Medical Imaging and Applications

Master Thesis, June 2024

NeuroSculpt: Forecasting Brain Structure 9 Years Ahead Using Structural MRI

Agustin CARTAYA, Asta Håberg

Department of Neuromedicine and Movement Science - NTNU, Trondheim, Norway

Abstract

As people age, their brains undergo various structural transformations, primarily involving tissue loss. Accelerated changes can lead to serious conditions such as dementia or Parkinson's disease. Early detection of such abnormal changes in healthy individuals is crucial, as it may allow for early interventions to mitigate these consequences. However, continuous Magnetic Resonance Imaging (MRI) studies, necessary for such detection, are both time-intensive and costly. Currently, several alternatives have been proposed to predict brain structural changes using advances in machine learning and deep learning. However, most focus on patients with neurodegenerative diseases and none specialize in healthy adult populations. In this study, we aimed to predict structural brain changes over a span of nine years in a healthy adult population. We used 3D T1-weighted MR images and explored two primary family of methods. The first family was based on Deformation Fields (DFs), while the second employed deep learning techniques using Generative Adversarial Networks (GANs). DF-based methods were built on the hypothesis, that brain changes observed in one subset of individuals could predict changes in others within the same population. The GAN-based methods were inspired by advancements in predicting brain changes in infants and Alzheimer's disease patients. We evaluated the results of these methods using various assessment criteria, including image similarity, similarity of brain regions, and total brain atrophy. Our results indicated that DF-based techniques were more effective and stable than GANs, demonstrating a greater ability to capture subtle changes, particularly in the thalamus and cortex, as well as significant changes in the ventricles in line with our hypothesis. In contrast, GAN-based methods primarily predicted volumetric changes in the ventricles. This study provided a foundation for future research in brain change prediction, highlighting the effectiveness of DF-based methods and suggesting improvements for GAN approaches.

Keywords: Brain Aging, Deformation Fields, GANs

1. Introduction

1.1. Longitudinal Prediction

Longitudinal prediction involves anticipating how certain characteristics of an individual will change over time based on data collected at earlier moments or theoretical models that describe possible patterns of change (Caruana et al., 2015). In neurology, this approach is crucial for forecasting the progression of neurodegenerative diseases such as Alzheimer's, Parkinson's, or Multiple Sclerosis, enabling treatments before clinical symptoms become evident and slowing disease progression (Arya et al., 2023; Coll et al., 2023; Li et al., 2019). However, despite its benefits, longitudinal prediction faces several challenges, as the accuracy of predictions heavily depends on the quality and quantity of available data, which is not always easy to obtain, especially in the medical domain. (Bandettini, 2012; Bernal et al., 2021; Modat et al., 2014).

1.2. Brain Changes with Aging

As the brain ages, significant structural and functional changes occur that primarily affect cognition (Schulz et al., 2022). On a large scale, grey matter (GM) and white matter (WM), which contain neuronal cell bodies and long-distance synapses, respectively, undergo atrophy, being replaced by cerebrospinal fluid (CSF) (Ge et al., 2002). Some studies indicate that certain brain structures are more susceptible to aging-related changes (Choi et al., 2022; Fujita et al., 2023; Raz et al., 2005).

Structures such as the hippocampus, thalamus, and cortex, crucial for memory, sensory information transmission, and complex cognitive functions, show significant atrophy. These changes are reflected in the expansion of the ventricles, which dilate to compensate for brain volume loss, and the increase of CSF around the brain due to the reduction in the height between sulci and gyri (Kaye et al., 1992).

Brain aging varies between healthy individuals and those with neurodegenerative diseases (Habes et al., 2016). In healthy individuals, structural changes are generally slower and more subtle, influenced by genetics and lifestyle (Mulugeta et al., 2022). In contrast, in patients with diseases like Alzheimer's, atrophy is more accelerated and follows specific, well-documented patterns (Pini et al., 2016). Consequently, numerous predictive models for neurodegenerative diseases have been developed (Arya et al., 2023).

To study these age-related brain changes, Magnetic Resonance Imaging (MRI) has been used as a fundamental tool due to its ability to provide detailed visualization of brain structures (Vemuri et al., 2015). Particularly, T1-weighted (T1w) MR images are especially useful for anatomical visualization, offering good resolution and contrast between GM, WM and CSF (Chen et al., 2018). These images also allow the observation of the subcortical structures sensitive to aging (Duan et al., 2020), thereby facilitating the monitoring of structural changes associated with aging and neurodegenerative diseases.

1.3. Methods for Longitudinal Brain Prediction

Over the last decades, advances in machine learning have offered a powerful tool in longitudinal neurological studies, allowing the quantification of brain aging in patients with neurodegenerative diseases (Zapaishchykova et al., 2024). Currently, two families of methods are most commonly used to infer longitudinal brain changes:

- The first and most used is based on Deformation Fields (DFs). A DF is a fundamental element in the area of non-rigid registrations (Crum et al., 2004) and is based on a vector field that indicates how each pixel (or voxel in 3D images) of a moving image M should be displaced to align it with a fixed image F.
- The second, more recent and based on advances in deep learning, uses generative adversarial networks (GANs) (Goodfellow et al., 2014). A GAN consists of two neural networks: a Generator that creates images from an input and a Discriminator that evaluates their realism, competing with each other to continuously improve.

In the context of longitudinal brain prediction, methods of the first family seek to infer a DF that explains structural changes over time, which can then be applied to the initial brain images to obtain their evolution using image registration. Meanwhile, methods in the second family train a GAN for image-to-image translation (Isola et al., 2018a) using historical data (e.g., initial and future images), and then predict the brain's evolution given the initial image.

1.4. Predicting Brain Changes in Healthy Populations

While the majority of research focuses on structural brain changes caused by neurodegenerative diseases (Camara et al., 2006; Rachmadi et al., 2019; Ravi et al., 2019; Xia et al., 2021), there is significant value in extending these predictive models to healthy populations. Predictive models tailored for healthy individuals could offer insights into normal aging trajectories, identify atypical changes indicative of early disease onset, and highlight the impact of lifestyle and genetic factors on brain health (Hedman et al., 2012). Moreover, such models could facilitate early interventions, potentially mitigating the risk of developing neurodegenerative conditions (Rachmadi et al., 2019). However, predicting brain changes in healthy populations presents challenges, such as the variability of aging processes due to the influence of individual's sociodemographic, health, genetics and lifestyle factors (Mulugeta et al., 2022) and the need for extensive longitudinal data (Bethlehem et al., 2021). This naturally leads to the question: Is it possible to predict brain changes in healthy populations?

1.5. Objective of the Master's Thesis

The objective of this project is to address the previous question and, specifically, to attempt to predict structural brain changes over a nine-year period in healthy adults using 3D T1w MR images. with participants having an average age of 60 years at the time of the initial scan (baseline) and 69 years at the time of the second scan (follow-up).

To achieve our objective, we implemented various methods based on the two main families of longitudinal prediction mentioned earlier:

- **DF-Based Methods:** These methods are based on inferring a DF that captures the necessary volumetric changes to register the baseline and, consequently, predict the follow-up scan. First, we create a dataset of deformation atlases by registering baseline to follow-up and obtaining the resulting DFs from a subset of our population. Then, we implement four different methods based on variants of multi-atlas techniques (Iglesias and Sabuncu, 2014) to combine the obtained deformation atlases and create the desired DF.

- **GAN-Based Methods:** In this family, the methods are based on training a GAN with baseline and follow-up scans from a subset of our population, allowing it to learn the longitudinal changes. Then, from the baseline of a new individual, the GAN can predict the follow-up. To achieve this, we implemented four different GANs based on the architectures proposed by Peng et al. (2021), Huang et al. (2022) and Choi et al. (2020) and adapted them to our objective.

Finally, we conducted a statistical analysis to determine the best method of each family and overall. We used various comparison metrics between the predicted and expected images, based on image similarity, similarity of brain structures relevant to aging (Choi et al., 2022; Fujita et al., 2023), and total brain atrophy using the Brain Parenchymal Fraction (BPF) (Rudick et al., 1999).

2. State of the art

During our review of the state of the art, our primary focus was on longitudinal brain changes, where most of the works we found employed DF-based techniques or GAN-based techniques, primarily for predicting brain atrophy. Additionally, we expanded our search to facial aging studies as they also presented innovative techniques in longitudinal prediction.

2.1. DF-Based Approaches

The prediction of brain atrophy in patients with Alzheimer's or other neurodegenerative diseases has been extensively researched in recent years, primarily using models that aim to infer a DF with specific volumetric changes. Smith et al. (2003) presented a biomechanical model using finite element method and applied thermal loads to induce expansion or contraction in the desired tissues by a DF. Camara et al. (2006) expanded this approach with a thermoelastic model and added acquisition artifacts to the generated image for greater realism. Karacali and Davatzikos (2006) and Sharma et al. (2010) presented models that minimize an energy function, penalizing the deviation between the desired volumetric loss and that inferred from the Jacobian of the DF, preserving brain topology and allowing free movement of CSF. Modat et al. (2014) employed multimodal registrations to obtain a set of velocity fields describing actual brain changes, subsequently combining them to generate DFs specific to each type of disease. Khanal et al. (2017, 2016) developed a biophysical model to generate a DF based on Stokes equations from fluid mechanics, but with a non-zero mass source term to allow the deformation of each tissue based on its prescribed atrophy. Da Silva et al. (2020) used deep neural networks to predict the DF from an atrophy map. In a subsequent work, Da Silva et al. (2021) presented a more comprehensive model that infers the atrophy map from the patient's medical data. More recently, Bernal et al. (2021) proposed a cascade U-Net (Ronneberger et al., 2015) approach to generate controlled synthetic volumes based on probability maps of altered tissues.

Many of these methods propose quite accurate prediction results. However, except for Modat et al. (2014) and Da Silva et al. (2021), these results depend on prespecified atrophy maps. This reliance can be limiting because intermediary scans between baseline and follow-up are needed to construct these maps and observe specific changes for each patient. Given that our dataset does not contain intermediary scans, we propose DF-based models that infer changes based on inter-individual similarity rather than relying on atrophy maps.

2.2. GAN-Based Approaches

Recent research using GANs has demonstrated their utility in predicting the progression of neurodegenerative diseases and aging in MRIs. Rachmadi et al. (2019) proposed DEP-GAN to predict the evolution of white matter hyperintensities in patients with small vessel disease. This model combines GAN with Irregularity Maps to generate Disease Evolution Maps. Similarly, Ravi et al. (2019) and Xia et al. (2021) presented models to predict the evolution of atrophy in brain MRI as a function of age and Alzheimer's disease status. The former proposed DaniNet, a model that combines a conditional deep autoencoder with a GAN, integrating biological constraints to predict realistic synthetic images. The latter developed a network that does not require longitudinal data for training, using identity-preserving losses to maintain subject-specific features in the predicted images. More recently, Gadewar et al. (2023) employed a style-transfer-based architecture to predict brain changes in subjects aged 60 to 79, using multiple age and sex-specific domains. In the field of infant brain development, Peng et al. (2021) and Huang et al. (2022) focused on longitudinal prediction of structural and contrast changes in infants over the first year of life. The first work introduced MPGAN, which combines a feature extractor with a GAN to generate high-quality images using perceptual loss. The second work addressed the problem differently with MGAN, a GAN-based network that uses spatial and frequency information from the baseline to predict metamorphic changes.

All these approaches underscore the capability of GANs for predicting brain changes, but they present several limitations. First, training with 2D slices (Gadewar et al., 2023; Rachmadi et al., 2019; Ravi et al., 2019; Xia et al., 2021), which in most cases is not a choice but rather unavoidable due to lack of computational resources, may result in the loss of inherent 3D information in structural MRI. Second, although training

without longitudinal data is innovative (Gadewar et al., 2023; Xia et al., 2021), it lacks mechanisms to verify the results and guide the network toward individual-specific predictions. Finally, most of the presented works validate their results using global image metrics, which do not detect subtle structural brain changes, mainly in subcortical regions, which are important in brain aging.

