Medical Image Analysis 86 (2023) 102789

Contents lists available at ScienceDirect

Medical Image Analysis

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

SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining

Benjamin Billot [a,*], Douglas N. Greve [b], Oula Puonti [c], Axel Thielscher [c,d], Koen Van Leemput [b,d], Bruce Fischl [b,e,f], Adrian V. Dalca [b,e], Juan Eugenio Iglesias [a,b,e], for the ADNI[1]

[a] Centre for Medical Image Computing, University College London, UK
[b] Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, USA
[c] Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital, Denmark
[d] Department of Health Technology, Technical University of, Denmark
[e] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA
[f] Program in Health Sciences and Technology, Massachusetts Institute of Technology, USA

ARTICLE INFO

Keywords:
Domain randomisation
Contrast and resolution invariance
Segmentation
CNN

ABSTRACT

Despite advances in data augmentation and transfer learning, convolutional neural networks (CNNs) difficultly generalise to unseen domains. When segmenting brain scans, CNNs are highly sensitive to changes in resolution and contrast: even within the same MRI modality, performance can decrease across datasets. Here we introduce SynthSeg, the first segmentation CNN robust against changes in contrast and resolution. SynthSeg is trained with synthetic data sampled from a generative model conditioned on segmentations. Crucially, we adopt a *domain randomisation* strategy where we fully randomise the contrast and resolution of the synthetic training data. Consequently, SynthSeg can segment real scans from a wide range of target domains without retraining or fine-tuning, which enables straightforward analysis of huge amounts of heterogeneous clinical data. Because SynthSeg only requires segmentations to be trained (no images), it can learn from labels obtained by automated methods on diverse populations (e.g., ageing and diseased), thus achieving robustness to a wide range of morphological variability. We demonstrate SynthSeg on 5,000 scans of six modalities (including CT) and ten resolutions, where it exhibits unparallelled generalisation compared with supervised CNNs, state-of-the-art domain adaptation, and Bayesian segmentation. Finally, we demonstrate the generalisability of SynthSeg by applying it to cardiac MRI and CT scans.

1. Introduction

1.1. Motivation

Segmentation of brain scans is of paramount importance in neuroimaging, as it enables volumetric and shape analyses (Hynd et al., 1991). Although manual delineation is considered the gold standard in segmentation, this procedure is tedious and costly, thus preventing the analysis of large datasets. Moreover, manual segmentation of brain scans requires expertise in neuroanatomy, which, even if available, suffers from severe inter- and intra-rater variability issues (Warfield et al., 2004). For these reasons, automated segmentation methods have been proposed as a fast and reproducible alternative solution.

Most recent automated segmentation methods rely on convolutional neural networks (CNNs) (Ronneberger et al., 2015; Milletari et al., 2016; Kamnitsas et al., 2017b). These are widespread in research, where the abundance of high quality scans (i.e., at high isotropic resolution and with good contrasts between tissues) enables CNNs to obtain accurate 3D segmentations that can then be used in subsequent analyses such as connectivity study (Müller et al., 2011).

However, supervised CNNs are far less employed in clinical settings, where physicians prefer 2D acquisitions with a sparse set of high-resolution slices, which enables faster inspection under time constraints. This leads to a huge variability in image orientation (axial, coronal, or sagittal), slice spacing, and in-plane resolution. Moreover, such 2D scans often use thick slices to increase the signal-to-noise

* Corresponding author.
E-mail address: benjamin.billot.18@ucl.ac.uk (B. Billot).

[1] Data used in this article are partly from the Alzheimer's Disease Neuroimaging Initiative database (http://adni.loni.usc.edu). Investigators in the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis of this report. A complete listing of investigators at: adni.loni.usc.edu/wp-content/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.media.2023.102789
Received 26 February 2022; Received in revised form 20 January 2023; Accepted 22 February 2023
Available online 25 February 2023

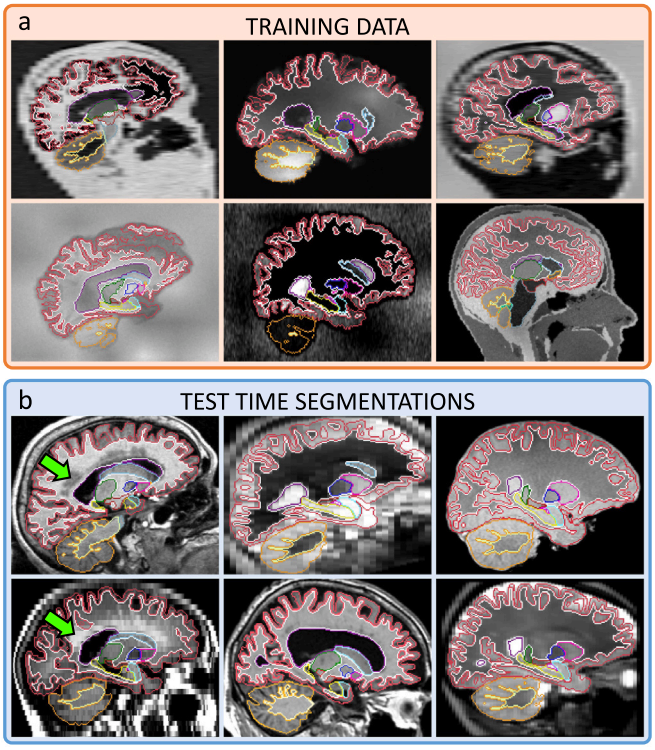

Fig. 1. (a) Representative samples of the synthetic 3D scans used to train SynthSeg for brain segmentation, and contours of the corresponding ground truth. (b) Test-time segmentations for a variety of contrasts and resolutions, on subjects spanning a wide age range, some presenting large atrophy and white matter lesions (green arrows). All segmentations are obtained with the same network, without retraining or fine-tuning.

ratio, thus introducing considerable partial voluming (PV). This effect arises when several tissue types are mixed within the same voxel, resulting in averaged intensities that are not necessarily representative of underlying tissues, often causing segmentation methods to underperform (Van Leemput et al., 2003). Additionally, imaging protocols also span a huge diversity in sequences and modalities, each studying different tissue properties, and the resulting variations in intensity distributions drastically decrease the accuracy of supervised CNNs (Chen et al., 2019).

Overall, the lack of tools that can cope with the large variability in MR data hinders the adoption of quantitative morphometry in the clinic. Moreover, it precludes the analysis of vast amounts of clinical scans, currently left unexplored in picture archiving and communication systems (PACS) in hospitals around the world. The ability to derive morphometric measurements from these scans would enable neuroimaging studies with sample sizes in the millions, and thus much higher statistical power than current research studies. Therefore, there is a clear need for a fast, accurate, and reproducible automated method, for segmentation of brain scans of any contrast and resolution, and that can adapt to a wide range of populations.

1.2. Contributions

In this article, we present SynthSeg, the first neural network to segment brain scans of a wide range of contrasts and resolutions, without having to be retrained or fine-tuned (Fig. 1). Specifically, SynthSeg is trained with synthetic scans sampled on the fly from a generative model inspired by the Bayesian segmentation framework, and is thus never exposed to real scans during training. Our main contribution is the adoption of a domain *randomisation strategy* (Tobin et al., 2017), where all the parameters of the generative model (including orientation, contrast, resolution, artefacts) are fully randomised. This exposes the network to vastly different examples at each mini-batch, and thus forces it to learn domain-independent features. Moreover, we apply a random subset of common preprocessing operations to each example (e.g., skull stripping, bias field correction), such that SynthSeg can segment scans with or without preprocessing.

With this domain randomisation strategy, our method only needs to be trained once. This is a considerable improvement over supervised CNNs and domain adaptation strategies, which all need retraining or fine-tuning for each new contrast or resolution, thus hindering clinical applications. Moreover, training SynthSeg is greatly facilitated by the fact that it only requires a set of anatomical label maps to be trained (and no real images, since all training scans are synthetic). Furthermore, these maps can be obtained automatically (rather than manually), since the training scans are directly generated from their ground truths, and are thus perfectly aligned with them. This enables us to greatly improve the robustness of SynthSeg by including automated training maps from highly diverse populations.

Overall, SynthSeg yields almost the accuracy of supervised CNNs on their training domain, but unlike them, exhibits a remarkable generalisation ability. Indeed, SynthSeg consistently outperforms state-of-the-art domain adaptation strategies and Bayesian segmentation on all tested datasets. Moreover, we demonstrate the generalisability of SynthSeg by obtaining state-of-the-art results in cross-modality cardiac segmentation.

This work extends our recent articles on contrast-adaptiveness (Billot et al., 2020a) and PV simulation at a specific resolution (Billot et al., 2020b; Iglesias et al., 2021), by building, for the first time, robustness to both contrast and resolution without retraining. Our method is thoroughly evaluated in four new experiments. The code and trained model are available at https://github.com/BBillot/SynthSeg as well as in the widespread neuroimaging package FreeSurfer (Fischl, 2012).

2. Related works

Contrast-invariance in brain segmentation has traditionally been addressed with Bayesian segmentation. This technique is based on a generative model, which combines an anatomical prior (often a statistical atlas) and an intensity likelihood (typically a Gaussian Mixture Model, GMM). Scans are then segmented by "inverting" this model with Bayesian inference (Wells et al., 1996; Fischl et al., 2002). Contrast-robustness is achieved by using an unsupervised likelihood model, with parameters estimated on each test scan (Van Leemput et al., 1999; Ashburner and Friston, 2005). However, Bayesian segmentation requires approximately 15 min per scan (Puonti et al., 2016), which precludes its use in time-sensitive settings. Additionally, its accuracy is limited at low resolution (LR) by PV effects (Choi et al., 1991). Indeed, even if Bayesian methods can easily model PV (Van Leemput et al., 2003), inferring high resolution (HR) segmentations from LR scans quickly becomes intractable, as it requires marginalising over all possible configurations of HR labels within each LR supervoxel. While simplifications can be made (Van Leemput et al., 2003), PV-aware Bayesian segmentation may still be infeasible in clinical settings.

Supervised CNNs prevail in recent medical image segmentation (Milletari et al., 2016; Kamnitsas et al., 2017b), and are best represented by the UNet architecture (Ronneberger et al., 2015). While these networks obtain fast and accurate results on their training domain, they do not generalise well to unseen contrasts (Karani et al., 2018) and resolutions (Ghafoorian et al., 2017), an issue known as the "domain-gap" problem (Pan and Yang, 2010). Therefore, such networks need to be retrained for any new combination of contrast and resolution, often requiring new costly labelled data. This problem can partly be ameliorated by training on multi-modality scans with modality dropout (Havaei et al., 2016), which results in a network able to individually segment each training modality, but that still cannot be applied to unseen domains.

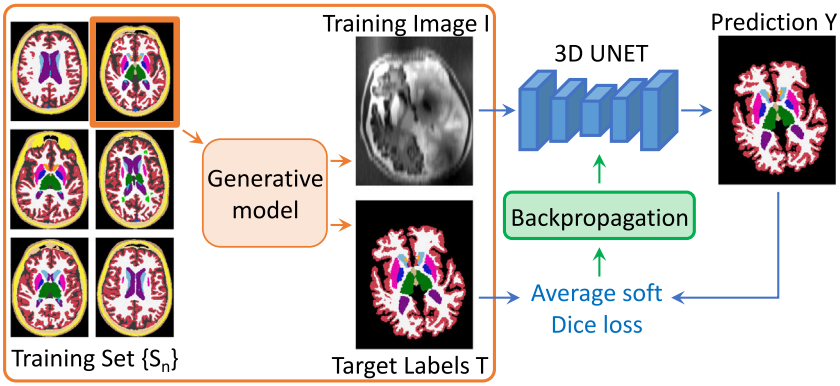

Fig. 2. Overview of a training step. At each mini-batch, we randomly select a 3D label map from a training set $\{S_n\}$ and sample a pair $\{I, T\}$ from the generative model. The obtained image is then run through the network, and its prediction Y is used to compute the average soft Dice loss, that is backpropagated to update the weights of the network.

Data augmentation improves the robustness of CNNs by applying simple spatial and intensity transforms to the training data (Zhang et al., 2020). While such transforms often relies on handcrafted (and thus suboptimal) parameters, recent semi-supervised methods, such as adversarial augmentation, explicitly optimise the augmentation parameters during training (Zhang et al., 2017; Chaitanya et al., 2019; Chen et al., 2022). Alternatively, contrastive learning methods have been proposed to leverage unsupervised data for improved generalisation ability (Chaitanya et al., 2020; You et al., 2022b). Overall, although these techniques improve generalisation in intra-modality applications (Zhao et al., 2019), they generally remain insufficient in cross-modality settings (Karani et al., 2018).

