



King's Research Portal

DOI:

[10.1016/j.media.2024.103278](https://doi.org/10.1016/j.media.2024.103278)

Document Version

Publisher's PDF, also known as Version of record

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Fernandez, V., Pinaya, W. H. L., Borges, P., Graham, M. S., Tudosiu, P.-D., Vercauteren, T., & Cardoso, M. J. (2024). Generating multi-pathological and multi-modal images and labels for brain MRI. *Medical Image Analysis*, 97, Article 103278. <https://doi.org/10.1016/j.media.2024.103278>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Generating multi-pathological and multi-modal images and labels for brain MRI

Virginia Fernandez ^{*}, Walter Hugo Lopez Pinaya, Pedro Borges, Mark S. Graham, Petru-Daniel Tudosi, Tom Vercauteren, M. Jorge Cardoso

Department of Biomedical Engineering and Imaging Sciences, King's College London, Strand, London, WC2R 2LS, United Kingdom

ARTICLE INFO

MSC:

92C55

68T45

68T01

Keywords:

Generative modelling

Medical imaging segmentation

Brain MRI

ABSTRACT

The last few years have seen a boom in using generative models to augment real datasets, as synthetic data can effectively model real data distributions and provide privacy-preserving, shareable datasets that can be used to train deep learning models. However, most of these methods are 2D and provide synthetic datasets that come, at most, with categorical annotations. The generation of paired images and segmentation samples that can be used in downstream, supervised segmentation tasks remains fairly uncharted territory. This work proposes a two-stage generative model capable of producing 2D and 3D semantic label maps and corresponding multi-modal images. We use a latent diffusion model for label synthesis and a VAE-GAN for semantic image synthesis. Synthetic datasets provided by this model are shown to work in a wide variety of segmentation tasks, supporting small, real datasets or fully replacing them while maintaining good performance. We also demonstrate its ability to improve downstream performance on out-of-distribution data.

1. Introduction

Medical imaging segmentation involves the delineation of one or multiple regions of interest (ROIs) in a medical image via the assignation of an ROI class to relevant voxels in the image. The task, which requires radiological expertise, is traditionally performed manually and is not only time-consuming but can also give rise to inter- and intra-observer variability (Renard et al., 2020). Recent studies have shown that automated segmentation algorithms using Convolutional Neural Networks (CNN) can be a good surrogate for manual segmentation (Isensee et al., 2021), as they are fast and effective (Goodfellow et al., 2015).

1.1. Limitations with deep learning for image segmentation

Nonetheless, the application of CNN-based methods clinically encounters a significant limitation. If trained with scarce data, they can lead to overfitting on the training dataset, making them biased and ineffective towards unseen distributions (Goodfellow et al., 2015; Kaissis et al., 2020). Whereas large and comprehensive datasets are available in computer vision, this is not the case in medical imaging.

First, medical images are not as easy to acquire, all medical imaging modalities require costly equipment that needs expert manipulation. Secondly, medical images are Protected Health Information (PHI) and

are subject to strict data protection policies (ESR, 2017), which results in most imaging datasets remaining private, without the possibility of being shared across institutions. Thirdly, available public datasets usually comply with a specific clinical criteria, excluding subjects who do not fulfil certain criteria, resulting in biases (Ricci Lara et al., 2022). Lastly, because most state-of-the-art segmentation methods are supervised (Anwar et al., 2018), manually drawn segmentations are required for training. As stated above, these are hard to obtain, typically resulting in available segmentations being constrained to a single task, often resulting in overly specific, small, and biased datasets.

1.2. What approaches can we use to overcome the shortfall of medical imaging datasets?

One of the approaches is to make the segmentation models robust to incomplete, small or biased datasets. Approaches using multi-task learning to tackle incomplete datasets (Dorent et al., 2021), domain adaptation (Guan and Liu, 2022) or randomisation (Billot et al., 2023) can be used to improve gaps between data distributions. However, most of these approaches tend to produce models tailored to specific tasks. A more generic strategy is to use deep learning-based data augmentation, which increases the size of training datasets along relevant directions of variability (Chlap et al., 2021). Traditional data augmentation is

^{*} Corresponding author.

E-mail address: virginia.fernandez@kcl.ac.uk (V. Fernandez).

<https://doi.org/10.1016/j.media.2024.103278>

Received 20 November 2023; Received in revised form 4 July 2024; Accepted 15 July 2024

Available online 18 July 2024

1361-8415/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

defined by the user and typically consists of an aggregated and randomised set of simple mathematical transformations on the images. Still, it does not necessarily model real variations in the input data distributions and is limited by the user's knowledge about the latter.

Generative modelling is a class of unsupervised learning algorithms that models an input data distribution and does not require the user to have prior knowledge about its characteristics (Goodfellow et al., 2015). Most state-of-the-art generative models are deep learning-based, such as generative adversarial networks (GANs), first introduced in Goodfellow et al. (2014), variational auto-encoders (VAEs) (Kingma and Welling, 2014), or the state-of-the-art diffusion models (Rombach et al., 2021). One main advantage of using deep learning-based generative models is that their latent representation of the data distribution typically allows for stochastic sampling, resulting in a source of infinite synthetic data. However, most of the state-of-the-art studies in medical imaging focus on only generating images, which makes them unsuitable for training supervised methods such as state-of-the-art nnU-Net (Isensee et al., 2021), due to the absence of corresponding labels also being generated.

In magnetic resonance imaging (MRI) of the brain, where acquired contrasts are chosen to highlight the specific physiological and/or pathological traits required by the study being conducted, there is an argument to be made in favour of generative models that are capable of providing physiological and pathological multi-modal images and associated labels, as they will have the potential to augment datasets for a wide range of brain segmentation tasks.

2. Related works

Recent works in the field of computer vision have shown that synthetic data generated using a generative model can improve the performance of downstream classification tasks (Azizi et al., 2023), including in the medical imaging domain (Motamed et al., 2021; Khader et al., 2023). Using different base architectures, including VAEs, GANs, Vector-Quantised VAEs (van den Oord et al., 2017) or the state-of-the-art diffusion models (Ho et al., 2020), numerous works have been published on the topic of data augmentation via generative modelling for medical imaging (Kazerouni et al., 2023), even leading to open-source libraries and databases such as MEDIGAN (Osuala et al., 2022) or 'MONAI Generative Models' (Pinaya et al., 2023).

Despite this notable progress, some challenges remain, especially when generating brain MRI images for segmentation. First, brain MRI images are volumetric, unlike most natural images used in computer vision. State-of-the-art brain segmentation algorithms are 3D (Isensee et al., 2021), but the availability of 3D generative models is limited due to the large computational resources required to train them. To this date, relatively few examples in the literature have deployed 3D generative models for brain MRI, such as Tudosi et al. (2022), Sun et al. (2022), Khader et al. (2023) or Pinaya et al. (2022), even though these examples remain constrained to healthy data or single pathologies.

The second challenge is that of conditioning. Many published works do not offer the user any tool to constrain the nature of generated images, leading to models limited to a modality or pathological profile (Sun et al., 2022). The publication of text encoders and their coupling to diffusion models has enabled high-quality text-based conditioning (Chambon et al., 2022), which has proven useful for X-ray imaging. Diffusion models have also been coupled to age or ventricular size conditioning (Pinaya et al., 2022). More appealing in the context of using synthetic images for segmentation is the existence of semantic synthesis models, such as modified diffusion models which accept masks (Dorjsembe et al., 2023), or SPADE normalisation (Park et al., 2019), that allows for conditioning on a semantic map. SPADE has been applied in various medical imaging 2D applications (Qasim et al., 2020; Fernandez et al., 2022; Stojanovski et al., 2023), resulting in synthetic images paired to segmentations that can be used to supplement or

replace real datasets during the training of supervised segmentation algorithms. These methods are, nonetheless, limited to 2D.

One additional problem is the source of labels. Brain segmentation models are usually tailored to the dataset task (e.g. tumour segmentation) and rarely comprehend physiological and multi-pathological information. This narrow scope limits the applicability of SPADE-based methods, which require a rich semantic coverage of the images. In our previous work (Fernandez et al., 2022), we developed a generative model, brainSPADE, for paired 2D multi-modal brain MRI images and corresponding labels, by coupling a diffusion model-based label generator to a modified SPADE network. In that work, we showed the potential of applying healthy and diseased synthetic images to segmentation tasks, improving segmentation performance when real datasets are small, or addressing out-of-distribution issues. brainSPADE was, however, restricted to 2D images; and a different label generator had to be trained to allow for labels with specific pathological profiles.

3. Contributions

In this work, we extend brainSPADE (Fernandez et al., 2022) by making the following additions:

- In addition to re-implementing the 2D approach, with 2D label and image generators, we extend our method to 3D by implementing a 3D label generator, resulting in 3D labels that can be forwarded, slice by slice, to the 2D image generator, then assembled into volumetric images.¹ We minimise the impact of generating each image slice independently by incorporating losses that increase the consistency of the contrast of images corresponding to different slices, and run an ablation study on these losses and additions.
- We implement conditioning mechanisms in the 2D and 3D label generators to allow for control over which pathologies are present within the synthetic labels.
- Apply generated 2D and 3D datasets to a wide range of segmentation tasks, including out-of-distribution segmentation and pathology segmentation, showing that synthetic data from our model can be used to train segmentation networks that generalise well to real data.

4. Materials and methods

4.1. Data

To incorporate multiple pathologies and contrasts into our model training, we used a wide range of brain MRI datasets to train the label and image generators and to run the validation experiments. Dataset sizes were chosen to test for model resilience to scales of data:

- **Southall and Brent Revisited V3 (SABRE v3)** (Jones et al., 2020).
- **Alzheimer's Disease Neuroimaging Initiative 2 (ADNI 2)**² We used an internal pre-processed subset.
- **Brain Segmentation Tumour Challenge (BraTS) 2020**: we split BraTS into subjects from the Cancer Imaging Archive ("BraTS-TCIA") site and the rest of the sites ("BraTS-OTHER").

¹ Note that, even though the resulting datasets are 3D, the image generator employs a 2D model.

² The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Table 1

Summary of data used for this work. Abbreviations: LG: label generator, IG: image generator, EXP: segmentation experiments; GD-E: gadolinium-enhanced; NE: non-enhanced tumour.

Dataset	Split	n° subjects	Modalities	Pathologies	Used in...
SABRE	Train	583	T1, FLAIR, T2	WMH	LG, IG, EXP
	Test	54	T1, FLAIR, T2	WMH	EXP
ADNI-2	–	98	T1, FLAIR	WMH	IG
ABIDE	–	29	T1	–	EXP
OASIS	–	71	FLAIR	WMH	EXP
BraTS	BraTS-TCIA	129	T1, FLAIR, T2	NE tumour, GD-E tumour and oedema	LG, IG, EXP
	BraTS-OTHER	31	T1, FLAIR, T2	NE tumour, GD-E tumour and oedema	EXP

- **Open Access Series of Imaging Studies (OASIS)** (LaMontagne et al., 2019).
- **Autism Brain Imaging Data Exchange (ABIDE)** (Di Martino et al., 2014): we used the subset of subjects from the Yale site.

These datasets were registered via an affine transform to the ICBM T1 1 mm isotropic template using ANTsPy (<https://github.com/ANTsX/ANTsPy>), resulting in $(193 \times 229 \times 193, 1 \text{ mm}^3)$ volumes. T1 images were used for the registration, and the resulting transform matrix was applied to the other modalities and segmentations.

Table 1 summarises the characteristics of the datasets used in our experiments. We used GIF (Cardoso et al., 2012) to obtain generated probabilistic labels for cerebrospinal fluid (CSF), grey matter (GM), white matter (WM), deep grey matter (DGM) and brainstem. Lesion segmentations were provided with the datasets and fused with the above-mentioned healthy tissue maps. Each healthy tissue being in a different channel, we concatenated the lesion masks on additional channels, setting the values of healthy tissues to zero for voxels where the lesion was 1.

The 2D models we used were trained on slices taken from these 3D datasets. The slicing process is detailed in the appendix (Appendix B.1).

4.2. Model and training pipeline

Our model, brainSPADE, generates brain semantic label maps, including healthy and pathological regions, and corresponding multi-modal MR images. Its 2D version, brainSPADE2D was first introduced in Fernandez et al. (2022), and further modified in this work by adding conditioning on the label generator and modifying the losses in which both label and image generators were optimised. It was originally implemented as a fully 2D model comprising a label and an image generator. In this work, we extended the label generator to 3D, coupling it to a 2D image generator and resulting in 3D labels and images. We name this version “brainSPADE3D”. A diagram illustrates how brainSPADE works in Fig. 1. The architecture consists of a two-stage approach: first, a semantic label map generator based on a latent diffusion model (LDM), made up of a spatial variational autoencoder and a diffusion model operating in its latent space; second, a semantic VAE-GAN producing a realistic-looking brain image corresponding to the input semantic map and following the input contrast.

4.2.1. Label generator

Unlike natural images, segmentations lack textural information, making them challenging for standard generative models such as GANs, as they are prone to instability (Goodfellow et al., 2014). We used state-of-the-art latent diffusion models (LDM) to overcome these limitations. Diffusion models (DMs) sample from a Gaussian noise distribution and denoise the sample via a Markov Chain process. Recent work has shown the generative potential of these models, as well as their training stability (Rombach et al., 2021). However, their major drawback is that they operate in the image space, making the training process slow and computationally greedy. LDMs are diffusion models that operate in a latent space: they are coupled to a spatial variational auto-encoder (S-VAE) that downsamples inputs into a compressed space. LDMs translate

Gaussian noise samples into meaningful latent space samples that the auto-encoder can decode.

Our label generator is entirely based on the one from Fernandez et al. (2022), which is, in turn, based on Rombach et al. (2021). The code for the 2D label generator model was built from that of our previous work, while the one from the 3D label generator was built using MONAI Generative Models (Pinaya et al., 2023). Both follow the architecture depicted in appendix Figs. A.15 and A.16.

S-VAE: For brainSPADE2D, we trained a spatial VAE (S-VAE) with two downsamplings and a latent space of dimension (48, 64, 3). For brainSPADE3D, we trained a spatial VAE with three downsamplings and a latent space of dimension (20, 22, 14, 8). Note that more downsamplings were required in the case of brainSPADE3D because of computational constraints. Both 2D and 3D VAEs were optimised with the following loss:

$$L_{VAE} = \lambda_{KLD} \mathcal{L}_{KLD} \left(E_{lg}(I) \parallel \mathcal{N}(0, \mathbf{I}) \right) + \lambda_{perc} L_{perc}(l, \hat{l}) + \lambda_{adv} L_{adv} \left(D(I), D(\hat{l}) \right) + \lambda_{f_l} L_{f_l}(l, \hat{l}) + \lambda_{nr} L_{nr}(l, \hat{l}) \quad (1)$$

where l is the input label map, $\hat{l}$ is the reconstructed segmentation, $\mathcal{L}_{KLD}$ is the Kullback–Leibler divergence, E_{lg} is the encoder of the VAE, L_{perc} a perceptual loss, and L_{adv} a patch-GAN adversarial loss, obtained by forwarding synthetic and real segmentations through a Patch-GAN discriminator D ; L_{f_l} is a focal reconstruction loss ($\gamma = 0.3$) (Lin et al., 2018); L_{nr} is a loss that minimises the focal loss between the ground truth label and the label decoded from a latent vector to which random Gaussian noise has been added (we selected $\mu = 0; \sigma = 0.05$, as we found that, empirically, these values led to realistic noise levels). The inclusion of perceptual and adversarial losses in the training has shown a significant quality improvement in the reconstructions (Rombach et al., 2021). The network backbones used for the perceptual loss were set to VGG19 for the 2D model and MED3D (Chen et al., 2019) for the 3D model. Loss weights λ were tuned empirically and are reported in the supplementary materials (Appendix B.2).

DM: The 2D and 3D DM models were based on the same time-conditioned U-Net from Rombach et al. (2021). Following the objective function from Ho et al. (2020), we used a fixed noise variance between each of the 1000 timesteps and a reparametrised approach that predicts the noise added between time steps. Given a timestep t , the loss used is:

$$L_{dm} = L_2 \left(D \left(\sqrt{\bar{\alpha}_t} E(l) + \sqrt{1 - \bar{\alpha}_t} \epsilon \right), \epsilon \right) \quad (2)$$

where L_2 is an L2 loss, D corresponds to applying the diffusion model, ϵ is a sample from a normal distribution $\mathcal{N}(0, 1)$ of mean 0 and variance 1, $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ is the cumulative variance product until time step t , with $\alpha_t = 1 - \beta_t$, β_t being the variance at time step t , and $E(l)$ is the latent representation of label l .

