QUANTIFYING WHITE MATTER HYPERINTENSITY AND BRAIN VOLUMES IN HETEROGENEOUS CLINICAL AND LOW-FIELD PORTABLE MRI

Pablo Laso [1]

[1] Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands

ABSTRACT

Brain atrophy and white matter hyperintensity (WMH) are closely related to burdening diseases like stroke or multiple sclerosis. Automated segmentation and quantification is desirable but existing methods require high-resolution MRI with good signal-to-noise ratio (SNR). This precludes application to clinical and low-field portable MRI (pMRI) scans, thus hampering large-scale tracking of atrophy and WMH progression, especially in underserved areas where pMRI has huge potential. Here we present a method that segments white matter hyperintensity and 36 brain regions from scans of any resolution and contrast (including pMRI) without retraining. We show results on six public datasets and on a private dataset with paired high- and low-field scans (3T and 64mT), where we attain strong correlation between the WMH (ρ=.85) and hippocampal volumes (ρ=.89) estimated at both fields. Our method is publicly available as part of FreeSurfer, at: http://surfer.nmr.mgh.harvard.edu/fswiki/WMH-SynthSeg.

1. INTRODUCTION

White matter hyperintensity (WMH) on magnetic resonance imaging of the human brain is associated with stroke, cognitive decline, and cardiovascular disease. WMH is frequently detected in brain MRI scans in the elderly population: for example, in a recent observational study with adult patients with a vascular risk factor being evaluated for a non-stroke complaint, more than half of the subjects had WMH [1]. In addition, WMH is a hallmark of multiple sclerosis (MS), a disease that creates a demyelination process that may lead to disability [2]. The MS disease process is correlated with other neurodegeneration, leading to abnormally high atrophy rates in different brain regions [3]. Therefore, closer monitoring of WMH and its progression is desirable at a larger scale.

Inexpensive portable MRI (pMRI) technology is becoming increasingly available and has huge potential for imaging WMH in the community at large scale. For example, the low-field (64mT) Swoop system (Hyperfine Inc) produces images that have good agreement with their high-field counterparts when WMH are scored by a radiologist [1]. A crucial component of large-scale deployment is automated segmentation and quantification of WMH and brain regions, as manual identification and tracing of regions of interest (ROIs) in 3D is not only impractical, but also irreproducible.

The ability to quantify WMH and brain anatomy is also very desirable for clinical MRI. As opposed to research MRI, which is typically isotropic, clinical scans often comprise a smaller number of slices acquired in 2D. These take less time to review by a clinician and are less susceptible to motion artifacts. Precise quantitative analysis of these scans would allow closer tracking of atrophy and WMH progression.

A large array of methods exist for segmenting brain anatomy and WMH. Representative classical methods include: FreeSurfer [4] and FSL [5] for brain ROIs; LST [6] and BIANCA [7] for WMH; or SAMSEG [8, 9], which segments both. Machine learning techniques, often using convolutional neural networks (CNNs), include: QuickNat [10] or FastSurfer [11], for brain ROIs; or [12, 13] for WMH. These methods are designed for conventional high-field MRI (1.5-3T), and often have requirements in terms of resolution (typically 1mm isotropic), pulse sequence (often T1-weighted for anatomy, FLAIR for WMH), or both. Therefore, they struggle with the huge variability in orientation (axial, coronal, sagittal), resolution, and contrast of clinical MRI in real scenarios. This problem is exacerbated in pMRI, where the low field imposes limitations in signal-to-noise ratio (SNR) that are compensated with large voxel sizes, and where the geometry of the scanner often leads to severe signal loss away from its center. While domain adaptation [14] can mitigate these problems to some extent, a CNN than can handle any MRI contrast and resolution without retraining is highly desirable.

Here we present WMH-SynthSeg, a CNN that segments WMH and brain anatomy from scans of any resolution and contrast, including low-field pMRI. WMH-SynthSeg builds on our previous work on domain randomization [15, 16] to achieve such agnosticity. Compared with our previous method for simultaneous segmentation of WMH and anatomy [17], WMH-SynthSeg: *(i)* does not require retraining; *(ii)* uses a specific WMH model and a composite loss to improve sensitivity and specificity; *(iii)* adapts to low-field MRI; and *(iv)* uses multi-task learning for enhanced robustness. We show that, as a result, WMH-SynthSeg can robustly segment WMH and anatomy from clinical and pMRI.

2. LITERATURE REVIEW

2.1. Classical methods

Several studies have already outlined the feasibility of LF or portable MRI (pMRI) in the clinical field for moderate to severe MS lesions [18, 19]. The latter proved the sensitivity of existing algorithms initially designed for 3T MRI on LF MRI. In it, [20] designed MIMoSA, an algorithm that was used for automatic segmentation: a 3T data automated pipeline utilizing the coupling of shared information between modalities, to generate probability maps of white matter lesions. MIMoSA yielded superior segmentation results compared to that of [21], known as OASIS; and LesionTOADS, by [22], a segmentation algorithm that combines fuzzy c-means with the integration of topological constraints and a statistical atlas.

[8] developed SAMSEG, a methods that utilizes a generative approach for automated segmentations by inverting a forward probabilistic model. The method operates on multi-contrast brain MRI scans, represented by an intensity matrix. The corresponding labels are estimated by sampling from a segmentation prior and a likelihood function. The segmentation process involves inferring the unknown labels from the observed intensities under the generative model. The segmentation prior and likelihood used in SAMSEG, along with the resulting model, are summarized for obtaining automated segmentations.

SAMSEG was originally designed to handle brain segmentation without white matter lesions. Using a contrast-adaptive method, [9] builds on SAMSEG to propose a method for simultaneous segmentation of white matter lesions and normal-appearing neuroanatomical structures from multi-contrast brain MRI scans of MS patients. The method integrates a novel model for white matter lesions into a previously validated generative model for whole-brain segmentation, allowing adaptation to different scanners and imaging protocols without retraining.

Other non-DL-based algorithms include LST (as part of SPM) and Bianca (FSL). LST is designed by [6] for the automatic segmentation of brain lesions, particularly in multiple sclerosis (MS) patients. The authors present a robust and efficient algorithm that utilizes a combination of tissue probability maps and intensity-based features to accurately identify and segment lesions. Their approach demonstrates promising results in terms of lesion detection and volume estimation, providing valuable tools for studying MS pathology. Respectively, [7] designed Bianca, i.e., a tool developed within the FSL (FMRIB Software Library) software package for the automated segmentation of white matter hyperintensities (WMH) in brain MRI scans. Bianca utilizes a Bayesian model to incorporate spatial information and intensity features to accurately identify and quantify WMH. The authors demonstrate the effectiveness of Bianca in various datasets, showing its potential for studying WMH in clinical and research settings. The tool offers a valuable resource for investigating the impact of WMH on brain health and cognitive function.

2.2. Deep Learning methods

The aforementioned non-DL methods often rely on well-established algorithms that leverage specific domain knowledge or incorporate explicit assumptions about the data. These methods can be effective when the task at hand is well-defined, the data characteristics are well-understood, and the features used for segmentation are carefully designed.