In three of our proposed GAN-based models, we address the challenge of loss of 3D information by employing 3D models and reducing image bit-depth to conserve memory. We overcame the second challenge by leveraging our dataset's longitudinal images. We meticulously evaluate model performance and guide training through tailored loss functions designed for individualized longitudinal changes. Furthermore, we present results specific to different brain regions and evaluate them using different metrics.

2.3. Facial Aging

Studies on facial aging propose a different and innovative approach that can be adapted for longitudinal brain prediction, as demonstrated by Ravi et al. (2019) and Gadewar et al. (2023). Among the most notable methods found are those by Antipov et al. (2017) and Choi et al. (2020), which propose GAN-based models. The former proposed Age-cGAN, which generates aged images while preserving the individual's identity. The process uses an encoder to find an optimal latent vector allowing the generator to reconstruct the image; then, the age category in the generator's input is changed to produce the image with the desired age. To ensure identity preservation, a pretrained facial recognition network is used. In the second method, they proposed StarGAN-v2, a network that can transform images from one domain to another with diversity and variability. It implements a style encoder that extracts features (e.g., hairstyle and facial characteristics) from an image A and a generator that adds those features to an image B. Some more recent works implemented diffusion models (Sohl-Dickstein et al., 2015). In Chen and Lathuilière (2023), they used a model that inverts the input image to a latent noise and performs local age-guided text and attention control editing to achieve precise and realistic transformations. In another method proposed by Banerjee et al. (2023), a latent diffusion model with contrastive and biometric losses is used, preserving identity and achieving realistic and high-fidelity age modifications.

These approaches offer different sources of inspiration for longitudinal prediction. However, all these methods rely on 2D images and must be adapted to work with 3D MRI scans, which could be challenging due to misaligned slices. To overcome this limitation in our fourth GAN-based model, we ensure accurate alignment between baseline and follow-up during preprocessing. Additionally, we implemented a dataloader capable of handling inter-individual slice alignment.

3. Material and methods

3.1. Data

In our study, we used a total of 703 individuals from the Nord-Trøndelag Health Study (HUNT) (Åsvold et al., 2022), a longitudinal study involving a healthy population from Nord-Trøndelag, Norway, since 1984. Our study focuses solely on using the 3D T1w MR images obtained during the third wave (HUNT3) (Håberg et al., 2016) in 2009 to predict images from the fourth wave (HUNT4) collected in 2018. The HUNT3 images were obtained using a 1.5T General Electric scanner with a resolution of $1.25 \times 1.25 \times 1.20\,\mathrm{mm}^3$, while the HUNT4 images were acquired using a 3T General Electric scanner with an isotropic resolution of 1 mm. Appendix A provides more information about HUNT3 and HUNT4 T1w MR scans. In this study, we randomly divided the dataset into two main sets for training and testing, with 620 and 83 individuals respectively. Depending on the method employed, validation subsets were also taken from the training set.

3.2. Preprocessing

Given that the baseline and follow-up were obtained nine years apart and with different magnetic field strengths, we harmonized the whole dataset applying a preprocessing. This was performed using FreeSurfer tools (Fischl, 2012) and its deep learning implementation FastSurfer (Henschel et al., 2020).
We began the preprocessing by converting the images to 1mm isotropic MP-RAGE format using the SyntSR tool. This was done for both HUNT3 and HUNT4 images, as employing this network also facilitated bias field correction and contrast standardization, as indicated in the original work (Iglesias et al., 2023, 2021). Then, we aligned the individuals to the MNI-ICBM 152 2009c space (Fonov et al., 2011, 2009) using affine registration with mri_robust_register (Reuter et al., 2010). To ensure that each individual's baseline was adequately aligned with their follow-up, we first registered the baseline to the MNI space and then registered the follow-up to its corresponding registered baseline scan. Finally, we performed skull stripping using SynthStrip (Hoopes et al., 2022), followed by normalization to extract only the brain region within an intensity range of $[0, 1]$. During preprocessing, we obtained two brain masks, with and without the cerebellum, and three types of tissue segmentation. The first segmentation included the 3 primary tissues: CSF, GM, and WM. The second segmentation delineated 35 tissues, incorporating subcortical structures, while the third segmentation encompassed 95 tissues, including both subcortical structures and various cortical regions. The final size in voxels of

the resulting baseline and follow-up images, along with their segmentations and masks, was $193 \times 229 \times 193$. Figure 1 shows the complete preprocessing pipeline and the results of the obtained images.

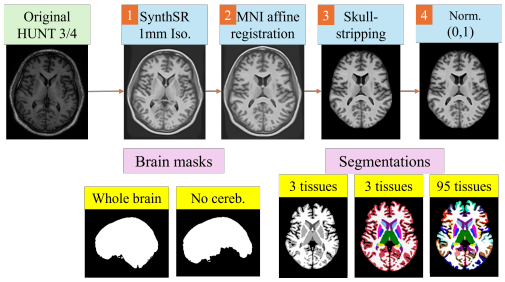

Figure 1: **Preprocessing Pipeline:** Steps performed during preprocessing and the obtained brain masks and segmentations.

Notation and Main Objective

Hereafter, we will refer to the training set for the baseline scans as TX_0 and for the follow-up scans as TX_1, while the test set is referred to as X_0 for the baseline and X_1 for the follow-up scans. Our primary objective is to find $\hat{x}_1$, the best possible approximation of $x_1 \in X_1$, based on the corresponding baseline $x_0 \in X_0$. To achieve this, we employed several methods derived from the two main families of longitudinal brain prediction, which are detailed in the subsequent sections.

3.3. DF-Based Methods

Hypothesis — Our primary hypothesis for this family of methods is that the brain changes of an individual from a specific population could be predicted using the brain changes of other individuals from the same population.

The first step to verify our hypothesis was an inter-individual statistical analysis. We evaluated the similarity in both baseline and follow-up scans to determine if individuals with similar brain structures at baseline maintained this similarity at follow-up in our dataset. For each individual I_0, we identified the individual I_1 with the highest baseline similarity to I_0, and checked if I_1 remained the most similar to I_0 at follow-up or was among the top N most similar individuals. Table 1 and Fig. 2 shows the results of this analysis. To compute the similarity between individuals, we tested two metrics: the Structural Similarity Index (SSIM) introduced by Wang et al. (2004), which ranges from -1 to 1, where 1 indicates perfect similarity, 0 indicates no similarity, and -1 indicates perfect anti-correlation; and the mean Dice coefficient across the three main tissues, which ranges from 0 to 1, where 1 indicates perfect overlap. For two tissues A and B, the Dice coefficient is defined as follows:

$$dice = \frac{2|A \cap B|}{|A| + |B|}$$

We tested these two metrics to capture different aspects of brain structure similarity. SSIM provides a global assessment of structural information and visual quality, while the Dice coefficient focuses on tissue correspondence.

Table 1: Inter individual similarity consistency (%)

Metric	MS	Top3	Top5	Top10	Top15
SSIM	67	93	97	99	100
Mean $dice_3$	58	82	91	96	99

Probability that I_1 is the most similar (MS) to I_0 at follow-up or is among the topN most similar individuals using different similarity metrics.

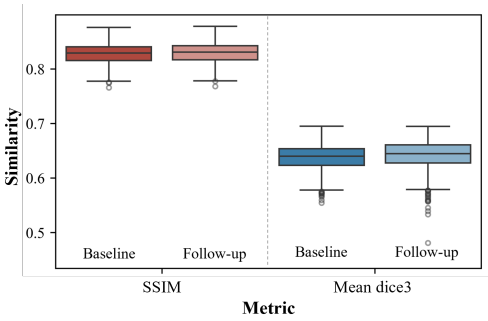

Figure 2: **Inter-Individual Similarities.** Similarity between I_0 and I_1 at baseline and follow-up for all individuals using different metrics.

Results in Table 1 demonstrated that I_1 was consistently identified as the most similar individual to I_0 at follow-up with a probability of 67% using SSIM and 58% using the mean Dice coefficient. Additionally, I_1 was among the top5 with over 90% probability using both metrics. Furthermore, as shown in Figure 2 the similarity between I_0 and I_1 remained stable from baseline to follow-up. These results confirmed that individuals with similar brain structures at baseline maintained this similarity at follow-up in our dataset and motivated us to proceed with the second part of the hypothesis evaluation.

For this second part, we obtained a dataset of DFs from the training data, which we called TDf with tdf as one of its elements. This was achieved by applying non-rigid registrations to the images of TX_0 towards their corresponding images in TX_1 using Elastix (Klein et al., 2009). For these registrations, we used B-Spline transformations with advanced normalized cross-correlation as the similarity metric and a pyramidal approach. Additional information about the registration can be found in Appendix B.

Next, we calculated an average DF from n tdf_i, $i \in [1, n]$, and used it to register the images of X_0. The obtained registered scans indicated an improvement in all

individuals compared to the initial differences between the baseline and follow-up scans. The results showed a mean improvement of 4.1% for the Dice coefficient of the CSF, as well as 0.7% and 0.5% for the GM and WM, respectively. The image similarity based on the SSIM also improved by 0.8%. More details about these results can be found in the Results Section 4.2. These findings confirmed that the use of an average DF, based on a subset of a population, can effectively infer some brain changes in the remaining population, corroborating our initial hypothesis. This prompted us to develop our DF-based methods explained in the following sections.

Objective — Our objective with the following four methods is to infer $\hat{df}$, a DF that explains the longitudinal volumetric changes, allowing us to register x_0 to obtain $\hat{x}_1$. We base these methods on multi-atlas techniques and an adaptation of the K-Nearest Neighbors algorithm to combine the elements of TDf and obtain $\hat{df}$.

3.3.1. Similar Images

Here, we attempted to use image similarity to infer $\hat{df}$. Initially, we calculated the similarity s between x_0 and each $tx_0 \in TX_0$, selecting the n most similar tx_{0i} with their corresponding tdf_i, $i \in [1, n]$. Subsequently, we weighted the tdf_i with their respective normalized s_i and computed their average, resulting in $\hat{df}$ (see Equation (1)). In the implementation of the method, we used L1 normalization. Additionally, we tested different similarity metrics and values for n to evaluate their impact on the final predictions.

$$\hat{df} = \frac{\sum_{i=1}^{n} s_i \cdot tdf_i}{\sum_{i=1}^{n} s_i} \tag{1}$$

3.3.2. Similar Images with Registration

In this method, we followed a similar approach to the previous one, but with one key difference: after identifying the n tdf_i, we registered them to the x_0 space before computing the weighted average (see Equation (2)). We adopted this approach because we considered that obtaining a more precise alignment of the starting point of each vector from a given tdf_i with respect to the image x_0 might result in a more accurate deformation of certain tissues. To register a tdf_i to the x_0 space and obtain $tdf_{i_{x0}}$, we first applied a registration of tx_i to x_0 to obtain the necessary deformation, and subsequently applied it to the corresponding tdf_i. All registrations were made using Elastix, and we tested two different types of registration, affine and B-spline.

$$\hat{df} = \frac{\sum_{i=1}^{n} s_i \cdot tdf_{i_{x0}}}{\sum_{i=1}^{n} s_i} \tag{2}$$

3.3.3. Similar Patches

In this approach, we aimed to infer $\hat{df}$ by patches to capture more anatomical variability. First, we obtained m overlapping uniform patches p of size w that covered the entire x_0. Similarly, we proceeded with all tx_0 and their corresponding tdf, generating tp and $tdfp$ respectively. Then, given a patch p_j, $j \in [1, m]$, we calculated the similarity s between p_j and each tp_j. Next, we selected the n most similar tp_j and finally we computed the weighted average of their corresponding $tdfp_j$ to obtain $\hat{dfp}_j \in \hat{df}$. This process was repeated for each p_j to reconstruct the complete $\hat{df}$ (see Equation (3)). During reconstruction, we used a spline-based method to address overlapping, which helped minimize artifacts in the overlapping areas. In this approach, we set $w = 32$ and an overlap of 50%, both values were experimentally favorable. During the evaluation of the method, we used different values of k and n to assess their effects on the final prediction.

$$\hat{df} = \bigoplus_{j=1}^{m} \hat{dfp}_j \tag{3}$$

Where $\bigoplus$ denotes the operation of patch concatenation with overlap, and each $\hat{dfp}_j$ is constructed as follows:

$$\hat{dfp}_j = \frac{\sum_{i=1}^{n} s_{ji} \cdot tdfp_{ji}}{\sum_{i=1}^{n} s_{ji}}$$

The similarity metric used between patches is based on a weighted Dice coefficient with k tissues, as explained in the following equation:

$$s_j = \frac{\sum_{q=1}^{k} (w_q + a1_q + a2_q) \cdot dice_q(p_j, tp_j)}{2 + \sum_{q=1}^{k} w_q} \tag{4}$$

Where $\text{dice}_q(x, y)$ is the Dice coefficient for tissue q between the segmentation with k tissues of x and y; $a1_q$ and $a2_q$ are the areas of tissue q with respect to the patch size; and w_q is a weigh given to each tissue.

