# A 3D generative model of pathological multi-modal MR images and segmentations

 $\begin{array}{c} \label{eq:2.1} \mbox{Virginia Fernandez}^{1[0000-0001-5984-197X]}, \mbox{Walter Hugo Lopez} \\ \mbox{Pinaya}^{1[0000-0003-3739-1087]}, \mbox{Pedro Borges}^{1[0000-0001-5357-1673]}, \mbox{Mark S.} \\ \mbox{Graham}^{1[0000-0001-6435-5079]}, \mbox{Tom Vercauteren}^{1[0000-0003-1794-0456]}, \mbox{ and M.} \\ \mbox{Jorge Cardoso}^{1[0000-0003-1284-2558]} \end{array}$ 

King's College London, London WC2R 2LS, UK

Abstract. Generative modelling and synthetic data can be a surrogate for real medical imaging datasets, whose scarcity and difficulty to share can be a nuisance when delivering accurate deep learning models for healthcare applications. In recent years, there has been an increased interest in using these models for data augmentation and synthetic data sharing, using architectures such as generative adversarial networks (GANs) or diffusion models (DMs). Nonetheless, the application of synthetic data to tasks such as 3D magnetic resonance imaging (MRI) segmentation remains limited due to the lack of labels associated with the generated images. Moreover, many of the proposed generative MRI models lack the ability to generate arbitrary modalities due to the absence of explicit contrast conditioning. These limitations prevent the user from adjusting the contrast and content of the images and obtaining more generalisable data for training task-specific models. In this work, we propose brainSPADE3D, a 3D generative model for brain MRI and associated segmentations, where the user can condition on specific pathological phenotypes and contrasts. The proposed joint imaging-segmentation generative model is shown to generate high-fidelity synthetic images and associated segmentations, with the ability to combine pathologies. We demonstrate how the model can alleviate issues with segmentation model performance when unexpected pathologies are present in the data.

## 1 Introduction

In the past decade, it has been shown that deep learning (DL) has the potential to ease the work of clinicians in tasks such as imaging segmentation [10], an otherwise time-consuming task that requires expertise in the imaging modality and anatomy. Nonetheless, the performance and generalisability of DL algorithms is linked to how extensive and unbiased the training dataset is [14]. While large image datasets in computer vision are widely available [9], this is not the case for medical imaging because images are harder to acquire and share, as they are subject to tight data regulations [23]. In addition, most state-of-the-art segmentation algorithms are supervised and require labels as well as images, and because obtaining these requires substantial time and expertise, they tend to

#### 2 Fernandez et al.

focus on a specific region of interest, making dataset harmonisation harder. In brain MRI, where studies are tailored to a pathology and population of interest, obtaining a large, annotated, multi-modal and multi-pathological dataset is challenging. An option to overcome this is to resort to domain randomisation methods such as SynthSeg [5], but their performance in the presence of highly variable pathologies such as tumours has not been tackled. Alternatively, data augmentation via deep generative modelling, an unsupervised DL branch that learns the input data distribution, has been applied in recent years to enrich existing medical datasets, producing realistic, usable synthetic data with the potential to complement [3] or even replace [11] real datasets, using architectures such as generative adversarial networks (GANs) and the more recent diffusion models (DMs) [26,21]. One of the major roadblocks, though, when it comes to applying synthetic data to segmentation tasks is that of producing labelled data. Conditioning can give the user some control over the generated phenotypes, such as age [21]. However, to our knowledge, only a handful of works deliver segmentations to accompany the data. In the case of published models generating data based on real labels [22], we must consider that labels are not usually shared, as they are sometimes considered protected health information, due to the risk of patient re-identification [27], and due to the above-mentioned difficulty to produce. Therefore, it may be beneficial that the labels themselves are also algorithmically generated from a stochastic process. Few works in the literature provide synthetic segmentations [4,12,11], especially enclosing healthy and multiple diseased regions. Among the latter, these models are limited to 2D, and they hardly allow the user to modulate their content (e.g., selecting the subject's pathology or age), limiting their applicability.