Domain adaptation explicitly seeks to bridge a given domain gap between a source domain with labelled data, and a specific target domain without labels. A first solution is to map both domains to a common latent space, where a classifier can be trained (Kamnitsas et al., 2017a; Dou et al., 2019; Ganin et al., 2017; You et al., 2022a). In comparison, generative adaptation methods seek to match the source images to the target domain with image-to-image translation methods (Sandfort et al., 2019; Huo et al., 2019; Zhang et al., 2018). Since these approaches are complementary, recent methods propose to operate in both feature and image space, which leads to state-to-the-art results in cross-modality segmentation (Chen et al., 2019; Hoffman et al., 2018). In contrast, state-of-the-art results in intra-modality adaptation are obtained with test-time adaptation methods (Karani et al., 2021; He et al., 2021), which rely on light fine-tuning at test-time. More generally, even though domain adaptation alleviates the need for supervision in the target domain, it still needs retraining for each new domain.

Synthetic training data can be used to increase robustness by introducing surrogate domain variations, either generated with physics-based models (Jog and Fischl, 2018), or adversarial generative networks (Frid-Adar et al., 2018; Chartsias et al., 2018), possibly conditioned on label maps for improved semantic content (Mahmood et al., 2020; Isola et al., 2017). These strategies enable to generate huge training datasets with perfect ground truth obtained by construction rather than human annotation (Richter et al., 2016). However, although generated images may look remarkably realistic, they still suffer from a "reality gap" (Jakobi et al., 1995). In addition, these methods still require retraining for every new domain, and thus do not solve the lack of generalisation of neural networks. To the best of our knowledge, no current learning method can segment medical scans of any contrast and/or resolution without retraining.

Domain randomisation is a recent strategy that relies on physics-based generative models, which, unlike learning-based methods, offer full control over the generation process. Instead of handcrafting (Jog and Fischl, 2018) or optimising (Chen et al., 2022) this kind of generative model to match a specific domain, Domain Randomisation (DR) proposes to considerably enlarge the distribution of the synthetic data by *fully randomising* the generation parameters (Tobin et al., 2017). This learning strategy is motivated by converging evidence that augmentation beyond realism leads to improved generalisation (Bengio et al., 2011; Zhao et al., 2019). If pushed to the extreme, DR yields highly unrealistic samples, in which case real images are encompassed within the landscape of the synthetic training data (Tremblay et al., 2018). As a result, this approach seeks to bridge all domain gaps in a given semantic space, rather than solving this problem for each domain gap separately. So far, DR has been used to control robotic arms (Tobin et al., 2017), and for car detection in street views (Tremblay et al., 2018). Here we combine DR with a generative model inspired by Bayesian segmentation, in order to achieve, for the first time, segmentation of brain MRI scans of a wide range of contrasts and resolutions without retraining.

3. Methods

3.1. Generative model

SynthSeg relies on a generative model from which we sample synthetic scans to train a segmentation network (Billot et al., 2020a,b; Iglesias et al., 2021). Crucially, the training images are all generated on the fly with fully randomised parameters, such that the network is exposed to a different combination of contrast, resolution, morphology, artefacts, and noise at each mini-batch (Fig. 2). Here we describe the generative model, which is illustrated in Fig. 3 and exemplified in Supplement 1.

3.1.1. Label map selection and spatial augmentation

The proposed generative model assumes the availability of N training label maps $\{S_n\}_{n=1}^N$ defined over discrete spatial coordinates (x, y, z) at high resolution r_{HR}. Let all label maps take their values from a set of K labels: $S_n(x, y, z) \in \{1, \dots, K\}$. We emphasise that these training label maps can be obtained manually, automatically (by segmenting brain scans with an automated method), or even can be a combination thereof – as long as they share the same labelling convention.

The generative process starts by randomly selecting a segmentation S_i from the training dataset (Fig. 3a). In order to increase the variability of the available segmentations, S_i is deformed with a random spatial transform ϕ, which is the composition of an affine and a non-linear transform.

The affine transformation ϕ_{aff} is the composition of three rotations $(\theta_x, \theta_y, \theta_z)$, three scalings (s_x, s_y, s_z), three shearings (sh_x, sh_y, sh_z), and three translations (t_x, t_y, t_z), whose parameters are sampled from uniform distributions:

$$\theta_x, \theta_y, \theta_z \sim \mathcal{U}(a_{rot}, b_{rot}), \tag{1}$$

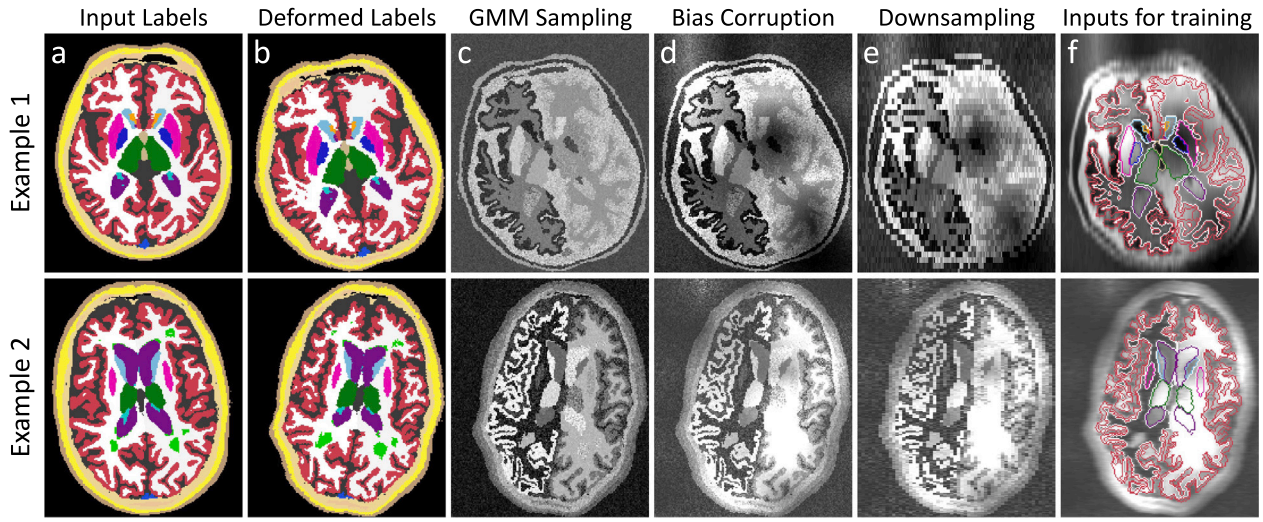

Fig. 3. Intermediate steps of the generative model: (a) we randomly select an input label map from the training set, which we (b) spatially augment in 3D. (c) A first synthetic image is obtained by sampling a GMM at HR with randomised parameters. (d) The result is then corrupted with a bias field and further intensity augmentation. (e) Slice spacing and thickness are simulated by successively blurring and downsampling at random LR. (f) The training inputs are obtained by resampling the image to HR, and removing the labels we do not wish to segment (e.g., extra-cerebral regions).

$$s_x, s_y, s_z \sim \mathcal{U}(a_{sc}, b_{sc}), \tag{2}$$

$$sh_x, sh_y, sh_z \sim \mathcal{U}(a_{sh}, b_{sh}), \tag{3}$$

$$t_x, t_y, t_z \sim \mathcal{U}(a_{tr}, b_{tr}), \tag{4}$$

$$\phi_{\text{aff}} = \text{Aff}(\theta_x, \theta_y, \theta_z, s_x, s_y, s_z, sh_x, sh_y, sh_z, t_x, t_y, t_z), \tag{5}$$

where $a_{rot}, b_{rot}, a_{sc}, b_{sc}, a_{sh}, b_{sh}, a_{tr}, b_{tr}$ are the predefined bounds of the uniform distributions, and $\text{Aff}(\cdot)$ refers to the composition of the aforementioned affine transforms.

The non-linear component ϕ_{nonlin} is a diffeomorphic transform obtained as follows. First, we sample a small vector field of size $10 \times 10 \times 10 \times 3$ from a zero-mean Gaussian distribution of standard deviation σ_{SVF} drawn from $\mathcal{U}(0, b_{\text{nonlin}})$. This field is then upsampled to full image size with trilinear interpolation to obtain a stationary velocity field (SVF). Finally, we integrate this SVF with a scale-and-square approach (Arsigny et al., 2006) to yield a diffeomorphic deformation field that does not produce holes or foldings:

$$\sigma_{SVF} \sim \mathcal{U}(0, b_{\text{nonlin}}), \tag{6}$$

$$\text{SVF}' \sim \mathcal{N}_{10\times10\times10\times3}(0, \sigma_{\text{SVF}}), \tag{7}$$

$$\text{SVF} = \text{Resample}(\text{SVF}'; r_{HR}), \tag{8}$$

$$\phi_{\text{nonlin}} = \text{Integrate}(\text{SVF}). \tag{9}$$

Finally, we obtain an augmented map L by applying ϕ to S_i using nearest neighbour interpolation (Fig. 3b):

$$L = S_i \circ \phi = S_i \circ (\phi_{\text{aff}} \circ \phi_{\text{nonlin}}). \tag{10}$$

3.1.2. Initial HR synthetic image

After deforming the input segmentation, we generate an initial synthetic scan G at HR by sampling a GMM conditioned on L (Fig. 3c). For convenience, we regroup all the means and standard deviations of the GMM in $M_G = \{\mu_k\}_{1\leq k\leq K}$ and $\Sigma_G = \{\sigma_k\}_{1\leq k\leq K}$ respectively. Crucially, in order to randomise the contrast of G, all the parameters in M_G and Σ_G are sampled at each mini-batch from uniform distributions of range $\{a_\mu, b_\mu\}$ and $\{a_\sigma, b_\sigma\}$, respectively. We highlight that Σ_G jointly models tissue heterogeneities as well as the thermal noise of the scanner. G is then formed by independently sampling at each location (x, y, z) the distribution indexed by $L(x, y, z)$:

$$\mu_k \sim \mathcal{U}(a_\mu, b_\mu), \tag{11}$$

$$\sigma_k \sim \mathcal{U}(a_\sigma, b_\sigma), \tag{12}$$

$$G(x, y, z) \sim \mathcal{N}(\mu_{L(x,y,z)}, \sigma^2_{L(x,y,z)}). \tag{13}$$

3.1.3. Bias field and intensity augmentation

We then simulate bias field artefacts to make SynthSeg robust to such effects. We sample a small volume of shape 4^3 from a zero-mean Gaussian distribution of random standard deviation σ_B. We then upsample this small volume to full image size, and take the voxel-wise exponential to obtain a smooth and non-negative field B. Finally, we multiply G by B to obtain a biased image G_B (Fig. 3d), where the previous exponential ensures that division and multiplication by the same factor are equally likely (Van Leemput et al., 1999; Ashburner and Friston, 2005):

$$\sigma_B \sim \mathcal{U}(0, b_B), \tag{14}$$

$$B' \sim \mathcal{N}_{4\times4\times4}(0, \sigma_B^2), \tag{15}$$

$$B = \text{Upsample}(B'), \tag{16}$$

$$G_B(x, y, z) = G(x, y, z) \times \exp[B(x, y, z)]. \tag{17}$$

Then, a final HR image I_{HR} is produced by rescaling G_B between 0 and 1, and applying a random Gamma transform (voxel-wise exponentiation) to further augment the intensity distribution of the synthetic scans. This transform enables us to skew the distribution while leaving intensities in the $[0, 1]$ interval. In practice, the exponent is sampled in the logarithmic domain from a zero-mean Gaussian distribution of standard deviation σ_γ. As a result, I_{HR} is given by:

$$\gamma \sim \mathcal{N}(0, \sigma_\gamma^2), \tag{18}$$

$$I_{HR}(x, y, z) = \left(\frac{G(x, y, z) - \min_{x,y,z} G}{\max_{x,y,z} G - \min_{x,y,z} G} \right)^{\exp(\gamma)}. \tag{19}$$

3.1.4. Simulation of resolution variability

In order to make the network robust against changes in resolution, we now model differences in acquisition direction (i.e., axial, coronal, sagittal), slice spacing, and slice thickness. After randomly selecting a direction, the slice spacing r_{spac} and slice thickness r_{thick} are respectively drawn from $\mathcal{U}(r_{HR}, b_{res})$ and $\mathcal{U}(r_{HR}, r_{spac})$. Note that r_{thick} is bound by r_{spac} as slices very rarely overlap in practice.