DL-based algorithms, nonetheless, achieved state-of-the-art performance in various image analysis tasks, such as segmentation [23], including MS lesions segmentation in neuro MRI [24]. The paper by [25] introduces the V-Net, a fully convolutional neural network designed for volumetric medical image segmentation. The network architecture incorporates 3D convolutions and skip connections to effectively segment medical images. The authors demonstrate the effectiveness of the V-Net on various medical imaging tasks, highlighting its potential for accurate and efficient segmentation.

Similarly, [26] presents an efficient multi-scale 3D convolutional neural network (CNN) with a fully connected conditional random field (CRF) for accurate brain lesion segmentation. The proposed network combines multiple scales of information to improve segmentation performance. The authors demonstrate the effectiveness of their approach on brain lesion segmentation tasks, achieving high accuracy. The integration of the fully connected CRF further enhances the segmentation results.

Other recent studies also include new approaches such as single image super-resolution. In it, [27] trained a convolutional neural network using pairs of noisy low-resolution images and noise-free high-resolution images.

2.2.1. Domain adaptation

However, the performance of DL methods heavily relies on the availability of large, labeled datasets for training. Furthermore, DL methods struggle to adapt to different images and resolution. This challenge is known as **domain gap** and it is precisely this lack of generalisability that keeps many algorithms back from being successfully implemented.

[28] combined a probabilistic, atlas-based approach with unsupervised DL in an attempt to overcome the domain gap challenge. A probabilistic atlas is a volume where each voxel is assigned a vector indicating the prior probability that such segmentation label be observed at that specific point in space. The employment of prior knowledge has been also exploited in similar studies. [29] introduced the concept of template transformer networks. [30] use a stacked convolutional autoencoder to capture the underlying anatomy and represent such as a statistical distributions that the network will follow.

2.2.2. Data Augmentation

Several authors have shown the relevance of data augmentation for variance-robust learning [31, 32]. In the context of one-shot image segmentation, [33] use a set of unlabeled examples along with a single labeled example to learn spatial and appearance transformations that will be later used for image synthesis. These synthetic images are subsequently used to form dataset big enough for successful supervised training.

Similarly, albeit armed with a more substantial dataset, both [16] and [15] use a randomization strategy in which they use synthetic MRI images to gain generalisability for their algorithms, i.e., SynthSR and SynthSeg, respectively.

2.2.3. Synthetic images

In the context of LF brain segmentation, SynthSR is a method that utilizes DL to reconstruct high-resolution images from low-resolution MRI scans, thereby enhancing resolution in MR images to address the limitations of LF MRI systems. By generating high-resolution images from their low-resolution counterparts, SynthSR facilitates the visualization of fine details while also improving the accuracy of brain structure segmentation in LF MRI. Consequently, this advancement enables a more precise brain segmentation. However, it was not trained taking into account MS lesions, hence leading to the removal the characteristic signs of the pathology [16].

Similarly, SynthSeg also emerges as a valuable tool. SynthSeg is a deep learning-based method that focuses on the accurate segmentation of brain structures. By leveraging the power of deep learning algorithms, SynthSeg claims to effectively segment brain regions in LF MRI images, compensating for the challenges associated with lower image quality and reduced spatial resolution. However, when fed our LF images, we found that this approach fails and therefore does not fully aid in precise delineation of brain structures without the image having been previously processed by SynthSR. Consequently, MS lesions are erased prior to SynthSeg-based segmentation and cannot facilitate an improved analysis and detection of such neurological condition in LF settings like ours [15]. Figure 1 shows the training process for SynthSeg. Label generation maps are fed into a generative model that will create a synthetic image used for training the 3D Unet. The predicted label map is compared against the ground truth label segmentation map, and the loss thereby computed is back-propagated to adjust the weights of the network during training.

Recent research conducted by some of our collaborators at UCL (London, UK) led to the implementation of a novel algorithm for brain segmentation, named MindGland. It incorporates an ensemble of eleven NNs which collectively produce a majority-voted label map. While initially not intended for LF MRI applications, our findings indicate its exceptional performance in capturing LF MR lesions, without needing any previous processing. The algorithm demonstrates promising potential for accurately delineating MS lesions in the context of LF MRI, highlighting its relevance and applicability within this domain. It however shorts fall to produce a complete segmentation of the brain as in SynthSeg, delineating a smaller number of sections and mixing some labels together.

2.2.4. Low-field domain

Regarding models performance in the LF domain, most of them, like SAMSEG or LST, were not developed for that aim, so lower or null performance can be expected.

Other methods, such as MindGland, and WMH-SynthSeg (ours), have been trained on synthetic data, also simulating very adverse conditions that make them more adaptable and suitable to run on LF scans. The base algorithm we are building on top of, i.e., SynthSeg, however, cannot perform on LF. Figure 2 (b) shows the product of SynthSeg on a LF scan (a). Super-resolving by means of SynthSR (d) is recommended in the documentation as a prior step to segmentation, in an aim to overcome this challenge posed by the low-resolution of a LF MRI scan. SynthSR brings the LF image onto the HF domain, preserving its anatomical characteristics. However, it is not designed to maintain WM lesions, which are removed from the image and therefore cannot be used for our ultimate goal, that is, segmenting WMH -although it does aid to capture the rest of the brain structures (e).

MindGland, on the other hand, is able to run on the same LF scan that SynthSeg failed (c). Figure 2 shows the prediction label map produced by MindGland on WM lesions. Since no prior steps were taken and it was run on the original LF image, WMH are preserved and may be segmented. This makes MindGland an algorithm worth examining in the context of pMRI and WMH quantification.

2.2.5. Architectures

The aforementioned methods illustrate the importance of data pre-processing in DL. As a more topological approach, however, several algorithms architectures can be highlighted within the scope of medical image segmentation.

The paper by [31] introduces the `U-Net` architecture, which is convolutional neural network designed for biomedical image segmentation, and considered most suitable. The U-Net model consists of an encoder-decoder structure with skip connections, allowing for the integration of both local and global information in the segmentation process. The authors demonstrate the effectiveness of U-Net on various biomedical image segmentation tasks, achieving accurate and detailed segmentation results. The U-Net architecture has since become widely used in the field of medical image analysis.