**Contributions:** in this work, we propose a 3D generative model of the brain that provides multi-modal brain MR images and corresponding semantic maps generated by giving the user the power to condition on the pathological phenotype of the synthetic subject. We showcase the benefits of using these synthetic datasets on a downstream white matter hyperintensity (WMH) segmentation task when the test dataset contains also contains tumours.

# 2 Methods

#### 2.1 Data

For training, we used the SABREv3 dataset consisting of 630 T1, FLAIR and T2 images [17], a subset of 66 T1 and FLAIR volumes from ADNI2 [16], and 103 T1, FLAIR and T2 volumes from a set of sites from BRATS [1,2,19]. Due to the large computational costs associated with training generative models, it is not tenable to train them using full resolution, full size 3D images: we circumvented this issue by, on one hand, mapping images to a 2mm isotropic MNI space, resulting in volumes of dimensions  $96 \times 128 \times 96$ , and on the other hand, by operating with patches, taken from 1mm data of dimensions  $146 \times 176 \times 112$ . Bronze-standard partial volume (PV) maps of the cerebrospinal fluid (CSF), grey matter (GM), white matter (WM), deep grey matter (DGM) and brainstem were

obtained using GIF [6], masking out tumours for BRATS. These healthy labels were overlaid with manual lesion labels provided with the datasets: WMH for the first two; and gadolinium-enhancing (GDE), non-enhancing (nGDE) tumour and edema for BRATS.

#### 2.2 Algorithm

Our pipeline consists of a conditional generator of semantic maps and an image generator, depicted in Fig. 1. It is based on the generative model proposed in [11]: a label generator, consisting of a latent diffusion model (LDM), is trained on the healthy tissue and lesion segmentations. Independently, a SPADE-like [20] network is trained on the PV maps and the multi-modal images. For the  $1mm^3$  data, patches of  $146 \times 176 \times 64$  had to be used for the image generator.

Label generator: The proposed label generator is based on a latent diffusion model (LDM), made up of a spatial variational auto-encoder (VAE) and a diffusion model (DM) operating in its latent space. The VAE of the  $2mm^3$ resolution model had 3 downsamplings, and the  $1mm^3$  had 4, resulting in latent spaces of shapes  $32 \times 24 \times 32$  and  $24 \times 24 \times 16$ , respectively. The VAE is trained using focal loss ( $\gamma = 3$ ), Kullback-Leibler distance (KLD) loss to stabilise the latent space, Patch-GAN adversarial loss and a perceptual loss based on the features of MED3D [7], implemented using MONAI [8]. For the diffusion model, we predict the velocity using the *v*-parametrization approach from [24] and optimise it via an  $l_2$  loss. We used T=1000 timesteps. We used a PNDM [18] scheduler to sample data, predicting only 150 timesteps. In addition, disease conditioning dc was applied using a cross-attention mechanism. For each subject j and disease type l, we have a label map  $M_{jl}$ , from which we produce a conditioning value  $dc_{il}$  reflecting the voxels labelled as l in the map, normalised by the maximum



**Fig. 1.** Architecture of our two-stage model: the left block corresponds to the label generator, and the right block to the image generator. Training and inference pathways are differentiated with black, red and dashed arrows.

4 Fernandez et al.

number of l voxels across the dataset, i.e.:

$$dc_{jl} = \frac{\sum_{n=1}^{N} M_{jl}}{\max_{j} \sum_{n=1}^{N} M_{jl}},$$
(1)

where the sum is across all voxels in the map. We trained the VAE for 250 epochs and the DM for 400, on an NVIDIA A100 DGX node.