Once all resolution parameters have been sampled, we first simulate slice thickness by blurring I_{HR} into I_σ with a Gaussian kernel that approximates the real slice excitation profile. Specifically, its standard

deviation σ_{thick} is designed to divide the power of the HR signal by 10 at the cut-off frequency (Billot et al., 2020b). Moreover, σ_{thick} is multiplied by a random coefficient α to introduce small deviations from the nominal thickness, and to mitigate the Gaussian assumption.

Slice spacing is then modelled by downsampling I_σ to I_{LR} at the prescribed low resolution r_{spac} with trilinear interpolation (Fig. 3e) (Van Leemput et al., 2003). Finally, I_{LR} is upsampled back to r_{HR} (typically 1 mm), such that the CNN is trained to produce crisp HR segmentations, regardless of the simulated resolution. This process can be summarised as:

$$r_{spac} \sim \mathcal{U}(r_{HR}, b_{res}), \tag{20}$$

$$r_{thick} \sim \mathcal{U}(r_{HR}, r_{spac}), \tag{21}$$

$$\alpha \sim \mathcal{U}(a_\alpha, b_\alpha), \tag{22}$$

$$\sigma_{thick} = 2\alpha \log(10)(2\pi)^{-1} r_{thick}/r_{HR}, \tag{23}$$

$$I_\sigma = I_{HR} * \mathcal{N}(0, \sigma_{thick}), \tag{24}$$

$$I_{LR} = \text{Resample}\left(I_\sigma; r_{spac}\right), \tag{25}$$

$$I = \text{Resample}\left(I_{LR}; r_{HR}\right). \tag{26}$$

3.1.5. Model output and segmentation target

At each training step, our method produces two volumes: an image I sampled from the generative model, and its segmentation target T. The latter is obtained by taking the deformed map L in (10), and resetting to background all the label values that we do not wish to segment (i.e., labels for the background structures, which are of no interest to segment). Thus, T has $K' \leq K$ labels (Fig. 3f).

We emphasise that the central contribution of this work lies in the adopted domain randomisation strategy. The values of the hyperparameters controlling the uniform priors (listed in Supplement 2) are tuned using a validation set, and are the object of a sensitivity analysis in Section 5.2.

3.2. Segmentation network and learning

Given the described generative model, a segmentation network is trained by sampling pairs $\{I, T\}$ on the fly. Here we employ a 3D UNet architecture (Ronneberger et al., 2015) that we used in previous works with synthetic scans (Billot et al., 2020a). Specifically, it consists of five levels, each separated by a batch normalisation layer (Ioffe and Szegedy, 2015) along with a max-pooling (contracting path), or upsampling operation (expanding path). All levels comprise two convolution layers with $3 \times 3 \times 3$ kernels. Every convolutional layer is associated with an Exponential Linear Unit activation (Clevert et al., 2016), except for the last one, which uses a softmax. While the first layer counts 24 feature maps, this number is doubled after each max-pooling, and halved after each upsampling. Following the UNet architecture, we use skip connections across the contracting and expanding paths. Note that the network architecture is not a focus of this work: while we employ a UNet (Ronneberger et al., 2015) (the most widespread network for medical images), it could in principle be replaced with any other segmentation architecture.

We use the soft Dice loss for training (Milletari et al., 2016):

$$\text{Loss}(Y, T) = 1 - \sum_{k=1}^{K'} \frac{2 \times \sum_{x,y,z} Y_k(x,y,z) T_k(x,y,z)}{\sum_{x,y,z} Y_k(x,y,z)^2 + T_k(x,y,z)^2}, \tag{27}$$

where Y_k is the soft prediction for label $k \in \{1, \ldots, K'\}$, and T_k is its associated ground truth in one-hot encoding. We use the Adam optimiser (Kingma and Ba, 2017) for 300,000 steps with a learning rate of 10^{-4}, and a batch size of 1. The network is trained twice, and the weights are saved every 10,000 steps. The retained model is then selected relatively to a validation set. In practice, the generative model and the segmentation network are concatenated within a single model, which is entirely implemented on the GPU in Keras (Chollet, 2015) with a Tensorflow backend (Abadi et al., 2016). In total, training takes around seven days on a Nvidia Quadro RTX 6000 GPU.

3.3. Inference

At test time, the input is resampled to r_{HR} with trilinear interpolation (such that the output of the CNN is at HR), and its intensities are rescaled between 0 and 1 with min–max normalisation (using the 1st and 99th percentiles). Preprocessed scans are then fed to the network to obtain soft predictions maps for each label. In practice, we also perform test-time augmentation (Moshkov et al., 2020), which slightly improved results on the validation set. Specifically, we segment two versions of each test scan: the original one, and a right-left flipped version of it. The soft predictions of the flipped input are then flipped back to native space (while ensuring that right-left labels end up on the correct side), and averaged with the predictions of the original scan. Once test-time augmentation has been performed, final segmentations are obtained by keeping the biggest connected component for each label. On average, inference takes ten seconds on a Nvidia TitanXP GPU (12 GB), including preprocessing, prediction, and postprocessing.

4. General experimental setup

4.1. Brain scans and ground truths

Our experiments employ eight datasets comprising 5000 scans of six different modalities and ten resolutions. The splits between training, validation, and testing are given in Table 1.

T1-39: 39 T1-weighted (T1) scans with manual labels for 30 structures (Fischl et al., 2002). They were acquired with an MP-RAGE sequence at 1 mm isotropic resolution.

HCP: 500 T1 scans of young subjects from the Human Connectome Project (Van Essen et al., 2012), acquired at 0.7 mm resolution, and that we resample at 1 mm isotropic resolution.

ADNI: 1500 T1 scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI.[2]) All scans are acquired at 1 mm isotropic resolution from a wide array of scanners and protocols. In contrast to HCP, this dataset comprises ageing subjects, some diagnosed with mild cognitive impairment (MCI) or Alzheimer's Disease (AD). As such, many subjects present strong atrophy patterns and white matter lesions.

T1mix: 1000 T1 scans at 1 mm isotropic resolution from seven datasets: ABIDE (Di Martino et al., 2014), ADHD200 (The ADHD-200 Consortium, 2012), GSP (Holmes et al., 2015), HABS (Dagley et al., 2017), MCIC (Gollub et al., 2013), OASIS (Marcus et al., 2007), and PPMI (Marek et al., 2011). We use this heterogeneous dataset to assess robustness against intra-modality contrast variations due to different acquisition protocols.

FSM: 18 subjects with T1 and two other MRI contrasts: T2-weighted (T2) and a sequence used for deep brain stimulation (DBS) (Iglesias et al., 2018). All scans are at 1 mm resolution.

MSp: 8 subjects with T1 and proton density (PD) acquisitions at 1 mm isotropic resolution (Fischl et al., 2004). These scans were skull stripped prior to availability, and are manually delineated for the same labels as T1-39.

FLAIR: 2393 fluid-attenuated inversion recovery (FLAIR) scans at $1 \times 1 \times 5$ mm axial resolution. These subjects are from another subset of the ADNI database, and hence also present morphological patterns related to ageing and AD. This dataset enables assessment on scans that are representative of clinical acquisitions with real-life slice selection profiles (as opposed to simulated LR, see below). Matching 1 mm T1 scans are also available, but they are not used for testing.

[2] The ADNI was launched in 2003 by the National Institute on Ageing, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, pharmaceutical companies and non-profit organisations, as a 5-year public–private partnership. The goal of ADNI is to test if MRI, PET, other biological markers, and clinical and neuropsychological assessment can analyse the progression of MCI and early AD, develop new treatments, monitor their effectiveness, and decrease the time and cost of clinical trials.

Table 1
Summary of the employed brain datasets. The 10 test resolutions are 1 mm^3 and 3/5/7mm in either axial, coronal, or sagittal direction. SynthSeg is trained solely on label maps (no intensity images).

Dataset	Subjects	Modality	Resolution
Training			
T1-39	20	SynthSeg: Label maps Baselines: T1	1 mm isotropic
HCP	500	SynthSeg: Label maps Baselines: T1	1 mm isotropic
ADNI	500	SynthSeg: Label maps Baselines: T1	1 mm isotropic
Validation			
T1–39	4	T1	all tested resolutions
FSM	3	T2, DBS	all tested resolutions
Testing			
T1–39	15	T1	all tested resolutions
ADNI	1,000	T1	1 mm isotropic
T1mix	1,000	T1	1 mm isotropic
FSM	15	T1, T2, DBS	all tested resolutions
MSp	8	T1, PD	all tested resolutions
FLAIR	2,393	FLAIR	5 mm axial
CT	6	CT	3 mm axial

CT: 6 computed tomography (CT) scans at $1 \times 1 \times 3$ mm axial resolution (West et al., 1997), with the aim of assessing SynthSeg on imaging modalities other than MRI. As for the FLAIR dataset, matching 1 mm T1 scans are also available.

In order to evaluate SynthSeg on more resolutions, we artificially downsample all modalities from the T1-39, FSM, and MSp datasets (all at 1 mm isotropic resolution) to nine different LR: 3, 5 and 7 mm spacing in axial, coronal, and sagittal directions. These simulations do not use real-life slice selection profiles, but are nonetheless very informative since they enable to study the segmentation accuracy as a function of resolution.

Except for T1-39 and MSp, which are available with manual labels, segmentation ground truths are obtained by running FreeSurfer (Fischl, 2012) on the T1 scans of each dataset, and undergo a thorough visual quality control to ensure anatomical correctness. FreeSurfer has been shown to be very robust across numerous independent T1 datasets and yields Dice scores in the range of 0.85–0.88 (Fischl et al., 2002; Tae et al., 2008). Therefore, its use as silver standard enables reliable assessment of Dice below 0.85; any scores above that level are considered equally good. Crucially, using FreeSurfer segmentations enables us to evaluate SynthSeg on vast amounts of scans with very diverse contrasts and resolutions, which would have been infeasible with manual tracings only.

4.2. Training segmentations and population robustness

As indicated in Table 1, the training set for SynthSeg comprises 20 label maps from T1-39, 500 from HCP, and 500 from ADNI. Mixing these label maps considerably increases the morphological variety of the synthetic scans (far beyond the capacity of the proposed spatial augmentation alone), and thus enlarges the robustness of SynthSeg to a wide range of populations. We emphasise that using automated label maps for training is possible because synthetic images are by design perfectly aligned with their segmentations. We highlight that the training data does not include any real scan.

Because SynthSeg requires modelling all tissue types in the images, we complement the training segmentations with extra-cerebral labels (Supplement 3) obtained with a Bayesian segmentation approach (Puonti et al., 2020). Note that these new labels are dropped with 50% chances during generation, to make SynthSeg compatible with skull stripped images. Moreover, we randomly "paste" lesion labels from FreeSurfer with 50% probability, to build robustness against white matter lesions (Supplement 4). Finally, we further increase the variability of the training data by randomly left/right flipping segmentations and cropping them to 160^3 volumes.

4.3. Competing methods

We compare SynthSeg against five other approaches:

T1 baseline (Zhang et al., 2020): A supervised network trained on real T1 scans (Table 1). This baseline seeks to assess the performance of supervised CNNs on their source domain, as well as their generalisation to intra-modality (T1) contrast variations. For comparison purposes, we use the same UNet architecture and augmentation scheme (i.e., spatial deformation, intensity augmentation, bias field corruption) as for SynthSeg.

nnUNet (Isensee et al., 2021)[3]: A state-of-the-art supervised approach, very similar to the T1 baseline, except that the architecture, augmentation, pre- and postprocessing are automated with respect to the (real) T1 input data.

Test-time adaptation (TTA) (Karani et al., 2021)[4]: A state-of-the-art domain adaptation method relying on fine-tuning. Briefly, this strategy uses three CNN modules: an image normaliser (five convolutional layers), a segmentation UNet, and a denoising auto-encoder (DAE). At first, the normaliser and the UNet are jointly trained on supervised data of a source domain, while the DAE is trained separately to correct erroneous segmentations. At test time, the UNet and DAE are frozen, and the normaliser is fine-tuned on scans from different target domains by using the denoised predictions of the UNet as ground truth.

SIFA (Chen et al., 2019)[5]: A state-of-the-art unsupervised domain adaptation strategy, where image-to-image translation and segmentation modules are jointly trained (with shared layers). SIFA seeks to align each target domain to the source data in both feature and image spaces for improved adaptation.

SAMSEG (Puonti et al., 2016): A state-of-the-art Bayesian segmentation framework with unsupervised likelihood. As such, SAMSEG is contrast-adaptive, and can segment at any resolution, albeit not accounting for PV effects. SAMSEG does not need to be trained as it solves an optimisation problem for test each scan. Here we use the version distributed with FreeSurfer 7.0, which runs in approximately 15 min.