In both the 2D and 3D models, we used cross-attention conditioning on the disease proportion (WMH, non-enhancing (NE) tumour, oedema and gadolinium-enhancing (GE) tumour). For each lesion type c and image i , the lesion proportion $P_{c,i}$ was obtained by computing the

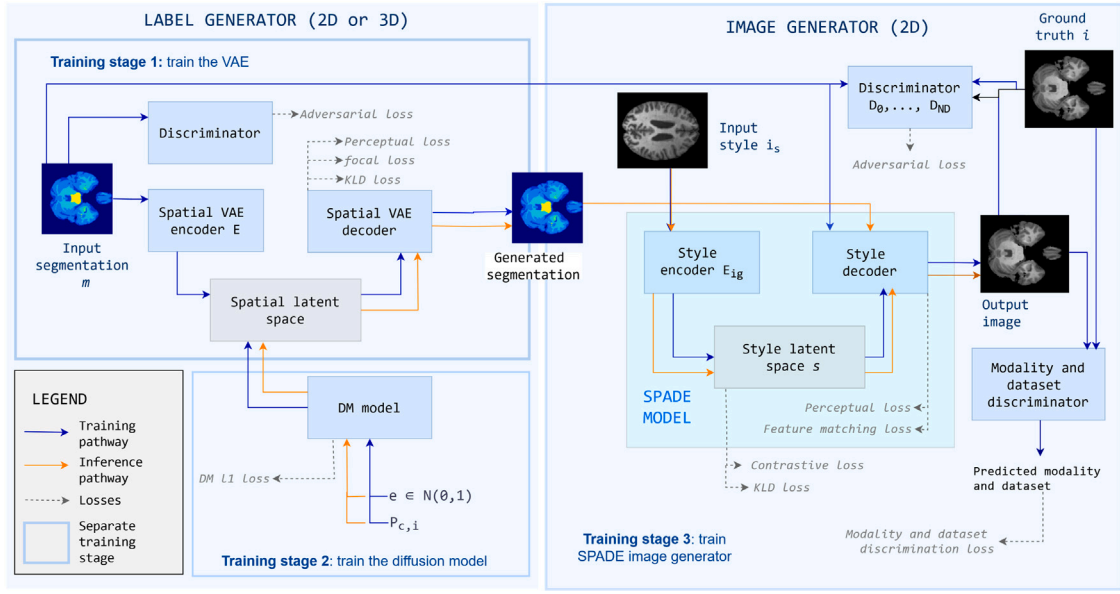


Fig. 1. Architecture of brainSPADE. Different training stages are enclosed in blue First, we train the label generator: in the first stage, we train the VAE to reconstruct the semantic maps; then, we train the diffusion model to generate latent representations of the VAE, conditioned by the pathology proportion $P_{c,i}$. Last, the SPADA image generator is trained on unpaired real labels and multi-modal images. At inference, the user specifies $P_{c,i}$ to produce a label map (2D or 3D, depending on the approach). This map is passed to the image generator (slice by slice in the case of 3D maps) along with a 2D style image, to produce a resulting 2D image (or a stack of 2D images that can be assembled together by concatenation). Blue arrows show the training pathway and orange ones show the inference pathway. Grey dotted arrows point to the training losses.

number of pixels/voxels labelled as disease in the map, normalised by the maximum across the dataset, i.e.

$$P_{c,i} = \frac{M_{c,i}}{\max_j(M_{c,j})}, \quad (3)$$

where $M_{c,i}$ is the sum of voxels in image i corresponding to lesion c and $\max_j$ is the maximum of this number across the dataset. In the case of the 2D model, we added an additional conditioning dimension to account for the axial slice position of our 2D semantic maps. The “relative slice number” is obtained by dividing the axial slice index (obtained during the slicing process, see Appendix B.1 for further details) by the maximum dimension across it (256 in the case of our MNI-registered images).

Training procedure: The 2D and 3D LDM models were trained on a set of 12 280 slices and 712 volumes, respectively, from SABRE and BraTS-TCIA (see Table 1). To strengthen the ability of our model to produce multi-pathological data, we selected a random set of 50 BraTS subjects and superimposed their tumour masks on the label maps of 50 randomly selected SABRE subjects, resulting in 50 new label maps containing both lesions. These labels were added to the 3D LDM training set. Even though BraTS and SABRE were registered to the same space, this did not always result in visually plausible images for all cases, but we hypothesised that these few samples would still help the model yield more consistent conditioning. To verify if the model could still extrapolate and produce unseen disease combinations without these extra labels, we trained an LDM with and without these 50 volumes.

4.2.2. Image generator

Original framework: The image generator is based on SPADE (Park et al., 2019). SPADE is a semantic synthesis VAE-GAN where the style and content pathways are separate. The encoder branch E_{ig} takes in an input style image i and encodes it into a compressed latent space representation s , while the decoder combines this latent style s and an input semantic map m driving the content of the output $\tilde{i}$. m is fed to the decoder via special normalisation blocks called SPADE blocks. Park et al. (2019) coupled it to a multi-scale patch-based discriminator $D = \{D_0, \dots, D_{N_D}\}$, where N_D is the number of individual Patch-GAN discriminators, each focusing on a different scale of the input. The original

loss used to train SPADE is an addition of (1) an adversarial (GAN) loss, $\mathcal{L}_{gan} = \sum_{j=1}^{N_D} HL(D_j(i), D_j(\tilde{i}))$, where HL is the Hinge loss; (2) the VAE Kullback-Leibler divergence (KLD) loss $\mathcal{L}_{KLD} = KLD(s \parallel \mathcal{N}(0,1))$; (3) a perceptual loss $\mathcal{L}_{perc} = \sum_{j=1}^M L_1(VGG_j(i), VGG_j(\tilde{i}))$, where VGG_j are M internal representations of a VGG19 network (Johnson et al., 2016) for the ground truth image i and the synthetic one $\tilde{i}$; (4) and a feature loss $\mathcal{L}_{feat} = \sum_{j=1}^{N_D} \sum_{k=1}^{k=F} L_1(D_{j,k}(i), D_{j,k}(\tilde{i}))$ based on the L_1 loss between F intermediate feature representations of the multi-scale discriminator for the real and synthetic images. $\mathcal{L}_{feat}$ aims to regularise the multi-scale discriminator.

Modified training framework: The original SPADE network was designed for computer vision applications, and hereby, we identified some limitations when it came to applying it to medical imaging that we tried to mitigate accordingly in our previous work Fernandez et al. (2022).

1. **Implicit latent space clustering:** Whereas the styles in computer vision are merely a qualitative concept (i.e., in a landscape dataset, a style can be the *dawn* or *night*), the contrasts in MRI are directly related to the physical tissue properties and scanner settings. Thus, we had to ensure that the same contrast from different slices of the same subject results in the same style code and that codes belonging to different acquisitions with the same modality are clustered together. Because the acquisition parameters, even for the same modality, can differ across datasets, resulting in slightly different contrasts, we also aimed to cluster the codes belonging to the same dataset together, within the modality cluster (e.g. within the ‘T1’ cluster, have a ‘T1-SABRE’ cluster a ‘T1-ADNI’ cluster etc.). For this, we implemented two losses. The first, a modality and dataset discrimination loss $\mathcal{L}_{mod-disc} = \lambda_{mod} BCE(m, \hat{m}) + \lambda_{dat} BCE(d, \hat{d})$, where m and d are the ground truth modality and dataset, $\hat{m}$ and $\hat{d}$ are the modality and datasets predicted on the generated images by a feed-forward network pretrained on the training dataset, and BCE is the binary cross-entropy loss. More information about this network characteristics and training are included in the implementation details in Appendix B.4. Secondly, we used a contrastive loss (Chen et al., 2020) $\mathcal{L}_{cont} = \cosim(E_{ig}(i), E_{ig}(\tilde{i}))$,

where E_{ig} is the encoder of the image generator, i the input style image, $\tilde{i}$ a randomly affine-augmented version of the input style image, and $\cosim$ the cosine similarity.

2. **Binary segmentations:** the original network took in binary categorical segmentations. In [Rusak et al. \(2020\)](#), it was shown that using probabilistic segmentations, where a pixel is attributed probabilities of belonging to more than one tissue, yields finer results for brain image synthesis. Therefore, we replaced the categorical labels with probabilistic ones, which are also output by GIF ([Cardoso et al., 2012](#)).
3. **Paired styles and segmentations:** Originally, during training, the input style image i matched the input segmentation m fed to the decoder, making it hard to rule out whether the style image has some influence on the decoder or the segmentation on the style encoder. During training, we enforced the separation of the style and content pathways by using unpaired style images and segmentation slices within the same volume, to force disentanglement and increase the consistency across slices by forcing any input style slice to convey only the contrast. Hence, we input image i_s to the encoder, which does not match map m , unlike ground truth image i , which also shares its contrast with i_s . We do this by drawing a different 2D image slice i_s from the same 3D volume from which we draw the ground truth image i , which is paired to its input map m .

The two new losses $\mathcal{L}_{mod-disc}$ and $\mathcal{L}_{cont}$ were added during training to the original SPADE losses. The weights used for these losses and the new ones are provided in the appendix ([Appendix B.3](#)).

To balance the training and avoid mode-collapse, we implemented an adversarial scheduler, which establishes whether the discriminator and generator are trained as a function of the discriminator accuracy averaged across the last 20 iterations. Over an upper threshold $T_{up} = 0.75$, only the generator is trained. On the other hand, only the discriminator is trained if its accuracy is below a lower threshold $T_{down} = 0.6$. Between the two, both networks are trained. These thresholds were chosen empirically. The discriminator accuracy was calculated as an average accuracy across discriminated patches and individual discriminators. Besides concatenating the semantic label to the ground truth/synthesised image before inputting it to the discriminator, we concatenate the image modality to give the discriminator more context about the variability existing within the inputs.

4.3. Dataset sampling

An overview of the differences between sampling from brainSPADE2D or brainSPADE3D can be seen in the appendix.

brainSPADE2D: In brainSPADE2D, the user can condition the output on (1) which pathologies are present as a proportion within the slice. In cases where the user does not want a lesion to be present, the conditioning value associated with that lesion is set to 0. Otherwise, we draw it from a uniform distribution bounded by the 1st and 3rd quartiles of the disease proportion across our training dataset; (2) the slice number, as a float value <1 ; (3) the desired contrast, by passing an axial image with the desired MRI modality to the encoder of the image generator.

brainSPADE3D: Similar to brainSPADE2D, the user can condition the output on the pathology proportion in a similar manner by specifying a relative proportion of the disease within the volume. After generating a 3D label, each of its axial slices is forwarded to the image generator, one by one. Resulting axial 2D images are assembled into an output 3D volume. To ensure a consistent intensity transition between neighbouring slices, we keep the style image constant across the process.

The sampling from the label generator was accelerated using a denoising diffusion implicit model (DDIM) sampler ([Song et al., 2020](#)); the number of intermediate time steps was set to 200 across all experiments unless stated otherwise.

4.4. Evaluation methodology

4.4.1. Label evaluation

There is no established standard for assessing label quality, as there is for images (SSIM, FID_{Rad} , etc.). As a surrogate of a label quality evaluation method, we compared the proportion of regions of interest within a volume or slice between our synthetic label maps and real datasets. For this purpose, we used 150 synthetic and real volumes (equivalent to ~ 15000 slices in 2D). Additionally, we studied the effect of conditioning on the output labels by observing and quantifying the different lesions when we vary the conditioning values.

4.4.2. Qualitative image evaluation

In addition to providing visual examples in 2D and 3D (reconstructed volumes by forwarding volumetric segmentations slice by slice), we generated a set of 250 slices spanning multiple modalities and diseases and shuffled them into a dataset of 250 real slices, then asked a human rater with >15 years of experience in neuroimaging to label them as real or fake.

4.4.3. Quantitative image evaluation

As a reference, we simulated images from the real and the synthetic labels using a similar approach to synthesis method proposed in [Billot et al. \(2023\)](#). For each modality m and label map L , we obtain a ‘statistics-derived’ (SD) image $SDI_{m,L}$:

$$SDI_{m,L} = BF \sum_{i=1}^R L_i \times \mathcal{N}(\mu_{m,d,i}, \sigma_{m,d,i}/K) \quad (4)$$

where BF is a random bias field operator implemented using MONAI (range: 0.05, 0.15, probability: 0.4), $\mathcal{N}$ is a Gaussian distribution of mean $\mu_{m,d,i}$ and standard deviation $\sigma_{m,d,i}/K$, K is a modality-dependent factor selected empirically to ensure good contrast between tissues, and R is the number of semantic regions. The modality means and standard deviations were obtained by computing the average intensity for semantic region i and dataset d on a bias field-corrected subset of our dataset (50 images per modality and dataset). The dataset selected for each simulated image was randomly selected. Example images are displayed in [Figs. 6, C.17 and C.18](#).

We simulate T1, FLAIR and T2 SD images from the test set of 1176 label slices. We also forward these labels through our image generator, using style images from the same test dataset. For the resulting SD and generated images, we calculate the Structural Similarity Index (SSIM) and the mean square error (MSE) with regard to their ground truth. To further assess the performance of the image generator, we calculate the Fréchet Inception Distance (FID) ([Heusel et al., 2017](#)) on the activations of the last functional layer of an InceptionV3 network trained on RadImageNet ([Mei et al., 2022](#)) between real training images and (we use two sets of real images gt_1 and gt_2) (1) synthetic 2D images generated from real labels ($gen_{real-labs}$) (2) SD images generated from real labels ($SD_{real-labs}$) (3) synthetic 2D images generated from synthetic 2D labels ($gen_{syn-labs-2D}$) (4) SD images generated from synthetic 2D labels ($SD_{syn-labs-2D}$) (5) synthetic 2D images generated from sliced synthetic 3D labels ($gen_{syn-labs3D}$) (6) SD images generated from sliced synthetic 3D labels ($SD_{syn-labs3D}$). We call this RadImageNet FID metric “ FID_{Rad} ”. As more samples are required for FID_{Rad} calculation, we use 6000 T1, FLAIR and T2 images for each of these groups. For reference, we also calculate the FID_{Rad} between different sets of 6000 real training images.

4.4.4. Data privacy

The premise of our work focuses on creating synthetic models and datasets that can be shared. This is only the case if, indeed, the resulting images cannot be traced back to the original subjects. A validation phase should confirm this fact, but there are a number of impracticalities surrounding such a test:

- There is no specific agreement on what is considered acceptable regarding privacy in synthetic datasets. Some regulatory bodies and institutions, such as the NHS Transformation Directorate in the United Kingdom, provide broad guidelines on what synthetic data should comply with, but the research community is still far from specific metrics or methods. The literature on privacy and Generative Modelling is mainly theoretical and rarely provides practical examples (Sun et al., 2021).
- There is a diverse scenario in terms of potential data breaches. Model inversion attacks consist of retrieving the training data from the model weights. Membership attacks assume that the attacker has access to the training samples: a data breach happens when it is confirmed that those samples were indeed used to train the model (Sun et al., 2021). The few examples found in the literature for these focus on niche scenarios and thus are hardly applicable to our case.

There is, however, a consensus on the need to train these models on sufficient data to avoid overfitting (Carlini et al., 2023). We tried to avoid this by using substantial data augmentation in both the image and label generators. The full separation between the content and style streams imposed by our two-stage scenario further diminishes the probability of a sample being mapped back to a specific label-image pair. Despite the measures, we still wanted to assess whether it would be theoretically possible to know – at least – if the model is memorising the data. To do this, we focused on the following questions:

- Are the labels being memorised? We sample a certain number of labels and, for each of them, we retrieve the nearest neighbours in terms of dice score (DS) in the training dataset and plot them.
- Is the image generator behaving differently when presented with an input style from the training distribution than it does when presented with one from an unseen distribution? For this, we use images and labels coming from three 500 slice sets: (1) in-distribution training data; (2) in-of-distribution test data; (3) out-of-distribution data (from OASIS); (4) out-of-distribution data (from BraTS-OTHER). We assess whether style codes belonging to these four categories can be clustered using a K-Means algorithm (initialised at random, 300 iterations). To keep the test fair, we ruled out tumour images from the ID sets, as the OD set does not have this phenotype.

4.4.5. Evaluation on segmentation tasks

The critical validation experiments must prove that our synthetic data can be employed to train segmentation models. For this, we propose several segmentation tasks in which we compare the performance of segmentation models trained, in part or exclusively, on different amounts of synthetic data to that of models trained on real data. To ensure reproducibility, we use 2D and 3D nnU-Net (Isensee et al., 2021), using the default settings, and training for 1000 epochs. All the models were tested on specific sets referred to in each experiment section. Two-sided Wilcoxon signed-rank tests were ran on the performance metrics using *scipy*, setting α to 0.05.

5. Results

5.1. Quality of the generated labels

Tissue proportions: To evaluate the quality of the generated 2D labels, we calculated the proportion of brain pixels of each semantic region as a function of the slice number for a set of 5000 generated healthy labels and compared it to the proportions from the training set. For the 3D label generator, we calculated the total proportion of brain pixels of each semantic region. Results are shown in Fig. 2. The 2D label generator manages to approximately follow the distribution per tissue and slice, of real data, showing that the slice conditioning works effectively. In the case of the 3D label generator, we observe

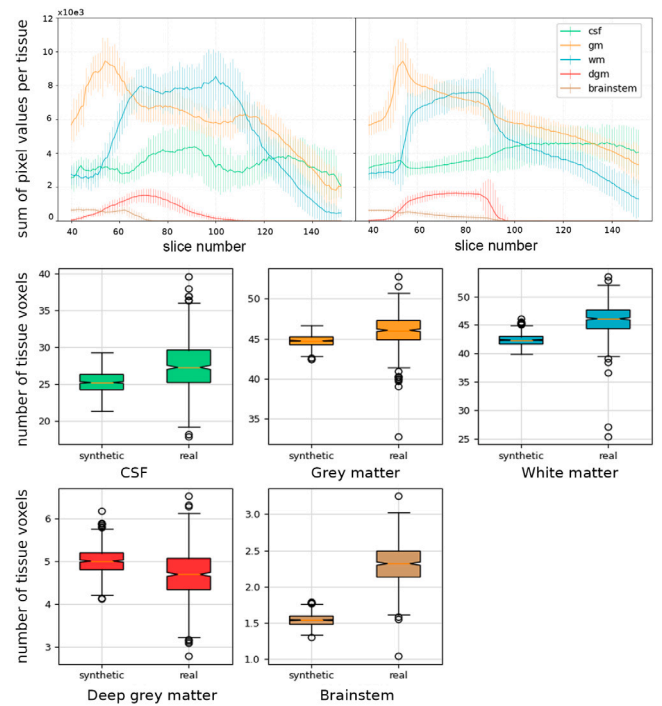


Fig. 2. Top: sum of pixel probabilities per tissue and slice for real (left) and synthetic (right) 2D label datasets, vertical bars indicating $2\times$ the standard deviation for each slice; bottom: sum of voxels for synthetic (left) and real (right) per tissue ($\times 10^3$).

that the standard deviation of all synthetic tissues is statistically lower than that of real ones, showing less variability overall. In addition, we can see a slight discrepancy in the means, with the synthetic brainstem mean being considerably lower than its real counterpart.