**Image generator**: We modified the SPADE model used in [11] to extend the 2D generator and multi-scale discriminator to 3D. The encoder, which should only convert the contrast of the input image to a style vector, was kept as a 2D network, as it was found to work well while being parsimonious. To ensure that the most relevant brain regions informed the style, we used sagittal slices instead of axial ones as done in [11], selecting them randomly from the central 20 slices of each input volume. We kept the original losses from [20] to optimise the network. We replaced the network on which the perceptual loss is calculated with MED3D [7], as its features are also in 3D and fine-tuned on medical images, which are more domain-pertinent. We ran a full ablation study on the losses introduced in [11]; we dropped the modality-dataset discriminator loss, as it did not lead to major improvements but kept the slice consistency loss. To train the  $1mm^3$  model, we used random patches of size 64 along the axial dimension. During inference, we used a sliding-window approach with a 5-slice overlap. We trained the networks for 350 epochs on an NVIDIA A100 GPU. Further details are provided in the supplementary materials. Code is available at https: //github.com/virginiafdez/brainSPADE3D rel.git.

#### 2.3 Downstream segmentation task

To compare the performance in segmentation tasks of our synthetic datasets, we performed several experiments using nnUNetv2 [15] as a strong baseline architecture, adjusting only the number of epochs until convergence. The partial volume maps were converted to categorical labels via an *argmax* operator.

## 3 Experiments

#### 3.1 Quality of the generated images and labels pairs

Without established baselines for paired 3D healthy and pathological labels and image pairs, we assess our synthetic data by comparing them to real data and showing how they can be applied to downstream segmentation tasks. Examples of generated labels and images are depicted in Fig. 2. In one case, we used a conditioning unseen by the model during training, WMH + all tumour layers: both  $1mm^3$  and  $2mm^3$  label generators show the capability of handling this unseen combination, resulting in both lesions being present in the resulting images. However, we observed a lower ability of extrapolating to unseen phenotypes in

the  $1mm^3$  model, with only about 37% of the labels inferred using such conditioning resulting in the desired phenotype being met, as opposed to the  $2mm^3$ model which manifested both lesion types in 100% of the generated samples.

Quality of the labels: As the label generator is stochastic, we cannot compute paired similarity metrics. Instead, we compare the number  $V_{i,j}$ , for image *i* and region *j*, of CSF, GM, WM, DGM and brainstem voxels across subjects between a synthetic dataset of 500 volumes and a subset of the training dataset of the same size, excluding tumour images, but allowing for low WMH values as these don't disrupt the anatomy of the brain. The number of voxels  $V_{i,j}$  is calculated as:  $V_{i,j} = \sum_{n=1}^{N} (i_{n==j})$ , where N is the number of pixels in the image. Table 1 reports mean values and standard deviations, demonstrating that our model generates labels with mean volumes similar to real data. By comparing the labels, we saw that the considerable discrepancy between  $V_{i,CSF}$  at  $1mm^3$  is due to a loss of details in the subarachnoid CSF, which is very thin, likely due to the high number of VAE downsamplings at  $1mm^3$  (visual comparison is available in supplementary Fig. 1). However, we observe a lower standard deviation in the tissue volumes, indicating that the label generator does not capture the natural variability of brain tissues.



**Fig. 2.** Example synthetic  $1mm^3$  and  $2mm^3$  isotropic labels and images generated using tumour+WMH (left) and WMH (right) conditioning. The augmented frame in the top left images shows the small WMH lesions near the ventricles.

#### 6 Fernandez et al.

Table 1. Mean number of voxels and standard deviation of brain regions across real and synthetic datasets. Every value has been multiplied by  $10^{-4}$ .