3.3.4. Similar Tissues

Here, we aimed to reconstruct $\hat{df}$ by tissues to allow variability and ensure that each individual tissue deforms consistently. To do this, we used the segmentation with k tissues $segk_0$ of x_0 as well as the segmentations $tsegk_0$ of tx_0 and reconstructed a unique DF for each tissue, subsequently combining them to form $\hat{df}$. This was done very similarly to the patch approach 3.3.3 but with tissue regions instead of patches (see Equation (5)). In this case, there was no overlapping since $segk_0$ contains mutually exclusive tissues. The used similarity metric between the tissues was the Dice coefficient, and during the evaluation, we used different values of k and n.

$$\hat{df} = \bigcup_{j=1}^{k} \hat{dfseg}_j \tag{5}$$

Where $\bigcup$ denotes the operation of tissue concatenation and each $\hat{dfseg}_j$ is constructed as follows:

$$\hat{dfseg}_j = \frac{\sum_{i=1}^{n} s_i \cdot tdfseg_{ji}}{\sum_{i=1}^{n} s_i}$$

3.4. GANs-Based Methods

Objective — Our primary objective with the following four methods is to train a GAN to predict TX_1 from TX_0, enabling it to learn to infer longitudinal structural changes. This way, given an x_0, the network's generator can predict $\hat{x}_1$. To this end, we implemented the architectures proposed by Peng et al. (2021), Huang et al. (2022), and Choi et al. (2020) and adapted them to our objective. In the following methods, we refer to $\hat{tx}_1$ as a predicted image during training and $tx_1 \in TX_1$ as the expected image.

3.4.1. MPGAN

In this approach, we used the multi-contrast perceptual adversarial network MPGAN proposed by Peng et al. (2021). Originally, this network was built to predict longitudinal changes in infant brains during the first year of life, which undergo quite different changes compared to adult aging brains (Huang et al., 2022). In the original paper, they proposed a simple architecture and a multimodal one; in our case, we only implemented the first one given our dataset.

Network Architecture — The MPGAN architecture consists of three main components: A Generator (G) using a U-Net architecture with residual blocks in both the encoder and decoder; a Discriminator (D) that is a classifier composed of convolutional layers followed by an output layer with sigmoid activation; and a pre-trained feature extractor (ϕ) to extract perceptual features. To build ϕ they used the encoder part of the architectured proposed by (Zhou et al., 2019) which is a U-Net model trained with 3D medical images.

Loss Functions — The original paper proposed three loss functions: An adversarial loss (L_{adv}), original to GANs (Goodfellow et al., 2014), which helps $\hat{tx}_1$ approach the distribution of TX_1. A voxel-wise reconstruction loss (L_{vr}), as introduced in Isola et al. (2018b), which ensures consistency between $\hat{tx}_1$ and tx_1 by penalizing voxel-to-voxel differences with an L1 loss. Finally, a perceptual loss (L_{p}), which helps produce sharper and more detailed images by penalizing the difference between the features extracted from $\hat{tx}_1$ and tx_1 using ϕ. The total loss function used is the following:

$$L_{\text{total}} = L_{\text{adv}} + \alpha L_{\text{vr}} + \beta L_{\text{p}} \tag{6}$$

Implementation Details — We used TensorFlow and built the proposed architecture from scratch following the instructions of the original paper, as a functional source code was not available. The Adam optimizer (Kingma and Ba, 2017) with an initial learning rate of 2e-4 was employed, and we applied a decay of 0.5 and a patience of 10 epochs based on the validation loss. The trade-off coefficients α and β were both set to 25, as proposed in the original paper.

To train the network, we used 80% of the images from TX_0 for training and 20% for validation in each epoch. The training process was conducted for a total of 100 epochs with a batch size of 1, applying early stopping with a patience of 10 to avoid overfitting. In order to save GPU memory and train the model using the complete 3D volumes, we used TensorFlow Mixed Precision, which employs both 16-bit and 32-bit floating-point types during training.

3.4.2. MPGAN + Segmentation Loss

Here, we used the same MPGAN network explained in the previous section 3.4.1, with the addition of a segmentation similarity constraint. This was done to increase the similarity of the three main brain tissues (CSF, GM, WM) between $\hat{tx}_1$ and tx_1. In this way, we ensured that global structures and specific tissue details remained consistent, improving the accuracy of segmentation and the structural quality of the generated images.

Loss Functions — To calculate the segmentation loss (L_{seg}), we used a dice-based loss between the segmentation with three tissues of $\hat{tx}_1$ and tx_1 as shown below:

$$\begin{aligned} L_{\text{seg}} = 1 - \frac{1}{3} \big(&\text{dice}_{\text{CSF}}(\hat{tx}_1, tx_1) \\ &+ \text{dice}_{\text{GM}}(\hat{tx}_1, tx_1) \\ &+ \text{dice}_{\text{WM}}(\hat{tx}_1, tx_1)\big) \end{aligned} \tag{7}$$

Where $\text{dice}_q(x, y)$ is the same as used in Equation (4). For tx_1, the segmentation with three tissues obtained during preprocessing was used. However, for $\hat{tx}_1$, we had to calculate the segmentation during training. To achieve this, we used a Gaussian Mixture Model with priors based on the mean and variance of the tissues from tx_1. This allowed to calculate a segmentation for CSF, GM, and WM quickly and easily, with the possibility of gradient propagation in the loss function. Finally, we modified the total loss function as follows:

$$L_{\text{total}} = L_{\text{adv}} + \alpha L_{\text{vr}} + \beta L_{\text{p}} + \gamma L_{\text{seg}} \tag{8}$$

Implementation Details — The implementation was similar to the one described in the previous section 3.4.1, with the only difference being that we adjusted

α, β, and γ to 25, 20, and 15 respectively. These values were found to provide the best results for the validation set.

3.4.3. MGAN

For this method, we used the metamorphic generative adversarial network (MGAN) proposed by Huang et al. (2022). Similar to Peng et al. (2021), the original objective was to predict longitudinal changes in infant brains during the first year of life. However, in this work a 3D patch-based approach using spatial and frequency domains to capture metamorphic changes is proposed.

Network Architecture — The MGAN architecture is based on a CycleGAN (Zhu et al., 2020) and consists of two generators and two discriminators. Each generator includes an encoder, a spatial-frequency transfer block (SFT), and a decoder. The SFT is a dual-branch structure that captures and transforms information in both spatial and frequency domains. For the spatial domain, residual modules in series are used, and for the frequency domain, a discrete wavelet transform (DWT) is applied, followed by residual modules in series and finally an inverse DWT. This allows the preservation of structural and contrast details of the tissues throughout the reconstruction. On the other hand, the discriminators have a U-shaped architecture and generate voxel-level quality probability maps, guiding the generators to focus on the most challenging regions. Both the discriminators and generators use deep supervision in the decoder to strengthen the gradient flow and promote the learning of useful representations at multiple scales (Karnewar and Wang, 2020). It is worth noting that due to the cyclical nature of the network, it would also be possible to predict the baseline from the follow-up, but we did not use this functionality.

Loss Functions — The loss functions used in the paper include an adversarial loss (L_{adv}), a paired loss (L_{p}), and a cyclic loss (L_{cyc}) at different resolutions. The L_{adv}, has the same objective as explained earlier. The L_{p} consists of several components: a quality loss (L_Q), which penalizes voxel-to-voxel differences with an L1 loss, using the discriminator results to focus on the more challenging regions to predict; a texture loss (L_T), which ensures that the texture of $\hat{tx}_1$ is similar to that of tx_1; and a frequency loss (L_F), which compares the wavelet representations between $\hat{tx}_1$ and tx_1 to preserve structural details. Finally, the cyclic loss (L_{cyc}), original to CycleGANs (Zhu et al., 2020), ensures cyclical consistency between the generated and real images, warranting that a transformed image, when reverted, is similar to the original. The total loss function implemented at each scale is the following:

$$L_{\text{total}} = L_{\text{adv}} + \alpha L_{\text{p}} + \beta L_{\text{cyc}} \tag{9}$$

Where:

$$L_{\text{p}} = L_Q + aL_T + bL_F \tag{10}$$

Implementation Details — We used TensorFlow and built the proposed architecture from scratch following the instructions of the original paper, as the source code was not available. The Adam optimizer (Kingma and Ba, 2017) with an initial learning rate of 1e-4 was employed, and we applied a decay of 0.5 and a patience of 10 epochs based on the validation loss. The trade-off coefficients for α, β, a, and b were set to 1, assuming these values were used in the original paper since they were not explicitly mentioned.

To train the network, we extracted patches of size $64 \times 64 \times 64$ with 50% overlap from the images of TX_0. These patches were selected to contain at least 15% brain tissue to avoid creating background-biased generators. The training process was conducted for a total of 10,000 epochs with a batch size of 1, ensuring that all patches from 80% of TX_0 were used for training, while the remaining 20% were reserved for validation.

3.4.4. StyleGAN

In this method, we used the StarGAN-V2 network proposed by Choi et al. (2020). This network was originally designed for style transfer between multiple domains with diversity in the generated images using a single Generator. In our case, we adapted the network similar to the work of Gadewar et al. (2023), to predict $\hat{x}_1$ from x_0 and a desired style s, taken from an element of TX_1.

Network Architecture — The StarGAN v2 architecture is based on four main elements: a generator (G), a mapping network (F), a style encoder (E), and a discriminator (D). G uses a U-Net-like architecture with an encoder, bottleneck, and decoder constructed with residual blocks. The style s is injected into the decoder during the image reconstruction using adaptive instance normalization (Huang and Belongie, 2017). F is a multitask multilayer perceptron that generates a style code s from a latent vector z and a domain y. In our implementation, z is a vector randomly sampled from a Normal Gaussian Distribution, and y is an integer indicating whether the style belongs to the baseline or the follow-up. E is a multitask encoder that, given an image and its corresponding domain, extracts the style code s. Finally, D is a multitask discriminator that differentiates between real and generated images of a domain y. In this context multitask refers to the fact that the network has different output branches, one for each domain y. It is worth noting that all the networks were trained simultaneously. Due to the network's design, it is also possible to predict the baseline from the follow-up. However,

similar to the previous method, we will not focus on that functionality.

Loss Functions — The proposed loss functions consist of an adversarial loss(L_{adv}) and a cyclic loss (L_{cyc}) with the same purpose as in the previous methods; a style reconstruction loss (L_{sty}) that forces the generator to use the style code s when generating the image, extracting and comparing the style of $t\hat{x}_1$ with the desired style; and a style diversification loss (L_{ds}) that encourages the production of diverse images by regularizing the generator to explore different styles. The total loss function used is the following:

$$L_{total} = L_{adv} + \lambda_{cyc}L_{cyc} + \lambda_{sty}L_{sty} + \lambda_{ds}L_{ds} \quad (11)$$

Implementation Details — We used the code proposed by the original paper implemented in PyTorch and adapted it to our dataset. The training parameters we used were exactly the same as those proposed in the original paper.