Visual quality: A visual assessment of real labels and images (for an example, see Fig. D.19) shows that the middle brain, where the probabilistic labels are noisier, results in many brainstem voxel probabilities to fall under 0.5 in favour of other tissues, such as white or deep grey matter, which explains the observed result. In addition, this visual assessment shows that, even though causing a minor impact on the metrics showcased by Fig. 2, there is some loss of sharpness in the 3D labels, especially in some cortical folds (which causes the outer CSF to disappear, explaining why its mean is also lower in Fig. 2) and the cerebellar cortex. The lack of sharpness can be explained by the blurring caused by autoencoders, which has been observed in the literature (Zhao et al., 2017). This is a limitation of using diffusion versus using latent space diffusion, the latter of which is essential given the memory constraints of 3D modelling. The lack of variability can be explained by limited capacity, or by the absence of any conditioning that enforces variability across healthy subjects, such as age, something that could be left for future enhancements and has already been applied to generative modelling for brain MR (Pinaya et al., 2022).

Impact of conditioning: To evaluate the impact of the lesion conditioning, we evaluate the size of output WMH and tumour (all three subclasses) lesions when we increase the value of their corresponding conditioning. To sample the labels we used DDIM (Song et al., 2020), with 200 steps. Fig. 3 shows that, visually, for both tumours and WMH, increasing the conditioning value results in bigger lesion sizes. This is supported by Fig. 4, where the number of voxels corresponding to a lesion tends to increase as we increase the conditioning value.

Note that we limited this experiment to the 3D model because, in the 2D label generator, the lesion proportion can be dependent on the slice number as well, which makes it hard to disentangle the lesion size and the specific slice that is generated and hard to observe the progression we can see in Fig. 3. More information about 2D lesion

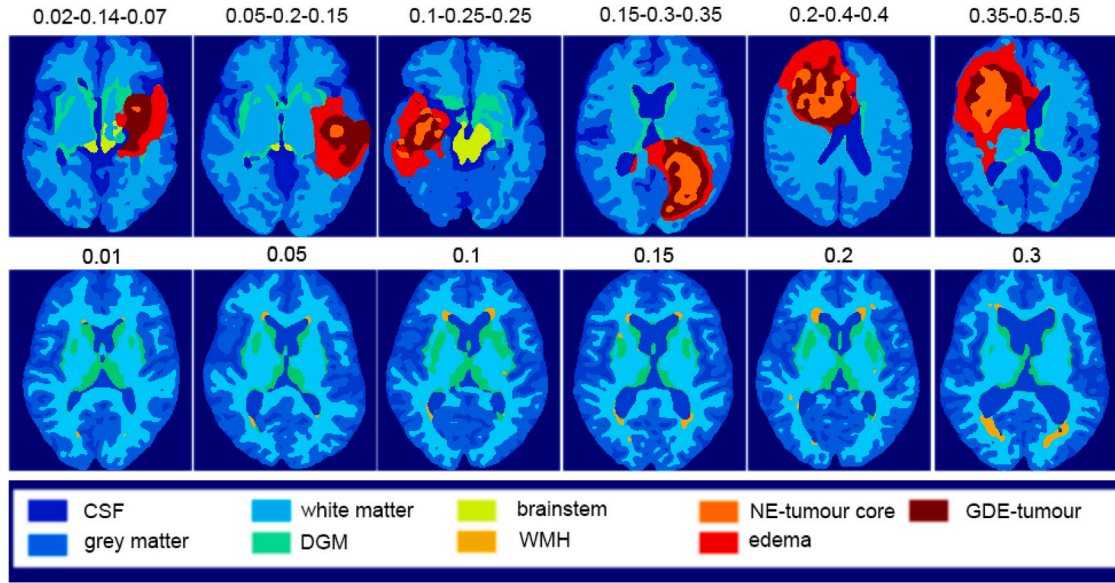


Fig. 3. Label maps sampled from the 3D label generator using increasing conditioning for the three tumour layers (top row: necrotic and non-enhancing tumour core conditioning value, peritumoral oedema conditioning value, and GD-enhanced tumour conditioning value, separated by hyphens) and WMH (bottom row). The axial slice chosen for tumours displays the maximum tumour size. For WMH, the same middle axial slice is shown for all samples.

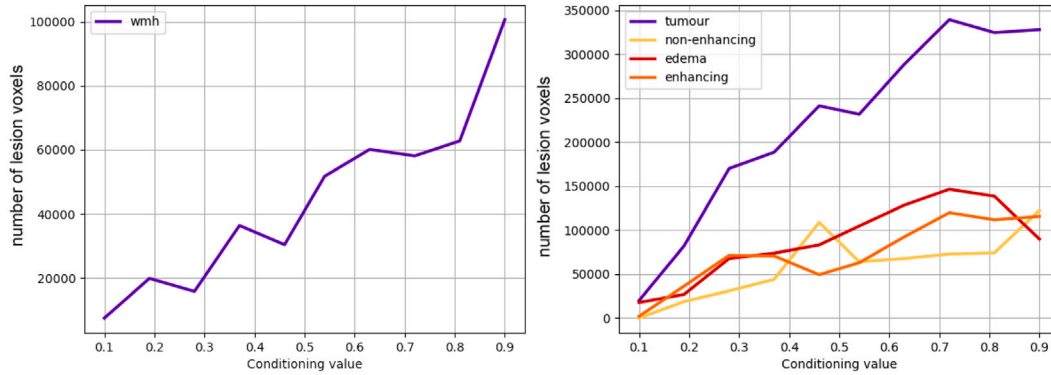


Fig. 4. Number of lesion voxels as a function of increasing conditioning values for WMH (left) and tumour (right).

conditioning analysis can be found in the appendix (see [Appendix D](#) and [Fig. D.20](#)).

Lesion extrapolation: We evaluated the ability of our 3D label generator to produce unseen conditioning combinations (WMH + all tumour layers). We used the model that had not seen any combination of these pathologies and used random values for all disease types. Approximately 41% of the sampled labels exhibited all of the lesions. Examples that show how the model is able to extrapolate can be seen in [Fig. 5](#). Due to the same reasons exposed in the previous section, we limited this test to the 3D model.

5.2. Quality of the generated images

Quantitative evaluation: The SSIM, MSE and FID_{Rad} metrics are reported in [Table 2](#), with example images displayed in [Fig. 6](#). For all cases (images derived from real, synthetic 2D and synthetic 3D labels) and all modalities, the images generated by our model yield a considerably higher SSIM, and lower MSE and FID_{Rad} than the SD ones. Nonetheless, the FID_{Rad} score obtained between different sets of real images $gt_1 - gt_2$ is generally at least an order of magnitude less than the rest of the calculated FID_{Rad} scores. To verify the impact of perturbations on FID_{Rad} , we smoothed one of our real images set (gt_2) by applying Gaussian smoothing ($\sigma = 1.0$) using an operator B implemented with MONAI. Even though the blur was subtle enough

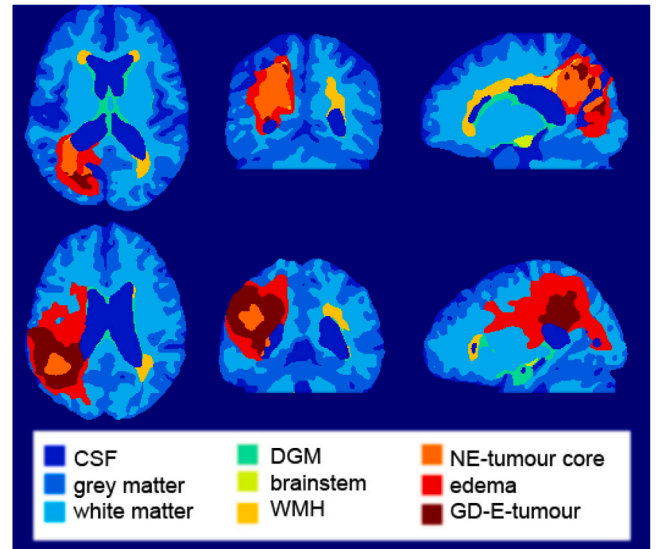


Fig. 5. Sample label maps generated using unseen conditioning combinations (WMH + tumour), showing how the model is able to generate both lesions, with WMH showing up in yellow, and the tumour layers in dark orange-red.

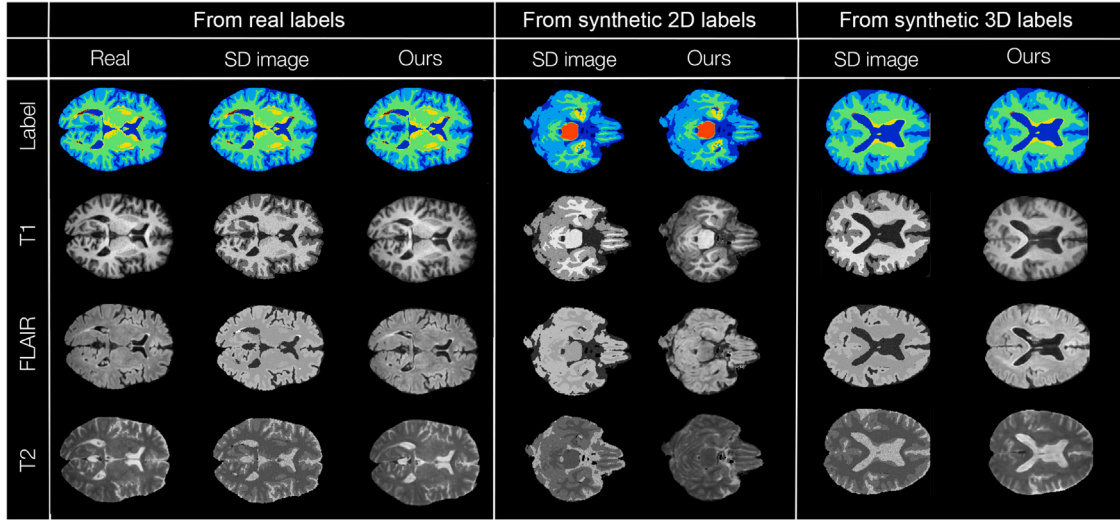


Fig. 6. Example generated and SD (statistics-derived) T1, FLAIR and T2 images generated using real labels (left section), synthetic 2D labels (middle section) and synthetic 3D labels (right section). Ground truth images are also depicted for the images derived from real labels.

Table 2

SSIM and MSE quality metrics between the generated - $gen_{real-labs}$ - and statistics-derived - $SD_{real-labs}$ - T1, FLAIR and T2 images using the real labels with regards to the ground truth test set; FID_{Rad} obtained between non-overlapping ground truth sets gt_1 and gt_2 , between gt_1 and a gt_2 subject to a smoothing operator B , between gt_1 and images generated from real labels $gen_{real-labs}$ and synthetic ones $gen_{syn-labs}$, and between gt_1 and images derived from real labels $SD_{real-labs}$ and synthetic ones $SD_{syn-labs}$.

Metric	Set	T1	FLAIR	T2
SSIM	$gen_{real-labs}$	0.985 _{0.012}	0.978 _{0.017}	0.982 _{0.015}
	$SD_{real-labs}$	0.849 _{0.071}	0.855 _{0.061}	0.884 _{0.056}
MSE($\times 10^2$)	$gen_{real-labs}$	0.337 _{0.096}	0.551 _{0.013}	0.604 _{0.012}
	$SD_{real-labs}$	0.939 _{0.337}	0.844 _{0.221}	0.810 _{0.121}
FID_{Rad}				
Set				
$gt_1 - gt_2$				
0.025				
$gt_1 - B(gt_2)$				
0.278				
$gt_1 - gen_{real-labs}$				
0.100				
$gt_1 - SD_{real-labs}$				
0.294				
$gt_1 - gen_{syn-labs-2D}$				
0.215				
$gt_1 - SD_{syn-labs-2D}$				
0.393				
$gt_1 - gen_{syn-labs-3D}$				
0.137				
$gt_1 - SD_{syn-labs-3D}$				
0.353				

that the perturbations were not noticeable with the naked eye, the FID_{Rad} score ($gt_1 - B(gt_2)$) was considerably affected, indicating that the slightest blur can change the FID_{Rad} to values that are comparable to ($gt_1 - gen_{syn-labs-2D}$, $gt_1 - gen_{syn-labs-3D}$), and even over ($gt_1 - gen_{real-labs}$) the ones observed for our generated images. Generally, synthetic labels deteriorate the FID_{Rad} , especially for FLAIR and T2 images.

Qualitative evaluation: Table 3 shows that the scores obtained by a human rater when attempting to discern real and synthetic images are at a chance level, which indicates that the 2D images are visually indistinguishable from real ones. We did not attempt the same test with 3D images since, due to the slight inter-slice interference persisting in the reconstructed volumes, it is obvious that the images are synthetic.

5.3. Ablation study

To assess the impact of the additional losses or features covered in Section 4.2.2, we ran an ablation study (for more details refer to Appendix C). We evaluated the consistency between generated neighbouring slices. Whereas the modality discrimination loss did not seem to qualitatively or quantitatively ameliorate consistency, the use of

Table 3

Accuracy, precision and recall per-modality obtained by an expert on classifying real or fake slices.

Contrast	Accuracy	Precision	Recall
T1	0.490	0.489	0.440
FLAIR	0.540	0.530	0.700
T2	0.420	0.433	0.520

unpaired styles and segmentations and the contrastive loss significantly improved the inter-slice structural similarity index and the visual consistency, for T1 and FLAIR images especially. This can be observed in Fig. 7 (FLAIR and T2 examples are available in the appendix - Figs. C.17 and C.18). Hence, we removed the modality and dataset discrimination losses $\mathcal{L}_{mod-disc}$ introduced in Section 4.2.2-Modified training framework in further trainings.

5.4. Code clustering

To assess the impact of the selected losses and added features ($\mathcal{L}_{cont}$ + using unpaired slices) on the style latent space, which aimed to cluster the style representations, or codes, of different 2D slices belonging to the same modality and dataset together, and remove any influence from the semantic content of these slices, we studied whether these styles were clustered as a function of their modality and dataset. For that, we first reduced the dimension of the 16-dimensional latent space codes to 2 dimensions using t-stochastic neighbour embedding (t-SNE, perplexity: 15, learning rate: 400). We encoded the images belonging to our test set introduced in Section 4.4.3, ran t-SNE and plotted the resulting 2D codes, for the models trained with contrastive loss only, with unpaired slice selection only, with both of these features, and with neither (see Fig. 8). Whereas the model trained on only the original SPADE losses displays some modality clustering, with blue (FLAIR), pink (T2) and green (T1) clusters being visible, there is more overlap between the clusters than in the others, notably between FLAIR and T1 images belonging to BRATS. In addition, our model achieves more clustering between codes belonging to the same dataset (within each modality), something that is likely translating into more style consistency.

5.5. Assessment of data memorisation

Memorised labels: As discussed in Section 4.4.4, we assessed whether the label generator was memorising samples. For this, we sample a series of labels using various types of conditioning, and perform

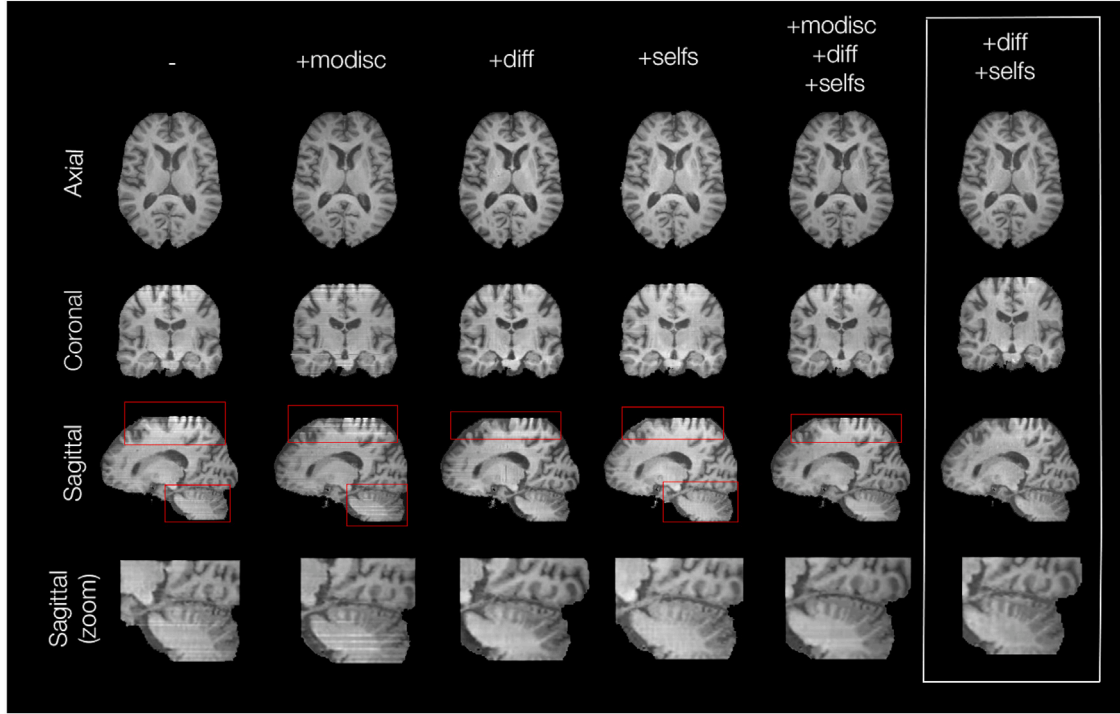


Fig. 7. Example reconstructed T1 volumes. From top to bottom: axial, coronal, sagittal slices and zoom on the cerebellum of the sagittal slices of the volumes reconstructed from the stack of synthetic T1 slices. From left to right: models using no additional losses (original SPADE network), $L_{mod-disc}$ loss only (+modisc), unpaired style and semantic slices (+diff), L_{cont} loss only (+selfs), all losses and unpaired style and semantic slices (+modisc, +diff, +selfs), and the statistically best one, using L_{cont} and unpaired style and semantics slices (+diff, +selfs).

