmoff, M., Mendonća, A.M., Campilho, A. (2018). End-to-end adversarial retinal image synthesis. *IEEE Transactions on Medical Imaging*, *37*(3), 781–791.

Cover, K.S., van Schijndel, R.A., van Dijk, B.W., Redolfi, A., Knol, D.L., Frisoni, G.B., Barkhof, F., Vrenken, H., Initiative, A.D.N., et al. (2011). Assessing the reproducibility of the SIENAX and SIENA brain atrophy measures using the ADNI back-to-back MP-RAGE MRI scans. *Psychiatry Research: Neuroimaging*, *193*(3), 182–190.

Crum, W.R., Camara, O., Hill, D.L.G. (2006). Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, *25*(11), 1451–1461.

de Boer, R., Vrooman, H.A., Ikram, M.A., Vernooij, M.W., Breteler, M.M., van der Lugt, A., Niessen, W.J. (2010). Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods. *NeuroImage*, *51*(3), 1047–1056.

Dice, L.R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

Ens, K., Wenzel, F., Young, S., Modersitzki, J., Fischer, B. (2009). Design of a synthetic database for the validation of non-linear registration and segmentation of magnetic resonance brain

images. In *Medical imaging 2009: image processing*, (Vol. 7259 p. 725933). International Society for Optics and Photonics.

Filippi, M., Rocca, M.A., Ciccarelli, O., De Stefano, N., Evangelou, N., Kappos, L., Rovira, A., Sastre-Garriga, J., Tintorè, M., Frederiksen, J.L., Gasperini, C., Palace, J., Reich, D.S., Banwell, B., Montalban, X., Barkhof, F. (2016). MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *The Lancet Neurology*, *15*(3), 292–303.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355.

Fox, N.C., Jenkins, R., Leary, S.M., Stevenson, V.L., Losseff, N.A., Crum, W.R., Harvey, R.J., Rossor, M.N., Miller, D.H., Thompson, A.J. (2000). Progressive cerebral atrophy in MS: a serial study using registered, volumetric MRI. *Neurology*, *54*(4), 807–812.

Freeborough, P.A., & Fox, N.C. (1997). The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. *IEEE Transactions on Medical Imaging*, *16*(5), 623–629.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, *321*, 321–331.

Ghafoorian, M., Karssemeijer, N., Heskes, T., Uden, I.W., Sanchez, C.I., Litjens, G., Leeuw, F.E., Ginneken, B., Marchiori, E., Platel, B. (2017). Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific Reports*, *7*(1), 5110.

Glorot, X., Bordes, A., Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the 14th international conference on artificial intelligence and statistics* (pp. 315–323).

Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M., Dickie, D., Wardlaw, J., et al. (2018). White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage: Clinical*, *17*, 918–934.

Haijma, S.V., Van Haren, N., Cahn, W., Koolschijn, P.C.M., Hulshoff Pol, H.E., Kahn, R.S. (2012). Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. *Schizophrenia Bulletin*, *39*(5), 1129–1138.

He, K., Zhang, X., Ren, S., Sun, J. (2015). Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034).

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Heimann, T., Van Ginneken, B., Styner, M.A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., et al. (2009). Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Transactions on Medical Imaging*, *28*(8), 1251–1265.

Hore, A., & Ziou, D. (2010). Image quality metrics: PSNR vs. SSIM. In *Proceedings of the 20th international conference on pattern recognition* (pp. 2366–2369).

Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, *30*(9), 1617–1634.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156.

Jenkinson, M., Bannister, P., Brady, M., Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841.

Jia, G., Heymsfield, S.B., Zhou, J., Yang, G., Takayama, Y. (2016). *Quantitative biomedical imaging: techniques and clinical applications*. BioMed Research International.

Karaçali, B., & Davatzikos, C. (2006). Simulation of tissue atrophy using a topology preserving transformation model. *IEEE Transactions on Medical Imaging*, *25*(5), 649–652.

Khanal, B., Ayache, N., Pennec, X. (2017). Simulating longitudinal brain mris with known volume changes and realistic variations in image intensity. *Frontiers in Neuroscience*, *11*, 132.

Kingma, D.P., & Ba, J. (2014). Adam: a method for stochastic optimization. coRR arXiv:1412.6980.

Krebs, J., e Delingette, H., Mailhé, B., Ayache, N., Mansi, T. (2019). Learning a probabilistic model for diffeomorphic registration. *IEEE Transactions on Medical Imaging*, *38*(9), 2165–2176.

Li, G. (1985). Robust regression. *Exploring Data Tables, Trends, and Shapes*, *281*, U340.