To train the network, we used 2D slices extracted from the sagittal plane of TX_0 and TX_1. The 2D slices were extracted to contain at least 15% brain tissue to avoid creating a background-biased generator. The training process was conducted for a total of 100,000 epochs with a batch size of 4. It is worth mentioning that the dataloader we designed ensured that the network was trained with slices aligned among individuals.

3.5. Post-Processing

After obtaining the results, we applied post-processing to remove artifacts introduced during prediction, enhance overall image quality, and obtain brain masks and segmentations to evaluate the results. This was performed differently for both families:

- **DF-Based Methods:** We applied the brain mask and normalized the brain area to eliminate edge artifacts caused by interpolation during the registration. To obtain the brain masks and segmentations the initial segmentations and brain masks were registered with Elastix using the inferred DF.

- **GAN-Based Methods:** Here, we first processed the images with SynthSR to eliminate common GAN artifacts (Lee et al., 2023) and correct errors in image reconstruction from 3D patches (MGAN 3.4.3) or 2D slices (StyleGAN 3.4.4). Then, we performed skull stripping, followed by normalization in the brain area to remove the skull and background added by SynthSR. Finally, to obtain the brain masks and segmentations we used FastSurfer.

Computational Resources

For preprocessing and postprocessing, we used FreeSurfer installed on a Linux Ubuntu 18 PC with an Intel(R) Core(TM) i7-7700 CPU and 32GB of RAM, and FastSurfer Docker-version on a Windows 11 PC with a Intel(R) Core(TM) i9-12900H CPU, 32GB of RAM, and an NVIDIA GeForce RTX 3060 GPU with 6GB. For training deep learning methods, we utilized the High Performance Computing cluster at NTNU (IDUN). Specifically, we used clusters with NVIDIA V100 16GB GPUs for models trained using 2D slices and patches, and clusters with NVIDIA A100 40GB GPUs for models trained with full 3D volumes.

4. Results

In this section, we present the predicted scans obtained for each method using the test set. These predictions are evaluated with respect to the actual follow-up scans to verify their exactitude. To help the reader have a comprehensive overview of the evaluation we performed to choose the best method for each family and overall, we have structured this section in three main parts: First, we present the initial similarity between the baseline and follow-up scans of each individual and use it as the lower bound (LB), as it is expected that the results from the implemented methods will surpass this. Second, we present the results for each family separately and choose the best among them. Finally, we compare the best results from each family, conducting a more exhaustive analysis to decide the overall best method.

Evaluation Metrics

During the evaluation of the results, we used various comparison metrics based on global image similarity, cerebral tissue segmentation, and brain atrophy.

To choose the best method for each family, we used SSIM and the mean Dice coefficient of the three main tissues. This allowed us to quickly and accurately select the best results based on global structure and tissue correspondence.

For the more detailed analysis, we used the Dice coefficient, the Absolute Symmetrized Percent Volume Change (ASPVC), the Volume Fraction (VF), and the Brain Parenchymal Fraction (BPF). ASPVC has been used in other analyses of structural changes as it provides a dimensionless measure of variability between tissues (Khanal et al., 2016). For two tissues A and B, ASPVC is defined as:

$$\text{ASPVC} = \frac{|A - B|}{0.5(A + B)} \cdot 100\%$$

On the other hand, VF helped us evaluate whether there was an increase or decrease in tissue volume. For

a tissue A, VP is defined as:

$$\text{VF} = \frac{A}{\text{Inter cranial volume}} \cdot 100\%$$

Finally, we used BPF to compute the atrophy inferred from the predicted scans, which is typically defined as the ratio of brain parenchymal volume to the intracranial volume. In our case, we computed the BPF using the GM and WM volumes (V_{GM}, V_{WM}), excluding cerebellum regions from the segmentation with three tissues, and the brain mask without cerebellum ($Bmask_{ncrb}$), as follows:

$$\text{BPF} = \frac{V_{GM} + V_{WM}}{Bmask_{ncrb}}$$

It is important to note that, for calculating both the BPF and the VF, we used the intracranial volume from the baseline scan to avoid potential segmentation errors. This approach is supported by extensive research demonstrating that total intracranial volume remains constant with aging (Brezova et al., 2014; Hansen et al., 2015; Pintzka et al., 2015).

Significance Evaluation

To determine if the results of our methods were significantly different from the LB, we performed a paired t-test and we considered p-values below 0.01 to be statistically significant.

4.1. Initial Similarity

We started by evaluating the initial similarity between baseline and follow-up scans, setting this as LB for our methods. Table 2 and Figure 3 illustrate these initial similarities, providing a foundation for subsequent analyses.

Table 2: Initial Similarities Between Baseline and Follow-up Scans

Initial	SSIM % ↑	Mean dice %		
		CSF ↑	GM ↑	WM ↑
LB	94.6 ± 1.0	84.3 ± 4.8	80.5 ± 2.6	90.0 ± 1.8

Initial similarity metrics between baseline and follow-up scans, including SSIM and mean Dice coefficient for CSF, GM, and WM. The shown values are the mean of the test set, and the ± values represent the standard deviation. ↑ indicates that higher values are better.

4.2. Family-Wise Results

4.2.1. DF-Based Results

Hypothesis results — Before implementing the DF-based family of methods we evaluated our primarily hypothesis with different values for n to verify its influence and modify this parameter in the actual methods. Table 3 shows the similarity of these results with the actual follow-up.

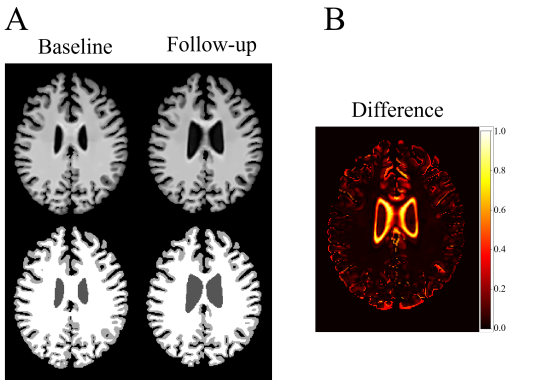

Figure 3: **Baseline and Follow-up Scans**. (A) shows the T1w scans in the first row and the segmentation of the three main tissues (CSF, GM, and WM) in the second row. (B) shows the difference image between the baseline and the follow-up; the lighter the color in a region, the more differences are present.

Table 3: Hypothesis Results - Similarities with Follow-up

n	SSIM % ↑	Mean dice%		
		CSF ↑	GM ↑	WM ↑
10	95.1 ± 0.8	87.8 ± 3.1	*80.8 ± 2.1	90.4 ± 1.3
100	95.3 ± 0.8	88.4 ± 3.0	81.2 ± 2.3	90.5 ± 1.4
200	95.3 ± 0.8	88.3 ± 3.0	81.3 ± 2.3	90.5 ± 1.5
300	95.3 ± 0.8	88.3 ± 3.0	81.3 ± 2.3	90.5 ± 1.5
620	95.4 ± 0.8	88.3 ± 3.1	81.3 ± 2.3	90.5 ± 1.5

Similarity metrics between hypothesis predictions and actual follow-up scans using different values for n. * indicates p-values > 0.01.

The results obtained indicate a slight improvement between $n = 10$ and $n = 100$ but for $n > 100$, the changes are extremely small or negligible.

DF-Based Methods Results — For each method in this family, we evaluated different settings. For the Similar Images method 3.3.1, we tested two similarity metrics (SSIM and the mean Dice coefficient) and three values for $n = [5, 10, 100]$. In the Similar Images with Registration method 3.3.2, we evaluated two types of registrations (affine and non-rigid using B-Splines) and set $n = 5$. It is worth mentioning that the B-spline registration parameters were chosen to prioritize faster registration times over exhaustive optimization. For the Similar Patches method 3.3.3, we used different numbers of tissues $k = [3, 95]$ to evaluate similarity between patches and two values for $n = [10, 100]$. The hyperparameter w in Equation 4 was set to 1 for all the tissues. Finally, for the Similar Tissues method 3.3.4, we tested different numbers of tissues to create the DF $k = [3, 95]$ and two values for $n = [10, 100]$. Table 4 shows the similarity with the follow-up for each method with their respective settings, and Figure 4 shows the predictions using the best setting for each method.

In this family of methods, all segmentation results were calculated using the registered segmentations. However, for the best method, the segmentation was re-

calculated from the predicted image using FastSurfer to avoid interpolation errors in discrete values caused by the registration. This result is also shown in Table 4 along with results obtained using ground truth deformations from the baseline to the follow-up scans through non-rigid registration with Elastix. This latter result could be interpreted as an upper bound (UB) for this family of methods.

Table 4: DF-Based Methods Results - Similarities with Follow-up

Method	SSIM % ↑	Mean dice %		
		CSF ↑	GM ↑	WM ↑
Images				
ssim 5	95.3 ± 0.7	89.7 ± 2.3	81.1 ± 2.1	90.7 ± 1.4
ssim 10	95.4 ± 0.7	89.8 ± 2.4	81.5 ± 2.2	90.9 ± 1.4
ssim 100	95.4 ± 0.8	89.6 ± 2.7	81.5 ± 2.3	90.9 ± 1.5
dice 5	95.3 ± 0.7	89.6 ± 2.8	81.1 ± 2.1	90.7 ± 1.3
dice 10	95.4 ± 0.7	89.8 ± 2.8	81.4 ± 2.1	90.9 ± 1.3
-dice 100	95.4 ± 0.7	89.7 ± 2.8	81.6 ± 2.2	90.9 ± 1.3
Images Reg				
aff dice 5	95.3 ± 0.7	89.8 ± 2.9	81.2 ± 2.1	90.7 ± 1.3
-bsp dice 5	95.4 ± 0.7	90.1 ± 3.0	81.8 ± 2.1	90.9 ± 1.3
Patches				
seg_3 10	95.5 ± 0.7	90.4 ± 2.8	82.1 ± 2.2	91.2 ± 1.4
seg_3 100	95.5 ± 0.8	90.1 ± 2.9	82.1 ± 2.2	91.1 ± 1.4
-seg_{96} 10	**95.5 ± 0.7**	**90.5 ± 2.8**	**82.2 ± 2.2**	**91.2 ± 1.4**
seg_{96} 100	95.5 ± 0.8	90.2 ± 2.9	82.1 ± 2.2	91.1 ± 1.4
Tissues				
seg_3 10	95.5 ± 0.7	90.3 ± 2.5	81.6 ± 2.2	91.1 ± 1.4
seg_3 100	95.5 ± 0.8	90.0 ± 2.7	81.7 ± 2.3	91.1 ± 1.4
-seg_{96} 10	95.5 ± 0.7	90.4 ± 2.6	81.7 ± 2.3	91.1 ± 1.4
seg_{96} 100	95.5 ± 0.8	90.1 ± 2.7	81.7 ± 2.3	91.1 ± 1.5
Best post	95.5 ± 0.7	92.2 ± 2.8	84.1 ± 2.4	92.2 ± 1.5
UB	97.1 ± 0.3	94.8 ± 1.4	86.9 ± 0.7	93.7 ± 0.3

Similarity metrics between DF-based methods predictions and follow-up scans using different settings for each method. In the table, the methods are referred to as Images, Images Reg, Patches, and Tissues for Similar Images, Similar Images with Registration, Similar Patches, and Similar Tissues methods, respectively. '-' indicates the best method of each family, and the overall best method is indicated in **bold**. Best post and UB refer to the best method with the recalculated segmentation and the Upper Bound, respectively. ↑ higher is better.