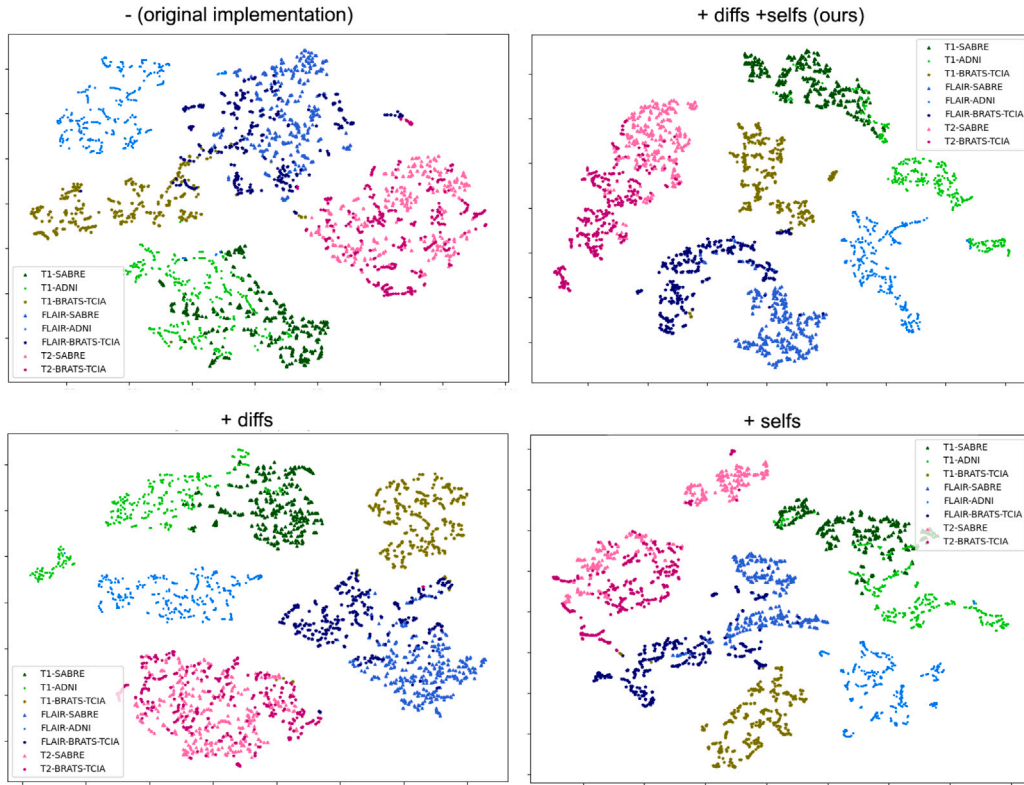


Fig. 8. t-SNE plots obtained on the test set codes for the model trained with no additional losses and paired style and label slices (top left), using both contrastive loss and unpaired slices (top right, '+diff +selfs'), using different slices (bottom left, '+diff') and contrastive loss (bottom right, '+selfs'). The colours denote different modalities and datasets.

qualitative comparisons between these samples and their four nearest neighbours (NN) in the training dataset, using Dice score as a ranking metric. Examples are shown in Fig. 9. All examined samples showed differences with regard to their NNs, suggesting that the model is likely not memorising the data. The closest similarities are those between the generated tumours (for example, the second row in Fig. 9). This is in line with the observations made by Carlini et al. (2023), where diffusion models were more likely to memorise out-of-distribution images. Indeed, even though we had a substantial amount of tumour volumes, they still have a relatively small prevalence within the training dataset — especially in 2D, where only a small fraction of slices from a volume containing a tumour will actually have a tumour. In addition, in 2D, the (slice number, lesion) pairing might be causing the subset of images falling under the specific conditioning to be even smaller.

Memorised image styles: As discussed earlier, we aimed to see whether the style encoder of our image generator is overfitting to the training distribution, in a way that makes it possible to retrieve sample slices that were part of the training set. For this, we obtained the style latent representations for in-distribution training and test data (ID-train, ID-test), and OASIS and BraTS-OTHER out-distribution data (OD, OD-tumours), reduced them to 2D with t-SNE, and used K-Means to cluster them. Fig. 10 shows that the K-Means clusters (represented by coloured frames) differ from the distribution clusters, with the exception of the one corresponding to BraTS-OTHER, for both modalities. This showcases that the model is not memorising training data, as it does not exhibit different behaviour for training and unseen test ID data, and OD data. Even if the style representations associated with BraTS-OTHER (OD-tumours) are clearly separable from the rest, this still does not allow retrieving latents associated with training samples from the total set of latents.

5.6. Segmentation experiments

5.6.1. Segmentation of healthy regions

This task focuses on segmenting five healthy tissues: CSF, GM WM, DGM and brainstem. We train the following nnU-Net models:

- $H_{ID-real-2D}$ Trained on ~15 000 label slices from SABRE and T1 images from the SABRE dataset (resulting from slicing 150 volumes)
- $H_{ID-real-3D}$ Trained on 150 label volumes and T1 images from the SABRE dataset.
- $H_{ID-syn-2D}$ Trained on ~15 000 synthetic label slices and T1 images generated using SABRE slices as style input.
- $H_{ID-syn-3D}$ Trained on 150 synthetic healthy label volumes and T1 images generated using SABRE slices as style input.

The term “ID” - “in-of-distribution” - encompasses synthetic data produced using (1) seen phenotypes (e.g. healthy, exclusive WMH or tumour lesions) and (2) seen styles (input style images belonging to datasets and modalities that have been seen during training). We tested the models on a test set of 30 T1 volumes from a hold-out set of SABRE. For the 2D models, we calculated the Dice metric after aggregating the individually segmented slices. Dice scores on each tissue are reported in Table 4. Although models trained on real data perform significantly better (p -value < 0.05) than their synthetic counterparts, we observe a comparable performance for all five tissues. Note that the higher performance of 2D real models over the 3D is likely due to a capacity issue, as the ratio between the number of model parameters and available samples was much higher for the training dataset.

5.6.2. Addressing out-distribution data with synthetic data

To investigate the ability of our synthetic data to address out-distribution domains, we focused on two tasks:

- near out-distribution (n-OD): testing synthetic data on 30 T1 volumes from ABIDE (see Table 1), provided we only have real T1 volumes from SABRE.

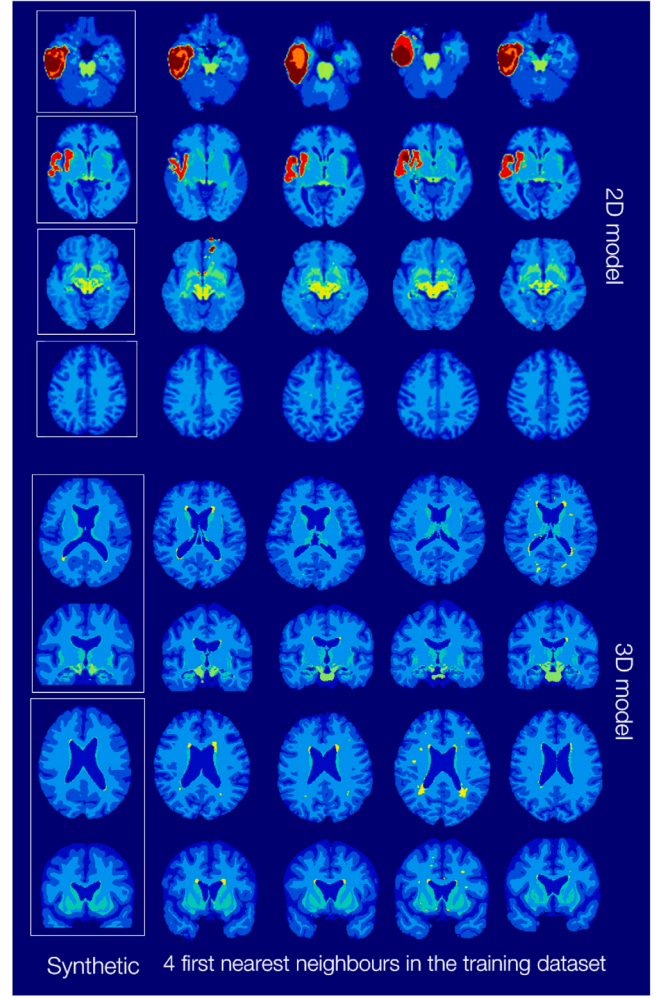


Fig. 9. Sample 2D and 3D synthetic labels, and their four nearest neighbours in the training dataset. For the 3D images, axial and coronal views are provided.

Table 4

Mean dice scores obtained on 2D and 3D healthy ID segmentation experiments. An asterisk indicates statistical significance.

Region	$H_{ID-real-2D}$	$H_{ID-syn-2D}$	$H_{ID-real-3D}$	$H_{ID-syn-3D}$
CSF	0.952 _{0.007} *	0.930 _{0.013}	0.933 _{0.008} *	0.906 _{0.013}
GM	0.956 _{0.004} *	0.933 _{0.010}	0.935 _{0.006} *	0.913 _{0.012}
WM	0.969 _{0.004} *	0.952 _{0.010}	0.961 _{0.004} *	0.943 _{0.006}
DGM	0.868 _{0.019} *	0.811 _{0.040}	0.805 _{0.026} *	0.746 _{0.026}
Brainstem	0.952 _{0.021} *	0.934 _{0.017}	0.931 _{0.019} *	0.884 _{0.011}

- far out-distribution (f-OD): testing synthetic data on 30 FLAIR volumes from OASIS, provided we only have T1 volumes from SABRE.

The regions to segment are the same as in Section 5.6.1. In the first case, we are exploring cross-dataset style transfer, whereas in the second, we address cross-modality style transfer. We sampled 2D (15 000 slices) and 3D (150 volumes) datasets conditioning on the styles of the target domains. We test in-distribution (ID) models $H_{ID-real-2D}$, $H_{ID-real-3D}$, $H_{ID-syn-2D}$ and $H_{ID-syn-3D}$ from Section 5.6.1 in both n-OD and f-OD sets. In addition, we trained $H_{n-OD-syn-2D}$, $H_{n-OD-syn-3D}$, $H_{f-OD-syn-2D}$ and $H_{f-OD-syn-3D}$ on the synthetic 2D and 3D n-OD and f-OD datasets. For reference, we also trained $H_{n-OD-real-2D}$, $H_{n-OD-real-3D}$, $H_{f-OD-real-2D}$ and $H_{f-OD-real-3D}$ on a small set of 10 T1 volumes/1090 slices from the n-OD and f-OD. A summary of these experiments and models is reported in Table 5.

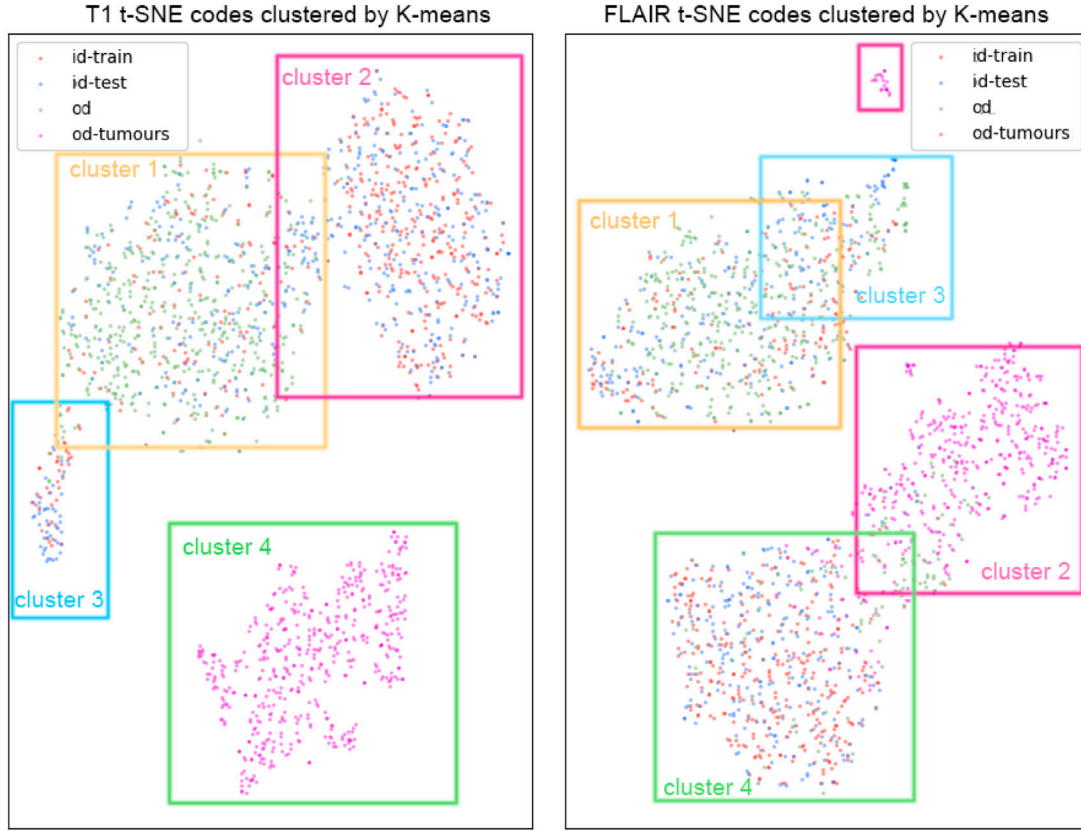


Fig. 10. t-SNE-reduced codes for T1 (left) and FLAIR (slices) for ID-train data, ID-test data, OD data and OD data coming from BraTS-OTHER. K-means clusters are shown as coloured frames. Clusters names are assigned arbitrarily. K-Means clusters of T1 and FLAIR codes.

Table 5

Summary of models trained on OD experiment. S: synthetic/R: real. Note that the real datasets do not need a “style” per se, as they are obtained using the original images. The populated field “style” in this case refers to the dataset and modalities used to create these training sets.

2D model	3D model	R/S	style
Used in near-OD and far-OD experiments			
$H_{ID-real-2D}$	$H_{ID-real-3D}$	R	SABRE (T1)
$H_{ID-syn-2D}$	$H_{ID-syn-3D}$	S	SABRE (T1)
Used in near-OD experiment			
$H_{n-OD-real-2D}$	$H_{n-OD-real-3D}$	R	ABIDE (T1)
$H_{n-OD-syn-2D}$	$H_{n-OD-syn-3D}$	S	ABIDE (T1)
Used in far-OD experiment			
$H_{f-OD-real-2D}$	$H_{f-OD-real-3D}$	R	OASIS (FLAIR)
$H_{f-OD-syn-2D}$	$H_{f-OD-syn-3D}$	S	OASIS (FLAIR)

Results are visible in Table 6. For n-OD, $H_{n-OD-syn-2D}$ and $H_{n-OD-syn-3D}$ typically perform better for all regions, only excelled by the reference models $H_{n-OD-real-2D}$ and $H_{n-OD-real-3D}$. Interestingly, whereas $H_{ID-real-2D}$ achieves comparable performance for CSF, WM and DGM, in 3D, $H_{ID-syn-3D}$ seems to perform better, overall, than $H_{ID-real-3D}$, showing that, in 3D, synthetic data, even sampled using ID styles, seem to help the model generalise better to unseen distributions. For f-OD, $H_{f-OD-syn-2D}$ and $H_{f-OD-syn-3D}$ perform significantly better than the ID ones, especially for DGM and brainstem, as their appearance is very different, less pronounced in FLAIR images compared to T1.

5.6.3. Disease segmentation with synthetic data

White matter hyperintensities segmentation:

We sampled 2D and 3D datasets conditioning the labels on only WMH, and using the style from SABRE FLAIR images as style condition-

ing, resulting in datasets of size 15 000 and 150 for 2D and 3D respectively, that we used to train models $D_{WMH-syn-2D}$ and $D_{WMH-syn-3D}$. We compared the performance on real datasets $D_{WMH-real-2D}$ and $D_{WMH-real-3D}$ of the same sizes as their synthetic counterparts. Additionally, since synthetic data will be potentially of use when there is not sufficient real data, we explored the potential of synthetic data to complement small datasets. For this, we trained segmentation models $D_{WMH-small-2D}$ and $D_{WMH-small-3D}$ on a subset of five subjects (for 2D: all slices corresponding to these five subjects) from the real dataset. Lastly, we combined the small real datasets and the synthetic ones into hybrid models $D_{WMH-hybrid-2D}$ and $D_{WMH-hybrid-3D}$. Results are reported in Table 7.

For most metrics and in 2D and 3D, the models trained on large real datasets perform significantly better (p -value < 0.05). In 2D and 3D, the hybrid model is able to boost the Dice and recall with regards to the models trained on scarce data, in most cases being significantly better than them (p -value < 0.05).

Note that the synthetic model in 3D achieves significantly lower Dice, precision and recall than the other models, showcasing that a synthetic-real domain gap persists in 3D.