Some time later, [34] built on the aforementioned U-net by replacing all 2D operations by their 3D counterparts. The `3D U-Net` model has been shown to achieve competitive

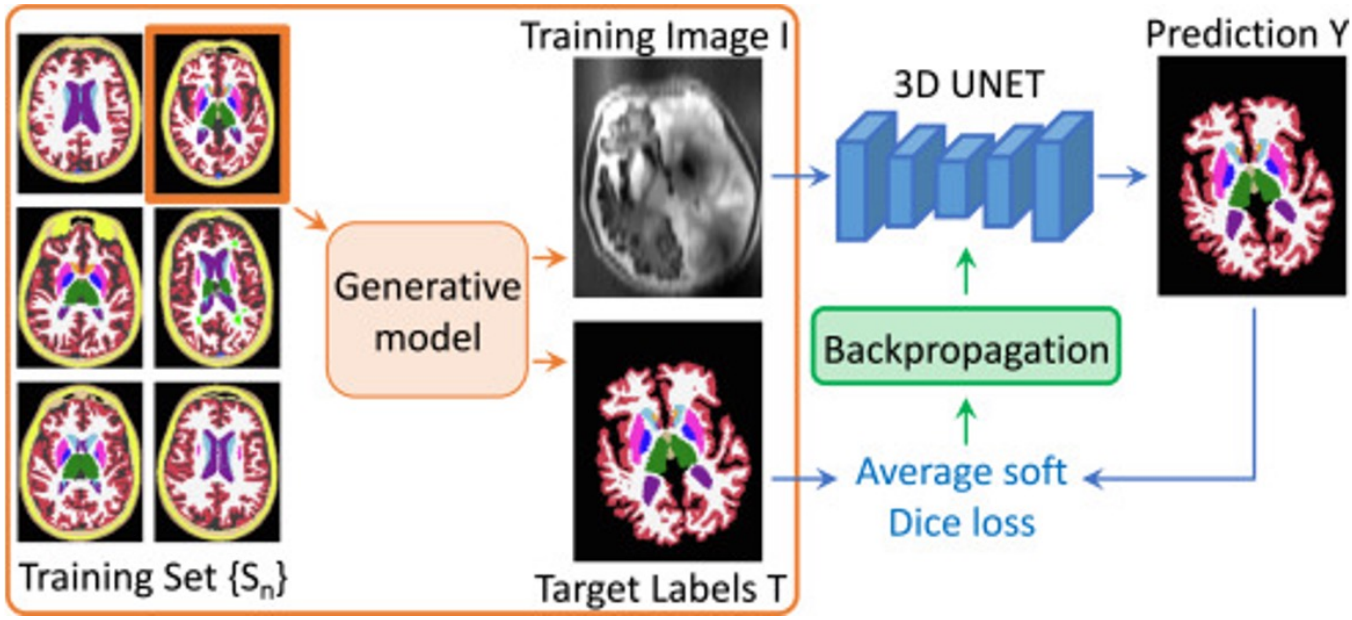

Fig. 1: How label maps are used for synthetic image generation and training of SynthSeg: Label maps (ground truth labels) are used for the generation of synthetic images. That synthetic image is fed into the 3-Unet. Its predicted label map is compared to the ground truth label map, and the error back-propagated. We note that SynthSeg uses DICE in its loss function.

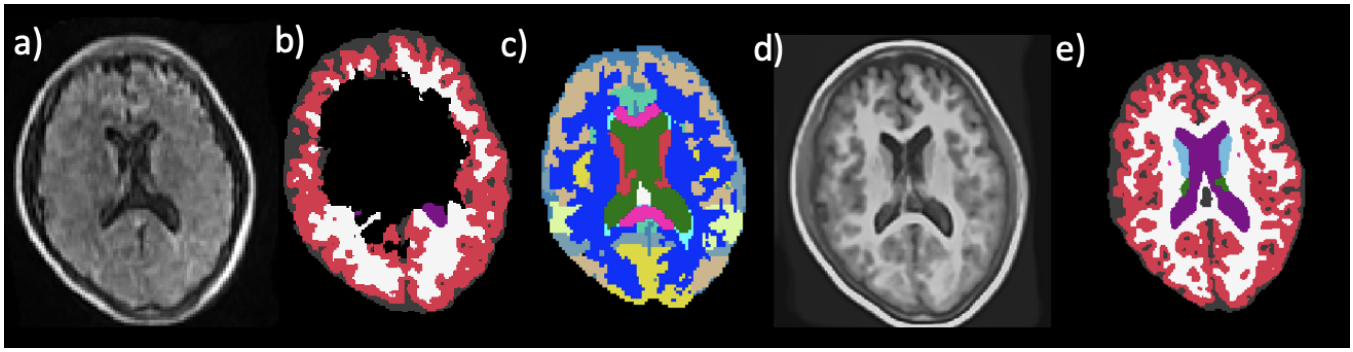

Fig. 2: (a) LF sample image from the MGH dataset. (b) Synthseg label map. (c) MindGland label map. (d) Super-resolved image by SynthSR. (e) Synthseg on (d), i.e., on the super-resolved image. It is observable how SynthSeg fails on a LF scan without it having been previously super-resolved. SynthSR super-resolution does, however, remove WM lesions from the image, thus turning it useless for our purpose of WMH segmentation. MindGland is able to run on LF scans without prior steps.

performance in various medical image segmentation tasks, even with limited annotated data. By leveraging the U-Net architecture and addressing the challenge of sparse annotations, this approach offers a promising solution for dense volumetric segmentation tasks in the medical imaging domain.

During the Medical Segmentation Decathlon (MSD) organized by [35], three DL algorithms achieved top scores. `K.A.V.athlon` method was proposed by [36]. It utilized an automated training and prediction process using image data and descriptions, combining V-Net and U-Net architectures with SE and residual blocks, employing various augmentations and DSC loss with the Adam optimizer, without any human intervention or parameter changes. No ensembling strategy was employed. [37] designed `NVDLMED`, i.e., a fully-supervised uncertainty-aware multi-view co-training strategy, utilizing 2D pre-trained models and three views to enhance robustness and generalization. They employed a 3D ResNet with anisotropic kernels and applied augmentation techniques. The ensemble consisted of three models trained on different views (coronal, sagittal, axial) to handle the ten tasks and utilized the DSC loss with the SGD optimizer.

Finally, the `nnU-Net` that was proposed by [38] took the first place at the MSD challenge -and kept wining other challenges during the following years ([39]). It is designed to handle a hugely wide range of medical image segmentation tasks and modalities with minimal manual intervention

-including brain, hypothalamus, lung, or heart. The architecture incorporates 3D U-Net and various data augmentation techniques, achieving state-of-the-art performance in different medical imaging challenges. There are several parameters that are fixed, such as the optimizer, the loss function, or the architecture template; rule-based parameters that are set according to the specific data fingerprint, such as the preprocessing, batch size, or topology; as well as empirical parameters that are chosen by cross-validation, such as the the optimal ensemble of models and the final post-processing.

The nnU-net is not the only modification of the well-known U-net architecture. [40] proposed the `MultiResUNet` as a modified and improved version upon the already state-of-the-art U-Net model, designed for multimodal biomedical image segmentation. The proposed architecture incorporates residual blocks and dense connections, which enhance information flow and enable better feature representation. These modifications improve the overall performance of the model by capturing and leveraging the complementary information present in multiple modalities. The experiments conducted in the paper demonstrate that MultiResUNet outperforms the traditional U-Net architecture in terms of segmentation accuracy and robustness in the context of multimodal biomedical image analysis.

Building in the idea of the U-net, [41] combined the strengths of both U-Net and Transformer into what they named `UNETR`. They incorporate the self-attention mechanism of Transformers into the U-Net framework, enhancing the modeling capacity and capturing long-range dependencies within the 3D medical images. The UNETR model leverages self-attention mechanisms to capture long-range dependencies and enhance information flow within the network. It is demonstrated on the paper that UNETR achieves state-of-the-art results in 3D medical image segmentation tasks, surpassing other existing methods.