As shown in Table 4 and Figure 4, all the results improved with respect to the LB and exhibit p-values < 0.01. The results of the methods are very similar when varying their hyperparameters. Despite this similarity, the patch-based and tissue-based methods show slight improvements over the others, particularly in CSF and GM for the patch-based method with 96 tissues and $n = 10$, which led us to select it as the best DF-based method.

4.2.2. GAN-Based Results

The next experiments we performed were using the GAN-based family. For the MPGAN 3.4.1 and MPGAN + Segmentation Loss 3.4.2 methods, we performed inference on the whole volume by feeding the network with the baseline scans. For the MGAN

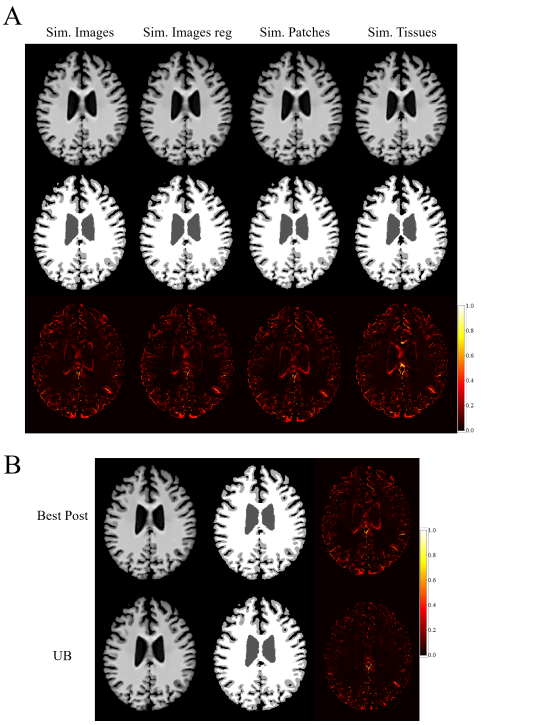

Figure 4: **DF-based Methods Predictions**. (A) Predictions of the DF-based methods using their best settings, including the segmentations of the three main tissues and the difference images with respect to the follow-up scans. (B) Prediction of the best method with the segmentation recalculated and prediction using ground truth deformation.

method 3.4.3, we extracted patches of size 64x64x64 with 50% overlap from the entire baseline scans, generating predictions for each patch. These patches were then assembled back together to create the complete volume. Similar to the method described in section 3.3.3, a spline-based method was used to handle the overlapping between patches and reduce the artifacts at the borders. Finally, for the StyleGAN method 3.4.4, since it required a style image to make the prediction, we selected the most similar baseline scan from our training set for each baseline scan in the test set and used its corresponding follow-up as the style. For each pair of images, we extracted slices from the sagittal plane and generated predictions for each slice. These slices were then stacked back together to reconstruct the complete volume.

After obtaining the predictions, we applied postprocessing to all the methods and then calculated the similarity results with the follow-up scans. These results are shown in Table 5, and the predicted images are displayed in Figure 5.

As shown in Table 5 and Figure 5, the predictions of most methods show results worse than the LB, with the MPGAN + Segmentation Loss method 3.4.2 being the only one that improved all metrics with a p-value < 0.01. Therefore, we selected it as the best GAN-based

Table 5: GAN-Based Methods Results - Similarities with Follow-up

Method	SSIM % ↑	Mean dice%		
		CSF ↑	GM ↑	WM ↑
MPGAN	92.1 ± 0.9	88.1 ± 3.1	72.2 ± 2.0	86.2 ± 1.6
MPGAN+seg	**94.9 ± 0.7**	**90.7 ± 3.2**	**81.4 ± 2.2**	**91.0 ± 1.3**
MGAN	92.6 ± 0.9	89.2 ± 2.9	72.7 ± 2.0	87.1 ± 1.5
StyleGAN	92.8 ± 0.8	*82.4 ± 7.9	75.2 ± 1.8	87.1 ± 1.1

Similarity metrics between GAN-based methods predictions and follow-up scans. The best method is indicated in **bold**, and values lower than the LB are underlined. * indicates p-values > 0.01. ↑ higher is better.

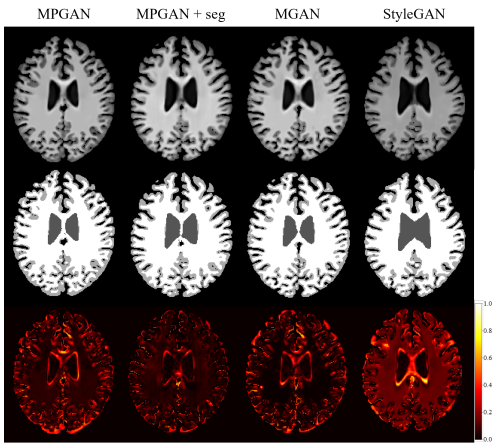

Figure 5: **GAN-based Methods Predictions.** Predicted T1w images, segmentations of the three main tissues, and difference images with respect to the follow-up scans.

method.

4.3. Best Methods Evaluation

Tissue Based Analysis — After selecting the best methods from each family, we conducted analyses based on cortical and subcortical structures to assess the ability of the predictions to capture subtle details. The structures selected for this analysis are presented in Figure 6. First, we assessed the volumetric changes of each structure using the VF to verify if the volumetric expansions or contractions were as expected. These results are shown in Table 6. Subsequently, we assessed the overlap and the volume differences between the structures from the predicted image and the actual follow-up using the Dice coefficient and ASPVC. These results are shown in Table 7 and 8.

Atrophy Analysis — We also performed an analysis based on the BPF to verify if the brain atrophy in the predicted images was similar to that in the actual follow-ups. During this analysis, we divided the test set into three groups with high, medium, and low BPF. This division allowed us to evaluate the predicted results for each group and verify the methods' performance. The results of this evaluation are presented in Figure 7.

Visual Results — Finally, we performed a visual inspection to determine if the computed metrics were consistent with the predicted scans. Figure 8 shows the predicted images of three individuals from each atrophy group. In this inspection, we also took into account the obtained segmentation highlighting the cortical and subcortical structures studied, as well as the difference image between the predictions and the actual follow-up.

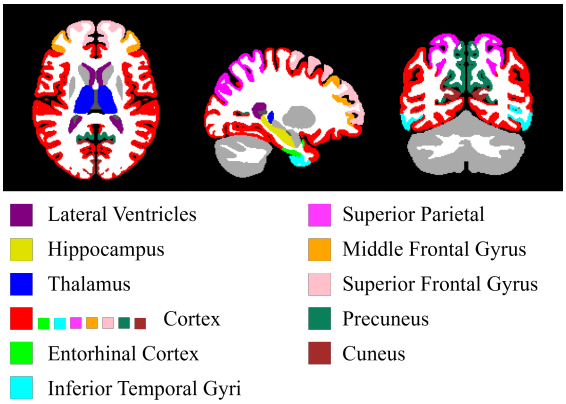

Figure 6: **Brain Structures Relevant to Brain Aging.** Eleven structures known to undergo marked changes with aging, used in this work to evaluate the accuracy of the predictions.

5. Discussion

Our research focused on determining the feasibility of predicting structural brain changes in healthy adults of around 60 years old over a nine-year period using 3D T1w MR images. We aimed to compare the accuracy of DF-based methods and GAN-based methods in predicting brain changes, evaluate their predictions in terms of image similarity, regional brain changes accuracy, and overall atrophy measured by the BPF, and assess their reliability in capturing the subtle and variable changes associated with healthy aging. As the results indicate, predicting brain changes during aging in a healthy population is indeed feasible, thereby answering our first research question. For almost all metrics, the best DF-based method outperformed the best GAN-based method. This suggests that DF-based methods remain superior for predicting longitudinal changes, as supported by our literature review.

5.1. Best Methods Comparison

Both visual and metrics results revealed that the DL and GAN methods effectively captured the volumetric changes of the ventricles. Similarly, the DF method accurately predicted changes (p-value < 0.01) in brain structures known to undergo marked changes in aging, particularly the thalamus and cortex (Choi et al., 2022; Fujita et al., 2023; Raz et al., 2005), as can be seen in

Table 6: Volume Fraction - Best methods Results

Metric	lat. vent. ↑	hippo.	thala.	cortex	ent. cortex	inf. temp.	sup. par.	mid. fron.	sup. fron.	precuneus	cuneus
Baseline	1.83±0.8	0.69±0.1	1.07±0.1	37.5±0.8	0.26±0.0	2.25±0.1	1.57±0.2	1.77±0.1	3.85±0.2	1.41±0.1	0.68±0.1
Follow-up	2.50±1.2	0.68±0.1	1.04±0.1	36.7±0.9	0.26±0.0	2.14±0.1	1.54±0.2	1.68±0.1	3.70±0.2	1.44±0.1	0.70±0.1
Best-DF	2.47±1.0	0.70±0.1	1.03±0.1	37.3±0.8	0.29±0.0	2.24±0.1	1.54±0.1	1.74±0.1	3.75±0.2	1.40±0.1	0.68±0.1
Best-GAN	2.69±1.1	0.66±0.1	0.98±0.1	38.6±0.8	0.28±0.0	2.41±0.1	1.59±0.2	1.77±0.1	3.98±0.2	1.48±0.1	0.68±0.1

Volume Fraction (VF) of the selected tissues for the baseline and follow-up scans and the predictions of the best methods. ↑ indicates that the volume should increase with respect to the baseline; if no arrow is present, the volumes are expected to decrease. Underlined values indicate that the volume change is not possible with respect to the baseline.

Table 7: Dice Coefficient % - Best Methods Results

Metric	lat. vent.	hippo.	thala.	cortex	ent. cortex	inf. temp.	sup. par.	mid. fron.	sup. fron.	precuneus	cuneus
LB	82.6±5.4	90.9±2.7	89.5±3.5	79.3±2.6	79.9±4.6	79.2±2.3	73.2±5.4	70.2±6.6	75.4±4.9	80.5±2.9	77.3±3.5
Best-DF	**91.3±3.2**	*__91.1±2.3__	**93.4±1.9**	**83.0±2.5**	*__81.3±3.9__	**82.6±2.3**	**75.9±5.4**	**78.4±4.3**	**82.4±7.0**	**81.6±2.8**	*__78.0±3.5__
Best-GAN	89.7±3.8	89.5±2.6	91.8±1.6	80.1±2.2	78.3±3.8	79.9±2.5	71.4±5.2	76.5±4.5	80.3±6.7	*78.5±2.9	*72.4±3.8

Dice coefficients of the selected tissues between the best methods and the follow-up scans. The initial Dice coefficient of the selected tissues between the baseline and follow-up scans is also shown as the Lower Bound (LB). The highest Dice coefficients between the best methods are indicated in **bold** (these values should also be higher than the LB). Values lower than the LB are underlined. * indicates p-values greater than 0.01. Note that higher Dice coefficients indicate better performance.

Table 8: Absolute Symmetrized Percent Volume Change - Best Methods Results

Metric	lat. vent.	hippo.	thala.	cortex	ent. cortex	inf. temp.	sup. par.	mid. fron.	sup. fron.	precuneus	cuneus
UB	29.5±12	2.68±2.4	3.60±2.8	2.41±1.12	4.60±4.7	5.15±2.2	3.61±2.7	5.29±2.7	4.12±2.0	2.45±1.8	3.38±2.5
Best-DF	**10.2±7.9**	3.28±2.8	**3.12±2.5**	**1.99±1.06**	10.4±6.1	**5.10±2.4**	*__2.99±2.1__	**3.87±2.4**	**1.79±1.7**	2.81±2.0	**3.23±2.5**
Best-GAN	13.6±9.9	3.29±2.2	*5.94±3.5	5.21±1.45	7.70±6.1	*__12.0±3.1__	*4.59±3.5	5.44±2.7	7.37±3.0	*3.54±2.6	*3.83±2.9

Absolute Symmetrized Percent Volume Change (ASPVC) of the selected tissues between the best methods and the follow-up scans. The initial ASPVC of the selected tissues between the baseline and follow-up scans is also shown as the Upper Bound (UB). The lowest ASPVC values are indicated in **bold** (these values should also be lower than the UB). Values lower than the UB are underlined. * indicates p-values greater than 0.01. Note that Lower ASPVC values indicate better performance.