Lin, M., Chen, Q., Yan, S. (2013). Network in network. CoRR arXiv:1312.4400, pp. 1–10.

Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L. (2010). Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, *22*(12), 2677–2684.

Nakamura, K., Guizard, N., Fonov, V.S., Narayanan, S., Collins, D.L., Arnold, D.L. (2014). Jacobian integration method increases the statistical power to measure gray matter atrophy in multiple sclerosis. *NeuroImage: Clinical*, *4*, 10–17.

Nakamura, K., Eskildsen, S.F., Narayanan, S., Arnold, D.L., Collins, D.L., Initiative, A.D.N., et al. (2018). Improving the SIENA performance using BEaST brain extraction. *PloS One*, *13*(9), e0196945.

Nyúl, L.G., Udupa, J.K., Zhang, X. (2000). New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging*, *19*(2), 143–150.

Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage*, *56*(3), 907–922.

Rocca, M.A., Battaglini, M., Benedict, R.H., De Stefano, N., Geurts, J.J., Henry, R.G., Horsfield, M.A., Jenkinson, M., Pagani, E., Filippi, M. (2017). Brain MRI atrophy quantification in MS: from methods to clinical application. *Neurology*, *88*(4), 403–413.

Rovira, À., Wattjes, M.P., Tintoré, M., Tur, C., Yousry, T.A., Sormani, M.P., De Stefano, N., Filippi, M., Auger, C., Rocca, M.A., et al. (2015). Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—clinical implementation in the diagnostic process. *Nature Reviews Neurology*, *11*(8), 471–482.

Roy, S., Carass, A., Prince, J. (2013). Magnetic resonance image example-based contrast synthesis. *IEEE Transactions on Medical Imaging*, *32*(12), 2348–2363.

Rudick, R.A., Fisher, E., Lee, J.-C., Simon, J., Jacobs, L., Multiple Sclerosis Collaborative Research Group, et al. (1999). Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting MS. *Neurology*, *53*(8), 1698–1698.

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., Rovira À., Lladó, X. (2019). Multiple Sclerosis Lesion Synthesis in MRI using an encoder-decoder U-NET. IEEE Access.

Sharma, S., Rousseau, F., Heitz, F., Rumbach, L., Armspach, J.P. (2013). On the estimation and correction of bias in local atrophy estimations using example atrophy simulations. *Computerized Medical Imaging and Graphics*, *37*(7–8), 538–551.

Shin, H.C., Tenenholtz, N.A., Rogers, J.K., Schwarz, C.G., Senjem, M.L., Gunter, J.L., Andriole, K., Michalski, M. (2018). Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International workshop on simulation and synthesis in medical imaging* (pp. 1–11): Springer.

Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P., Federico, A., De Stefano, N. (2002). Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage*, *17*(1), 479–489.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M. (2015). Striving for simplicity: the all convolutional net. In *ICLR (workshop track)* (pp. 1–14).

Steenwijk, M.D., Geurts, J.J.G., Daams, M., Tijms, B.M., Wink, A.M., Balk, L.J., Tewarie, P.K., Uitdehaag, B.M.J., Barkhof, F., Vrenken, H., et al. (2016). Cortical atrophy patterns in multiple sclerosis are non-random and clinically relevant. *Brain*, *139*(1), 115–126.

Storelli, L., Rocca, M.A., Pagani, E., Van Hecke, W., Horsfield, M.A., De Stefano, N., Rovira, A., Sastre-Garriga, J., Palace, J., Sima, D., et al. (2018). Measurement of whole-brain and gray matter atrophy in multiple sclerosis: assessment with MR imaging. *Radiology*, *288*(2), 554–564.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al. (2015). Going deeper with convolutions. In *IEEE conference on computer vision and pattern recognition*.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI* (Vol. 4 p. 12).

Trottier, L., Gigu, P., Chaib-draa, B., et al. (2017). Parametric exponential linear unit for deep convolutional neural networks. In *16th IEEE international conference on machine learning and applications* (pp. 207–214).

van Erp, T.G., Hibar, D.P., Rasmussen, J.M., Glahn, D.C., Pearlson, G.D., Andreassen, O.A., Agartz, I., Westlye, L.T., Haukvik, U.K., Dale, A.M., et al. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry*, *21*(4), 547.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wei, W., Poirion, E., Bodini, B., Durrleman, S., Colliot, O., Stankoff, B., Ayache, N. (2018). FLAIR MR image synthesis by using 3D fully convolutional networks for multiple sclerosis. In *ISMRM-ESMRMB 2018-joint annual meeting* (pp. 1–6).

Zhang, Y., Brady, M., Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57.