Interestingly, [42] proposed replacing multiplication operations in convolutional neural networks (CNNs) by their addition-based counterparts. An Addernet achieves so by computing the l_1 distance between the feature input and the filter, rather than a conventional multiplication operation. Albeit initially meant for minimizing computational resources, this approach does promote clustering of features -rather than an angle-based division of the samples in the feature space, as would produce a conventional CNN. Although the success of clustering-based feature mapping depends on the nature of the dataset, the specific task, and the network architecture, Addernets might be able to offer benefits in promoting clustering behavior and it might be an approach worth exploring.

3. RESEARCH QUESTIONS

After our research on different methods for brain segmentation, we noticed that most algorithms fail to generalize, especially to the LF domain. In our analysis of algorithms specifically tailored for the LF setting, three main approaches were examined. SynthSeg, one of the algorithms considered, demonstrated improved performance on LF MRI when preceded by prior processing using SynthSR. However, SynthSR may not be suitable for LF MS lesions due to its erasing effect on them. Notably, the novel algorithm developed at UCL, i.e., MindGland, showcased promising results in MS lesion detection while exhibiting relatively lower segmentation performance for other brain areas.

These findings shed light on the intricacies and trade-offs associated with algorithm selection for LF MRI analysis, emphasizing the need for further tailored solutions to the specific context of LF MS. Some of the questions we aim to answer while implementing a solution are:

- What methods lead to most accurate WMH segmentation, both in HF and LF?
- Can synthetic images solve a new domain gap, i.e., LF MRI?
- Can we use some of the innovations by MindGland to further develop SynthSeg into a better brain segmentation algorithm?

4. METHODS

4.1. Data preparation

WMH segmentation was performed with regards to, usually, the T1 image. Other MRI sequence images were, however, not in the same space. We perform a robust registration to align the rest of each subject's images to the T1 image used for delineation of the lesion. We then convert those images to a reference image with a voxel size of 1x1x1. The affine transformation matrix and reference image are used to resample and register the images. By both registering and converting the unregistered MRI images, we ensure proper alignment and voxel size consistency with the mask for further analysis.

After that, lesions are smoothed with a Gaussian filter within the boundaries of the feasibly affected anatomical areas. Setting threshold was a difficult and arbitrary decision, so a different threshold was randomly set for each image. Finally, we take k-means to have some labels for the rest of the image, i.e., external structures (mainly the skull) that are still not labelled since they are not part of the brain but that will appear in real MRI images. Albeit meaningless, those labels make sure that a more realistic synthetic MRI image is generated. Thus, the NN does not collapse during testing when the whole head is fed as input, and not just the brain (labelled area of the image).

4.2. Synthetic training data

WMH-SynthSeg relies on a synthetic MRI generator similar to [16], which requires a training dataset with N 1mm isotropic T1-weighted (T1w) scans $\{I_n\}$ and corresponding

3D segmentations $\{S_n\}$; these are defined on the same 1mm isotropic grid and include labels for brain ROIs and WMH.

Figure 3 shows the process of image synthesis: At every iteration during training: *(i)* a random pair (I_n, S_n) is selected; *(ii)* (I_n, S_n) are augmented with affine and non-linear deformation; *(iii)* a Gaussian mixture model conditioned on the deformed labels is sampled independently at every voxel, with means and variances that are randomly sampled from uniform distributions – except for the WMH class (details below); *(iv)* the Gaussian image is corrupted by a random smooth bias field; *(v)* random orientation and resolution are simulated to synthesize a lower resolution scan; and *(vi)* the low-resolution scan is upsampled to the original 1mm isotropic grid. The output of the generator comprises: the upsampled synthetic scan I^{syn}, the deformed segmentation S, the deformed real image I, and the bias field B. All these outputs are defined on the original 1mm isotropic grid.

During synthetic image generation with [16], several parameters can be adjusted on the problem at hand. The main parameters that are used in this project are:

- **patch size**: This parameter defines the size of the patch. It determines the spatial dimensions, such as width, height, and depth, of the output images. Bigger patches should lead to better performance since they provide more context, but might unintentionally also cause overfitting. We settled a size of [160, 160, 160]. We saw no difference when modifying these values.
- **max rotation**, **max shear**, **max scaling**: These parameters control the maximum amount of rotation, shear, and scaling that can be applied to the input image during the generation process. They introduce variations in orientation, shape, and size, respectively, to create diverse synthetic images. We set each parameter to 15, 0.2, and 0.2, respectively.
- **nonlin scale min**, **nonlin scale max**, **nonlin std max**: These parameters influence the non-linear deformation applied to the input image. They control the range of scaling and the maximum standard deviation of the deformation field, which contributes to generating realistic deformations in the synthetic images. We set each parameter to 0.03, 0.06, and 4, respectively.
- **bf scale min**, **bf scale max**, **bf std min**, **bf std max**: These parameters control the generation of bias field in the synthetic images. They specify the range of scaling and the minimum and maximum standard deviation of the bias field, which simulate intensity variations across the image due to non-uniformities in the imaging system. We set each parameter to 0.02, 0.04, 0.1, and 0.6, respectively.
- **gamma std**: This parameter was set to 0.1. It controls the standard deviation of gamma correction applied to the synthetic images. Gamma correction adjusts the image contrast and can introduce subtle variations in the intensity values.
- **min noise std**, **max noise std**: These parameters define the range of noise standard deviation applied to the synthetic images. Adding noise introduces randomness and variations in the pixel values, contributing to a more realistic and diverse dataset. We define the such range between 5 and 15.
- **deform one hots**: This parameter determines whether to deform the one-hot encoded labels along with the input image. It determines whether the labels are deformed in accordance with the applied deformations to maintain their spatial consistency.
- **integrate deformation fields**: This parameter controls the integration of deformation fields. It determines whether the deformation fields from different transformations are accumulated, resulting in more complex and intricate deformations.
- **parcial volume**: This parameter influences the the slice thickness, specially useful for clinical scenarios where doctors prefer thick 2D scans (faster and higher SNR) that might mix several tissue types, leading to underperformance by NN segmentation methods.

The new generator we propose has 4 key improvements compared with [16]:
(i) The mean intensity of the WMH class does not follow the same distribution as the other classes (i.e., $\mathcal{U}[0, 255]$). Instead, we simulate WMH hyperintensity in T2-like sequences (including FLAIR) and hypointensity in T1w-like sequences. This is done as follows: when the white matter (WM) mean is high (over 128), we constrain the WMH mean to be lower than the WM mean (T1w-like). Conversely, when the WM mean is below 128, we constrain the WMH mean to be greater than the WM mean (T2-like). Examples of two FLAIR synthetic images are shown in Figure 4.
(ii) The standard deviation of the noise (Gaussian variances) and bias field strength is twice as large as in [16], to accommodate the lower SNR and stronger signal losses of pMRI.
(iii) The generator produces not only I^{syn} but also a deformed image I and a bias field B that will be used as regression targets by the CNN in a multi-task learning setting. This boosts the robustness of the CNN as shown in the experiments.
(iv) The sampling scheme for the random resolution covers a wider spectrum of acquisitions. 25% of the time, we generate 1mm isotropic images, to support high-resolution scans. Another 25% we generate clinical scans of random orientation with 1mm in-plane resolution and random slice spacing between 2.5mm and 8.5mm. 25% of the scans mimic the resolution of the stock sequences that the Hyperfine Swoop ships with (axial with $\sim$1.5mm in plane and 5mm spacing). The final 25% simulates more isotropic scans acquired at low field, with random voxel sizes between 2-5 mm in every direction.