The predictions of all methods are postprocessed as in Section 3.3, except for nnUNet, which uses its own postprocessing. We use the default implementation for all competing methods, except for a few minor points that are listed in Supplement 5. All learning-based methods are trained twice, and models are chosen relatively to the validation set. Segmentations are assessed by computing (hard) Dice scores and the 95th percentile of the surface distance (SD95, in millimetres).

5. Experiments and results

Here we present four experiments that evaluates the accuracy and generalisation of SynthSeg. First, we compare it against all competing methods on every dataset. Then, we conduct an ablation study on the proposed method. The third experiment validates SynthSeg in a proof-of-concept neuroimaging group study. Finally, we demonstrate the generalisability of our method by extending it to cardiac MRI and CT.

[3] https://github.com/MIC-DKFZ/nnUNet
[4] https://github.com/neerakara/test-time-adaptable-neural-networks-for-domain-generalization
[5] https://github.com/cchen-cc/SIFA

Table 2
Mean Dice scores and 95th percentile surface distances (SD95) obtained by all methods for every dataset. The best score for each dataset is in bold, and marked with a star if significantly better than all other methods at a 5% level (two-sided Bonferroni-corrected non-parametric Wilcoxon signed-rank test). Supervised methods cannot segment non-T1 modalities, and domain adaptation strategies are not tested on the source domain.

		T1–39	ADNI	T1mix	FSM-T1	MSp-T1	FSM-T2	FSM-DBS	MSp-PD	FLAIR	CT
T1 baseline	Dice	0.91	0.83	0.86	0.84	0.82	–	–	–	–	–
	SD95	1.31	2.63	2.14	2.09	3.55	–	–	–	–	–
nnUNet (Isensee et al., 2021)	Dice	0.91	0.82	0.84	0.84	0.81	–	–	–	–	–
	SD95	1.31	2.8	2.32	2.11	3.71	–	–	–	–	–
TTA (Karani et al., 2021)	Dice	–	0.83	0.87	0.87	0.85	0.82	0.71	0.8	0.71	0.46
	SD95	–	2.26	1.73	1.72	2.14	2.35	4.48	3.71	3.95	19.43
SIFA (Chen et al., 2019)	Dice	–	0.8	0.82	0.84	0.84	0.82	0.82	0.74	0.73	0.62
	SD95	–	3.03	2.24	2.21	2.57	2.32	2.09	4.41	3.30	4.51
SAMSEG (Puonti et al., 2016)	Dice	0.85	0.81	0.86	0.86	0.83	0.82	0.81	0.81	0.64	0.71
	SD95	1.85	3.09	1.77	1.81	2.47	2.21	2.34	2.99	3.67	3.36
SynthSeg (ours)	Dice	0.88	**0.84**	0.87	**0.88**	**0.86***	**0.86***	**0.86***	**0.84***	**0.78***	**0.76***
	SD95	1.5	**2.18***	**1.69***	**1.59***	**1.89***	**1.83***	**1.81***	**2.06***	**2.35***	**3.29***

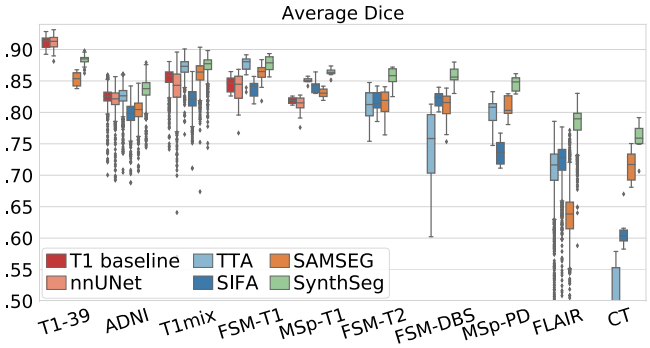

Fig. 4. Box plots showing Dice scores obtained by all methods for every dataset. For each box, the central mark is the median; edges are the first and third quartiles; and outliers are marked with ♦.

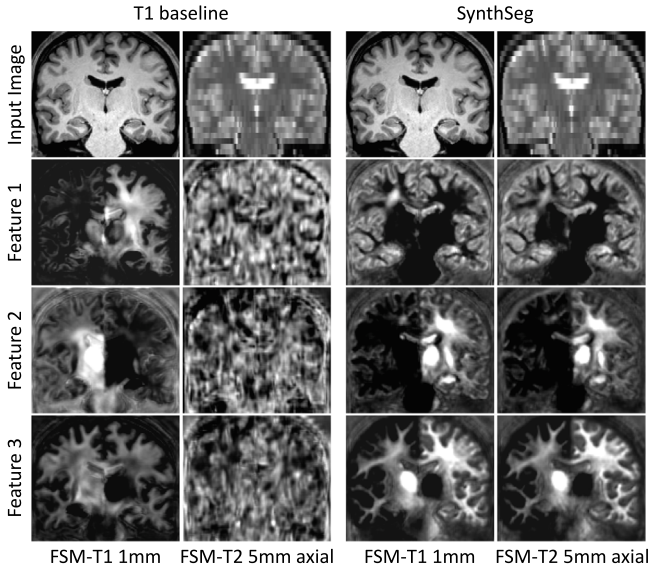

Fig. 5. Representative features of the last layer of the network for two scans of different contrast and resolution for the same subject. While the T1 baseline only produces noise outside its training domain, SynthSeg learns a consistent representation across contrasts and resolutions.

5.1. Robustness against contrast and resolution

In this experiment, we assess the generalisation ability of SynthSeg by comparing it against all competing methods for every dataset at native resolution (Fig. 4, Table 2).

Remarkably, despite SynthSeg has never been exposed to a real image during training, it reaches almost the same level of accuracy as supervised networks (T1 baseline and nnUNet) on their training domain (0.88 against 0.91 Dice scores on T1-39). Moreover, SynthSeg generalises better than supervised networks against intra-modality contrast variations, both in terms of mean (average difference of 2.5 Dice points with the T1 baseline for T1 datasets other than T1-39) and robustness (much higher lower-quartiles for SynthSeg). Crucially, the employed DR strategy yields very good generalisation, as SynthSeg sustains a remarkable accuracy across all tested contrasts and resolutions, which is infeasible with supervised networks alone. Indeed, SynthSeg outputs high-quality segmentations for all domains, even for FLAIR and CT scans at LR (Fig. 6). Quantitatively, SynthSeg produces the best scores for all nine target domains, six of which with statistical significance for Dice and nine for SD95 (Table 2). This flexibility is exemplified in Fig. 5, where SynthSeg produces features that are almost identical for a 1 mm T1 and a 5 mm axial T2 of the same subject, the latter being effectively super-resolved to 1 mm.

Although the tested domain adaptation approaches (TTA, SIFA) considerably increase the generalisation of supervised networks, they are still outperformed by SynthSeg for all contrasts and resolutions. This is a remarkable result since, as opposed to domain adaptation strategies, SynthSeg does not require any retraining. We note that fine-tuning the TTA framework makes it more robust than supervised methods for intra-modality applications (noticeably higher lower-quartiles), but its results can substantially fluctuate for larger domain gaps (e.g., on FSM-DBS, FLAIR, and CT). This is partly corrected by SIFA (improvement of 14.92 mm in SD95 for CT), which is better suited for larger domain gaps (Karani et al., 2021), albeit some abrupt variations (e.g., MSp-PD). In comparison, SAMSEG yields much more constant results across MR contrasts at 1 mm resolution (average Dice score of 0.83). However, because it does not model PV, its accuracy greatly declines at low resolution: Dice scores decrease to 0.71 on the 3 mm CT dataset (5 points below SynthSeg), and to 0.64 on the 5 mm FLAIR dataset (14 points below SynthSeg).

To further validate the flexibility of the proposed approach to different resolutions, we test SynthSeg on all artificially downsampled data (Table 1), and we compare it against T1 baselines retrained at each resolution, as well as SAMSEG. The results show that SynthSeg maintains a very good accuracy for all tested resolutions (Fig. 7). Despite the considerable loss of information at LR and heavy PV effects, SynthSeg only loses 3.8 Dice points between 1 mm and 7 mm slice spacing on average, mainly due to thin structures like the cortex (Fig. 8). Meanwhile, SAMSEG is strongly affected by PV, and loses 7.6 Dice points across the same range. As before, the T1 baselines obtain remarkable results on scans similar to their training data, but generalise poorly to unseen domains (i.e., FSM-T1 and MSp-T1), where

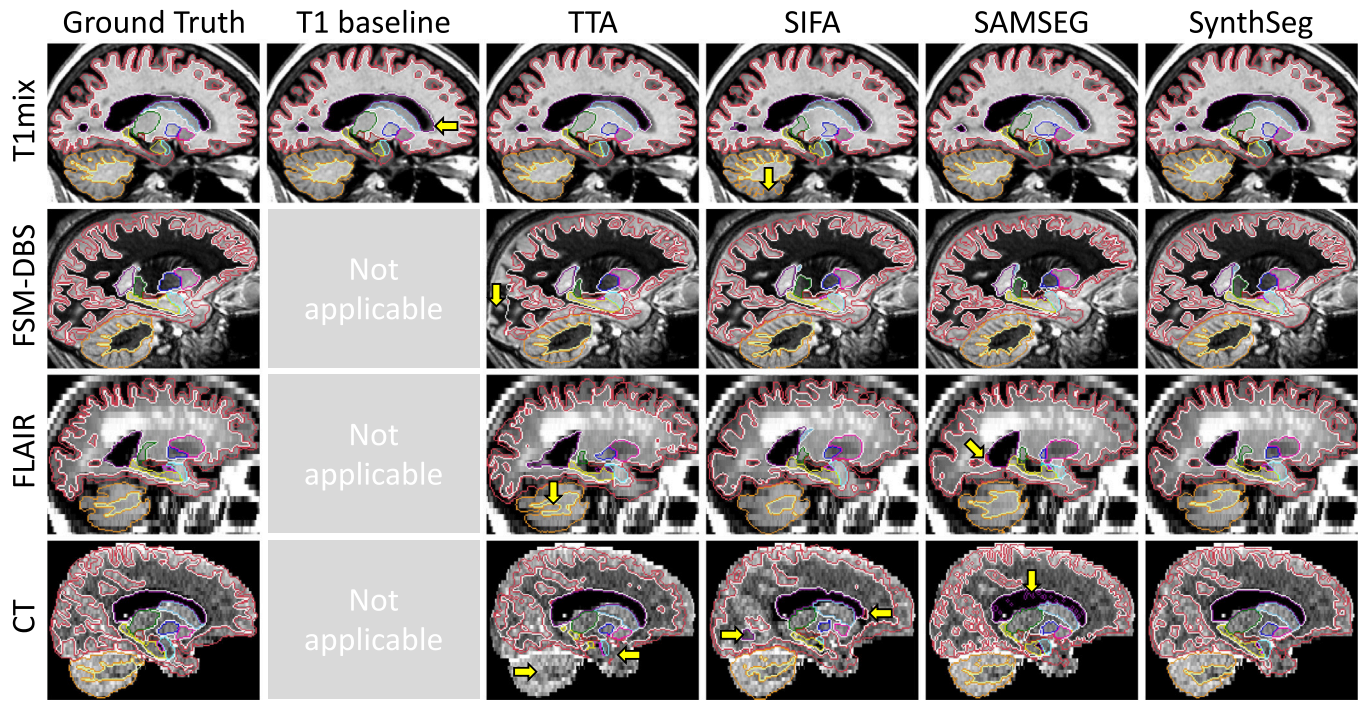

Fig. 6. Sample segmentations from the first experiment. Major segmentation mistakes are indicated with yellow arrows. SynthSeg produces very accurate segmentations for all contrasts and resolutions. The T1 baseline makes small errors outside its training domain and cannot be applied to other modalities. While the TTA approach yields very good segmentations for T1mix, its results degrade for larger domain gaps, where it is outperformed by SIFA. Finally, SAMSEG yields coherent results for scans at 1 mm resolution, but is heavily affected by PV effects at low resolution.

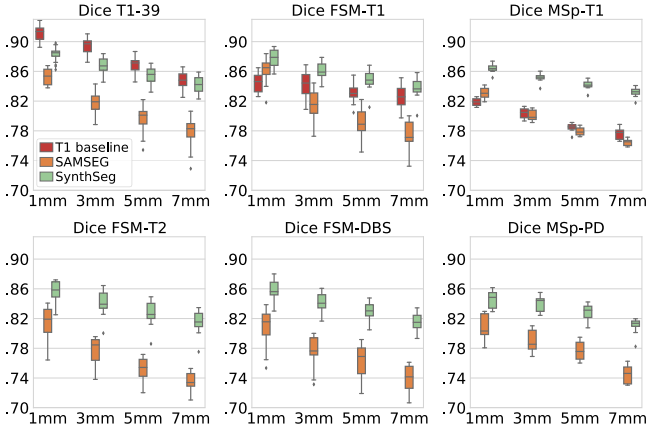

Fig. 7. Dice scores for data downsampled at 3, 5, or 7 mm in either axial, coronal, or sagittal direction (results are averaged across directions).