Dataset	CSF	$\mathbf{G}\mathbf{M}$	WM	DGM	Brainstem
Real $(1mm^3)$	$20.342_{2.830}$	$37.634_{1.747}$	$44.872_{2.270}$	$4.301_{0.581}$	$1.238_{0.269}$
Synthetic $(1mm^3)$	$14.141_{0.964}$	$43.583_{0.735}$	$42.029_{1.728}$	$6.713_{0.592}$	$1.066_{0.102}$
Real $(2mm^3)$	$4.790_{0.573}$	$9.879_{0.585}$	$7.436_{0.512}$	$0.453_{0.080}$	$0.363_{0.036}$
Synthetic $(2mm^3)$	$4.641_{0.163}$	$8.564_{0.297}$	$7.008_{0.198}$	$0.587_{0.978}$	$0.343_{0.013}$

Quality of the generated images: To assess the performance of our image generator, we use a hold-out test set of PV maps and corresponding T1, FLAIR and T2 images, to compute the structural similarity index (SSIM) between the ground truth images and the image generated when the real PV map and a slice from the ground truth were used as inputs to the models. The mean SSIMs obtained are summarised in table 2, showing that the model performs similarly across different contrasts. Discrepancies between real and synthetic images can be explained by the stochasticity of the style encoder.

Synthetic data for healthy region segmentation: We assess the performance of our generated pairs on a CSF, GM, WM, DGM and brainstem segmentation task. We train an nnU-Net model,  $M_{healthy}$  on the T1 volumes of the real subset of 500 subjects mentioned earlier, and  $M'_{healthy}$  on the 500 synthetic T1 and label pairs, then test both models on a hold-out test set of 30 subjects from the SABRE. The Dice scores on all regions are reported in table 3. Although  $M_{healthy}$  performs better in all regions, the  $M'_{healthy}$  trained on purely synthetic data demonstrated a competitive performance for all regions except the DGM. DGM is a complex anatomical region comprising several small structures with intensities ranging between those typical for GM and WM. Thus, the PV map value for a voxel in this region will split its probability between DGM, WM and GM rather than favouring just one class, which is problematic for nnU-Net, as it requires categorical inputs to train the model, resulting in noisy ground truth labels that cause a larger distribution shift for the region. Examples of these noisy training and test labels are showcased in supplementary Fig. 2.

#### 3.2 WMH segmentation in the presence of tumour lesions

Aim: The main aim of this work is to show how synthetic data can increase the performance of segmentation models when training datasets are biased towards a specific phenotype. We focus on WMH segmentation. Our target dataset is a

**Table 2.** SSIM values obtained between real and synthetic images for T1, FLAIR and T2 contrasts for both models, generated using real PV maps.

$1mm^3$			$2mm^3$		
T1	FLAIR	T2	T1	FLAIR	Τ2
$0.842_{0.083}$	$0.798_{0.082}$	$0.794_{0.075}$	$0.922_{0.010}$	$0.910_{0.030}$	$0.909_{0.025}$

**Table 3.** Mean Dice score and standard deviation obtained for the models trained on real and synthetic data. Asterisks denote significantly better performance.

Resolution	Model	CSF	GM	WM	DGM	Brainstem
$1mm^3$	$M_{healthy}$	$0.957_{0.005}*$	$0.959_{0.003}*$	$0.971_{0.003}$ *	$0.875_{0.015}$ *	$0.958_{0.021}*$
$1mm^3$	$M_{healthy}'$	$0.884_{0.014}$	$0.912_{0.009}$	$0.936_{0.005}$	$0.684_{0.034}$	$0.874_{0.036}$
$2mm^3$	$M_{healthy}$	$0.947_{0.057}^{*}$	$0.958_{0.046}*$	$0.968_{0.039}*$	$0.887_{0.065}*$	$0.962_{0.024}*$
$2mm^3$	$M_{healthy}^{\prime}$	$0.869_{0.057}$	$0.895_{0.052}$	$0.931_{0.047}$	$0.703_{0.100}$	$0.905_{0.025}$

subset of 30 unseen volumes from BRATS (a different set of sites from those seen by the generative model) containing WMH lesions and tumours. We hypothesise that a WMH segmentation model trained on images that do not contain tumours will label these as WMH, as tumours and WMH share some intensity similarities in the FLAIR contrast typically used to segment WMH [25]. With the proposed generative model, we can generate synthetic data containing subjects with *both* tumours and WMH, which should make the training model robust to cases where both diseases are present, therefore reducing false positives.