Tables 6, 7 and 8, demonstrating its capability to predict subtle changes in brain structures undergoing volume loss during aging. However, for other critical brain structures in aging, such as the hippocampus, entorhinal cortex, and precuneus, the DF method was unable to predict volume changes. This discrepancy may be due to the small size of these structures compared to the previous two, making them more challenging to predict. Additionally, there may be some segmentation errors as we observed unrealistic increases in volume in the actual follow-up scan in the precuneus and cuneus (see Table 6). In contrast, the GAN method did not show consistent predictions for any of these regions across the three metrics used, indicating a lack of sensitivity for brain structures other than the ventricles.

In the BPF analysis, the average results indicated a decrease in BPF in predictions made with the DF method, suggesting that brain atrophy was captured. However, in the GAN method, BPF tended to remain the same or even increase, which is unlikely in the aging brain of healthy individuals over a nine-year period (Fujita et al., 2023). The analysis showed that in the group with low BPF, the GANs results deviate much more from real predictions than the prediction by DL. This indicated that the method is less sensitive in individuals with accelerated brain aging. In contrast, predictions using the DF method showed that it was robust for all three BDF groups. These results were expected in the case of DF-based methods because, if an individual has low BPF (i.e., marked brain atrophy), the DF methods apply changes based on individuals with similarly low BPF, as these would be the most similar, thereby maintaining this trend in the prediction. The same principle applies to individuals with other BPF levels. The limitations of GANs may stem from the network's bias towards subjects with medium BPF fractions. Figure 7 shows that the means of the different groups in the GAN method are close to each other compared to the DF method or baseline/follow-up scans. A solution for this problem could be adding a hyperparameter to the network indicating that the individual has a high, medium, or low BPF at baseline, forcing the network to maintain appropriate BPF levels in predictions. This strategy has already been implemented in some studies predicting brain changes in patients with Alzheimer's disease (Ravi et al., 2019; Xia et al., 2021).

5.2. DF-Based Methods Analysis

A main finding in this family of methods was the validation of our hypothesis, that it is possible to use brain changes from known individuals to predict brain changes in others. The proof of our hypothesis is primarily shown in Table 3, but it can also be seen in Table 4 and Figure 4. Comparing the results of the best post-processed method with the upper bound indicates that registering with a DF obtained from individuals with similar structures yields results close to registering with the ground truth deformations.

As observed in Table 4, variations in results by chang-

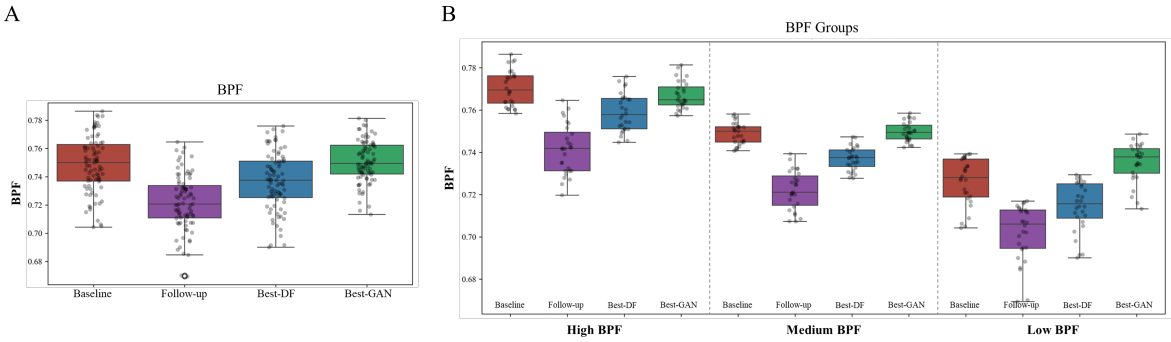

Figure 7: **Brain Parenchyma Fraction (BPF) - Best Methods Results.** (A) BPF of all individuals for the Baseline (Red), Follow-Up (Purple), Best DF-based Result (Blue), and Best GAN-based Result (Green). (B) BPF divided into groups by percentiles based on the BPF of the baseline: High BPF includes the 0-33 percentile (28 individuals), Medium BPF includes the 33-66 percentile (28 individuals), and Low BPF includes the 66-100 percentile (27 individuals).

ing hyperparameters were minimal, mainly affecting CSF and GM results slightly. However, the differences between local and non-local methods were much more pronounced, highlighting the potential of non-local methods to capture individual deformations and better adapt to the variability between individuals (Iglesias and Sabuncu, 2014), rather than calculating a global deformation average for the entire brain.

Despite the tissue-based method potentially being a more targeted approach for brain images, it did not yield better results than the patch-based method. This could be because the deformation field was obtained by non-overlapping tissues, leading to implausible deformations at the edges of each tissue due to abrupt changes that affects the inferred DF (Karacali and Davatzikos, 2006). A possible future solution could be to individually enlarge each tissue so they overlap and then calculate an average at their edges, which would avoid these abrupt deformations.

Another important point is that the B-spline method produced good visual results with low values for n. This suggests that performing a more exhaustive B-spline registration and slightly increasing n could yield even better metrics. However, this would come with a significantly higher computational cost compared to other methods due to the extra registrations.

5.3. GAN-Based Methods Analysis

For this family of methods, one of the most significant finding was that the segmentation layer in the MPGAN+seg method outperformed the results of the MPGAN and all other GAN methods (see Table 5). Moreover, this was the only GAN method that did not worsen the lower bound. This demonstrated that guiding GANs with tissue losses is an effective approach for improving the accuracy of predictions in brain changes (Zhang et al., 2018).

An unexpected result was that training with 3D patches using MGAN yielded slightly better results than training with the full volume using MPGAN. More notably, the use of 2D slices with StyleGAN achieved superior results in both GM and the global image metric compared to the previous two methods. These results could be due to several reasons, but it is likely that one of the main factors was that training with smaller inputs allowed the creation of deeper networks that captured more image features and made more detailed predictions (Brown et al., 2020)

As seen in the results provided by the StyleGAN network, this network tried to preserve the individuals' identity, but there were still some notable changes in the overall brain shape that do not usually happen in brain aging of healthy individuals (see Figure 5). These problems were not found in the other GAN methos that used baseline-to-follow-up training with longitudinal images, allowing better maintenance of the global structure and the individual's identity (Peng et al. (2021), Huang et al. (2022))

5.4. Limitations of the Best Method

Despite the promising results by the best DF method, it still had some limitations.

First, the volume changes are restricted to possible variations within the population, making it impossible to capture individuals with changes outside this range. This limits the ability to observe abrupt changes, as most individuals in our population exhibit smaller changes.

Another limitation is that the dataset deformations have a specific resolution of 193x229x193, making it impossible to apply this method to new images with different dimensions without rescaling, which can lead to loss of detail. This issue can potentially be addressed by creating multi-resolution deformation datasets or by using deep learning techniques to resize the images, thereby reducing the loss of information (Umirzakova et al., 2023).

Finally, as mentioned initially, there was an inability

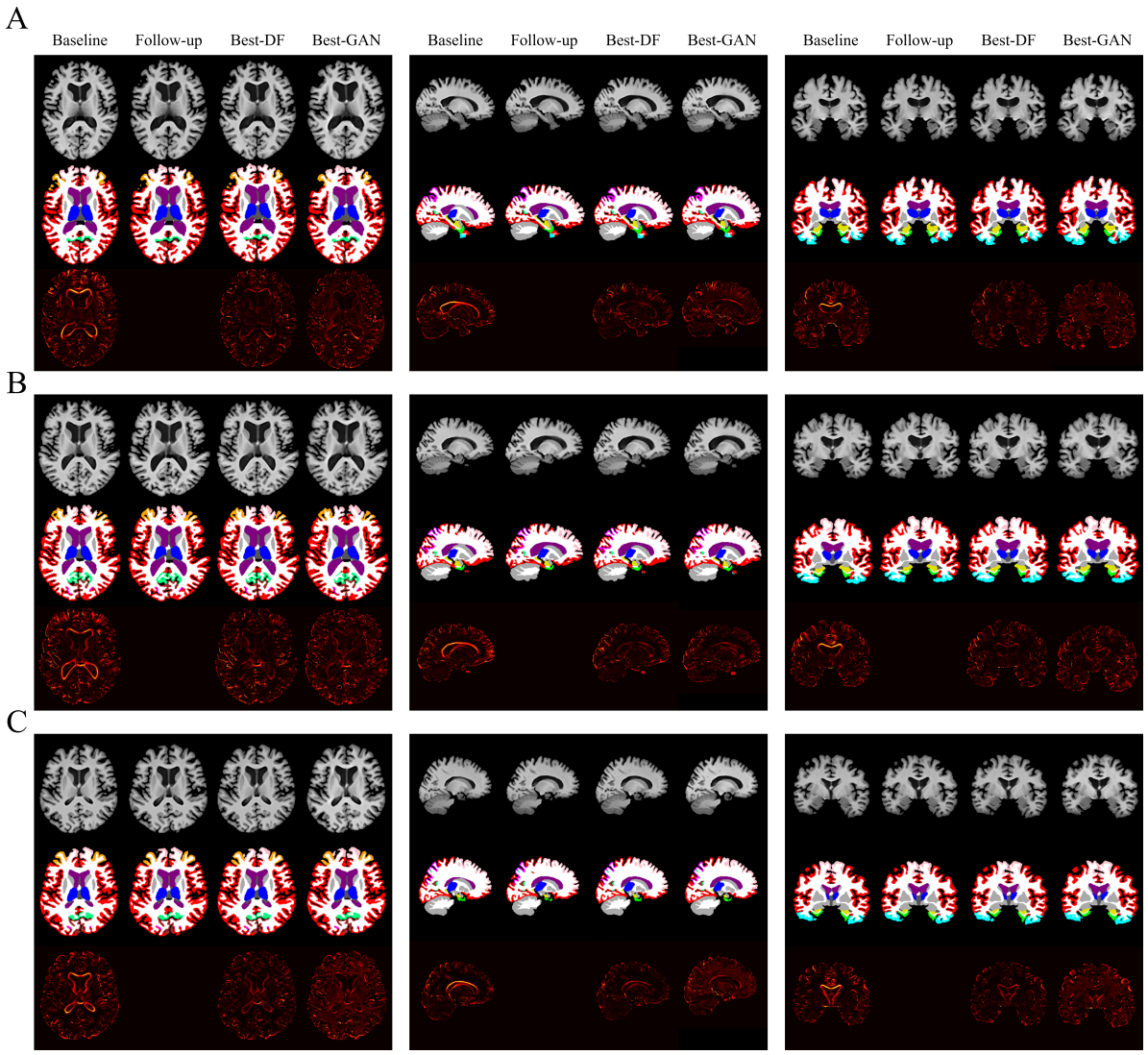

Figure 8: **Best Methods Predictions.** Baseline, Follow-Up, Best DF-based Prediction, and Best GAN-Based Prediction for three individuals. The axial plane is shown on the left, the sagittal plane in the center, and the coronal plane on the right. Each plane contains the prediction, the segmentation of the selected tissues (see Figure 6), and the difference with respect to the Follow-Up. (A) Individual with low BPF, (B) Individual with mean BPF, (C) Individual with high BPF.

to accurately predict changes in small brain regions such as the hippocampus, entorhinal cortex. This is a significant drawback, as these regions are crucial for the in-depth study of structural brain changes in aging (Fujita et al., 2023).

These three are the main limitations, although we know there may be others since this method has not been tested with images from other datasets.