Tumour: We sampled 2D and 3D datasets conditioning on the three tumour labels from the BraTs challenge and obtained T1, FLAIR and T2 images using the styles of our target distribution (the set of other sites from BraTs). We trained nnU-nets to segment all three tumour tissues: peritumoral oedema, necrotic and non-enhancing tumour core, and GD-enhanced tumour on our synthetic 2D and 3D datasets of sizes 15 000 and 150, respectively, resulting in models $D_{tum-syn-2D}$ and $D_{tum-syn-3D}$. We compared with real counterparts of equivalent sizes $D_{tum-real-2D}$ and $D_{tum-real-3D}$. Additionally, we explored the effect of combining a small labelled target dataset with our synthetic datasets. For this, we use a holdout set of six BraTS-OTHER subjects (in 2D, resulting in 1023

Table 6

Results for the 2D and 3D near and far out-distribution experiments. * indicates significantly better performance among all models, ** indicates significantly better performance among all models minus the reference ones (real-OD).

Dimension	Model	CSF	GM	WM	DGM	Brainstem
Near out-of-distribution						
2D	$H_{ID-real-2D}$	0.806 _{0.049} **	0.877 _{0.045}	0.919 _{0.031} *	0.700 _{0.286}	0.799 _{0.147}
	$H_{ID-syn-2D}$	0.768 _{0.032}	0.877 _{0.027}	0.918 _{0.022}	0.690 _{0.241}	0.793 _{0.172}
	$H_{n-OD-syn-2D}$	0.800 _{0.028}	0.892 _{0.020} **	0.924 _{0.019} *	0.789 _{0.048} **	0.854 _{0.051} **
	$H_{n-OD-real-2D}$	0.895 _{0.018} *	0.940 _{0.007} *	0.961 _{0.006} *	0.889 _{0.014} *	0.918 _{0.034} *
3D	$H_{ID-real-3D}$	0.630 _{0.079}	0.786 _{0.043}	0.817 _{0.043}	0.636 _{0.182}	0.641 _{0.308}
	$H_{ID-syn-3D}$	0.789 _{0.038} **	0.861 _{0.025} **	0.906 _{0.020}	0.770 _{0.042}	0.680 _{0.240}
	$H_{n-OD-syn-3D}$	0.771 _{0.034}	0.857 _{0.018}	0.908 _{0.017} *	0.780 _{0.038} **	0.723 _{0.245} *
	$H_{n-OD-real-3D}$	0.911 _{0.019} *	0.947 _{0.008} *	0.967 _{0.007} *	0.906 _{0.014} *	0.898 _{0.049} *
Far out-of-distribution						
2D	$H_{ID-real-2D}$	0.663 _{0.104}	0.459 _{0.049}	0.486 _{0.177}	0.055 _{0.056}	0.014 _{0.043}
	$H_{ID-syn-2D}$	0.659 _{0.111}	0.497 _{0.036}	0.373 _{0.179}	0.092 _{0.043}	0.036 _{0.090}
	$H_{f-OD-syn-2D}$	0.740 _{0.106} *	0.772 _{0.064} *	0.804 _{0.076} *	0.625 _{0.105} *	0.841 _{0.100} *
	$H_{f-OD-real-2D}$	0.783 _{0.114} **	0.812 _{0.071} **	0.843 _{0.083} **	0.678 _{0.137} **	0.867 _{0.092} **
3D	$H_{ID-real-3D}$	0.683 _{0.104}	0.546 _{0.058}	0.583 _{0.093}	0.093 _{0.063}	0.062 _{0.129}
	$H_{ID-syn-3D}$	0.697 _{0.119}	0.468 _{0.083}	0.472 _{0.137}	0.043 _{0.037}	0.012 _{0.038}
	$H_{f-OD-syn-3D}$	0.732 _{0.093} *	0.756 _{0.059} *	0.790 _{0.062} *	0.583 _{0.094} *	0.824 _{0.129} *
	$H_{f-OD-real-3D}$	0.825 _{0.112} **	0.837 _{0.078} **	0.870 _{0.084} **	0.693 _{0.128} **	0.858 _{0.113} **

slices), and train segmentation models $D_{tum-small-2D}$ and $D_{tum-small-3D}$. We combine our synthetic datasets with the latter into datasets that we used to train models $D_{tum-hybrid-2D}$ and $D_{tum-hybrid-3D}$. All models were tested on a test BraTS-OTHER set of 26 subjects. Using the same evaluation as in [Isensee et al. \(2021\)](#), we merged the labels and computed the Dice score on (1) the whole tumour (oedema + necrotic and non-enhancing tumour core + GD-enhanced tumour) and (2) the tumour core (necrotic and non-enhancing tumour core + GD-enhanced tumour). We left out the calculation of the performance on the segmentation of solely the GD-tumour, as, since we did not use the GD-enhanced T2 images from the BraTs dataset, none of the available contrasts offer sufficient distinction between both tumour tissues. Results are reported in [Table 8](#).

In 2D, the model $D_{tum-small-2D}$ performed significantly worse than its counterparts on segmenting the whole tumour (p -value < 0.05). Nonetheless, no significant difference was found in the task of segmenting the tumour core due to the high variance of the performance across the test set. If we look at the Dice distributions for 2D (see [Fig. 11](#)), we can see that, for the whole tumour segmentation, the hybrid model has a lower range and higher median than its fully synthetic counterpart and $D_{tum-real-2D}$.

In 3D, we observe a similar trend as in 2D, with $D_{tum-hybrid-3D}$ being close to $D_{tum-real-3D}$ in performance for both the whole tumour and the tumour core, significantly better than $D_{tum-syn-3D}$ and $D_{tum-small-3D}$ for the whole tumour. As in 2D, the high standard deviation of the performances on the tumour core diminishes any statistical significance.

5.6.4. Impact of synthetic dataset size segmentation performance

To observe the impact of synthetic dataset size on the performance of segmentation models, we ran a separate healthy and lesion segmentation experiment varying the synthetic dataset size. For 2D models, we sampled datasets of size 50 000, 5000, 500 and 50; for 3D models, we sampled 1000, 100 and 10 pairs of segmentations and T1 images for healthy segmentation, FLAIR images for WMH segmentation and T1, FLAIR and T2 images for tumour segmentation. The segmentation models were trained, and then tested on real data, in the same fashion as in Sections 5.6.1 and 5.6.3. These models were named labelled H_{dim-S} for healthy segmentation, $D_{dim-S-WMH}$ and $D_{dim-S-tum}$ for WMH and tumour segmentation, with dim being the dimension, 2D or 3D, and S being the size (50K, 5K, 500, 1000, 100 or 10). Because the segmentation of healthy tissues is easier in 3D, even with small dataset sizes, we turned off data augmentation for healthy 3D models. The results obtained for healthy segmentation are reported in [Table 9](#), and

Table 7

Mean dice score, precision and recall obtained on 2D and 3D WMH segmentation. * indicates statistical difference in performance (if more than one model displays it, no statistical difference in performance was found between them); ** indicates statistically lower than model(s) labelled with ‘*’, but better than the rest.

Model	Dice	Precision	Recall
$D_{WMH-real-2D}$	0.787 _{0.070} *	0.831 _{0.103} *	0.761 _{0.089} *
$D_{WMH-small-2D}$	0.749 _{0.076}	0.829 _{0.102} *	0.695 _{0.101}
$D_{WMH-hybrid-2D}$	0.759 _{0.073} **	0.814 _{0.095}	0.723 _{0.100} **
$D_{WMH-syn-2D}$	0.745 _{0.078}	0.764 _{0.118}	0.746 _{0.099} *
$D_{WMH-real-3D}$	0.806 _{0.070} *	0.824 _{0.98} **	0.801 _{0.095} *
$D_{WMH-small-3D}$	0.720 _{0.081}	0.850 _{0.109} *	0.637 _{0.110}
$D_{WMH-hybrid-3D}$	0.747 _{0.075} **	0.819 _{0.113} **	0.702 _{0.109} **
$D_{WMH-syn-3D}$	0.626 _{0.113}	0.604 _{0.129}	0.679 _{0.147} **

Table 8

Dice and standard deviation on Dice for the 2D and 3D tumour segmentation experiments; * indicates statistical difference in performance (if more than one model displays it, no statistical difference in performance was found between them).

Model	Dice (whole)	Dice (core)
$D_{tum-real-2D}$	0.879 _{0.059} *	0.559 _{0.275}
$D_{tum-small-2D}$	0.799 _{0.445}	0.455 _{0.243}
$D_{tum-syn-2D}$	0.872 _{0.100} **	0.559 _{0.306}
$D_{tum-hybrid-2D}$	0.876 _{0.114} *	0.535 _{0.323}
$D_{tum-real-3D}$	0.893 _{0.048} *	0.686 _{0.268}
$D_{tum-small-3D}$	0.832 _{0.128}	0.556 _{0.234}
$D_{tum-hybrid-3D}$	0.884 _{0.039}	0.606 _{0.313}
$D_{tum-syn-3D}$	0.867 _{0.061}	0.589 _{0.316}

those for lesion segmentation in [Table 10](#). Almost consistently across dimensions and tasks, the two largest datasets (1000–100 for 3D and 50 000 and 5000 for 2D) perform significantly better than the rest, except for tumour core segmentation and CSF in 3D, with the largest dataset outperforming the second largest for most tasks, notably for brainstem (2D and 3D) and DGM (3D), WMH (2D and 3D) and whole tumour (3D) showing general – but mild – improvement of segmentation scores with dataset size, especially for more complex tasks, and for 2D cases, where more samples mean more slice coverage and, likely, more relevant samples. Interestingly, significantly best performance was shared on multiple occasions by the largest and second-largest datasets, suggesting that the expressive power of the generative model might wane slightly as the synthetic dataset size exceeds that of the dataset used to train the generative model.

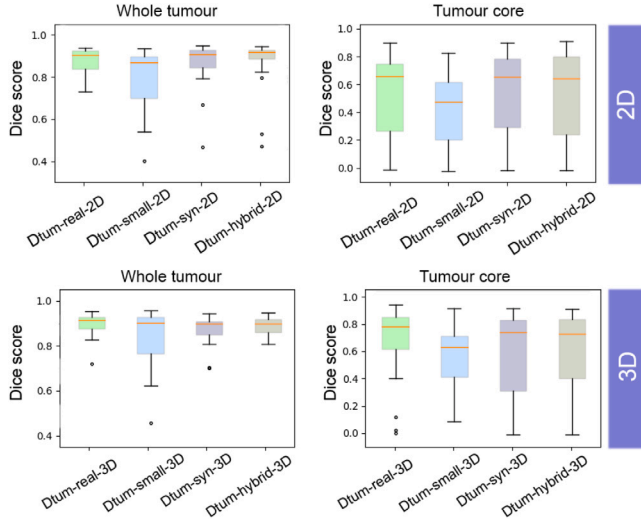


Fig. 11. Dice distributions of the 2D (top) and 3D (bottom) tumour segmentation experiments on the whole tumour (left) and core (right).

Table 9

Mean Dice scores obtained on 2D and 3D healthy region segmentation experiments with varying dataset sizes. A single asterisk indicates that the model performs significantly better (p -value < 0.05) than the rest; while double asterisks indicate that the model performs significantly better than any other model but worse than the single asterisk models. When more than one model is better, these models were not performing significantly differently between themselves.

Model	CSF	GM	WM	DGM	Brainstem
H_{2D-50K}	0.932 _{0.012} *	0.935 _{0.010} *	0.952 _{0.010} *	0.810 _{0.038} *	0.934 _{0.018} *
H_{2D-5K}	0.931 _{0.012} **	0.934 _{0.011} *	0.952 _{0.010} *	0.810 _{0.039} *	0.933 _{0.017} **
H_{2D-500}	0.927 _{0.012}	0.930 _{0.011}	0.949 _{0.010}	0.787 _{0.047}	0.922 _{0.016}
H_{2D-50}	0.925 _{0.010}	0.927 _{0.011}	0.945 _{0.012}	0.754 _{0.063}	0.896 _{0.018}
$H_{3D-1000}$	0.869 _{0.017}	0.885 _{0.010} **	0.927 _{0.006} *	0.734 _{0.034} *	0.871 _{0.014} *
H_{3D-100}	0.885 _{0.015} *	0.891 _{0.009} *	0.922 _{0.007} **	0.720 _{0.044} **	0.859 _{0.015} **
H_{3D-10}	0.873 _{0.019} **	0.883 _{0.009}	0.915 _{0.005}	0.711 _{0.027}	0.850 _{0.016}

Table 10

Mean Dice scores obtained on 2D and 3D lesion segmentation experiments with varying dataset sizes. A single asterisk indicates that the model performs significantly better (p -value < 0.05) than the rest; while double asterisks indicate that the model performs significantly better than any other model but worse than the single asterisk models. When more than one model is better, these models were not performing significantly differently between themselves.

Tumour segmentation task		
Model	Whole tumour	Tumour core
$D_{2D-50K-tum}$	0.857 _{0.105} *	0.561 _{0.256} *
$D_{2D-5K-tum}$	0.869 _{0.081} *	0.512 _{0.289} **
$D_{2D-500-tum}$	0.849 _{0.089}	0.249 _{0.194}
$D_{2D-50-tum}$	0.081 _{0.099}	0.074 _{0.125}
$D_{3D-1000-tum}$	0.859 _{0.069} *	0.577 _{0.329}
$D_{3D-100-tum}$	0.849 _{0.073} **	0.549 _{0.278}
$D_{3D-10-tum}$	0.794 _{0.089}	0.562 _{0.201}
WMH segmentation task		
Model	WMH	
$D_{2D-50K-WMH}$	0.746 _{0.074} *	
$D_{2D-5K-WMH}$	0.736 _{0.080} **	
$D_{2D-500-WMH}$	0.726 _{0.089}	
$D_{2D-50-WMH}$	0.704 _{0.108}	
$D_{3D-1000-WMH}$	0.616 _{0.116} *	
$D_{3D-100-WMH}$	0.593 _{0.130} **	
$D_{3D-10-WMH}$	0.573 _{0.130}	

5.6.5. Testing on synthetic data

So far, we have evaluated whether synthetic data can be used to train models and achieve acceptable performance on real data. In this

Table 11

Dice score and standard deviation obtained on testing models trained on real data on synthetic test datasets for healthy region segmentation (CSF, GM, WM, DGM and brainstem), tumour (whole tumour and tumour core) and WMH. Note that, because synthetic 2D data cannot be assembled into 3D volumes, the Dice scores for $H_{1D-real-2D}$, $D_{tum-real-2D}$ and $D_{WMH-real-2D}$ were calculated on 2D slices, contrarily to tests on 2D models ran for other experiments.

Healthy					
Model	CSF	GM	WM	DGM	Brainstem
$H_{1D-real-2D}$	0.942 _{0.020}	0.940 _{0.009}	0.953 _{0.017}	0.819 _{0.160}	0.793 _{0.117}
$H_{1D-real-3D}$	0.913 _{0.015}	0.905 _{0.009}	0.927 _{0.007}	0.805 _{0.014}	0.892 _{0.006}
Tumour			WMH		
Model	Dice (whole)	Dice (core)	Model	Dice (WMH)	
$D_{tum-real-2D}$	0.663 _{0.361}	0.358 _{0.394}	$D_{WMH-real-2D}$	0.383 _{0.347}	
$D_{tum-real-3D}$	0.677 _{0.205}	0.527 _{0.266}	$D_{WMH-real-3D}$	0.632 _{0.071}	

small section, we test a series of models trained on real 2D and 3D data on synthetic data, to further evaluate the quality of the synthetic label and image pairs. For this, we test the following models of the previous sections on synthetic test sets:

- $H_{1D-real-2D}$ (see Section 5.6.1), we test it on 500 2D synthetic T1 images and their CSF, GM, WM, DGM and brainstem labels.
- $H_{1D-real-3D}$ (see Section 5.6.1), we test it on 50 3D synthetic T1 images and their CSF, GM, WM, DGM and brainstem labels.
- $D_{WMH-real-2D}$ (see Section 5.6.3), we test it on 500 2D synthetic FLAIR images and their WMH lesions.
- $D_{WMH-real-3D}$ (see Section 5.6.3), we test it on 50 3D synthetic FLAIR images and their WMH lesions.
- $D_{tum-real-2D}$ (see Section 5.6.3), we test it on 500 2D synthetic T1, FLAIR and T2 images and their tumour layers segmentations.
- $D_{tum-real-3D}$ (see Section 5.6.3), we test it on 50 3D synthetic T1, FLAIR and T2 images and their tumour layers segmentations.

For 2D models, we cannot obtain Dice scores on aggregated 3D volumes, as we have done in previous experiments, because synthetic 2D labels are not part of a 3D volume. To have a reference for lesions and tumours, where volume Dice tends to make a difference over 2D slice Dice, we calculated the Dice obtained by the above-mentioned models on 2D slices from test sets of real data, yielding a Dice of 0.563_{0.297} for WMH, and 0.782_{0.321} and 0.451_{0.386} on the whole tumour and tumour core. The results of the test on synthetic data can be seen in Table 11. Visual examples of synthetic test images and segmentations are depicted in Fig. 12. Values obtained for 3D models can be compared to those reported in Tables 4, 8 and 7. Both in 3D and in 2D, we do not observe a considerable difference when it comes to segmenting healthy data. For lesions, models achieve a Dice that is 15 to 25% lower on synthetic pairs than on real pairs. Despite the observed disparity between real and synthetic data pairs – especially for lesion 3D data, which can be noted visually by looking at the images from Fig. 12 –, these results show that, overall, synthetic images are consistent with their labels.