We ran this experiment with  $2mm^3$  isotropic data, as the phenotype conditioning worked better (see 3.1) in this model. From a stack of 500 real FLAIR volumes from the SABRE dataset, and a stack of 500 FLAIR synthetic volumes generated from synthetic labels conditioned on both tumours and WMH, we train several models  $M_{R_{PR}S_{PS}}$ , varying the % proportions PR and PS of real and synthetic data respectively. In addition, even if the premise of this work is that users do not have access to real data containing tumours, we train model  $M_{R_{WMH}R_{tum}}$  on the real FLAIR volumes from the SABRE dataset, and the BRATS tumour volumes used to train the synthetic model, leaving the training labels empty, as no prior WMH segmentations are available for BRATS. We calculate the Dice score on WMH on a hold-out test set of 30 subjects from the SABRE dataset. As we do not have WMH ground truth labels for our test set from BRATS, but we have tumour labels, we compute the proportion of tumour pixels incorrectly labelled as WMH and note this metric  $FP_{tum}$ .

**Table 4.** Mean Dice score, precision and recall obtained on the SABRE test dataset by all the WMH segmentation models we trained, and  $FP_{tum}$  ratio on the BRATS holdout dataset. Asterisks indicate statistical significance.

Model	Dice (PD) $\uparrow$	precision (PD) $\uparrow$	recall (PD) $\uparrow$	$FP_{tum}$ ( <b>BRATS</b> ) $\downarrow$
$M_{R_{100}S_0}$	$0.728_{0.281}$ *	$0.761_{0.236}$	$0.751_{0.132}^{*}$	$0.325_{0.232}$
$M_{R_{75}S_{25}}$	$0.713_{0.208}$	$0.745_{0.231}$	$0.743_{0.137}$	$0.902_{0.187}$
$M_{R_{50}S_{50}}$	$0.716_{0.211}*$	$0.742_{0.227}$	$0.754_{0.132}$ *	$0.108_{0.209}$
$M_{R_{25}S_{75}}$	$0.722_{0.210}$	$0.755_{0.225}$	$0.722_{0.138}$	$0.079_{0.171}$
$M_{R_5S_{95}}$	$0.642_{0.210}$	$0.716_{0.243}$	$0.628_{0.146}$	$0.023_{0.078}$
$M_{R_0S_{100}}$	$0.362_{0.147}$	$0.726_{0.294}$	$0.263_{0.104}$	$0.001_{0.002}$ *
$M_{R_{WMH}R_{tum}}$	$0.710_{0.233}$	$0.742_{0.250}$	$0.741_{0.131}$	$0.026_{0.140}*$



Fig. 3. Sample WMH predictions on the BRATS dataset (top) and the SABRE test set (bottom) for all our models, in red. The leftmost column shows the tumour mask for the BRATS dataset (in blue) and the ground truth WMH for SABRE.

**Results**: Results are reported in table 4. Example segmentations and predicted WMH masks on both test sets are depicted in Fig. 3.  $M_{R_{100}S_0}$  achieved the top Dice on WMH for its in-domain test set, but it has one of the worse  $FP_{tum}$  scores on the BRATS set. While  $M_{R_0S_{100}}$  achieved a low Dice on the SABRE test set, it had a  $FP_{tum}$  score that is significantly lower than that of  $M_{tum-real}$ . All the models trained on a combination of real and synthetic data achieve a competitive  $FP_{tum}$  without compromising the WMH Dice. While all models have a comparable precision,  $M_{R_0S_{100}}$  has low recall; caused by an underestimation of WMH, as seen in Fig. 3. Interestingly, besides  $M_{R_0S_{100}}$  and  $M_{R_5S_{95}}$ , the edematous area of the tumours still gets partially segmented as WMH.  $M_{R_{WMH}R_{tum}}$  achieves very good Dice, precision, recall and  $FP_{tum}$  metrics; but, while examining the WMH segmentations on the BRATS dataset, all the segmentations were empty, as shown in Fig. 3, which indicates that, because the WMH training labels for BRATS were empty, the model has mapped the appearance and/or phenotype of the BRATS dataset to an absence of WMH.