5.5. Challenges of the Project

One of the main challenges of this project was attempting to predict the evolution of structural brain change with only two scans, assuming that the baseline scan had enough information in it to predict the follow-up. However, despite demonstrating that similar participants experienced similar brain aging, there were still specific changes in the brain of each participant that could only be calculated by having more time points between the baseline and the follow-up scans to measure the magnitude of changes for individuals in each brain region.

Another challenge was that the time between scans was quite long (around nine years). This causes much more variability between participants, as brain deformation is heavily affected by each individual's sociodemographic, health, genetics and lifestyle, and over nine years, many changes can occur (Mulugeta et al., 2022).

Another major challenge, was that most brain changes were quite subtle for most individuals. This led to very similar baseline and follow-up scans, making the visual evaluation of brain volume changes difficult.

5.6. Future Work

Future research could explore the integration of both strategies by introducing DF priors into GANs to guide volumetric changes. Additionally, incorporating diverse medical data from electronic health records or blood tests could further enhance the accuracy of these methods.

Enhancing GANs with segmentations that include a broader range of tissues, particularly those exhibiting significant changes during aging, could yield improved results. This could be accomplished by integrating an additional tissue segmentation network (Yu et al., 2022) and incorporating a loss function based on these tissues. However, this approach would necessitate substantially higher computational resources and result in slower training times.

Moreover, during this master's thesis, in collaboration with the computer science department, we experimented with a 2D diffusion model using autoencoders. The results were comparable to those obtained with the MGAN method but demonstrated greater stability during training. This suggests that future work focused on diffusion models holds significant promise.

6. Conclusions

This study investigated the prediction of structural brain changes in healthy adults over a nine-year period using 3D T1-weighted MRI images, comparing DF-based and GAN-based methods.

DF-based methods, based on the hypothesis that brain changes in some individuals can be used to predict changes in others individual from the same population, utilized multi-atlas techniques to combine volumetric changes from a subset of the population. Regional patch-based methods were the most effective.

We implemented four GAN methods based on recent work predicting brain structure changes in infants and patients with Alzheimer's disease, adapting them to our research questions. These methods aimed to train GANs to learn aging-related brain changes. However, most GAN methods were inaccurate in their predictions, with the exception of one model to which we added segmentation constraints.

Comparing the best methods from each family, DF-based methods outperformed GAN-based methods in nearly all metrics, capturing subtle changes in the thalamus and cortex. GAN methods predicted ventricular changes but lacked sensitivity for other structures. DF-based methods struggled with small regions like the hippocampus. DF-based methods were robust in predicting brain atrophy across varying BPF, while GAN methods were less accurate, especially for low BPF individuals.

This study provides a foundation for future research in brain change prediction, highlighting the effectiveness of DF-based methods and suggesting improvements for GAN methods. Future work could explore combining DF and GAN approaches, incorporating additional medical data, guiding GANs with more comprehensive segmentations, and exploring diffusion models.

Acknowledgments

I would like to express my gratitude to Professors Xavier Llado and Arnau Oliver for their support during this project and their valuable advice on implementing the methods. I also wish to thank Professor Gabriel Kiss for allowing me to use the High Performance Computing cluster IDUN for deep learning training and for providing guidance on this family of methods. Additionally, I am grateful to Karl Hofseth, a fellow master's student, for testing diffusion models as an alternative approach. Finally, I would also like to extend my thanks to Dr. Live Eikenes for granting me access to the electronic materials needed to use FreeSurfer.

References

Antipov, G., Baccouche, M., Dugelay, J.L., 2017. Face aging with conditional generative adversarial networks arXiv:1702.01983.

Arya, A., Verma, S., Chakarabarti, P., Chakrabarti, T., Elngar, A., Kamali, A.M., Nami, M., 2023. A systematic review on machine learning and deep learning techniques in the effective diagnosis of alzheimer's disease. Brain Informatics 10. doi:10.1186/s40708.023.00195.7.

Bandettini, P., 2012. Twenty years of functional mri: The science and the stories. NeuroImage 62, 575–88. doi:10.1016/j.neuroimage.2012.04.026.

Banerjee, S., Mittal, G., Joshi, A., Hegde, C., Memon, N., 2023. Identity-preserving aging of face images via latent diffusion models arXiv:2307.08585.

Bernal, J., Valverde, S., Kushibar, K., Cabezas, M., Oliver, A., Lladó, X., Alzheimer's Disease Neuroimaging Initiative, 2021. Generating longitudinal atrophy evaluation datasets on brain magnetic resonance images using convolutional neural networks and segmentation priors. Neuroinformatics 19, 477–492. doi:10.1007/s12021-020-09499-z.

Bethlehem, R.A., Seidlitz, J., White, S., Vogel, J., Anderson, K., Adamson, C., Adler-Wagstyl, S., Alexopoulos, G., Anagnostou, E., Areces Gonzalez, A., Astle, D., Auyeung, B., Ayub, M., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S., Benegal, V., Beyer, F., Alexander-Bloch, A., 2021. Brain charts for the human lifespan doi:10.1101/2021.06.08.447489.

Brezova, V., Moen, K.G., Skandsen, T., et al., 2014. Prospective longitudinal mri study of brain volumes and diffusion changes during the first year after moderate to severe traumatic brain injury. NeuroImage: Clinical 5, 128–140. doi:10.1016/j.nicl.2014.03.012. published 2014 Mar 28.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. arXiv:2005.14165.

Camara, O., Schweiger, M., Scahill, R., Crum, W., Sneller, B., Schnabel, J., Ridgway, G., Cash, D., Hill, D., Fox, N., 2006. Phenomenological model of diffuse global and regional atrophy using finite-element methods. IEEE Transactions on Medical Imaging 25, 1417–30. doi:10.1109/TMI.2006.880588.

Caruana, E., Roman, M., Hernández-Sánchez, J., Solli, P., 2015. Longitudinal studies. Journal of Thoracic Disease 7, E537–40. doi:10.3978/j.issn.2072-1439.2015.10.63.

Chen, X., Lathuilière, S., 2023. Face aging via diffusion-based editing arXiv:2309.11321.

Chen, Y., Almarzouqi, S.J., Morgan, M.L., Lee, A.G., 2018. T1-weighted image, in: Schmidt-Erfurth, U., Kohnen, T. (Eds.), Encyclopedia of Ophthalmology. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1747–1750. doi:10.1007/978.3.540.69000.9.1228.

Choi, E.Y., Tian, L., Su, J.H., Radovan, M.T., Tourdias, T., Tran, T.T., Trelle, A.N., Mormino, E., Wagner, A.D., Rutt, B.K., 2022. Thalamic nuclei atrophy at high and heterogenous rates during cognitively unimpaired human aging. NeuroImage 262, 119584. URL: https://www.sciencedirect.com/science/article/pii/S1053811922006991, doi:https://doi.org/10.1016/j.neuroimage.2022.119584.

Choi, Y., Uh, Y., Yoo, J., Ha, J.W., 2020. Stargan v2: Diverse image synthesis for multiple domains arXiv:1912.01865.

Coll, L., Pareto, D., Carbonell-Mirabent, P., Álvaro Cobo-Calvo, Arrambide, G., Ángela Vidal-Jordana, Comabella, M., Castilló, J., Rodríguez-Acevedo, B., Zabalza, A., Galán, I., Midaglia, L., Nos, C., Salerno, A., Auger, C., Alberich, M., Río, J., Sastre-Garriga, J., Oliver, A., Montalban, X., Àlex Rovira, Tintoré, M., Lladó, X., Tur, C., 2023. Deciphering multiple sclerosis disability with deep learning attention maps on clinical mri. NeuroImage: Clinical 38, 103376. doi:10.1016/j.nicl.2023.103376.

Crum, W., Hartkens, T., Hill, D., 2004. Non-rigid image registration: Theory and practice. The British Journal of Radiology 77 Spec No 2, S140–53. doi:10.1259/bjr/25329214.

Da Silva, M., Garcia, K., Sudre, C.H., Bass, C., Cardoso, M.J., Robinson, E., 2020. Biomechanical modelling of brain atrophy through deep learning arXiv:2012.07596.

Da Silva, M., Sudre, C.H., Garcia, K., Bass, C., Cardoso, M.J., Robinson, E.C., 2021. Distinguishing healthy ageing from dementia: a biomechanical simulation of brain atrophy using deep networks arXiv:2108.08214.

Duan, Y., Lin, Y., Rosen, D., Du, J., He, L., Wang, Y., 2020. Identifying morphological patterns of hippocampal atrophy in patients with mesial temporal lobe epilepsy and alzheimer disease. Frontiers in Neurology 11. doi:10.3389/fneur.2020.00021.

Fischl, B., 2012. Freesurfer. NeuroImage 62, 774–781. doi:10.1016/j.neuroimage.2012.01.021.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., 2011. Unbiased average age-appropriate atlases for pediatric studies. NeuroImage 54, 313–327. doi:10.1016/j.neuroimage.2010.07.033.

Fonov, V.S., Evans, A.C., McKinstry, R.C., Almli, C.R., Collins, D.L., 2009. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage 47, S102. doi:10.1016/S1053.8119(09)70884.5. organization for Human Brain Mapping 2009 Annual Meeting.

Fujita, S., Mori, S., Onda, K., Hanaoka, S., Nomura, Y., Nakao, T., Yoshikawa, T., Takao, H., Hayashi, N., Abe, O., 2023. Characterization of brain volume changes in aging individuals with normal cognition using serial magnetic resonance imaging. JAMA network open 6, e2318153. doi:10.1001/jamanetworkopen.2023.18153.

Gadewar, S., Zhu, A., Somu, S., Ramesh, A., Ba Gari, I., Thomopoulos, S., Thompson, P., Nir, T., Jahanshad, N., 2023. Normative aging for an individual's full brain mri using style gans to detect localized neurodegeneration, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023, pp. 387–395. doi:10.1007/978.3.031.45676.3.39.

Ge, Y., Grossman, R.I., Babb, J.S., Rabin, M.L., Mannon, L.J., Kolson, D.L., 2002. Age-related total gray matter and white matter changes in normal adult brain. part i: Volumetric mr imaging analysis. American Journal of Neuroradiology 23, 1327–1333.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks arXiv:1406.2661.

Habes, M., Janowitz, D., Erus, G., Toledo, J.B., Resnick, S.M., Doshi, J., Van der Auwera, S., Wittfeld, K., Hegenscheid, K., Hosten, N., Biffar, R., Homuth, G., Völzke, H., Grabe, H.J., Hoffmann, W., Davatzikos, C., 2016. Advanced brain aging: relationship with epidemiologic and genetic risk factors, and overlap with alzheimer disease atrophy patterns. Translational Psychiatry 6, e775. doi:10.1038/tp.2016.39.

Hansen, T., Brezova, V., Eikenes, L., Håberg, A., Vangberg, T., 2015. How does the accuracy of intracranial volume measurements affect normalized brain volumes? sample size estimates based on 966 subjects from the hunt mri cohort. AJNR. American journal of neuroradiology 36. doi:10.3174/ajnr.A4299.

Hedman, A.M., van Haren, N.E., Schnack, H.G., Kahn, R.S., Hulshoff Pol, H.E., 2012. Human brain changes across the life span: A review of 56 longitudinal magnetic resonance imaging studies. Human Brain Mapping 33, 1987–2002. doi:https://doi.org/10.1002/hbm.21334.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M., 2020. Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline. NeuroImage 219, 117012. doi:10.1016/j.neuroimage.2020.117012.

Hoopes, A., Mora, J.S., Dalca, A.V., Fischl, B., Hoffmann, M., 2022. Synthstrip: skull-stripping for any brain image. NeuroImage 260, 119474. doi:10.1016/j.neuroimage.2022.119474.

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization. arXiv:1703.06868.

Huang, Y., Ahmad, S., Han, L., Wang, S., Wu, Z., Lin, W., Li, G., Wang, L., Yap, P.T., 2022. Longitudinal prediction of postnatal brain magnetic resonance images via a metamorphic generative adversarial network arXiv:2208.04825.