4.3. Model Architecture and Training

WMH-SynthSeg uses the architecture of the original 3D U-net designed by Özgün Çiçek et al (depicted in Figure 5),

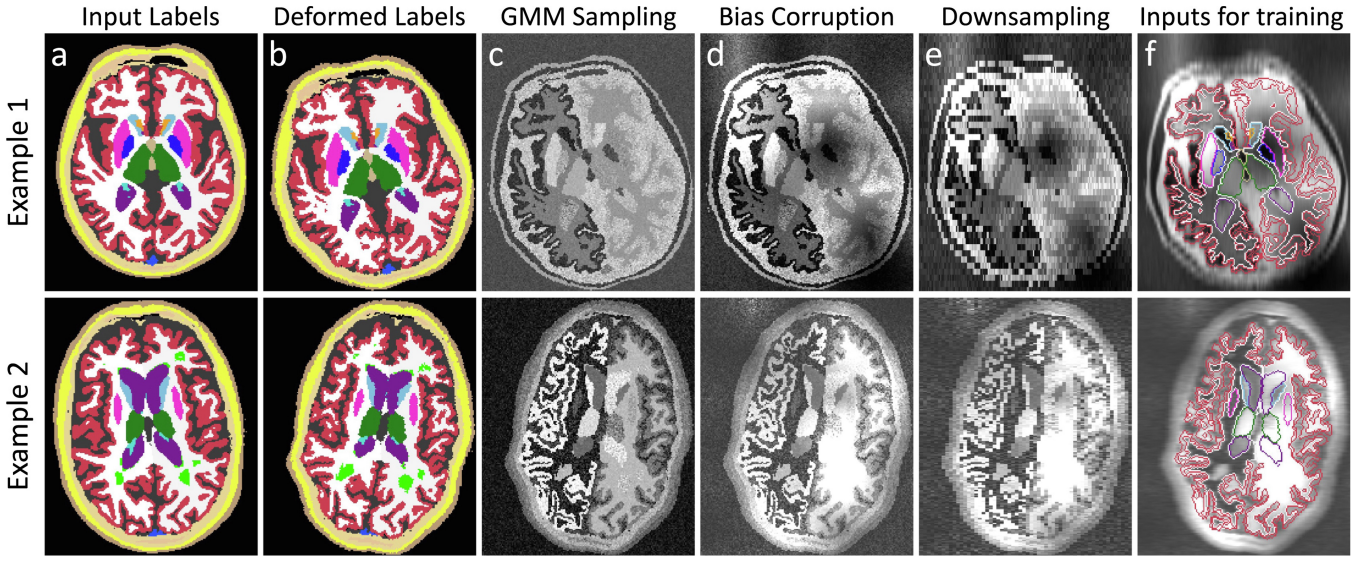

Fig. 3: SynthSeg training process.

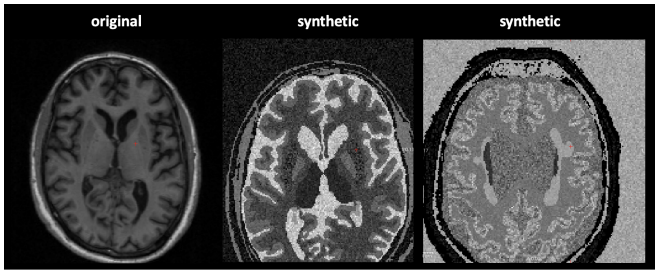

Fig. 4: Original FLAIR image from the ADNI dataset, and two synthetic images produced from its label segmentation map, in axial orientation. Since the WM intensity value that was randomly sampled falls in the lower range of the WM intensity values distribution, the WMH intensity values are forced to be brighter, as they would show on a real FLAIR image.

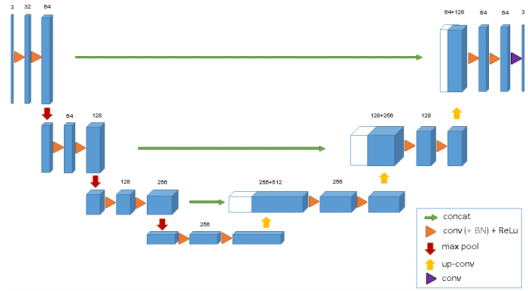

Fig. 5: Architecture of a 3D U-net [34].

which was an expansion of the original model, in 2D, by Ronneberger [31]. Our architecture is composed by five levels, 64 feature maps per level, and group normalization [43]. Each level has two convolutions (kernel size: 3x3x3) followed by ReLU activations. The final layer has $L + 2$ channels: the first L correspond to the labels and are fed to a softmax layer to produce soft segmentations; the last two correspond to the predicted bias field and high-resolution T1w intensities.

Training uses the Adam optimizer to minimize a loss function consisting of four terms with equal weight: the cross-entropy and Dice scores between the predicted and ground truth segmentations; the average ℓ_1 error of the predicted T1w intensities (normalized such that the median intensity of the WM is 1); and the ℓ_1 error of the predicted bias field (in logarithmic scale):

$$\mathcal{L} = CE(S, \hat{S}) - AvDice(S, \hat{S})) + |I - \hat{I}| + |\log B - \log \hat{B}|,$$

where $\hat{S}$, $\hat{I}$, and $\hat{B}$ are the predictions for the segmentation, T1w intensities, and bias field, respectively.

We note that, while training with Dice may be more common in segmentation, combining it with cross-entropy has two advantages. First, it provides a more informative gradient in the first iterations of training, when the gradient of the Dice loss is rather flat. And second, it explicitly penalizes false positives in scans without WMH – in which the Dice score for the WMH is zero independently of the prediction. In addition, including I and B in the loss increases the robustness of the method, as shown by the experiments below.

At test time, the input scan is resampled to 1mm isotropic resolution and fed to the CNN. Test-time augmentation is performed by left-right flipping the image, flipping the output back, and averaging with the non-flipped version. The first L channels of the output yield the final segmentation; the outputs corresponding to the bias field and the T1w intensities

are a potentially useful by-product, but are disregarded here.

Our framework is implemented in PyTorch, and its validation loss typically converges is ∼100,000 iterations using minibatches of size 1 with 160x160x160 voxel cubes.

5. EXPERIMENTS AND RESULTS

5.1. Datasets

We used nine different datasets in our experiments, some just for training ("Tr"), some for testing ("Te"), and some for both using cross validation ("Tr/Te"). The advantage of using a large number of different datasets is two-fold. Firstly, the total number of images, both for training and testing significantly increases. Secondly, we have different datasets, containing images with different characteristics. Thereby, we ensure that we are building a more robust algorithm and obtaining more reliable results, since more samples were used for both procedures but also, and most importantly, different and independent datasets were used for testing -thus reducing even further the risk of overfitting.

HCP [44] (Tr): 897 1mm isotropic scans of young subjects from the Human Connectome Project. We used FreeSurfer to automatically segment the anatomy into 36 ROIs. Since they are very young, we only consider healthy patients in this dataset, i.e., without WM lesions. Still, they are helpful for the generation of synthetic images used for overall brain segmentation during training.