## 4 Discussion & Conclusion

This work presents a label and multi-contrast brain MRI 3D image generator that can supplement real datasets in segmentation tasks for healthy tissues and pathologies. The synthetic data provided by our model can boost the precision and robustness of WMH segmentation models when tumours are present in the target dataset, showing the potential for having content and styledisentangled generative models that can combine the phenotypes seen in their training datasets. While disentanglement is covered in [11], 3D can help produce data usable in scenarios where the context of neighbouring slices is meaningful, such as segmenting small lesions. In addition, disease conditioning, which was not implemented in [11], can be challenging in 2D, as some diseases depend on the axial location, such as WMH. Our current set-up has, however, some limitations. First, there is a caveat in using  $2mm^3$  isotropic data or patching at  $1mm^3$ . Even so, diffusion models for high resolution 3D images have to operate in a latent space that causes loss of semantic variability (See 3.1) and small details, affecting the downstream segmentation task, partly because a gap appears in the image synthesis process between synthetic and real labels. Further work should attempt to harmonise the semantic synthetic and real domains. Although one of the causes of this limitation is capacity, the latest advances in diffusion models show that higher performance and resolution can be achieved [13], potentially leading to better labels and, effectively, less domain shift between real and synthetic domains. Secondly, conditioning on variables such as age or ventricle size could also translate into more variability across the generated volumes [21], overcoming the limitation in tissue variability observed in table 1. The method can be scaled to more pathologies and tasks, as model sharing allows for fine-tuning on more pathological labels, thus making segmentation models more generalisable to the diverse phenotypes of real brain MR data.

## References

- Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific data 4 (sep 2017). https://doi.org/10.1038/SDATA.2017.117, https:// pubmed.ncbi.nlm.nih.gov/28872634/
- Bakas, S., et al.: Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. Sandra Gonzlez-Vill 124 (nov 2018), https://arxiv.org/ abs/1811.02629v3
- Barile, B., Marzullo, A., Stamile, C., Durand-Dubief, F., Sappey-Marinier, D.: Data augmentation using generative adversarial neural networks on brain structural connectivity in multiple sclerosis. Computer methods and programs in biomedicine 206 (jul 2021). https://doi.org/10.1016/J.CMPB.2021.106113, https: //pubmed.ncbi.nlm.nih.gov/34004501/
- Basaran, B.D., Matthews, P.M., Bai, W.: New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation. Frontiers in neuroscience 16 (oct 2022). https://doi.org/10.3389/FNINS.2022.1007453, https: //pubmed.ncbi.nlm.nih.gov/36340756/
- Billot, B., Magdamo, C., Arnold, S.E., Das, S., Iglesias, J.E.: Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. Proceedings of the National Academy of Sciences 120(9), e2216399120 (sep 2022). https://doi.org/10.1073/PNAS.2216399120/SUPPL\_FILE/PNAS.2216399120.SAPP.PDF, http://arxiv.org/abs/2209.02032
- Cardoso, M.J., Modat, M., Wolz, R., Melbourne, A., Cash, D., Rueckert, D., Ourselin, S.: Geodesic information flows: spatially-variant graphs and their application to segmentation and fusion. IEEE transactions on medical imaging 34(9), 1976–1988 (2015)
- Chen, S., Ma, K., Zheng, Y.: MED3D: Transfer Learning for 3D Medical Image Analysis https://github.com/Tencent/MedicalNet.
- 8. Consortium, M.: MONAI: Medical Open Network for AI (mar 2020)