SynthSeg is clearly superior. Moreover, the gap between them on the training data progressively narrows with decreasing resolution, until it almost vanishes at 7 mm, thus making SynthSeg particularly useful for LR scans.

5.2. Ablations on DR and training label maps

We now validate several aspects of our method, starting with the DR strategy. We first focus on the intensity profiles of the synthetic scans by training four variants: *(i)* SynthSeg-R, which is resolution-specific; *(ii)* SynthSeg-RC, which we retrain for every new combination of contrast and resolution by using domain-specific Gaussian priors for the GMM and resolution parameters (Billot et al., 2020b, 2021), *(iii)* a variant using slightly tighter GMM uniform priors ($\mu \in [10, 240]$, $\sigma \in [1, 25]$, instead of $\mu \in [0, 255]$, $\sigma \in [0, 35]$); and *(iv)* a variant with even tighter priors ($\mu \in [50, 200]$, $\sigma \in [1, 15]$). SynthSeg-R and SynthSeg-RC assess the effect of constraining the synthetic intensity profiles to look more realistic, whereas the two last variants study the sensitivity of the chosen GMM uniform priors. Finally, we train three more networks by ablating the lesion simulation, bias field, and spatial augmentation, respectively.

Fig. 9 shows that, crucially, narrowing the distributions of the generated scans in SynthSeg-R and SynthSeg-RC to simulate a specific contrast and/or resolution, leads to a consistent decrease in accuracy: despite retraining them on each target domain, they are on average lower than SynthSeg by 1.4 and 2.6 Dice points respectively. Interestingly, the variant with slightly tighter GMM priors obtains scores almost identical to the reference SynthSeg, whereas further restricting these priors (at the risk of excluding intensities encountered in real scans) leads to poorer performance (2.1 fewer Dice points on average). Finally, the bias and deformation ablations highlight the impact of those two augmentations (loss of 3.7 and 4.3 Dice points, respectively), whereas ablating the lesion simulation mainly affects the ADNI and FLAIR datasets, where the ageing subjects are more likely to present lesions (average loss of 3.9 Dice points).

In a second set of experiments, we evaluate the effect of using different numbers of segmentations during training. Hence, we retrain SynthSeg on increasing numbers of label maps randomly selected from T1-39 ($N \in \{1, 5, 10, 15, 20\}$, see Supplement 6). Moreover, we include the version of SynthSeg trained on all available maps, to quantify the effect of adding automated segmentations from diverse populations. All networks are evaluated on six representative datasets (Fig. 10). The results reveal that using only one training map already attains decent Dice scores (between 0.68 and 0.80 for all datasets). As expected, the accuracy increases when adding more maps, and Dice scores plateau at $N = 5$ (except for MSp-PD, which levels off at $N = 10$). Interestingly, Fig. 10 also shows that SynthSeg requires fewer training examples than the T1 baseline to converge towards its maximum accuracy.

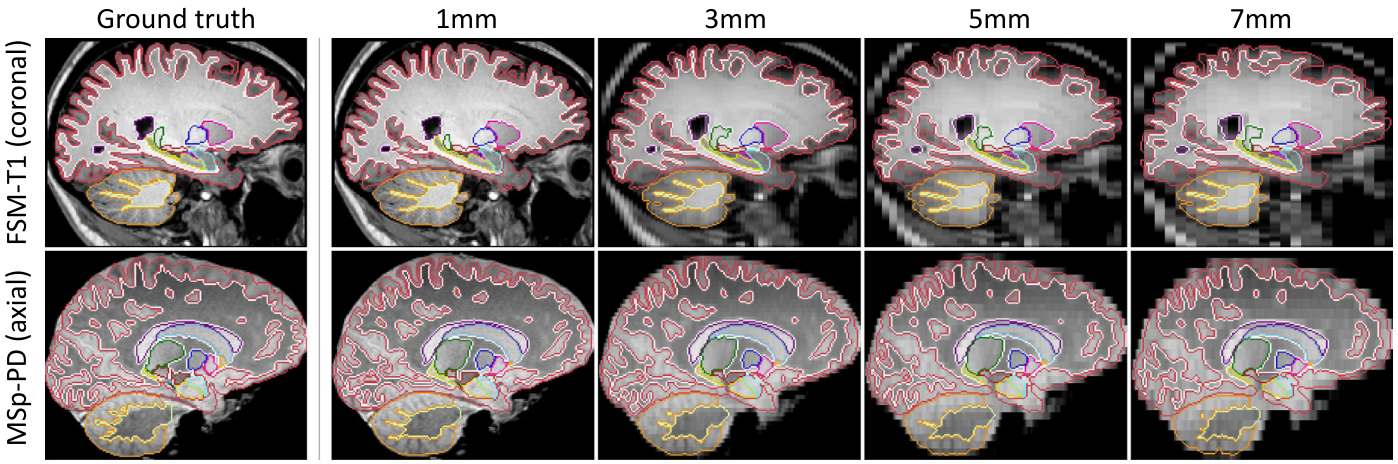

Fig. 8. Examples of segmentations obtained by SynthSeg for two scans artificially downsampled at decreasing LR. SynthSeg presents an impressive generalisation ability to all resolutions, despite heavy PV effects and important loss of information at LR. However, we observe a slight decrease in accuracy for thin and convoluted structures such as the cerebral cortex (red) or the white cerebellar matter (dark yellow).

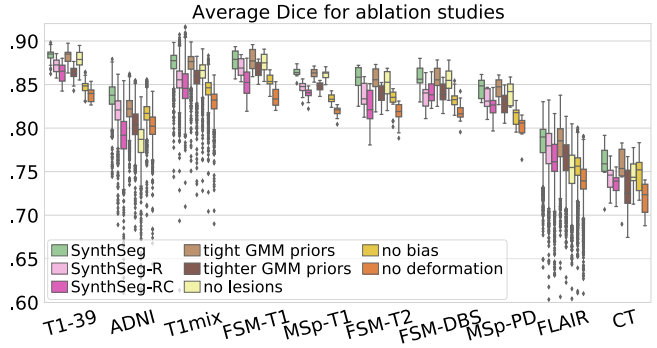

Fig. 9. Mean Dice scores obtained for SynthSeg and ablated variants.

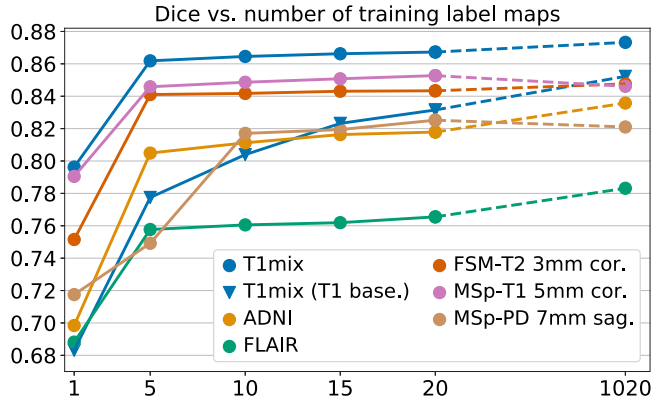

Fig. 10. Dice vs. number of training label maps for SynthSeg (circles) on representative datasets. The last points are obtained by training on all available labels maps (20 manual plus 1000 automated). We also report scores obtained on T1mix by the T1 baseline (triangles).

Meanwhile, adding a large amount of training automated maps enables us to improve robustness to morphological variability, especially for the ADNI and FLAIR datasets with ageing and diseased subjects (Dice scores increase by 1.9 and 2.0 points respectively). To confirm this trend, we study the 3% of ADNI subjects with the largest ventricular volumes (relatively to the intracranial volume, ICV), whose morphology substantially deviates from the 20 manual training maps. For these ADNI cases (30 in total), the average Dice score increases by 4.7 Dice points for the network trained on all label maps compared with the one trained on manual maps only. This result further demonstrates the gain in robustness obtained by adding automated label maps to the training set of SynthSeg (Fig. 11).

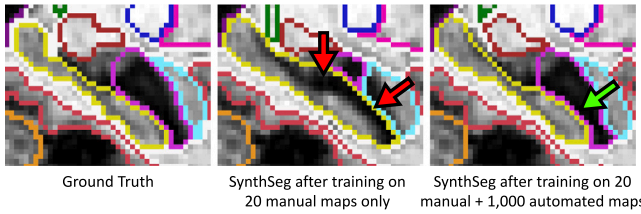

Fig. 11. Close-up on the hippocampus for an ADNI testing subject with atrophy patterns that are not present in the manual training segmentations. Hence, training SynthSeg on these manual maps only leads to limited accuracy (red arrows). However, adding a large number of automated maps from different populations to the training set enables us to improve robustness against morphological variability (green arrow).

5.3. Alzheimer's disease volumetric study

In this experiment, we evaluate SynthSeg in a proof-of-concept volumetric group study, where we assess its ability to detect hippocampal atrophy related to AD (Chupin et al., 2009). Specifically, we study whether SynthSeg can detect similar atrophy patterns for subjects who have been imaged with different protocols. As such, we run SynthSeg on a separate set of 100 ADNI subjects (50 controls, 50 AD), all with 1 mm isotropic T1 scans as well as FLAIR acquisitions at 5 mm axial resolution.

We measure atrophy with effect sizes in predicted volumes between controls and diseased populations. Effect sizes are computed with Cohen's d (Cohen, 1988):

$$d = \frac{\mu_C - \mu_{AD}}{s}, \quad s = \sqrt{\frac{(n_C - 1)s_C^2 + (n_{AD} - 1)s_{AD}^2}{n_C + n_{AD} - 2}}, \tag{28}$$

where μ_C, s_C^2 and μ_{AD}, s_{AD}^2 are the means and variances of the volumes for the two groups, and n_C and n_{AD} are their sizes. Hippocampal volumes are computed by summing the corresponding soft predictions, thus accounting for segmentation uncertainties. All measured volumes are corrected for age, gender, and ICV (estimated with FreeSurfer) using a linear model.

In addition to SynthSeg, we evaluate the performance of SAMSEG, and all Cohen's d are compared to a silver standard obtained by running FreeSurfer on the T1 scans (Fischl, 2012). The results, reported in

Table 3
Effect size (Cohen's d) obtained by FreeSurfer (Ground Truth, GT), SAMSEG and SynthSeg for hippocampal volumes between controls and AD patients for different types of scans.

Contrast	Resolution	FreeSurfer (GT)	SAMSEG	SynthSeg
T1	1 mm^3	1.38	1.46	1.40
FLAIR	5 mm axial		0.53	1.24

Table 3, reveal that both methods yield a Cohen's d close to the ground truth for the HR T1 scans. We emphasise that, while segmenting the hippocampus in 1 mm T1 scans is of modest complexity, this task is much more difficult for 5 mm axial FLAIR scans, since the hippocampus only appears in two to three slices, and with heavy PV. As such, the accuracy of SAMSEG greatly degrades on FLAIR scans, where it obtains less than half the expected effect size. In contrast, SynthSeg sustains a high accuracy on the FLAIR scans, producing a Cohen's d much closer to the reference value, which was obtained at HR.

5.4. Extension to cardiac segmentation

In this last experiment, we demonstrate the generalisability of SynthSeg by applying it to cardiac segmentation. With this purpose, we employ two new datasets: MMWHS (Zhuang et al., 2019), and LASC13 (Tobon-Gomez et al., 2015). MMWHS includes 20 MRI scans with in-plane resolutions from 0.78 to 1.21 mm, and slice spacings between 0.9 and 1.6 mm. MMWHS also contains 20 CT scans of non-overlapping subjects at high resolution (0.28-0.58 mm in-plane, 0.45-0.62 mm slice spacing). All these scans are available with manual labels for seven regions (see Table 4). On the other hand, LASC13 includes 10 MRI heart scans at $1.25 \times 1.25\times$ 1.37 mm axial resolution, with manual labels for the left atrium only. We form the training set for SynthSeg by randomly drawing 13 label maps from MMWHS MRI. For consistency, these training segmentations are all resampled at a common 1 mm isotropic resolution. Finally, the validation set consists of two more scans from MMWHS, while all the remaining scans are used for testing.