5.6.6. Label extrapolation in 3D: combining pathologies into the same subject

One of the main motivations for having a generative model of semantic labels and images is to extrapolate to unseen phenotypes. If we take the model $D_{WMH-real-3D}$ from Section 5.6.3, and now test it on the hold-out set from BraTS used in the previous tumour segmentation experiment, we obtain a high number of false positive voxels (see column two, top two rows in Fig. 13). By leveraging our model's ability to produce paired pathological labels and images, we produce a synthetic 3D dataset containing 150 labels and images, randomising WMH lesions and tumour size conditioning. Example images and labels can be seen in the appendix. We then used different combinations of the real dataset (containing WMH lesions only) and the synthetic

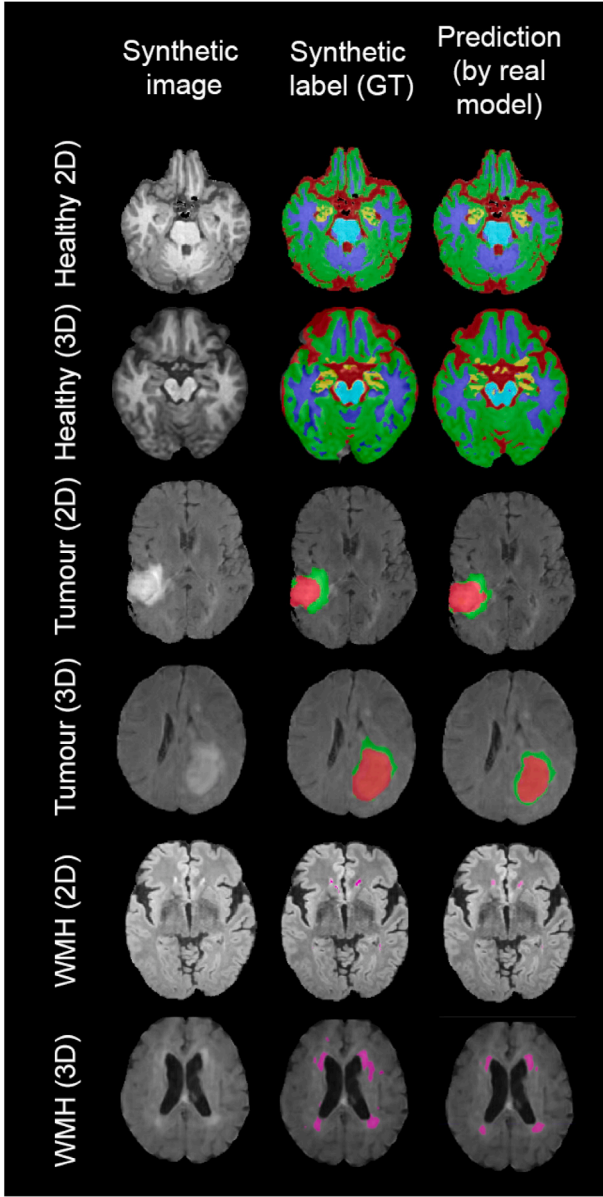


Fig. 12. Results from experiment 5.6.5: example synthetic 2D and 3D (axial slice from 3D volume) test images, ground truth synthetic labels of healthy regions – CSF (red), GM (green), WM (blue), DGM (yellow) and brainstem (light blue) – tumour core (red) and oedema (green) and WMH (magenta), and predictions obtained by models trained on real data.

one, training models on datasets $R_{R\%}S_{S\%}$, where $R\%$ is the percentage of real data and $S\%$ the percentage of synthetic data. We also train model $R_{WMH}R_{TUM}$ on our real dataset and the BraTS dataset used for training $D_{tum-real-3D}$. Because we do not have ground truth WMH segmentations for the BraTS images, we leave the labels blank.

To evaluate the performance of these WMH segmentation models, we test them on two different datasets: the hold-out SABRE dataset used in Section 5.6.3 for WMH, and the hold-out BraTS dataset used in Section 5.6.3 for tumours, BraTS-OTHER (see Table 1). For the first, we compute the Dice score with the ground truth WMH segmentations, but for the second, as we do not have these masks, we compute the number of tumour voxels mislabelled as WMH, normalised by the tumour size, FP_{tum} , given by:

$$\frac{1}{\sum_{i=0}^N I_{TUM-seg,i}} \sum_{i=0}^N I_{WMH-seg,i} \cap I_{TUM-seg,i} = 1.0 \quad (5)$$

Table 12

Performance of segmentation models trained on real, synthetic and hybrid datasets containing WMH lesions and WMH+tumour lesions. Dice is reported for the SABRE WMH hold-out set, and FP_{tum} is reported for the BraTS-OTHER dataset; *: note that $R_{100}S_0$ corresponds to model $D_{WMH-real-3D}$ from Section 5.6.3.

Training dataset	WMH Dice (on SABRE)	FP_{tum} (on BraTS)
$R_{100}S_0$	0.806 _{0.070} *	0.468 _{0.220}
$R_{75}S_{25}$	0.802 _{0.072} *	0.143 _{0.215}
$R_{50}S_{50}$	0.807 _{0.073} *	0.141 _{0.217}
$R_{25}S_{75}$	0.802 _{0.072} *	0.143 _{0.216}
R_5S_{95}	0.810 _{0.069}	0.025 _{0.078}
R_0S_{100}	0.631 _{0.0100}	0.005 _{0.017} **
$R_{WMH}R_{TUM}$	0.801 _{0.071} *	0.000 _{0.000} *

where N is the number of voxels, $I_{WMH-seg}$ is the predicted WMH segmentation, and $I_{TUM-seg}$ is the ground truth tumour segmentation.

Results are available in Table 12. We can see that, whereas using only synthetic data leads to a significant decrease in the Dice score obtained on the SABRE WMH test set, the FP_{tum} ratio decreases significantly as we add synthetic data to the real dataset, without compromising the performance on the source SABRE dataset. The best Dice and FP_{tum} results are achieved by $R_{WMH}R_{TUM}$ though, which poses the question of whether synthetic data is really necessary. Nonetheless, when we look at the segmentations themselves on SABRE and BraTS (see Fig. 13), we see that $R_{WMH}R_{TUM}$ does not segment anything for the BraTS dataset, despite the presence of WMH lesions. Because the BraTS WMH masks used for training were blank, the model learnt to map the presence of tumours to the absence of WMH. The rest of the models, however, do segment WMH lesions that are, indeed, present on the BraTS dataset. Note that R_0S_{100} performs significantly worse on SABRE than the rest; stressing, as seen in previous experiments, that a gap between real and synthetic domain prevails.

6. Discussion

In this work, we have shown that synthetic labels and images can be used to supplement existing datasets, or serve as surrogate, in a wide range of segmentation tasks. Our approach, which proposes a two-stage disease-conditioned label generator and multi-modal semantic image generator, is able to generate paired datasets in which the user can control which pathologies are present, and which modalities are generated. When used to complement real datasets in segmentation tasks, comprised of both healthy tissues and pathologies, our model improved most of the performance metrics when real data is limited, yielding results that are close to those achieved when using a large amount of real data. Seemingly, when not given any real data, the synthetic datasets can still be used to train models that had fairly good performance on real data, even though it was significantly lower than that achieved by a model trained solely on large real datasets.

6.1. 2D versus 3D

As mentioned earlier, this work is an extension of the work presented in Fernandez et al. (2022) to 3D, where the labels are 3D and the images are reconstructed from generated 2D images. Here, we also enable disease conditioning, which gives the user more control over the subject's phenotype, resulting in a more useable, versatile model. When we look at the performance of 2D models trained on synthetic 2D data (which was computed on 3D images, by assembling segmented 2D slices), we do not see a lower performance than models trained on synthetic 3D data. Therefore, one could argue whether 3D is really necessary. Other than the existing limitation of assembling 2D slices into a volume, which causes intensity inconsistencies between slices, if we have a look at individual samples generated using synthetic labels, such as those depicted in Fig. 14 or Fig. 12, we see that the visual quality of 2D images generated from 2D labels seems to be better.

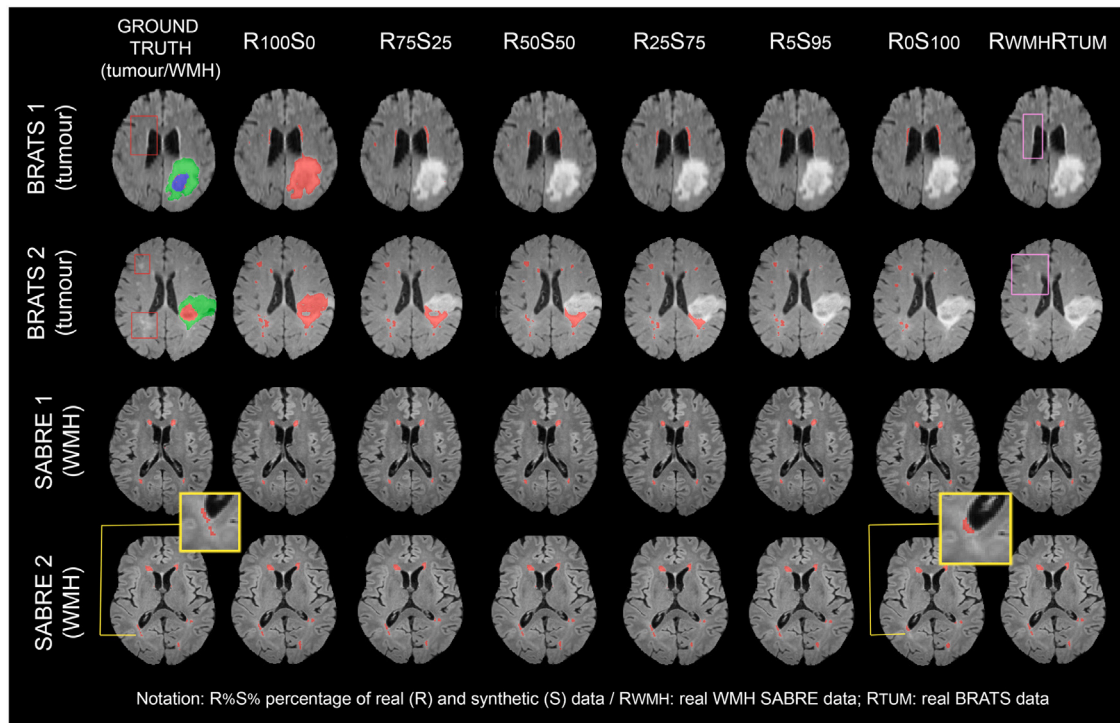


Fig. 13. Example segmentations obtained for experiment 5.6.6 on two BraTS-OTHER subjects (top two rows) and two SABRE subjects (bottom two rows), with their respective ground truth tumour (tumour sub-layers in red, green and blue)/WMH masks (in red) displayed, displayed in the first column, and predicted WMH masks in subsequent columns (in red). Yellow frames zoom on areas where R_0S_{100} undersegments WMH lesions. Red frames on the BraTS images show example WMH lesions (for which we do not have ground truth segmentations). Purple frames on the last column highlight the false negatives yielded by $R_{WMH}RTUM$.

Indeed, the slices resulting from the 3D labels appear to be less realistic than those produced by the fully 2D pipeline, which can fool a human rater. This can be explained by the availability of more data in 2D generative models (where each 3D image results in more than 200 slices), and by the possibility of having more parameters to model the label distribution. Through the enabling of data augmentation in the label generator training, and the use of techniques to fit the models to GPU, we have tried to push the model capacity and make it as generalisable as possible. Nonetheless, these come with challenges of their own. For example, augmentation can result in unrealistic generated labels, but it is necessary for 3D to avoid overfitting. This evidence of limited capacity is backed up by Fig. 2, where the 3D label generator produced a substantially lower variability across healthy tissues than that seen in real subjects.

However, 3D has advantages of its own. Even though we used conditioning in both 2D and 3D label generators, the 2D model struggled to produce them, as the presence of multiple lesions with different profiles and specific location areas does not make much sense in 2D. While we tried to mitigate this by adding a slice number conditioning, this only worked to a certain extent, with a wide range of conditioning values mapping to seemingly identical locations across the z -axis, and vice-versa. In 3D, on the other hand, conditioning works well, as shown by Fig. 3.

6.2. The synthetic-real domain gap

The segmentation experiments, as well as the visual assessment of images, show that, especially for 3D-rendered images, a clear domain gap still exists between synthetic and real domains. If we look at 3D images generated from real labels and compare them to those generated from synthetic labels, we also see a substantial difference (see Fig. 14), showing that this domain gap translates to the semantic domain as well. In addition, the presence of lesions, specifically tumour conditioning,

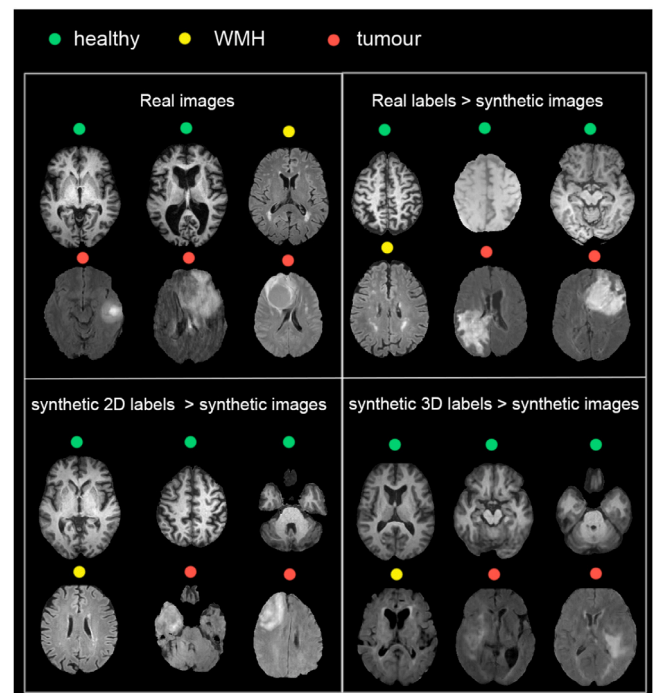


Fig. 14. Example healthy and diseased images coming from (top left): real distribution; (top right): images generated from real labels; (bottom left): images generated from synthetic 2D labels; (bottom right): images generated from synthetic 3D labels. T1 was chosen for healthy images to display better tissue contrast, and FLAIR for WMH and tumour images.

tends to deteriorate the resulting image quality, aggravating the impact of the synthetic labels on the image generator. A reason that might explain this is that BraTS is a heterogeneous dataset with a highly variable quality across the images (some of them having artifacts), and with tumours with very different appearance. This might lead to (1) the overall image generator being more unstable when tumours are present and (2) the image generator “overfitting” certain tumours, causing out-of-domain tumours to have a worse appearance. In addition, the fact that we are treating the different tumour layers as independent conditioning, and that, during sampling, we are randomising the value per layer, might lead to a combination of tumour layers that is too far from the real distribution, and cause the image generator to fail at producing a realistic-looking tumour. Future work should inspect this dependencies to achieve realistic tumour conditioning.

Indeed, it is natural that incorporating heterogeneous datasets and phenotypes deteriorates the image quality, especially in 3D, when the number of hyperparameters is much more limited. Good visual quality has been achieved in previous works in 3D, but they are typically limited to healthy subjects, specific modalities or constrained tasks (Tudosiu et al., 2022; Pinaya et al., 2022; Khader et al., 2023).

One could argue, however, whether it is actually desirable to yield images that are indistinguishable from the originals; as shown by domain randomisation methods (Billot et al., 2023), even unrealistic images can be beneficial for downstream tasks such as segmentation.

6.3. Privacy concerns

In Carlini et al. (2023), it is shown that diffusion models are more prone to memorisation than their GAN counterparts. To assess whether labels were being memorised or not, we looked at the closest neighbours of generated labels in the training dataset. We dimmed the synthetic labels sufficiently different from their closest real training counterparts to conclude that memorisation is not happening. Nonetheless, we are aware that the criteria to decide whether or not memorisation is occurring, currently, and across the literature, is subjective and non-standardised. In addition to the criteria used, the assessment method (in this case, we used Dice-based similarity) is also a choice of the author(s). Moreover, we saw that, whereas a clear difference exists in terms of healthy brain anatomy, synthetic tumours share more similarities with samples from the training dataset. This is discussed in Carlini et al. (2023): tumours, although widely present in our training set, have a more unique shape and location, and are closer to the concept of *outlier* and, therefore, more prone to be memorised. As we add pathological data to make our model more generalisable, we wonder whether our model will be able to not fall under data memorisation.

As argued in the experimental section, disentangling the semantic and contrast generation processes can help avoid memorisation, especially if different subsets of data are used to train each component. We showed, in Section 5.5, that the image encoder does not seem to be memorising the training distribution. However, entirely out-of-distribution data seems to fall within the same K-Means cluster, showcasing a different behaviour. This, nonetheless, still does not allow separate distributions. Since, nevertheless, we limited this experiment to healthy or WMH data, further work should extend this experiment to outliers.

6.4. Conclusion and further work

In this work, we propose a generative model for labels and multi-modal images, where the user has control over the lesions present and the desired contrasts. Our model is capable of providing privacy-preserving datasets that can be used in a wide range of segmentation tasks, either supplementing small, real datasets or replacing them if the end-user does not have access to any real paired image. This approach can, therefore, tackle data scarcity in medical imaging, easing the translation of segmentation models towards unseen modalities or phenotypes.

Code availability

The code for this project has been released on Github (<https://github.com/virginiafdez/brainSPADE23D.git>). Additionally to this repository, the MONAI Generative Models package must be installed from <https://github.com/Project-MONAI/GenerativeModels.git>.