- 10 Fernandez et al.
- Deng, J., Dong, W., Socher, R., Li, L.J., Kai Li, Li Fei-Fei: ImageNet: A largescale hierarchical image database pp. 248–255 (mar 2010). https://doi.org/10. 1109/CVPR.2009.5206848
- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J.: A guide to deep learning in healthcare. Nature Medicine 2019 25:1 25(1), 24–29 (jan 2019). https://doi.org/10.1038/ s41591-018-0316-z, https://www.nature.com/articles/s41591-018-0316-z
- Fernandez, V., Pinaya, W.H.L., Borges, P., Tudosiu, P.D., Graham, M.S., Vercauteren, T., Cardoso, M.J.: Can Segmentation Models Be Trained with Fully Synthetically Generated Data? In: Zhao, C., Svoboda, D., Wolterink, J.M., Escobar, M. (eds.) Simulation and Synthesis in Medical Imaging. pp. 79–90. Springer International Publishing, Cham (2022)
- Foroozandeh, M., Eklund, A.: Synthesizing brain tumor images and annotations by combining progressive growing GAN and SPADE (sep 2020). https://doi.org/ 10.48550/arxiv.2009.05946, https://arxiv.org/abs/2009.05946v1
- 13. Hoogeboom, E., Heek, J., Salimans, T.: simple diffusion: End-to-end diffusion for high resolution images
- 14. Ian Goodfellow, Yoshua Bengio, A.C.: Deep Learning Book. Deep Learning (2015). https://doi.org/10.1016/B978-0-12-391420-0.09987-X
- Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18(2), 203–211 (2021). https://doi.org/10.1038/ s41592-020-01008-z, https://doi.org/10.1038/s41592-020-01008-z
- Jack, C.R., et al.: The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. Journal of magnetic resonance imaging : JMRI 27(4), 685–691 (apr 2008). https://doi.org/10.1002/JMRI.21049, https://pubmed.ncbi.nlm.nih. gov/18302232/
- 17. Jones, S., Tillin, T., Park, C., Williams, S., Rapala, A., Al Saikhan, L., Eastwood, S.V., Richards, M., Hughes, A.D., Chaturvedi, N.: Cohort Profile Update: Southall and Brent Revisited (SABRE) study: a UK population-based comparison of cardiovascular disease and diabetes in people of European, South Asian and African Caribbean heritage. International Journal of Epidemiology 49(5), 1441–1442e (oct 2020). https://doi.org/10.1093/ije/dyaa135, https://doi.org/10.1093/ije/dyaa135
- Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo Numerical Methods for Diffusion Models on Manifolds (feb 2022). https://doi.org/10.48550/arxiv.2202.09778, https: //arxiv.org/abs/2202.09778v2
- Menze, B.H., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE transactions on medical imaging 34(10), 1993–2024 (oct 2015). https://doi.org/10.1109/TMI.2014.2377694, https://pubmed.ncbi.nlm.nih.gov/25494501https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4833122/
- Park, T., et al.: Semantic image synthesis with spatially-adaptive normalization. Proceedings of IEEE CVPR 2019-June, 2332–2341 (2019)
- Pinaya, W.H.L., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain Imaging Generation with Latent Diffusion Models. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D., Yuan, Y. (eds.) Deep Generative Models. pp. 117–126. Springer Nature Switzerland, Cham (2022)
- Qasim, A.B., Ezhov, I., Shit, S., Schoppe, O., Paetzold, J.C., Sekuboyina, A., Kofler, F., Lipkova, J., Li, H., Menze, B.: Red-GAN: Attacking class imbalance via conditioned generation. Yet another medical imaging perspective. (sep 2020), https://proceedings.mlr.press/v121/qasim20a.html