ADNI [45] (Tr): 1148 1mm isotropic scans from the ADNI. We used FreeSurfer to segment the anatomy and WMH. We do not consider WMH in this dataset either.

GE3T (Tr/Te): 20 cases with 1mm isotropic T1w and 1x1x3mm axial FLAIRs. This a subset of the WMH segmentation challenge [46]. We combined the automated FreeSurfer segmentation of the T1w with the manual delineations available for the FLAIRs into a single ground truth segmentation. This dataset contains meaningful WM lesions and is these are used both for synthetic image generation during training and DICE scoring of WMH during testing.

Singapore (Tr/Te): another subset of the WMH segmentation challenge with 20 cases from a separate site (same MRI acquisitions and labels).

Utrecht (Tr/Te): another subset with 20 cases from a third site.

ISBI [47] (Tr/Te): 15 1mm isotropic T1w scans (segmented with FreeSurfer) and 1x1x2mm axial FLAIRs with manually traced WMH (merged with the anatomy into one label map). Similarly to the three datasets from teh WMH segmentation challenge, this data is used for either training and testing, depending on the fold chosen for training.

FLI-IAM [48] (Tr/Te): T1w and FLAIR scans from 15 cases (from three different centers) with varying resolution but all close to 1mm isotropic. Consensus WMH tracings are available from 7 raters, which we merged with the FreeSurfer segmentations of the T1w scans. We use all data for either training or testing.

ADHD [49] (Te): 20 1mm isotropic T1w scans from typically developing control children and adolescents and no WMH. This datatset is used to test for FP. Since no WM lesions should be found, the lesion load predicted counts as an indicator of lower specificity.

MGH (Te): 12 MS patients from our hospital (MGH) with 1mm T1w and FLAIR, as well as pMRI axial T1w and FLAIR (in-plane resolution: 1.6-1.8mm; slice spacing: 5-6mm).

5.2. Competing methods

We compare our method with: *(i)* SAMSEG [8, 9], which is a Bayesian method that is adaptive to MRI contrast, and is (to our best knowledge) the only existing (published) method that can readily segment anatomy and WMH from scans acquired with any pulse sequence; *(ii)* LST-LPA [50], which yields great performance on FLAIR acquisitions but does not work on other MRI contrasts; and *(iii)* MindGland, which does not perform as well as other methods in HF -but it has a great generalizability that allows it to adapt to the LF realm. It also does not predict as many labels as SynthSeg. Other methods such as BIANCA [7], or LST-LGA [51] were not considered since they needed retraining or more than one inpout image, respectively.

We also consider two ablations of our method to assess the importance of its components: a version with just Dice in the loss (similar to [17] but with domain randomization), and a version without the prior on the mean of the WMH class. We note that LST and SAMSEG operate at the native resolution of the scan, whereas WMH-SynthSeg always produces a 1mm isotropic segmentation.

5.2.1. SAMSEG

The SAMSEG algorithm, as presented in [8] and [9], operates within a Bayesian framework, leveraging probabilistic modeling to facilitate the segmentation of anatomical structures and white matter hyperintensities (WMH) in medical images, particularly those acquired through magnetic resonance imaging (MRI).

1. **Bayesian Framework:** SAMSEG adopts a Bayesian approach, incorporating prior knowledge and uncertainties into the segmentation process, contributing to its robustness in handling diverse imaging scenarios.
2. **Adaptability to MRI Contrasts:** SAMSEG is characterized by its adaptability to various MRI contrasts. This feature enables the algorithm to effectively segment structures and WMH across different pulse sequences, ensuring versatility in its application.
3. **Probabilistic Modeling:** The algorithm utilizes probabilistic models to represent the distribution of pixel intensities corresponding to different tissue types. This proba-

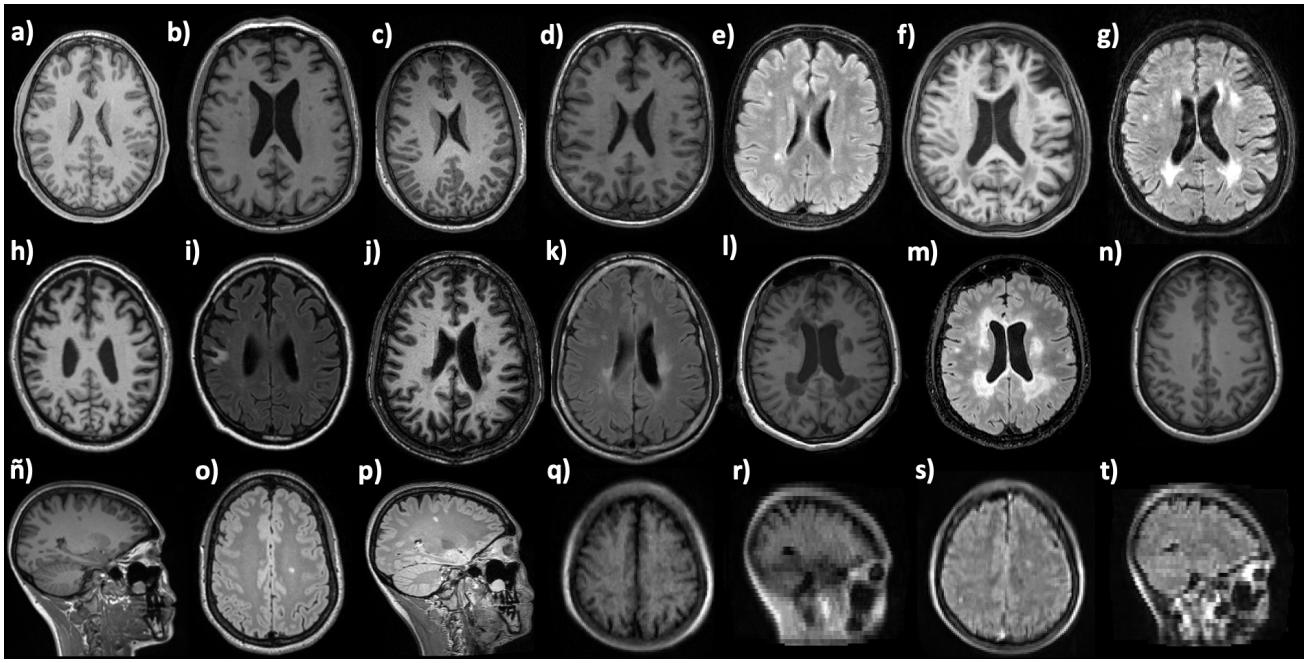

Fig. 6: Different images in each dataset: (a) HCP, (b) ADNI, (c) ADHD, (d) GE3T (T1), (e) GE3T (FLAIR), (f) Singapore (T1), (g) Singapore (FLAIR), (h) Utrecht (T1), (i) Utrecht (FLAIR), (j) ISBI (T1), (k) ISBI (FLAIR), (l) FLI-IAM (T1), (m) FLI-IAM (FLAIR), (n) MGH-HF (T1, axial), (ñ) MGH-HF (T1, sagittal), (o) MGH-HF (FLAIR, axial), (p) MGH-HF (FLAIR, sagittal), (q) MGH-LF (T1, axial), (r) MGH-LF (FLAIR, sagittal), (s) MGH-LF (T1, axial), (t) MGH-LF (FLAIR, sagittal). We notice each dataset may contain very different images to another, emphasizing the need for a robust algorithm.

bilistic modeling is crucial for capturing the inherent variability present in MRI images.