Håberg, A.K., Hammer, T.A., Kvistad, K.A., et al., 2016. Incidental intracranial findings and their clinical impact; the hunt mri study in a general population of 1006 participants between 50-66 years. PLoS One 11, e0151080. doi:10.1371/journal.pone.0151080. published 2016 Mar 7.

Iglesias, J., Sabuncu, M., 2014. Multi-atlas segmentation of biomedical images: A survey. Medical Image Analysis 24. doi:10.1016/j.media.2015.06.012.

Iglesias, J.E., Billot, B., Balbastre, Y., Magdamo, C., Arnold, S.E., Das, S., Edlow, B.L., Alexander, D.C., Golland, P., Fischl, B., 2023. Synthsr: A public ai tool to turn heterogeneous clinical brain scans into high-resolution t1-weighted images for 3d morphometry. Science Advances 9, eadd3607. doi:10.1126/sciadv.add3607.

Iglesias, J.E., Billot, B., Balbastre, Y., Tabari, A., Conklin, J., González, R.G., Alexander, D.C., Golland, P., Edlow, B.L., Fischl, B., 2021. Joint super-resolution and synthesis of 1 mm isotropic mp-rage volumes from clinical mri exams with scans of different orientation, resolution and contrast. NeuroImage 237, 118206. doi:10.1016/j.neuroimage.2021.118206.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2018a. Image-to-image translation with conditional adversarial networks arXiv:1611.07004.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2018b. Image-to-image translation with conditional adversarial networks. arXiv:1611.07004.

Karacali, B., Davatzikos, C., 2006. Simulation of tissue atrophy using a topology preserving transformation model. IEEE Transactions on Medical Imaging 25, 649–652. doi:10.1109/TMI.2006.873221.

Karnewar, A., Wang, O., 2020. Msg-gan: Multi-scale gradients for generative adversarial networks. arXiv:1903.06048.

Kaye, J.A., DeCarli, C., Luxenberg, J.S., Rapoport, S.I., 1992. The significance of age-related enlargement of the cerebral ventricles in healthy men and women measured by quantitative computed x-ray tomography. Journal of the American Geriatrics Society 40, 225–231. doi:10.1111/j.1532-5415.1992.tb02073.x.

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. doi:10.3389/fnins.2017.00132.

Khanal, B., Lorenzi, M., Ayache, N., Pennec, X., 2016. A biophys-

ical model of brain deformation to simulate and analyse longitudinal mris of patients with alzheimer's disease. NeuroImage 134. doi:10.1016/j.neuroimage.2016.03.061.

Kingma, D.P., Ba, J., 2017. Adam: A method for stochastic optimization. arXiv:1412.6980.

Klein, S., Staring, M., Murphy, K., Viergever, M., Pluim, J., 2009. Elastix: A toolbox for intensity-based medical image registration. IEEE transactions on medical imaging 29, 196–205. doi:10.1109/TMI.2009.2035616.

Lee, J., Mustafaev, T., Nishikawa, R., 2023. Impact of gan artifacts for simulating mammograms on identifying mammographically occult cancer. Journal of Medical Imaging 10. doi:10.1117/1.JMI.10.5.054503.

Li, S., Lei, H., Zhou, F., Gardezi, J., Lei, B., 2019. Longitudinal and multi-modal data learning for parkinson's disease diagnosis via stacked sparse auto-encoder, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 384–387. doi:10.1109/ISBI.2019.8759385.

Modat, M., Simpson, I., Cardoso, M.J., Cash, D., Toussaint, N., Fox, N., Ourselin, S., 2014. Simulating neurodegeneration through longitudinal population analysis of structural and diffusion weighted mri data, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, pp. 57–64. doi:10.1007/978.3.319.10443.0.8.

Mulugeta, A., Navale, S.S., Lumsden, A.L., Llewellyn, D.J., Hyppönen, E., 2022. Healthy lifestyle, genetic risk and brain health: A gene-environment interaction study in the uk biobank. Nutrients 14. doi:10.3390/nu14193907.

Peng, L., Lin, L., Lin, Y., Chen, Y.w., Mo, Z., Vlasova, R.M., Kim, S.H., Evans, A.C., Dager, S.R., Estes, A.M., McKinstry, R.C., Botteron, K.N., Gerig, G., Schultz, R.T., Hazlett, H.C., Piven, J., Burrows, C.A., Grzadzinski, R.L., Girault, J.B., Shen, M.D., Styner, M.A., 2021. Longitudinal prediction of infant mr images with multi-contrast perceptual adversarial learning. Frontiers in Neuroscience 15. doi:10.3389/fnins.2021.653213.

Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., Galluzzi, S., Marizzoni, M., Frisoni, G.B., 2016. Brain atrophy in alzheimer's disease and aging. Ageing Research Reviews 30, 25–48. doi:10.1016/j.arr.2016.01.002. brain Imaging and Aging.

Pintzka, C.W., Hansen, T.I., Evensmoen, H.R., Håberg, A.K., 2015. Marked effects of intracranial volume correction methods on sex differences in neuroanatomical structures: a hunt mri study. Frontiers in Neuroscience 9. doi:10.3389/fnins.2015.00238.

Rachmadi, M., Valdés-Hernández, M., Makin, S., Wardlaw, J., Komura, T., 2019. Predicting the evolution of white matter hyperintensities in brain mri using generative adversarial networks and irregularity map, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019, pp. 146–154. doi:10.1007/978.3.030.32248.9.17.

Ravi, D., Alexander, D.C., Oxtoby, N.P., 2019. Degenerative adversarial neuroimage nets: Generating images that mimic disease progression arXiv:1907.02787.

Raz, N., Lindenberger, U., Rodrigue, K.M., Kennedy, K.M., Head, D., Williamson, A., Dahle, C., Gerstorf, D., Acker, J.D., 2005. Regional brain changes in aging healthy adults: General trends, individual differences and modifiers. Cerebral Cortex 15, 1676–1689. doi:10.1093/cercor/bhi044.

Reuter, M., Rosas, H.D., Fischl, B., 2010. Highly accurate inverse consistent registration: A robust approach. NeuroImage 53, 1181–1196. doi:10.1016/j.neuroimage.2010.07.020.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241.

Rudick, R.A., Fisher, E., Lee, J.C., Simon, J., Jacobs, L., 1999. Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting ms. multiple sclerosis collaborative research group. Neurology 53, 1698–1704. doi:10.1212/wnl.53.8.1698.

Schulz, M., Mayer, C., Schlemm, E., Frey, B., Malherbe, C., Petersen, M., Gallinat, J., Kühn, S., Fiehler, J., Hanning, U., Twerenbold, R., Gerloff, C., Cheng, B., Thomalla, G., 2022. Association of age and structural brain changes with functional connectivity and executive function in a middle-aged to older population-based cohort. Frontiers in Aging Neuroscience 14. doi:10.3389/fnagi.2022.782738.

Sharma, S., Noblet, V., Rousseau, F., Heitz, F., Rumbach, L., Armspach, J.P., 2010. Evaluation of brain atrophy estimation algorithms using simulated ground-truth data. Medical Image Analysis 14, 373–89. doi:10.1016/j.media.2010.02.002.

Smith, A., Crum, W., Hill, D., Thacker, N., Bromiley, P., 2003. Biomechanical simulation of atrophy in mr images. Proceedings of SPIE 5032, 481–490. doi:10.1117/12.480412.

Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics arXiv:1503.03585.

Umirzakova, S., Mardieva, S., Muksimova, S., Ahmad, S., Whangbo, T., 2023. Enhancing the super-resolution of medical images: Introducing the deep residual feature distillation channel attention network for optimized performance and efficiency. Bioengineering 10. URL: https://www.mdpi.com/2306-5354/10/11/1332, doi:10.3390/bioengineering10111332.

Vemuri, P., Murray, M.E., Jack, C.R., 2015. Chapter 10 - neuroimaging in dementias, in: Rosenberg, R.N., Pascual, J.M. (Eds.), Rosenberg's Molecular and Genetic Basis of Neurological and Psychiatric Disease (Fifth Edition). fifth edition ed.. Academic Press, Boston, pp. 107–118. doi:10.1016/B978.0.12.410529.4.00010.3.

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, 600–612. doi:10.1109/TIP.2003.819861.

Xia, T., Chartsias, A., Wang, C., Tsaftaris, S.A., 2021. Learning to synthesise the ageing brain without longitudinal data. Medical Image Analysis 73, 102169. doi:10.1016/j.media.2021.102169.

Yu, X., Tang, Y., Zhou, Y., Gao, R., Yang, Q., Lee, H.H., Li, T., Bao, S., Huo, Y., Xu, Z., Lasko, T.A., Abramson, R.G., Landman, B.A., 2022. Characterizing renal structures with 3d block aggregate transformers. arXiv:2203.02430.

Zapaishchykova, A., Tak, D., Ye, Z., Liu, K.X., Likitlersuang, J., Vajapeyam, S., Chopra, R.B., Seidlitz, J., Bethlehem, R.A.I., Mak, R.H., Mueller, S., Haas-Kogan, D.A., Poussaint, T.Y., Aerts, H.J.W.L., Kann, B.H., 2024. Diffusion deep learning for brain age prediction and longitudinal tracking in children through adulthood. Imaging Neuroscience 2, 1–14. doi:10.1162/imag.a.00114.

Zhang, Z., Yang, L., Zheng, Y., 2018. Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9242–9251. doi:10.1109/CVPR.2018.00963.

Zhou, Z., Sodha, V., Siddiquee, M.M.R., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019. Models genesis: Generic autodidactic models for 3d medical image analysis arXiv:1908.06912.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2020. Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv:1703.10593.

Åsvold, B., Langhammer, A., Rehn, T., Kjelvik, G., Grøntvedt, T., Sørgjerd, E., Fenstad, J., Heggland, J., Holmen, O., Stuifbergen, M., Aalberg Vikjord, S., Brumpton, B., Skjellegrind, H., Thingstad, P., Sund, E., Selbæk, G., Mork, P., Rangul, V., Hveem, K., Krokstad, S., 2022. Cohort profile update: The hunt study, norway. International Journal of Epidemiology 52. doi:10.1093/ije/dyac095.

A. T1w MR Images Details

Table 9: MRI Sequence Parameters

Dataset	Matrix size	NSA	TR (ms)	TE (ms)	Flip-angle	Slice thickness (mm)	Gap (mm)	Overlap (mm)	FOV (mm)
HUNT3	192x192	1	10.2	4.1	10°	1.2	0	0	240
HUNT4	256x256	-	7.7	3.092	8°	1.0	0	0	256

Parameters of the MRI sequence for HUNT3 and HUNT4 dataset, including matrix size, number of signal averages (NSA), repetition time (TR), echo time (TE), flip-angle, slice thickness, gap, overlap, and field of view (FOV).

B. Used Parameter for Non-Rigid Registration

Table 10: Parameters Used in the B-Spline Transformation

Parámetro	Valor
UseDirectionCosines	true
Registration	MultiMetricMultiResolutionRegistration
Interpolator	BSplineInterpolator
ResampleInterpolator	FinalBSplineInterpolator
Resampler	DefaultResampler
FixedImagePyramid	FixedRecursiveImagePyramid
MovingImagePyramid	MovingRecursiveImagePyramid
Optimizer	AdaptiveStochasticGradientDescent
Transform	BSplineTransform
Metric	AdvancedNormalizedCorrelation, TransformBendingEnergyPenalty
FinalGridSpacingInVoxels	4 4 4
NumberOfHistogramBins	32
Metric0Weight	1.0
Metric1Weight	0.1
NumberOfResolutions	2
ImagePyramidSchedule	1 1 1 1 1 1
MaximumNumberOfIterations	1000
MaximumStepLength	0.117188
NumberOfSpatialSamples	2048
ImageSampler	Random
BSplineInterpolationOrder	1
FinalBSplineInterpolationOrder	3

Most important parameters used in the B-Spline registration to create the DF dataset used in the DF-based family.