Nevertheless, the training labels maps only model the target regions to segment, whereas SynthSeg requires labels for all the tissues present in the test images. Therefore, we enhance the training segmentations by subdividing all their labels (background and foreground) into finer sub-regions. This is achieved by clustering the intensities of the associated image with the Expectation Maximisation algorithm (Dempster et al., 1977). First, each foreground label is divided into two regions to model blood pools. Then, the background region is split into a random number of N regions ($N \in [3, 10]$), which aim at representing the surrounding structures with different levels of granularity (Supplement 6). All these label maps are precomputed to alleviate computing resources during training. We also emphasise that all sub-labels are merged back with their initial label for the loss computation during training. The network is trained twice as described in Section 3.2 with hyperparameters values obtained relatively to the validation set (see values in Supplement 8). Inference is then performed as in Section 3.3, except for the test-time flipping augmentation that is now disabled.

The results are reported in Table 4, and show that SynthSeg segments all seven regions with very high precision (all Dice scores are above 0.8). Moreover, it maintains a very good accuracy across all tested datasets, with mean Dice scores of 0.84 and 0.88 for MMWHS MRI and CT respectively. Interestingly, these scores are similar to the state-of-the-art results in cross-modality cardiac segmentation obtained by Chen et al. (2019) (Dice score of 0.82), despite not being *directly* comparable due to differences in resolution (2 mm for Chen et al. (2019), 1 mm for SynthSeg). Overall, segmenting all datasets at such a level of accuracy (Fig. 12) is remarkable for SynthSeg, since, as opposed to Chen et al. (2019), it is not retrained on any of them.

Table 4
Dice scores for seven cardiac regions: left atrium (LA), right atrium (RA), left ventricle (LV), right ventricle (RV), myocardium (MYO), ascending aorta (AA), and pulmonary artery (PA). LASC13 only has ground truth for LA.

	LA	LV	RA	RV	MYO	AA	PA
MMWHS MRI	0.91	0.89	0.9	0.84	0.81	0.86	0.86
MMWHS CT	0.92	0.89	0.86	0.88	0.85	0.94	0.84
LASC13	0.9	–	–	–	–	–	–

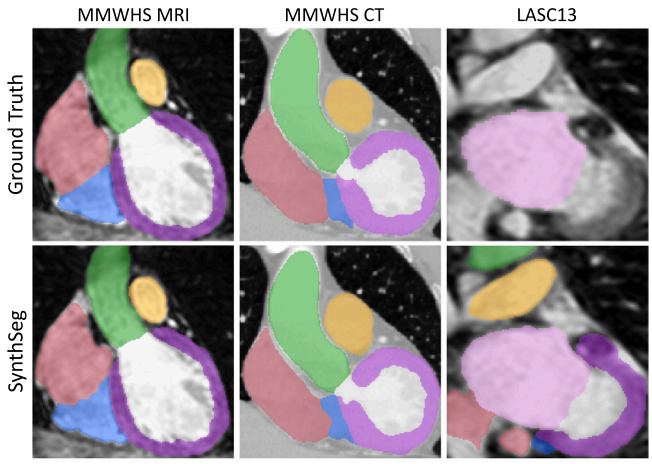

Fig. 12. Representative cardiac segmentations obtained by SynthSeg on three datasets, without retraining on any of them, and without using real images during training. LASC13 only has ground truth for LA (pink).

6. Discussion

We have proposed a method for segmentation of brain MRI scans that is robust against changes in resolution and contrast (including CT) without retraining or fine-tuning. Our main contribution lies in the adopted domain randomisation strategy, where a segmentation network is trained with synthetic scans of fully randomised contrast and resolution. By producing highly diverse samples that make no attempt at realism, this approach forces the network to learn domain-independent features.

The impact of the DR strategy is demonstrated by the domain-constrained SynthSeg variants, for which training contrast and resolution-specific networks yields poorer performance (Section 5.2). We believe this outcome is likely a combination of two phenomena. First, randomising the generation parameters enables us to mitigate the assumptions made when designing the model (e.g., Gaussian intensity distribution for each region, slice selection profile, etc.). Second, this result is consistent with converging evidence that augmenting the data beyond realism often leads to better generalisation (Tobin et al., 2017; Chaitanya et al., 2019; Bengio et al., 2011).

Additionally, SynthSeg enables to greatly alleviate the labelling labour for training purposes. First, it is only trained once and only requires a single set of anatomical segmentations (no real images), as opposed to supervised methods, which need paired images and labels for every new domain. Second, our results show that SynthSeg typically requires less training examples than supervised CNNs to converge to its maximum performance (Section 5.2). And third, parts of the training dataset can be acquired at almost no cost, by including label maps obtained by segmenting real brain scans with automated methods, and visually checking the results to ensure reasonable quality and anatomical plausibility. We highlight that while automated segmentations are generally not used for training (since they are prone to errors), this is made possible here by the fact that synthetic scans are, by design, perfectly aligned with their ground truths. We also emphasise that using automated segmentations to train SynthSeg is not only possible, but

recommended, as the inclusion of such segmentations greatly improves robustness against highly different morphologies caused by anatomical variability (e.g., ageing subjects).

Nevertheless, the employed Gaussian model imposes that the training label maps encompass tracings of all tissues present in the test scans. However, this is not a limitation in practice, since automated labels can be obtained for missing structures by simple intensity clustering. This strategy enabled us to obtain state-of-the-art results for cardiac segmentation, where the original label maps did not describe the complex distribution of finer structures (blood pools in cardiac chambers) and surrounding tissues (vessels, bronchi, bones, etc.). Moreover, our results show that SynthSeg can handle deviations from the Gaussian model within a given structure if they are mild (like the thalamus in brain MRI), or far away from the regions to segment (like the neck in brain MRI).

A limitation of this work is the high proportion of automated label maps used for evaluation. This choice was initially motivated by the wish to evaluate SynthSeg on a wide variety of contrasts and resolutions, which would have been infeasible with manual labels only. Nonetheless, we emphasise that a lot of testing datasets still use manual segmentations (T1-39, and MSp for brain segmentation; MMWHS and LASC13 for the heart experiment), and that the remaining datasets have all undergone thorough visual quality control. Importantly, SynthSeg has shown the same remarkable generalisation ability when evaluated with manual or automated ground truths. Finally, the conclusions of this paper are further reinforced by the indirect evaluation performed in Section 5.3, which demonstrates the accuracy and clinical utility of SynthSeg.

Thanks to its unprecedented generalisation ability, SynthSeg yields direct applications in the analysis of clinical scans, for which no general segmentation routines are available due to their highly variable acquisition procedures (sequence, resolution, hardware). Indeed, current methods deployed in the clinic include running FreeSurfer on companion 1 mm T1 scans and/or using such labels to train a supervised network (possibly with domain adaptation) to segment other sequences. However, these methods preclude the analysis of the majority of clinical datasets, where 1 mm T1 scans are rarely available. Moreover, training neural networks in the clinic is difficult in practice, since it requires corresponding expertise. In contrast, SynthSeg achieves comparable results to supervised CNNs on their training domain (especially at LR), and can be deployed much more easily since it does not need to be retrained.

7. Conclusion

In this article, we have presented SynthSeg, a learning strategy for segmentation of brain MRI and CT scans, where robustness against a wide range of contrasts and resolutions is achieved without any retraining or fine-tuning. First, we have demonstrated SynthSeg on 5000 scans spanning eight datasets, six modalities and 10 resolutions, where it maintains a uniform accuracy and almost attains the performance of supervised CNNs on their training domain. SynthSeg obtains slightly better scores than state-of-the-art domain adaptation methods for small domain gaps, while considerably outperforming them for larger domain shifts. Additionally, the proposed method is consistently more accurate than Bayesian segmentation, while being robust against PV effects and running much faster. SynthSeg can reliably be used in clinical neuroimaging studies, as it precisely detects AD atrophy patterns on HR and LR scans alike. Finally, by obtaining state-of-the-art results in cardiac cross-modality segmentation, we have shown that SynthSeg has the potential to be applied to other medical imaging problems.

While this article focuses on the use of domain randomisation to build robustness against changes in contrast and resolution, future work will seek to further improve the accuracy of the proposed method. As such, we will explore the use of adversarial networks to enhance the quality of the synthetic scans. Then, we plan to investigate the use of CNNs to "denoise" output segmentations for improved robustness, and we will examine other architectures to replace the UNet employed in this work. Finally, while the ablation of the lesion simulation in Section 5.2 is a first evidence of the robustness of SynthSeg to the presence of lesions, future work will seek to precisely quantify the performance of SynthSeg when exposed to various types of lesions, tumours, and pathologies.

The trained model is distributed with FreeSurfer. Relying on a single model will greatly facilitate the use of SynthSeg by researchers, since it eliminates the need for retraining, and thus the associated requirements in terms of hardware and deep learning expertise. By producing robust and reproducible segmentations of nearly any brain scan, SynthSeg will enable quantitative analyses of huge amounts of existing clinical data, which could greatly improve the characterisation and diagnosis of neurological disorders.

Declaration of competing interest

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

Data availability

Links to the code and public datasets used in this study have been specified in the manuscript.

Acknowledgements

This research is supported by the European Research Council (ERC Starting Grant 677697, project BUNGEE-TOOLS), the EPSRC, United Kingdom-funded UCL Centre for Doctoral Training in Medical Imaging (EP/L016478/1) and the Department of Health's NIHR-funded Biomedical Centre at UCL Hospitals. Further support is provided by Alzheimer's Research UK, United Kingdom (ARUKIRG2-019A003), the NIH BRAIN Initiative (RF1MH123195 and U01MH117023), the National Institute for Biomedical Imaging and Bioengineering (P41EB015896, 1R01EB023281, R01EB006758, R21EB018907, R01EB019956), the National Institute on Aging, United States (1R01AG070988, 1R56AG064027, 5R01A-G008122, R01AG016495, 1R01AG064027), the National Institute of Mental Health the National Institute of Diabetes and Digestive (1R21DK10827701), the National Institute for Neurological Disorders and Stroke (R01NS112161, R01NS0525851, R21NS072652, R01NS070963, R01NS0835-34, 5U01NS086625, 5U24NS10059103, R01NS105820), the Shared Instrumentation Grants (1S10RR023401, 1S10R-R019307, 1S10RR023043), the Lundbeck foundation (R3132019622), the NIH Blueprint for Neuroscience Research, United States (5U01MH093765).

The collection and sharing of the ADNI data was funded by the Alzheimer's Disease Neuroimaging Initiative (National Institutes of Health, United States Grant U01 AG024904) and Department of Defence (W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, United States, the National Institute of Biomedical Imaging and Bioengineering, and the following: Alzheimer's Association, United States; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research, Canada is providing funds for ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National

Institutes of Health. The grantee is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI is disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2023.102789.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., 2016. Tensorflow: A system for large-scale machine learning. In: Symposium on Operating Systems Design and Implementation. pp. 265–283.

Arsigny, V., Commowick, O., Pennec, X., Ayache, N., 2006. A log-Euclidean framework for statistics on diffeomorphisms. In: Medical Image Computing and Computer Assisted Intervention. pp. 924–931.

Ashburner, J., Friston, K., 2005. Unified segmentation. NeuroImage 26 (3), 839–851.

Bengio, Y., Bastien, F., Bergeron, A., Boulanger-Lewandowski, N., Breuel, T., Chherawala, Y., Cisse, M., et al., 2011. Deep learners benefit more from out-of-distribution examples. In: International Conference on Artificial Intelligence and Statistics. pp. 164–172.

Billot, B., Cerri, S., Van Leemput, K., Dalca, A., Iglesias, J.E., 2021. Joint segmentation of multiple sclerosis lesions and brain anatomy in MRI scans of any contrast and resolution with CNNs. In: IEEE International Symposium on Biomedical Imaging. pp. 1971–1974.

Billot, B., Greve, D., Van Leemput, K., Fischl, B., Iglesias, J.E., Dalca, A., 2020a. A learning strategy for contrast-agnostic MRI segmentation. In: Medical Imaging with Deep Learning. pp. 75–93.

Billot, B., Robinson, E., Dalca, A., Iglesias, J.E., 2020b. Partial volume segmentation of brain MRI scans of any resolution and contrast. In: Medical Image Computing and Computer Assisted Intervention. pp. 177–187.

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. In: Advances in Neural Information Processing Systems, vol. 33, pp. 12546–12558.

Chaitanya, K., Karani, N., Baumgartner, C., Becker, A., Donati, O., Konukoglu, E., 2019. Semi-supervised and task-driven data augmentation. In: Information Processing in Medical Imaging. pp. 29–41.

Chartsias, A., Joyce, T., Giuffrida, M., Tsaftaris, S., 2018. Multimodal MR synthesis via modality-invariant latent representation. IEEE Trans. Med. Imaging 37, 803–814.

Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A., 2019. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. Proc. AAAI Conf. Artif. Intell. 33, 65–72.

Chen, C., Qin, C., Ouyang, C., Li, Z., Wang, S., Qiu, H., Chen, L., Tarroni, G., Bai, W., Rueckert, D., 2022. Enhancing MR image segmentation with realistic adversarial data augmentation. Med. Image Anal. 82.

Choi, H., Haynor, D., Kim, Y., 1991. Partial volume tissue classification of multichannel magnetic resonance images-a mixel model. IEEE Trans. Med. Imaging 10, 395–407.

Chollet, F., 2015. Keras. https://keras.io.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., et al., 2009. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19 (6), 579–587.

Clevert, D.A., Unterthiner, T., Hochreiter, S., 2016. Fast and accurate deep network learning by exponential linear units (ELUs). arXiv:1511.07289 [cs].

Cohen, J., 1988. Statistical Power Analysis for the Behavioural Sciences. Routledge Academic.

Dagley, A., LaPoint, M., Huijbers, W., Hedden, T., McLaren, D., et al., 2017. Harvard aging brain study: Dataset and accessibility. NeuroImage 144, 255–258.

Dempster, A., Laird, N., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. 39 (1), 1–22.

Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F., et al., 2014. The autism brain imaging data exchange: Towards large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry 19 (6), 659–667.

Dou, Q., Ouyang, C., Chen, C., Chen, H., Glocker, B., Zhuang, X., Heng, P.A., 2019. PnP-AdaNet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. IEEE Access 7.

Fischl, B., 2012. FreeSurfer. NeuroImage 62, 774–781.

Fischl, B., Salat, D., Busa, E., Albert, M., et al., 2002. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. Neuron 33, 41–55.

Fischl, B., Salat, D., van der Kouwe, A., Makris, N., Ségonne, F., Quinn, B., Dale, A., 2004. Sequence-independent segmentation of magnetic resonance images. NeuroImage 23, 69–84.

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H., 2018. Synthetic data augmentation using GAN for improved liver lesion classification. In: IEEE International Symposium on Biomedical Imaging. pp. 289–293.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2017. Domain-adversarial training of neural networks. In: Domain Adaptation in Computer Vision Applications. In: Advances in Computer Vision and Pattern Recognition, pp. 189–209.

Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Fedorov, A., Abolmaesumi, P., Platel, B., Wells, W., 2017. Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 516–524.

Gollub, R., Shoemaker, J., King, M., White, T., Ehrlich, S., et al., 2013. The MCIC collection: A shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. Neuroinformatics 11 (3), 367–388.

Havaei, M., Guizard, N., Chapados, N., Bengio, Y., 2016. HeMIS: Hetero-modal image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 469–477.

He, Y., Carass, A., Zuo, L., Dewey, B., Prince, J., 2021. Autoencoder based self-supervised test-time adaptation for medical image analysis. Med. Image Anal. 102136.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T., 2018. CyCADA: Cycle-consistent adversarial domain adaptation. In: International Conference on Machine Learning. pp. 1989–1998.

Holmes, A., Hollinshead, M., O'Keefe, T., Petrov, V., Fariello, G., et al., 2015. Brain genomics superstruct project initial data release with structural, functional, and behavioral measures. Sci. Data 2 (1), 1–16.

Huo, Y., Xu, Z., Moon, H., Bao, S., Assad, A., Moyo, T., Savona, M., Abramson, R., Landman, B., 2019. SynSeg-Net: Synthetic segmentation without target modality ground truth. IEEE Trans. Med. Imaging 38 (4), 1016–1025.

Hynd, G., Semrud-Clikeman, M., Lorys, A., Novey, E., Eliopulos, D., Lyytinen, H., 1991. Corpus callosum morphology in attention deficit-hyperactivity disorder: Morphometric analysis of MRI. J. Learn. Disabil. 24 (3), 141–146.

Iglesias, J.E., Billot, B., Balbastre, Y., Tabari, A., Conklin, J., Gilberto González, R., Alexander, D.C., Golland, P., Edlow, B.L., Fischl, B., 2021. Joint super-resolution and synthesis of 1 mm isotropic MP-RAGE volumes from clinical MRI exams with scans of different orientation, resolution and contrast. NeuroImage 237.

Iglesias, J.E., Insausti, R., Lerma-Usabiaga, G., Bocchetta, M., Van Leemput, K., Greve, D., 2018. A probabilistic atlas of the human thalamic nuclei combining ex vivo MRI and histology. NeuroImage 183, 314–326.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456.

Isensee, F., Jaeger, P., Kohl, S., Petersen, J., Maier-Hein, K., 2021. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211.

Isola, P., Zhu, J., Zhou, T., Efros, A., 2017. Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134.

Jakobi, N., Husbands, P., Harvey, I., 1995. Noise and the reality gap: The use of simulation in evolutionary robotics. In: Advances in Artificial Life. pp. 704–720.

Jog, A., Fischl, B., 2018. Pulse sequence resilient fast brain segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 654–662.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Nori, A., Criminisi, A., Rueckert, D., Glocker, B., 2017a. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Information Processing in Medical Imaging. pp. 597–609.

Kamnitsas, K., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Rueckert, D., Glocker, B., 2017b. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.

Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E., 2018. A lifelong learning approach to brain MR segmentation across scanners and protocols. In: Medical Image Computing and Computer Assisted Intervention. pp. 476–484.

Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E., 2021. Test-time adaptable neural networks for robust medical image segmentation. Med. Image Anal. 68, 101907.

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

Mahmood, F., Borders, D., Chen, R., Mckay, G., Salimian, K., Baras, A., Durr, N., 2020. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. IEEE Trans. Med. Imaging 39 (11), 3257–3267.

Marcus, D., Wang, T., Parker, J., Csernansky, J., Morris, J., Buckner, R., 2007. Open access series of imaging studies: Cross-sectional MRI data in Young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19, 498–507.

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., et al., 2011. The parkinson progression marker initiative (PPMI). Progress Neurobiol. 95 (4), 629–635.

Milletari, F., Navab, N., Ahmadi, S., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision. pp. 565–571.

Moshkov, N., Mathe, B., Kertesz-Farkas, A., Hollandi, R., Horvath, P., 2020. Test-time augmentation for deep learning-based cell segmentation on microscopy images. Sci. Rep. 10, 182–192.

Müller, R., Shih, P., Keehn, B., Deyoe, J., Leyden, K., Shukla, D., 2011. Underconnected, but How? A survey of functional connectivity MRI studies in autism spectrum disorders. Cerebral Cortex 21 (10), 2233–2243.

Pan, S., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 45–59.

Puonti, O., Iglesias, J.E., Van Leemput, K., 2016. Fast and sequence-adaptive whole-brain segmentation using parametric Bayesian modeling. NeuroImage 143, 235–249.

Puonti, O., Van Leemput, K., Saturnino, G., Siebner, H., Madsen, K., Thielscher, A., 2020. Accurate and robust whole-head segmentation from magnetic resonance images for individualized head modeling. NeuroImage 219, 117044.

Richter, S.R., Vineet, V., Roth, S., Koltun, V., 2016. Playing for data: Ground truth from computer games. In: Computer Vision – ECCV. pp. 102–118.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 234–241.

Sandfort, V., Yan, K., Pickhardt, P., Summers, R., 2019. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. Sci. Rep. 9 (1), 16884.

Tae, W., Kim, S., Lee, K., Nam, E., Kim, K., 2008. Validation of hippocampal volumes measured using a manual method and two automated methods (FreeSurfer and IBASPM) in chronic major depressive disorder. Neuroradiology 50 (7), 569–581.

The ADHD-200 Consortium, 2012. The ADHD-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. Front. Syst. Neurosci. 6.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P., 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 23–30.

Tobon-Gomez, C., Geers, A.J., Peters, J., 2015. Benchmark for algorithms segmenting the left atrium from 3D CT and MRI datasets. IEEE Trans. Med. Imaging 34 (7), 1460–1473.

Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., Birchfield, S., 2018. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 969–977.

Van Essen, D., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., et al., 2012. The human connectome project: A data acquisition perspective. NeuroImage 62 (4), 2222–2231.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imaging 87–98.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 2003. A unifying framework for partial volume segmentation of brain MR images. IEEE Trans. Med. Imaging 22.

Warfield, S., Zou, K., Wells, W., 2004. Simultaneous truth and performance level estimation: An algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921.

Wells, W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R., 1996. Multi-modal volume registration by maximization of mutual information. Med. Image Anal. 1 (1), 35–51.

West, J., Fitzpatrick, J., Wang, M., Dawant, B., Maurer, C., Kessler, R., Maciunas, R., et al., 1997. Comparison and evaluation of retrospective intermodality brain image registration techniques. J. Comput. Assist. Tomogr. 21 (4), 554–568.

You, C., Xiang, J., Su, K., Zhang, X., Dong, S., Onofrey, J., Staib, L., Duncan, J., 2022a. Incremental learning meets transfer learning: Application to multi-site prostate MRI segmentation. In: Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health. pp. 3–16.

You, C., Zhou, Y., Zhao, R., Staib, L., Duncan, J., 2022b. SimCVD: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. IEEE Trans. Med. Imaging 41 (9), 2228–2237.

Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D., Xu, Z., 2020. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans. Med. Imaging 39 (7), 2531–2540.

Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D., Chen, D., 2017. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: Medical Image Computing and Computer-Assisted Intervention. pp. 408–416.

Zhang, Z., Yang, L., Zheng, Y., 2018. Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 242–251.

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J., Dalca, A., 2019. Data augmentation using learned transformations for one-shot medical image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 8543–8553.

Zhuang, X., Li, L., Payer, C., 2019. Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. Med. Image Anal. 58.

SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining
Supplementary materials

Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, and Juan Eugenio Iglesias

Supplement 1: Examples of generated synthetic scans

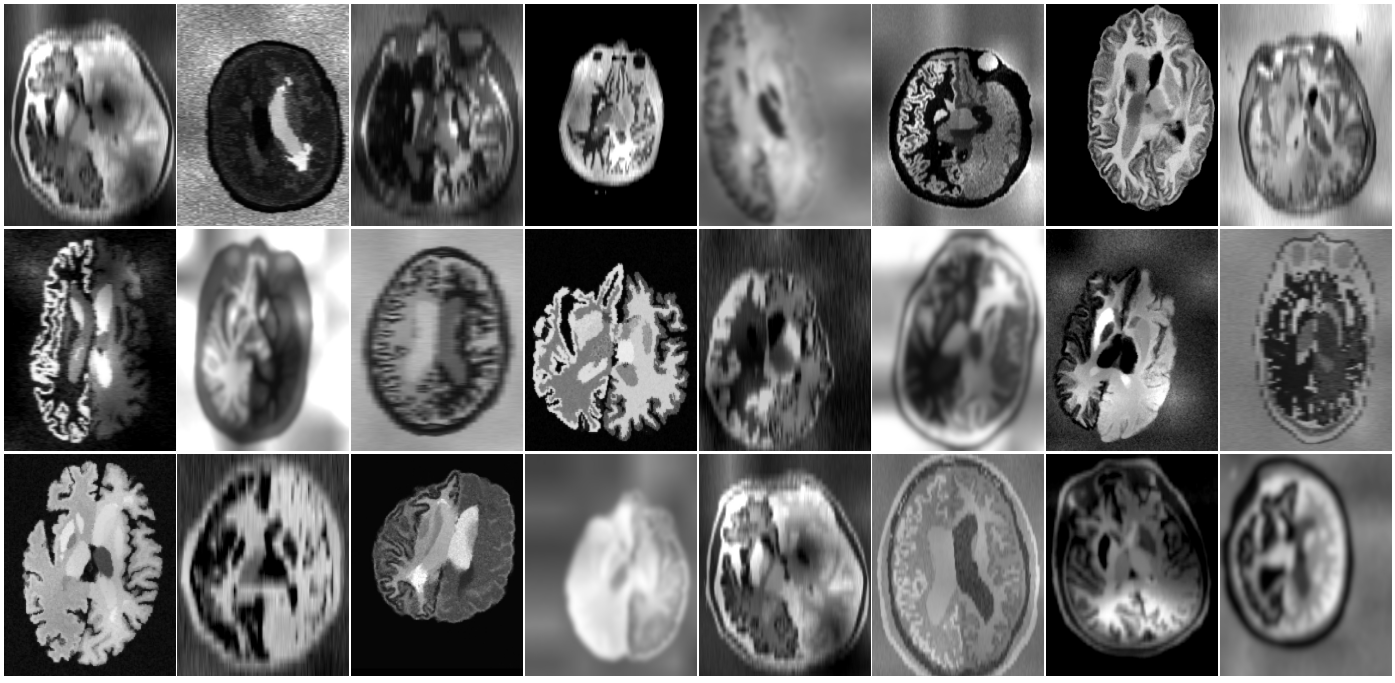

Fig. S1. Representative samples from the presented generated model. Synthetic scans present a considerable diversity in terms of contrasts and resolutions, but also in terms of sizes, shapes, bias fields, skull stripping, lesions, and anatomical morphology.