4. **Segmentation of Anatomical Structures and WMH:** SAMSEG is designed to segment both anatomical structures and WMH. The segmentation process involves the identification and classification of pixels or regions corresponding to different tissue types, including normal anatomy and pathological features.
5. **Adaptive Nature:** SAMSEG demonstrates an adaptive nature, dynamically adjusting its segmentation strategy based on the characteristics of the input MRI data. This adaptability enhances its performance across a variety of imaging conditions.
6. **Incorporation of Contrast Information:** SAMSEG incorporates information about the contrast properties of MRI images during the segmentation process. This may involve considering local intensity variations and adapting the segmentation strategy accordingly.
7. **Fast Implementation:** The algorithm is designed for efficiency and computational speed, as indicated by the term ”fast” in [8]. This characteristic is essential for practical applications in clinical settings, where real-time or near-real-time results are desirable.

5.2.2. LST

The Lesion Segmentation Toolbox (LST) is a comprehensive tool for lesion segmentation in medical images, featuring two primary algorithms: the Lesion Growth Algorithm (LGA) and the Lesion Prediction Algorithm (LPA).

- **Lesion Growth Algorithm (LGA)**.

At the core of the toolbox is the Lesion Growth Algorithm (LGA) proposed by [51]. LGA segments T2-hyperintense lesions using a combination of T1 and FLAIR images. The algorithm initially segments the T1 image into three main tissue classes (CSF, GM, and WM). This information is then combined with FLAIR intensities to calculate lesion belief maps. By thresholding these maps using a pre-chosen initial threshold (kappa), an initial binary lesion map is obtained. The map is further grown along voxels appearing hyperintense in the FLAIR image, resulting in a lesion probability map. Although an unsupervised algorithm, its drawback lies in the sensitivity to the choice of the initial threshold.

A disadvantage of this unsupervised algorithm is the choice of the initial threshold. Different kappa-values yield different segmentation results. Also, two images are required (T1 and FLAIR) for LGA to run; the reason why we will not ocnsider this method but, instead, LPA.

- **Lesion Prediction Algorithm (LPA)**.

An alternative to LGA is the Lesion Prediction Algorithm (LPA), introduced by [50]. LPA offers advantages over LGA, requiring only a FLAIR image and eliminating the need for user-set parameters. This algorithm, trained using logistic regression with data from 53 MS patients, generates lesion probability maps. The training involves a high-dimensional model with a novel approach for fitting large-scale regression models. The derived parameters are then used for lesion segmentation in new images, providing voxel-specific estimates of lesion probability.

LPA is generally faster and more sensitive than LGA, making it a favorable option. It demonstrates robustness across different scanners, but users are advised to consider training the algorithm with their data if significant differences exist. Further information on training procedures is available upon request.

5.2.3. MindGland

Contrarily to the aforementioned methods, MindGland rises as a Deep Learning-based approach for brain and WMH segmentation. Albeit a smaller number of brain areas can be segmented with MindGland than through SynthSeg, it does allow WM lesions segmentation. This algorithm shares similarities with SynthSeg, while also incorporating new additions while training that might pose an advantage.

Similarly to SynthSeg, MindGland is trained on synthetic data created by a similar generator to that used by SynthSeg. However, it does combine synthetic with real data, following a 2:1 ratio, where a synthetic image is chosen with twice as high probability as a real one during model training. Furthermore, CE (cross-entropy) is included in the loss function, which might help in the early stages of training, as well as in fighting FP (False Positives). The complete algorithm also employs an ensemble of 11 models and the prediction of each is merged via majority voting into a final product. A very low increase in performance is reported after the combination of different models. Finally, MindGland uses a significantly larger dataset than any of the aforementioned. MindGland has reportedly been trained on a dataset encompassing 20 thousands images, put together by UCL and its collaborating institutions over the past two decades.

5.3. Experimental setup

5.3.1. SynthSeg and MindGland together

An initial approach was to create a pipeline that uses both SynthSR and SynthSeg for super-resolution and anatomical segmentation, respectively, followed by MindGland for lesion segmentation. MindGland does a good job at segmenting WMH, but can only segment few anatomical structures compared to SynthSeg, and with lower accuracy. Combining the two brings both strengths together. Such pipeline was used at the beginning of the project for comparison purposes, but it involved too many re-sampling and recalculating steps to allow the image be fed to all different algorithms. It was also highly time-consuming. Not only does it involve several steps and algorithms, but also MindGland uses a ensemble method that computes eleven label maps that are then merged by Majority Voting. Figure 7 illustrates the flowchart that was designed: A LF image (e.g.: from the Yale dataset) is input into the pipeline. It must first be super-resolved for SynthSeg to be able to run. Their product is then used as a reference to re-sample the image, so that MindGland produces an image that has the same size too. Both posterior probabilities are merged by adding the lesion probabilities computed by MindGland into the total posterior probabilities by SynthSeg, for which the latter must be re-sampled, too, by means of the following formula

$$P'_n = P_n \cdot (1 - P_L)$$

where P_n are the posterior probabilities by SynthSeg in each of the N anatomical areas, P_L are the lesion probability predicted by MindGland, and P'_n the merged probabilities that are reported at the end of the pipeline. The highest probability of all labels (every anatomical area) is taken as the predicted label, i.e., the predicted anatomical area, for that voxel.

5.3.2. WMH-SynthSeg : An upgraded version of SynthSeg

We analyze the performance of our proposed method WMH-SynthSeg with three different experiments. The first experiment assesses the performance of the method directly with Dice scores. We first trained WMH-SynthSeg using GE3T and Singapore (using 15 scans for validation), and tested on ISBI, FLI-IAM, and Utrecht. We then reversed the roles to obtain Dice scores for GE3T and Singapore. We note that HCP and ADNI were also part of the training dataset in both folds. We note that training inputs are all synthetic and that the real images are only used as regression targets.

The second experiment assesses false positive rates (FPR) using young healthy controls from the ADHD dataset. Since WMH is not expected in these scans, we can use the estimated WMH loads as a proxy for FPR. The model in this experiment is trained with all the datasets from the first experiment.

The third experiment assesses the ability of the methods to segment pMRI data, using the same model as in the second experiment. We used the FreeSurfer segmentations of the high-field 1mm T1w scans as ground truth for the anatomy, and the LST segmentations of the high-field 1mm FLAIRs as ground truth for the WMH. Since accurate co-registration of low- and high-field scans is difficult due to nonlinear geometric distortions, we use the correlation between the ground truth and estimated ROI volumes to assess performance.