CRediT authorship contribution statement

Virginia Fernandez: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Walter Hugo Lopez Pinaya:** Software, Methodology, Data curation. **Pedro Borges:** Writing – review & editing, Methodology, Data curation. **Mark S. Graham:** Writing – review & editing, Software, Data curation. **Petru-Daniel Tudosiu:** Software. **Tom Vercauteren:** Writing – review & editing, Supervision. **M. Jorge Cardoso:** Writing – review & editing, Supervision, Resources, Project administration, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Virginia Fernandez reports financial support, administrative support, and travel were provided by EPSRC Centre for Doctoral Training in Smart Medical Imaging. Tom Vercauteren reports a relationship with Hypervision Surgical that includes: board membership and equity or stocks. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Datasets used in this work are either private and shareable upon request by the owners (references in the paper) or publicly available in the cited sources. The code is on Github (link on paper).

Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Smart Medical Imaging (EP/S022104/1) at King's College London. During their involvement in this work, M.J. Cardoso, Walter Hugo Lopez Pinaya, Pedro Borges, Mark S. Graham were funded by the Wellcome Flagship Programme (WT213038/Z/18/Z); Petru-Daniel Tudosiu was funded by the AI Centre and the EPSRC Doctoral Training Partnerships (EP/R513064/1). M.J. Cardoso and Tom Vercauteren were also funded by the Wellcome EPSRC Centre for Medical Engineering (CME) (WT203148/Z/16/Z, NS/A000049/1) and the National Institute for Healthcare and Research (NIHR) GSTT. Tom Vercauteren has received funding from NIHR, Wellcome, EPSRC, United Kingdom, Medical Research Council (MRC), European Commission, Royal Academy of Engineering (RAEng), Medtronic, United States, Intel, United States and Moon Surgical in the past three years.

We gratefully acknowledge NVIDIA corporation for the donation of the GPUs that were used for this work.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health, United States Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate;

Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organisation is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Data were provided, in part, by OASIS-3: Longitudinal Multimodal Neuroimaging; Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P30 AG066444, P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

Disclosures: Dr. Tom Vercauteren is co-founder and shareholder of Hypervision Surgical Ltd.

Appendix A. Architectures

The architecture for the spatial VAE and the diffusion model used for the label generator stage are depicted in Fig. A.15 (S-VAE) and Fig. A.16 (DM). The architecture of the image generator has been implemented from and is depicted in Park et al. (2019).

Appendix B. Implementation details

B.1. Obtention of 2D datasets for training

To train the 2D label generator and the image generator, we took slices from the original 3D volumes. These slices were sampled to ensure an equal balance between disease and non-diseased slices (we consider a slice as 'diseased' if it has at least 5 pixels associated with white matter hyperintensities (WMH) or any tumour sub-category). On the other hand, since uppermost and lowermost brain slices contain

mostly non-brain tissue, which is not relevant for our downstream segmentation tasks, we calculated the ratio between brain and total voxels for each slice and discarded the ones with a percentage lower than 10%. Before discarding these slices, we associated the slice index z to the sliced images. This number was then divided by 256 (the total number of axial slices in the MNI space) to obtain the relative slice number used to condition the 2D label generator.

B.2. Label generator

The weights used for the 2D and the 3D label generator are the following:

- $\lambda_{perc} = 0.002$ (2D) and 0.001 (3D)
- $\lambda_{adv} = 0.005$ (2D) and 0
- $\lambda_{fl} = 1.0$ (2D) and 1.0 (3D)
- $\lambda_{kld} = 0.000001$ (2D) and 0.0 (2D) and 0.5 (3D)
- $\lambda_{nr} = 0.0$ (2D) and 0.5 (3D)

Adam optimiser with $\beta_1 = 0.99$ and $\beta_2 = 0.999$ was used. For the S-VAE, we used a learning rate of $5 \cdot 10^{-5}$ for the generator (2D and 3D) and of 10^{-4} for the discriminator (2D). In the 2D case, an exponential decay with $\gamma = 0.999995$ was applied. For the LDMs, a learning rate of $2.5 \cdot 10^{-4}$ was used. The variance schedule was linear, starting at 0.0015 and ending at 0.0195.

The 2D and 3D S-VAEs were trained for 200 epochs, and the 2D and 3D DMs were trained for 130 and 275 epochs respectively, on an Nvidia DGX-A100 GPU.

B.3. Image generator

The weights used for the image generator were tuned empirically, and are the following:

- λ_{adv} (adversarial loss) = 1.0
- λ_{kld} (KLD loss) = 0.0005
- λ_{feat} (feature-matching loss) = 0.25
- λ_{perc} (perceptual loss) = 1.5
- $\lambda_{contrastive}$ (contrastive loss on the latent space) = 1.0
- $\lambda_{modisc-mod}$ (modality-discriminator loss) = 0.75
- $\lambda_{modisc-dat}$ (dataset-discriminator loss) = 0.3

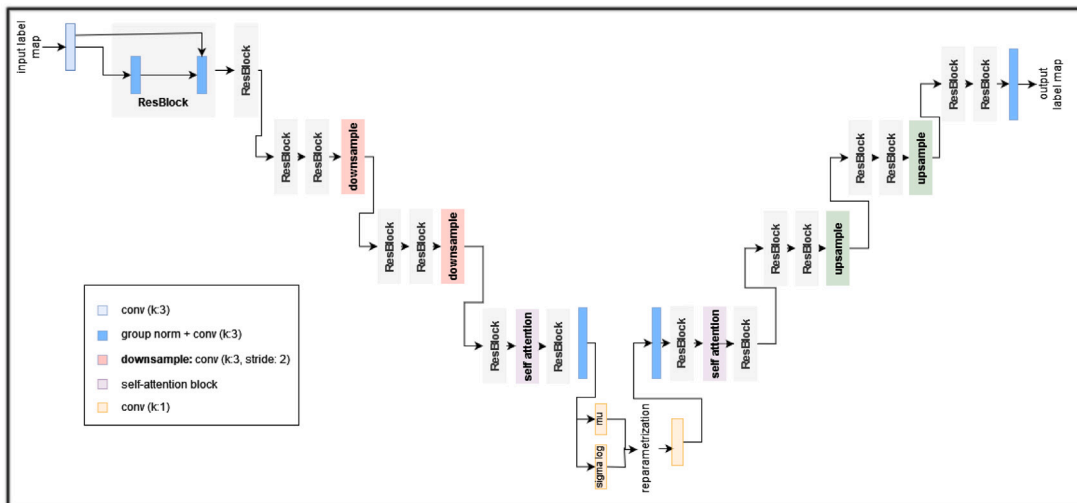


Fig. A.15. Architecture of the spatial VAE used to encode-decode the semantic label maps and bring them to a latent space.

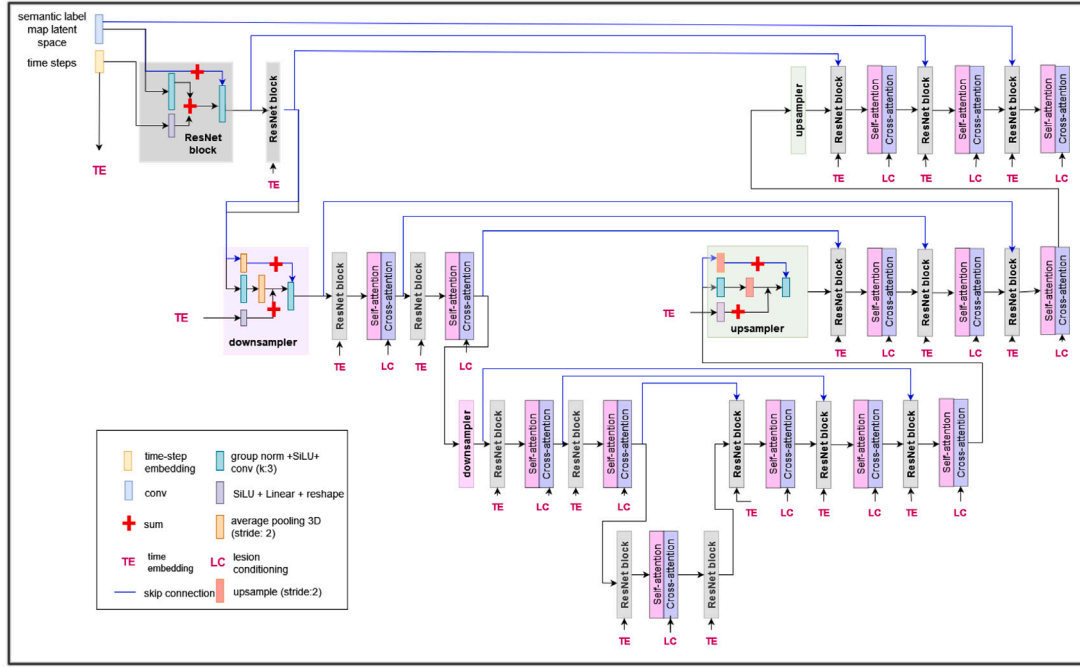


Fig. A.16. Architecture of the time-gated residual U-Net used to implement the diffusion model.

Adam optimiser with $\beta_1 = 0.5$ and $\beta_2 = 0.99$ was used. We trained the models on an Nvidia RTX8000 GPU for 200 epochs with a fixed learning rate of $2 \cdot 10^{-4}$ and then applied a linear decay for 50 more epochs. The training dataset size consisted of 18 000 slices.

B.4. Modality and dataset discriminator network

To train the modality and dataset classifier used to yield the *modisc* loss used for the image generator, we used a Densenet121 architectures with two parallel output layers, one to classify the modality (T1, FLAIR or T2) and the other one to classify the dataset (SABRE, BraTs or ADNI). These modalities and datasets were one-hot encoded. Binary-cross-entropy with logits (BCE with logits) was used to train this network, summing the BCE for each of the tasks, using Adam optimizer and an exponential decaying learning rate going from 0.01 to 0.0001 across 50 epochs. The data loading pipeline was equal to that used for the main image generator training. The final accuracy for modalities was 91.61% (AUC 0.986) and that for datasets was 78.33% (AUC 0.823). The network weights were frozen for the *modisc* loss calculation.

Appendix C. Ablation study

The ablation study consisted of training models on the following loss functions:

1. Original SPADE losses
2. Original SPADE losses + unpaired label and image slices
3. Original SPADE losses + modality and classifier discriminator loss
4. Original SPADE losses + contrastive loss
5. Original SPADE losses + unpaired label and image slices + modality and classifier discriminator loss + contrastive loss

Because we aim for an increased match between the contrasts of generated neighbouring slices, we evaluated the performance of these models on 12 3D volumes, reconstructed by forward-passing the segmentation slices through them. For each volume, we fixed the input

style image to the central slice of the image volume. We calculated the SSIM between neighbouring slices for all the slice pairs and volumes, and normalised them by the SSIM between neighbouring slices in the ground truth volumes. We assessed the reconstructed volumes visually, and we performed a statistical analysis of the SSIM values using paired one-sided T-tests with Bonferroni correction.

The paired statistical one-sided t-tests revealed that model 4 yields higher significant SSIM ratio than the others (p -value < 0.05 for all pairs). Model 2 is also significantly better than 5, 3 and 1 (p -value < 0.05) for all pairs. No statistical improvement is shown for the model using all the losses.

Visual examples for FLAIR and T2 contrasts can be seen in supplementary Figs. C.17 and C.18.

Appendix D. Quality of the generated labels

Tissue proportions: To get an intuition as to why we observed some major discrepancies between the average number of voxels per tissue in the labels sampled from the 3D label generator, compared to real labels, we drew some images and compared them side by side. Fig. D.19 shows a real and synthetic label.

Conditioning in 2D: As explained in the main results, we could not quantify or observe the effect of increasing the lesion conditioning in 2D as well as we did in 3D, due to the interplay between the slice and lesion conditionings. Fig. D.20 shows the lesion proportion per lesion and slice conditioning value. Whereas the tumour plots are rather noisy, we can see a pattern in WMH where each lesion value plot has a peak at a specific slice, with that slice number increasing as the lesion value increases. This points towards the idea that the lesion conditioning only works when the combination of slice and WMH conditioning matches that observed in the training dataset. For tumours, we do not see these patterns because they do not have an approximate constant appearance per slice as do WMH: their location and size varies across subjects. If we look at the visual examples, we clearly see that the slice conditioning, which worked in a healthy setting, as shown by Fig. 2, does not work for lesion conditioning values that are unlikely to happen. Overall, this

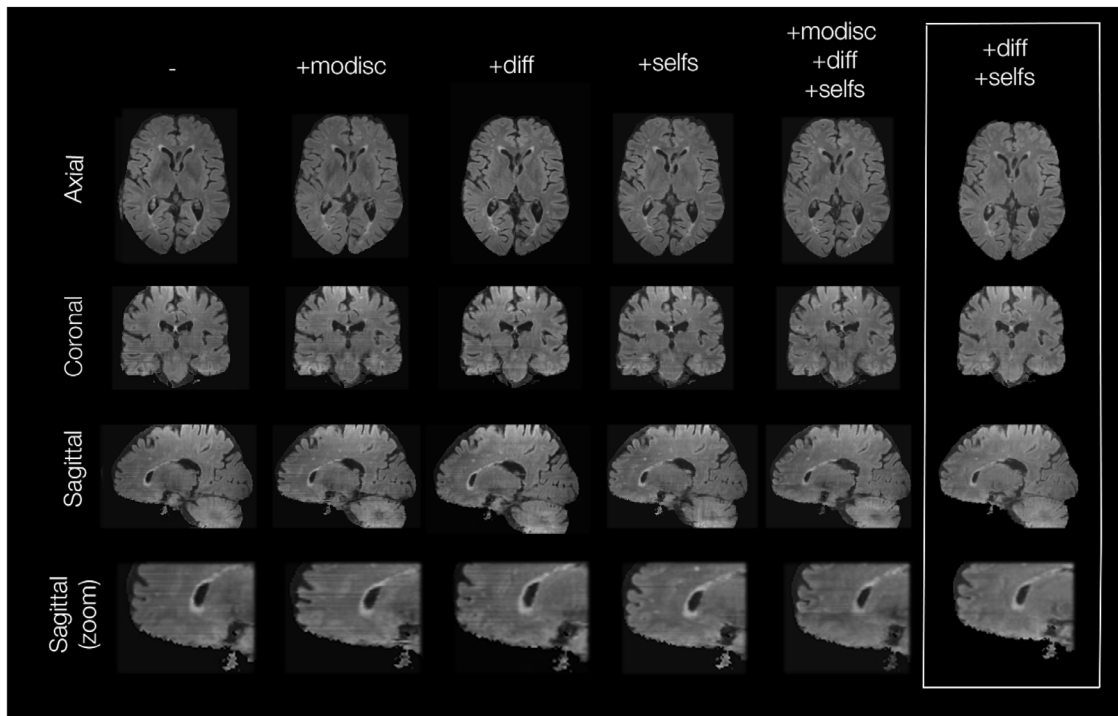


Fig. C.17. Example reconstructed FLAIR volumes. From top to bottom: axial, coronal, sagittal slices and zoom on the cerebellum of the sagittal slices of the volumes reconstructed from the stack of synthetic T1 slices. From left to right: models using no additional losses (original SPADE network), $L_{mod-disc}$ loss only, unpaired style and semantic slices, L_{cont} loss only, all losses and unpaired style and semantic slices, and the statistically best one, using L_{cont} and unpaired style and semantics slices.

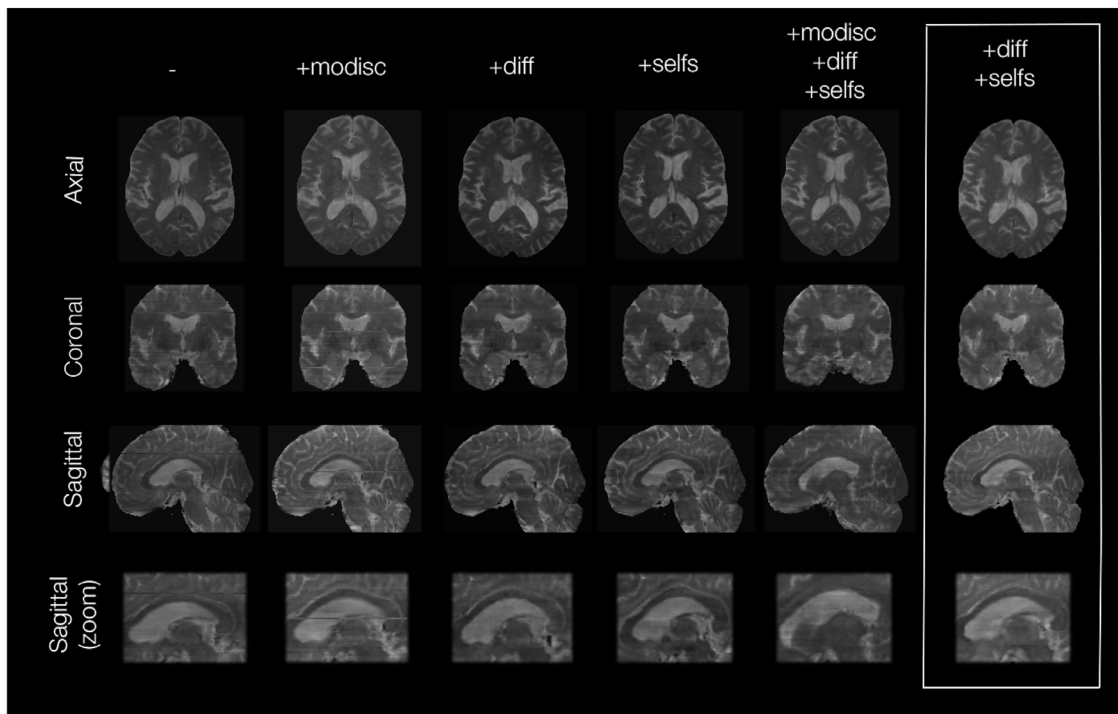


Fig. C.18. Example reconstructed T2 volumes. From top to bottom: axial, coronal, sagittal slices and zoom on the cerebellum of the sagittal slices of the volumes reconstructed from the stack of synthetic T1 slices. From left to right: models using no additional losses (original SPADE network), $L_{mod-disc}$ loss only, unpaired style and semantic slices, L_{cont} loss only, all losses and unpaired style and semantic slices, and the statistically best one, using L_{cont} and unpaired style and semantics slices.

proves that lesion extrapolation does not work in 2D. If we were to use a lesion and slice conditioned-model, these two elements should be

disentangled to ensure that they work independently to achieve good results.