Supplement 2: Values of the generative model hyperparameters

Table S1. Values of the hyperparameters controlling the generative model. Intensity parameters assume an input in the [0, 255] interval. Rotations are expressed in degrees, and spatial measures are in millimeters.

Hyperparameter	a_{rot}	b_{rot}	a_{sc}	b_{sc}	a_{sh}	b_{sh}	a_{tr}	b_{tr}	b_{nonlin}	a_{μ}	b_{μ}	a_{σ}	b_{σ}	b_B	σ_{γ}^2	r_{HR}	b_{res}	a_{α}	b_{α}
Value	-20	20	0.8	1.2	-0.015	0.015	-30	30	4	0	255	0	35	0.6	0.4	1	9	0.95	1.05

Supplement 3: List of label values used during training

Table S2. List of the labels used for image synthesis, prediction, and evaluation. Note that during generation, we randomly model skull stripping with 50% probability, by removing all extra-cerebral labels from the training segmentation. In addition, we also model *imperfect* skull stripping, with a further 50% chances (so 25% of the total cases), by removing all the extra-cerebral labels except the cerebro-spinal fluid (which surrounds the brain). Different contralateral labels are used for structures marked with $^{R/L}$.

Label	removed for skull stripping simulation	predicted	evaluated
Background	N/A	yes	no
Cerebral white matter$^{R/L}$	no	yes	yes
Cerebral cortex$^{R/L}$	no	yes	yes
Lateral ventricle$^{R/L}$	no	yes	yes
Inferior Lateral Ventricle$^{R/L}$	no	yes	no
Cerebellar white matter$^{R/L}$	no	yes	yes
Cerebellar grey matter$^{R/L}$	no	yes	yes
Thalamus$^{R/L}$	no	yes	yes
Caudate$^{R/L}$	no	yes	yes
Putamen$^{R/L}$	no	yes	yes
Pallidum$^{R/L}$	no	yes	yes
Third ventricle	no	yes	yes
Fourth ventricle	no	yes	yes
Brainstem	no	yes	yes
Hippocampus$^{R/L}$	no	yes	yes
Amygdala$^{R/L}$	no	yes	yes
Accumbens area$^{R/L}$	no	yes	no
Ventral DC$^{R/L}$	no	yes	no
Cerebral vessels$^{R/L}$	no	no	no
Choroid plexus$^{R/L}$	no	no	no
White matter lesions$^{R/L}$	no	no	no
Cerebro-spinal Fluid (CSF)	yes/no	no	no
Artery	yes	no	no
Vein	yes	no	no
Eyes	yes	no	no
Optic nerve	yes	no	no
Optic chiasm	yes	no	no
Soft tissues	yes	no	no
Rectus muscles	yes	no	no
Mucosa	yes	no	no
Skin	yes	no	no
Cortical bone	yes	no	no
Cancellous bone	yes	no	no

Supplement 4: Versions of the training label maps

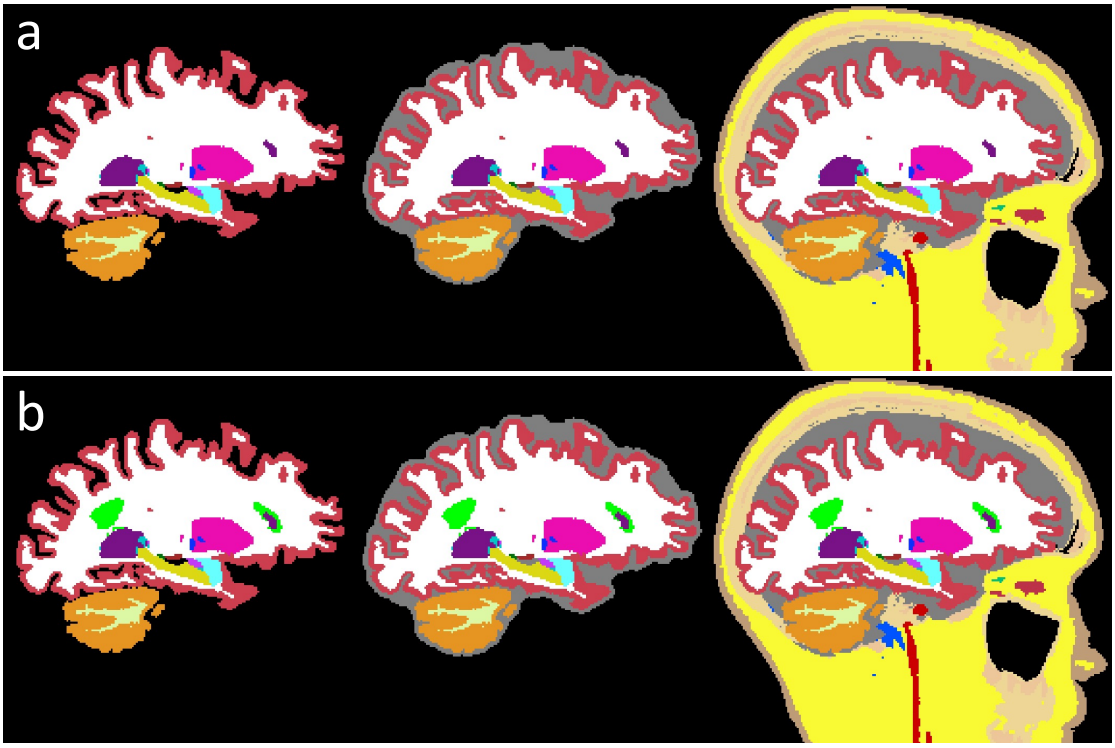

Fig. S2. Example of all versions of the label maps used during training. Each map is available (a) with, or (b) without lesion labels (bright green), and at different levels of skull stripping: perfect (left), imperfect (middle), or no skull stripping (right). Using these different versions of training label maps enables us to build robustness to white matter lesions and to (possibly imperfect) skull stripping. The lesion labels are obtained with FreeSurfer (Fischl, 2012) and directly "pasted" on the existing label maps. Training lesion labels are mainly located in the cerebral white matter, but occurrences are also found in the cerebellum, thalamus, pallidum, and putamen. Regarding the extra-cerebral labels, these are obtained with a Bayesian segmentation approach (Puonti et al., 2020).

Supplement 5: Modifications to the nnUNet, and TTA, and SIFA methods

All competing methods tested in this article are used with their default implementation, except for few minor differences that we list here. All the following modifications improve the scores obtained by the original implementations on the validation set.

nnUNet (Isensee et al., 2021)[6]: We now apply random flipping along the right/left axis (Dice score improvement of 0.09 on the validation set), as opposed to the original implementation where flipping was applied in any direction. This also mimics the augmentation strategy used for SynthSeg (see Section 4.2).

TTA (Karani et al., 2021)[7]: First, the image normaliser now uses five instead of three convolutional layers, which increases its learning capacity, especially in the case of large domain gaps (Dice improvement of 0.16 on the validation set). The second modification is relative to the training atlas that is used in Karani et al. (2021) as ground-truth during the first steps of the adaptation. Here, we add an offline step, where we rigidly register this atlas to the test scan with NiftyReg (Modat et al., 2010) We emphasise that this step was not done in the original implementation, since test scans in Karani et al. were already pre-aligned. Moreover, we also increase the number of steps during which the atlas is used by increasing the beta threshold from 0.25 to 0.4 (Karani et al., 2021) (Dice improvement of 0.06 on the validation set). Finally, we replace the existing data augmentation scheme by the same spatial, intensity and bias augmentations as for SynthSeg (Dice improvement of 0.03 on the validation set).

SIFA (Chen et al., 2019)[8]: For this method, we added an online data augmentation step during training, where we apply the same spatial, intensity and bias augmentations as for SynthSeg (Dice improvement of 0.18 on the validation set).

[6] https://github.com/MIC-DKFZ/nnUNet
[7] https://github.com/neerakara/test-time-adaptable-neural-networks-for-domain-generalization
[8] https://github.com/cchen-cc/SIFA

Supplement 6: Number of retraining for each value of N

Table S3. Number of label maps and associated retrainings used to assess performance against the amount of training subjects. Scores are then averaged across the retrainings. We emphasise that the number of retrainings is higher for low values of N to compensate for the greater variability in random subject selection. All training label maps are randomly taken from the manual segmentations of T1-39.

Number of training segmentations	1	5	10	15	20
Number of retrainings	8	5	4	3	2

Supplement 7: Training label maps for cardiac segmentation

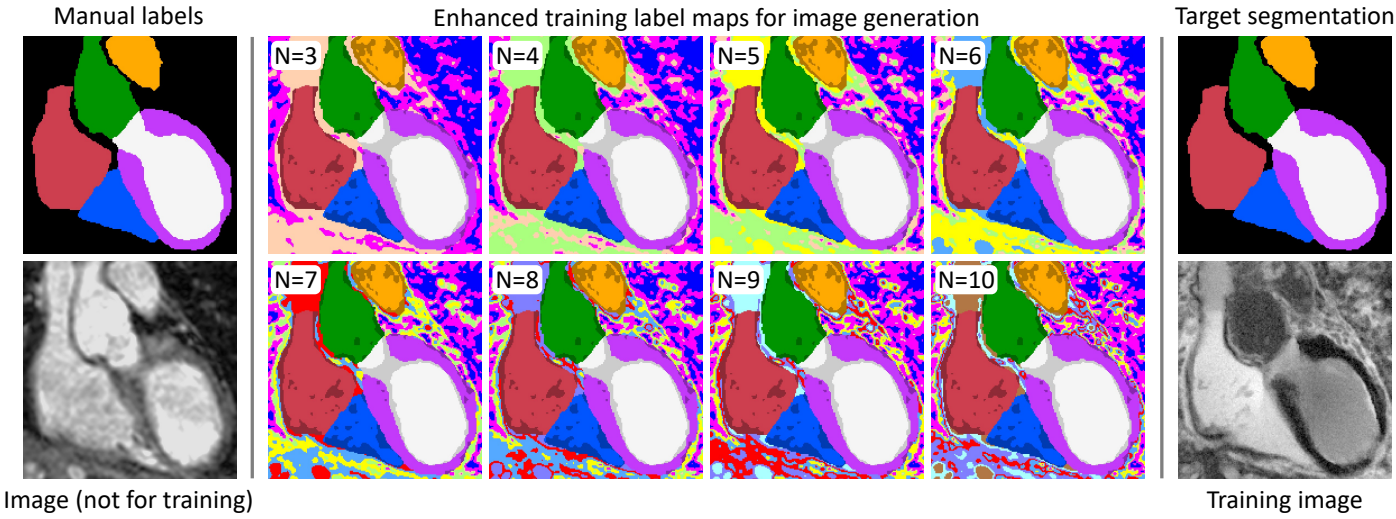

Fig. S3. Training label maps for extension to cardiac segmentation are obtained by combining three types of labels. We first start with manual delineations (top left). Second, we obtain labels for sub-regions (represented in the middle label maps by slightly shaded colours) of each of these foreground regions by clustering the associated intensities in the corresponding image (bottom left). Third, we obtain automated labels for the background structures (i.e., vessels, bronchi, bones, etc., for which no manual segmentations are available), by clustering the corresponding intensities into N classes ($N \in [3, 10]$), in order to model them with different levels of granularity. During training, one of these enhanced label maps is randomly selected to synthesise a training image (bottom right) by using the proposed generative model. Note that the target segmentation is reset to the initial manual labels.

Supplement 8: Values of the generative model hyperparameters used in the heart experiments

Table S4. Values of the hyperparameters controlling the generative model used in the heart experiments. As before, intensity parameters assume an input in the [0, 255] interval. Rotations are expressed in degrees, and spatial measures are in millimeters.

Hyperparameter	a_{rot}	b_{rot}	a_{sc}	b_{sc}	a_{sh}	b_{sh}	a_{tr}	b_{tr}	b_{nonlin}	a_{μ}	b_{μ}	a_{σ}	b_{σ}	b_B	σ_{γ}^2	r_{HR}	b_{res}	a_{α}	b_{α}
Value	-45	45	0.8	1.2	-0.02	0.02	-40	40	8	0	255	0	35	0.7	0.5	1	10	0.95	1.05