- Rieke, N., et al.: The future of digital health with federated learning. npj Digital Medicine 3(1), 119 (2020)
- 24. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models
- Sudre, C.H., Cardoso, M.J., Bouvy, W.H., Biessels, G.J., Barnes, J., Ourselin, S.: Bayesian Model Selection for Pathological Neuroimaging Data Applied to White Matter Lesion Segmentation. IEEE Transactions on Medical Imaging 34(10), 2079– 2102 (2015). https://doi.org/10.1109/TMI.2015.2419072
- Tudosiu, P.D., Pinaya, W.H.L., Graham, M.S., Borges, P., Fernandez, V., Yang, D., Appleyard, J., Novati, G., Mehra, D., Vella, M., Nachev, P., Ourselin, S., Cardoso, J.: Morphology-Preserving Autoregressive 3D Generative Modelling of the Brain. In: Zhao, C., Svoboda, D., Wolterink, J.M., Escobar, M. (eds.) Simulation and Synthesis in Medical Imaging. pp. 66–78. Springer International Publishing, Cham (2022)
- 27. Wachinger, C., et al.: BrainPrint: a discriminative characterization of brain morphology. NeuroImage **109**, 232–248 (apr 2015)

H.						
$\sim$						
Table 1: Label and image generator hyperparameters						
	Variational	autoencoder				
Loss v	veights	Other parameters				
Ratch GAN loss	0.1	epochs	250			
KLD loss	10 <sup>-8</sup>	train time	4h38 min			
reconstruction loss	1.0	optimiser	Adam			
perceptual loss	10	learning rate	vae: $2 \cdot 10^{-4}$ disc:			
4			$10^{-4}$			
Chardware	NVIDIA DGX A100	batch size	8			
Latent diffusion model						
epochs	400	training time	12h			
warm-up learning	$10^{-8}$	batch size	8			
rate						
base learning rate	$2.5 \cdot 10^{-5}$	loss	$l_1$			
perceptual loss	10	hardware	NVIDIA DGX A100			
	Image g	enerator				
Epochs	350	training time	2 weeks			
Thardware	single A100 GPU	optimiser	Adam			
learning rate	$2 \cdot 10^{-4}$	number of	3			
		discriminator				
Loss weights						
Gene	erator	Discriminator				
KLD	$10^{-5}$	feature loss	0.25			
perceptual loss	1.5	gen. and disc. tr	aining thresholds			
slice consistency	0.5	lower (D only)	0.65			
perceptual loss	10	lower (G only)	0.75			

A Training hyperparameters and augmentations

8 Nov 2023

Augmentations were implemented using MONAI (https://monai.io/).

Table 2: Transformations	applied to the	different modules.	The intensity trans-
forms, marked with (*) w	vere only applied	d to images, not la	abels.

Augmentation	VAE (ranges)	DM (ranges)	Image generator (ranges)
Random affine	rotation: [-0.05, 0.05], shear: [0.001,	rotation: [-0.1, 0.1], shear: [0.001,	rotation: [-0.05, 0.05], shear:
	0.05], scale:[0, 0.05], probability:	0.15], scale:[0, 0.3], probability: 0.15	[0.001, 0.05],  scale:[0, 0.05],
	0.15		probability: 0.33
Random bias field	-	-	intensity: (0, 0.005), probabil-
(*)			ity: 0.33
Random Gaus-	-	-	mean: 0.0, $\sigma$ range: [0.005,
sian noise (*)			0.015], probability: 0.33
Random contrast	-	-	$\gamma$ range: [0.9, 1.15], probability:
adjust (*)			0.33

# **B** Additional result figures



Fig. 1: Example CSF real and synthetic label channels for both the  $1mm^3$  and  $2mm^3$  models, showing the loss of detail in the subarachnoid space CSF on the  $1mm^3$  model. All images are unpaired.



Fig. 2: From left to right: example real label from the SABRE dataset used to train model  $M_{healthy}$  (see section 3.1), synthetic healthy label generated by our model used to train  $M'_{healthy}$ , ground truth sample from the test set of the SABRE dataset and corresponding  $M_{healthy}$  and  $M'_{healthy}$  outputs. Examples shown are for the  $2mm^3$  model.