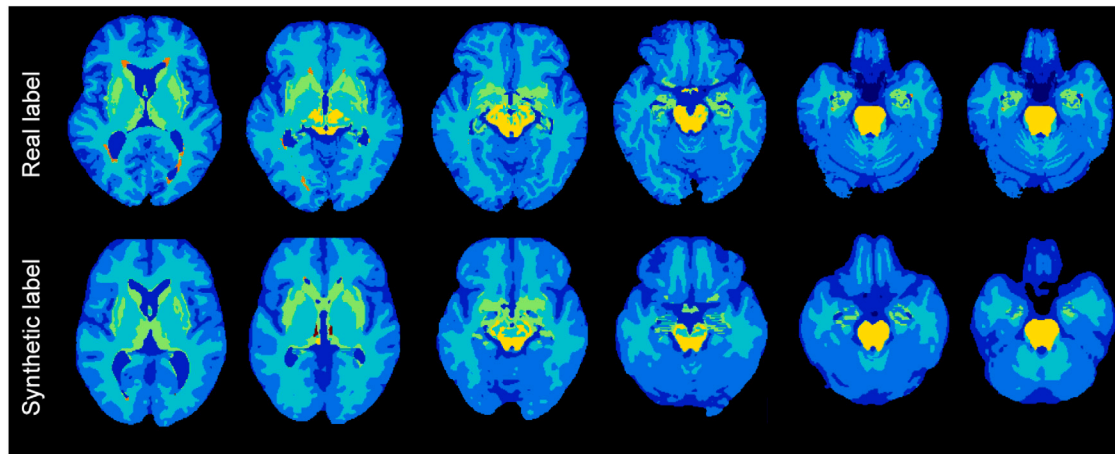


Fig. D.19. Set of axial slices from the argmaxed probabilistic map of a real (top) and synthetic label (bottom).

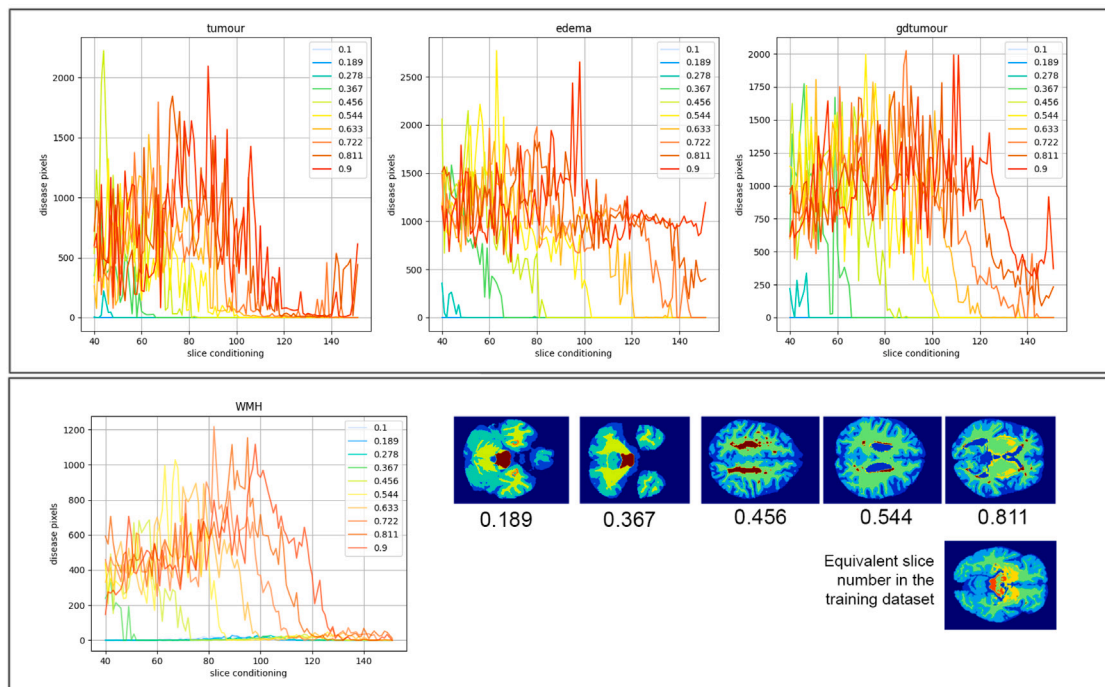


Fig. D.20. Proportion of lesion pixels as a function of slice conditioning and lesion conditioning values for tumour layers (top) and WMH (bottom). For tumour, each sample had the same value per tumour layer (i.e.: all three, oedema, non-enhancing and GD-enhancing tumour were set to 0.1, 0.189 etc.). On the bottom, example WMH conditioned images for a specific slice number which are displayed along with their WMH conditioning value. Below, an example of that slice number from the training dataset is displayed.

References

- Anwar, S.M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., Khan, M.K., 2018. Medical Image Analysis using Convolutional Neural Networks: A Review. *J. Med. Syst.* 42 (11), 1–13. <http://dx.doi.org/10.1007/S10916-018-1088-1/TABLES/4>, arXiv:1709.02250, URL <https://link.springer.com/article/10.1007/s10916-018-1088-1>.
- Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J., 2023. Synthetic Data from Diffusion Models Improves ImageNet Classification. arXiv:2304.08466, URL <https://arxiv.org/abs/2304.08466v1>.
- Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E., 2023. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Med. Image Anal.* 86, 102789. <http://dx.doi.org/10.1016/J.MEDIA.2023.102789>, arXiv:2107.09559.
- Cardoso, M.J., Wolz, R., Modat, M., Fox, N.C., Rueckert, D., Ourselin, S., 2012. Geodesic information flows. In: Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention. Vol. 15, pp. 262–270. http://dx.doi.org/10.1007/978-3-642-33418-4_33, Pt 2.
- Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Zurich, E., Balle, B., Ippolito, D., Wallace, E., Berkeley, U., 2023. Extracting Training Data from Diffusion Models. In: Proceedings of the 32nd USENIX Security Symposium. URL <https://www.usenix.org/conference/usenixsecurity23/presentation/carlini>.
- Chambon, P., Bluethgen, C., Delbrouck, J.-B., Van der Sluijs, R., Polacin, M., Chaves, J.M.Z., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A., 2022. RoentGen: Vision-Language Foundation Model for Chest X-ray Generation. arXiv:2211.12737, URL <https://arxiv.org/abs/2211.12737v1>.
- Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709.
- Chen, S., Ma, K., Zheng, Y., 2019. Med 3d: transfer learning for 3d medical image analysis. arXiv:1904.00625v4, URL <https://github.com/Tencent/MedicalNet>.
- Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A., 2021. A review of medical image data augmentation techniques for deep learning applications. *J. Med. Imaging Radiat. Oncol.* URL <https://doi.org/10.1111/1754-9485.13261>.
- Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D.A., Gallagher, L., Kennedy, D.P., Keown, C.L., Keyser, C., Lainhart, J.E., Lord, C., Luna, B., Menon, V., Minshew, N.J., Monk, C.S., Mueller, S., Müller, R.A., Nebel, M.B., Nigg, J.T., O'Hearn, K., Pelphrey, K.A., Peltier, S.J., Rudie, J.D., Sunaert, S., Thioux, M., Tyszka, J.M., Uddin, L.Q., Verhoeven, J.S.,

- Wenderoth, N., Wiggins, J.L., Mostofsky, S.H., Milham, M.P., 2014. The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* <http://dx.doi.org/10.1038/mp.2013.78>.
- Dorent, R., Booth, T., Li, W., Sudre, C.H., Kafiabadi, S., Cardoso, J., Ourselin, S., Vercauteren, T., 2021. Learning joint segmentation of tissues and brain lesions from task-specific hetero-modal domain-shifted datasets. *Med. Image Anal.* 67, 101862. <http://dx.doi.org/10.1016/j.media.2020.101862>, URL <https://www.sciencedirect.com/science/article/pii/S1361841520302267>.
- Dorjsembe, Z., Pao, H.-K., Odonchimed, S., Xiao, F., 2023. Conditional Diffusion Models for Semantic 3D Brain MRI Synthesis. *arXiv:2305.18453v4*, URL <http://arxiv.org/abs/2305.18453>.
- ESR, E.S.o.R., 2017. The new EU General Data Protection Regulation: what the radiologist should know. *Insights Imaging* 8 (3), 295–299, URL <https://doi.org/10.1007/s13244-017-0552-7>.
- Fernandez, V., Pinaya, W.H.L., Borges, P., Tudosiu, P.-D., Graham, M.S., Vercauteren, T., Cardoso, M.J., 2022. 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*. Springer International Publishing, Cham, pp. 79–90.
- Goodfellow, I.J., Bengio, Y., Courville, A., 2015. Deep Learning Book. Deep Learning <http://dx.doi.org/10.1016/B978-0-12-391420-0.09987-X>, arXiv:arXiv:1011.1669v3.
- Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS '14, MIT Press, Cambridge, MA, USA, pp. 2672–2680.
- Guan, H., Liu, M., 2022. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Trans. Biomed. Eng.* 69, 1173–1185.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Adv. Neural Inf. Process. Syst.* 2017-December, 6627–6638. <http://dx.doi.org/10.18034/ajase.v8i1.9>, arXiv:1706.08500, URL <https://arxiv.org/abs/1706.08500v6>.
- Ho, J., Jain, A., Abbeel, P., 2020. Denoising Diffusion Probabilistic Models. *Neural Inf. Process. Syst.*
- Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18 (2), 203–211, URL <https://doi.org/10.1038/s41592-020-01008-z>.
- Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), *Computer Vision – ECCV 2016*. Springer International Publishing, Cham, pp. 694–711.
- Jones, S., Tillin, T., Park, C., Williams, S., Rapala, A., Al Saikhan, L., Eastwood, S.V., Richards, M., Hughes, A.D., Chaturvedi, N., 2020. 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. *Int. J. Epidemiol.* 49 (5), 1441–1442E. <http://dx.doi.org/10.1093/IJE/DYAA135>, URL <https://pubmed.ncbi.nlm.nih.gov/33049759/>.
- Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F., 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* 2 (6), 305–311. <http://dx.doi.org/10.1038/s42256-020-0186-1>, 2020 2:6. URL <https://www.nature.com/articles/s42256-020-0186-1>.
- Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hachililoglu, I., Merhof, D., 2023. Diffusion models in medical imaging: A comprehensive survey. *Med. Image Anal.* 88, 102846. <http://dx.doi.org/10.1016/j.media.2023.102846>.
- Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarbuerger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baefler, B., Foersch, S., Stegmaier, J., Kuhl, C., Nebelung, S., Kather, J.N., Truhn, D., 2023. Denoising diffusion probabilistic models for 3D medical image generation. *Sci. Rep.* 13 (1), 1–12. <http://dx.doi.org/10.1038/s41598-023-34341-2>, 2023 13:1. URL <https://www.nature.com/articles/s41598-023-34341-2>.
- Kingma, D.P., Welling, M., 2014. Auto-Encoding Variational Bayes. *CoRR*, arXiv:1312.6.
- LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., Raichle, M.E., Cruchaga, C., Marcus, D., 2019. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. <http://dx.doi.org/10.1101/2019.12.13.19014902>, medRxiv, URL: <https://www.medrxiv.org/content/10.1101/2019.12.13.19014902v1> <https://www.medrxiv.org/content/10.1101/2019.12.13.19014902v1.abstract>.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2018. Focal Loss for Dense Object Detection. *arXiv:1708.02002*.
- Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., Wang, Y., Greenspan, H., Deyer, T., Fayad, Z.A., Yang, Y., 2022. RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. *Radiol. Artif. Intell.* e210315. <http://dx.doi.org/10.1148/ryai.210315>.
- Motamed, S., Rogalla, P., Khalvati, F., 2021. Data augmentation using Generative Adversarial Networks (GANs) for GAN-based detection of Pneumonia and COVID-19 in chest X-ray images. *Inform. Med. Unlocked* 27, 100779. <http://dx.doi.org/10.1016/j.imu.2021.100779>, URL <https://www.sciencedirect.com/science/article/pii/S2352914821002501>.
- van den Oord, A., Vinyals, O., Kavukcuoglu, K., 2017. Neural Discrete Representation Learning. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS '17*, Curran Associates Inc., Red Hook, NY, USA, pp. 6309–6318.
- Osuala, R., Skorupko, G., Lazrak, N., Garrucho, L., García, E., Joshi, S., Jouide, S., Rutherford, M., Prior, F., Kushibar, K., Diaz, O., Lekadir, K., 2022. medigan: a Python library of pretrained generative models for medical image synthesis. *J. Med. Imaging* 10 (06), arXiv:2209.14472v2, URL: <http://arxiv.org/abs/2209.14472> <http://dx.doi.org/10.1117/1.JMI.10.6.061403>.
- Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y., 2019. Semantic image synthesis with spatially-adaptive normalization. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2019-June, pp. 2332–2341. <http://dx.doi.org/10.1109/CVPR.2019.00244>, arXiv:1903.07291.
- Pinaya, W.H.L., Graham, M.S., Kerfoot, E., Tudosiu, P.-D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., da Costa, P.F., Patel, A., Chung, H., Zhao, C., Peng, W., Liu, Z., Mei, X., Lucena, O., Ye, J.C., Tsafaris, S.A., Dogra, P., Peng, A., Modat, M., Nachev, P., Ourselin, S., Cardoso, M.J., 2023. Generative AI for Medical Imaging: extending the MONAI Framework. *arXiv:2307.15208*, URL <https://arxiv.org/abs/2307.15208v1>.
- Pinaya, W.H.L., Tudosiu, P.-D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J., 2022. Brain Imaging Generation with Latent Diffusion Models. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D., Yuan, Y. (Eds.), *Deep Generative Models*. Springer Nature Switzerland, Cham, pp. 117–126.
- Qasim, A.B., Ezhov, I., Shit, S., Schoppe, O., Paetzold, J.C., Sekuboyina, A., Kofler, F., Lipkova, J., Li, H., Menze, B., 2020. Red-GAN: Attacking class imbalance via conditioned generation. Yet another medical imaging perspective.. URL <https://proceedings.mlr.press/v121/qasim20a.html>.
- Renard, F., Guedria, S., Palma, N.D., Vuillerme, N., 2020. Variability and reproducibility in deep learning for medical image segmentation. *Sci. Rep.* 10 (1), 1–16. <http://dx.doi.org/10.1038/s41598-020-69920-0>, 2020 10:1. URL <https://www.nature.com/articles/s41598-020-69920-0>.
- Ricci Lara, M.A., Echeveste, R., Ferrante, E., 2022. Addressing fairness in artificial intelligence for medical imaging. *Nat. Commun.* 13 (1), 1–6. <http://dx.doi.org/10.1038/s41467-022-32186-3>, 2022 13:1. URL <https://www.nature.com/articles/s41467-022-32186-3>.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv:2112.10752*.
- Rusak, F., Cruz, R.S., Bourgeat, P., Fookes, C., Frapp, J., Bradley, A., Salvado, O., 2020. 3D brain MRI GAN-based synthesis conditioned on partial volume maps. In: Burgos, N., Svoboda, D., Wolterink, J.M., Zhao, C. (Eds.), *Simulation and Synthesis in Medical Imaging: 5th International Workshop, SASHIMI 2020, Held in Conjunction with MICCAI 2020, Proceedings*. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, Switzerland, pp. 11–20. http://dx.doi.org/10.1007/978-3-030-59520-3_2, URL <https://eprints.qut.edu.au/205579/>.
- Song, J., Meng, C., Ermon, S., 2020. Denoising Diffusion Implicit Models. *ICLR 2021 - 9th International Conference on Learning Representations*. International Conference on Learning Representations, ICLR, *arXiv:2010.02502*, URL <https://arxiv.org/abs/2010.02502v4>.
- Stojanovski, D., Hermida, U., Lamata, P., Beqiri, A., Gomez, A., 2023. Echo from Noise: Synthetic Ultrasound Image Generation Using Diffusion Models for Real Image Segmentation. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14337 LNCS, Springer Science and Business Media Deutschland GmbH, pp. 34–43. http://dx.doi.org/10.1007/978-3-031-44521-7_4/FIGURES/5, URL https://link.springer.com/chapter/10.1007/978-3-031-44521-7_4.
- Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., Batmanghelich, K., 2022. Hierarchical amortized GAN for 3D high resolution medical image synthesis. *IEEE J. Biomed. Health Inf.* 26 (8), 3966–3975. <http://dx.doi.org/10.1109/JBHI.2022.3172976>.
- Sun, H., Zhu, T., Zhang, Z., Jin, D., Xiong, P., Zhou, W., 2021. Adversarial Attacks Against Deep Generative Models on Data: A Survey. *IEEE Trans. Knowl. Data Eng.* 35 (4), 3367–3388. <http://dx.doi.org/10.1109/TKDE.2021.3130903>, arXiv:2112.00247.
- 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., 2022. 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*. Springer International Publishing, Cham, pp. 66–78.
- Zhao, S., Song, J., Ermon, S., 2017. Towards Deeper Understanding of Variational Autoencoding Models. *arXiv:1702.08658*, URL <https://arxiv.org/abs/1702.08658v1>.