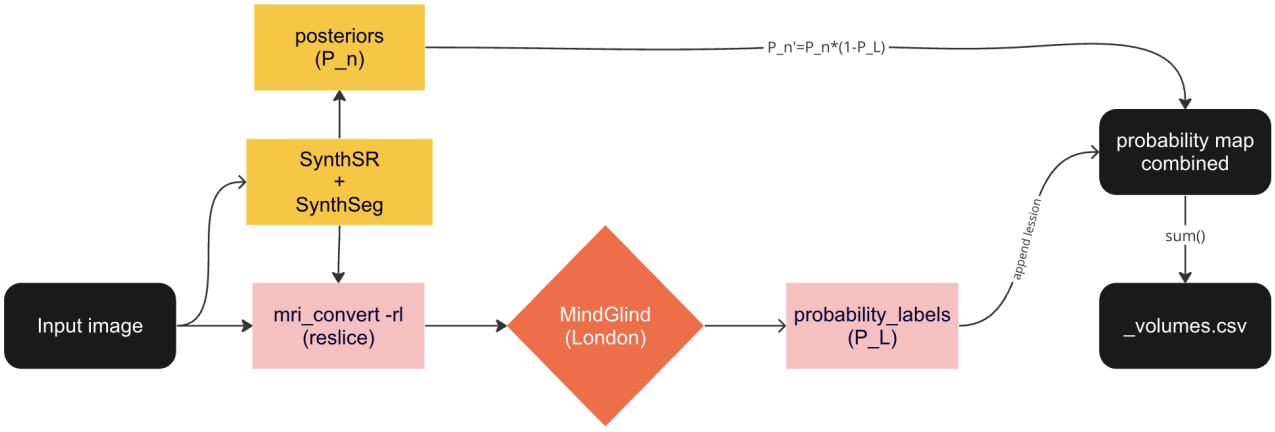

Fig. 7: Initial pipeline combining SynthSR, SynthSeg, and MindGland. The input image takes two paths: (i) It is run though SynthSR and SynthSeg; and (ii) it is re-sliced like the output of SynthSeg, and run through London. The probability map of image in path (i) is re-sampled to be able to insert the predicted lesion probability computed in path (ii). A single probability map is produced as the combination of both, and the volumes are thereof computed.

5.4. Results

The Synthseg+MindGland pipeline, albeit accurate and potentially useful, is overly time-consuming. Only MindGland (the final stage) takes five times longer than WMH-SynthSeg to run. Figure 8 shows the difference between MindGland and WMH-SynthSeg in WMH prediction. Both the ground truth (which might have been expanded by a Gaussian filter during data preprocessing) and MindGland seem to oversegment the lesion (shown as dark in the T1 image), whereas WMH-SynthSeg is able to outline the lesion in a much higher precision degree.

In this particular case depicted in Figure 8, the fact that we are stating that our model might be outperforming the ground truth might seemly make no sense. However, the model is firstly trained with several datasets and, second, with synthetic data. Since we generate our own image from the label map, this might remove the need for the mask to be extremely precise, since a new image will be created in which it actually is. In other words, each labelled brain area will have a different intensity, so the alignment of the WM-labelled and the WMH-labelled areas in the synthetic image will, by definition, perfectly match the label map.

Since the label map used for generation is identical (in terms of WM and WMH labels) to that used for segmentation (to compute the loss function and, sub-sequentially for error propagation), using synthetic data can make the model outperform the masks regarded as ground truth, outlined by radiologists. This demonstrates another major advantage of the use of synthetic data, and again, adds robustness to our model.

Table 1 shows the average Dice across the high-field

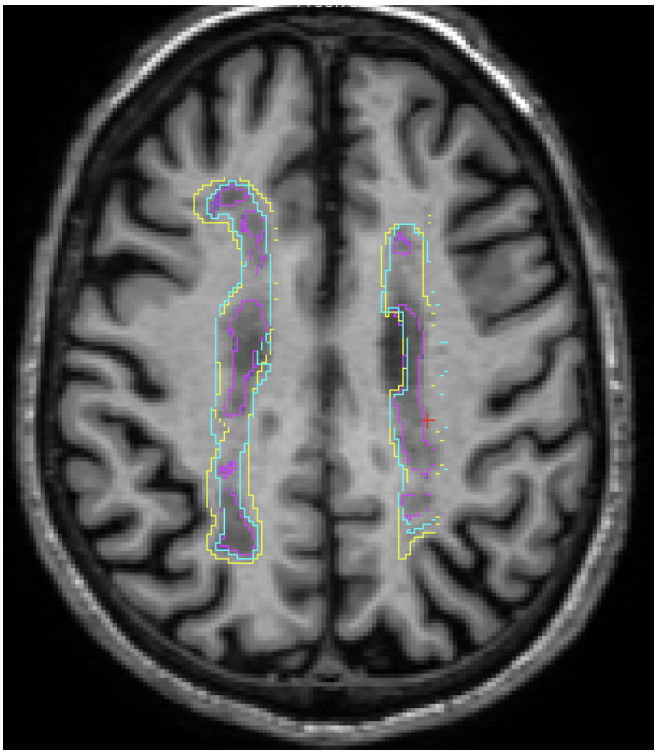

Fig. 8: Lesion prediction on a HF, T1 scan from Utrecht dataset by MindGland (blue), WMH-SynthSeg (pink), compared to the ground truth (yellow). We notice that both the ground truth and MindGland, in this case, seem to be oversegmenting the hypo-intense anatomical area denoting the WM lesion. Our model, however, is able to more accurately dealing the lesion, thanks to the robustness granted by the usage of synthetic data during training.

Method	T1w		FLAIR	
	Anat	WMH	Anat	WMH
LST (LPA)	N/A	N/A	N/A	0.57
SSAMSEG	0.81	0.46	0.72	0.56
WMH-SynthSeg (NoWMH-noCE-noMTL)	0.83	0.47	0.76	0.53
WMH-SynthSeg (NoWMH)	**0.85**	0.47	0.78	0.54
MindGland	N/A	0.49	N/A	0.56
WMH-SynthSeg (full)	**0.85**	**0.55**	**0.79**	**0.62**

Table 1: Average Dice scores for anatomy (averaged over 23 ROIs) and WMH, on high-field T1w and FLAIR scans. NoWMH-noXE-noMTL is the ablation without prior on the WMH mean, cross-entropy term in the loss, or multi-task learning (i.e., similar to [17]). NoWMH is the ablation without the prior on the mean of the WMH intensities.

Fig. 9: DICE scores by WMH-SynthSeg, with and wihtout WMH prior added during image synthesis, on high-field FLAIR scans.

datasets in the first experiment, for the WMH and for 23 representative brain ROIs: brainstem, and left/right cortex, WM, hippocampus, amygdala, thalamus, caudate, pallidum, putamen, accumbens, and cerebellum cortex and WM. WMH-SynthSeg outperforms the competing methods across the board.

The ablations show that cross-entropy and multi-task learning have a moderate positive impact on the segmentation of anatomy, whereas the prior on mean of WMH component greatly boosts the performance of the WMH segmentation. Figure 9 also shows the DICE score of the WMH-SynthSeg model with and without adding the WMH prior during synthetic image generation. In absolute terms, our new method yields competitive Dice scores for anatomy (Dice=.85 for isotropic T1w) and WMH (Dice=.62 in FLAIR, higher than SAMSEG, LST, and MindGland). We also highlight its capability to produce useful WMH segmentations from the T1w, with Dice scores as high as those of the competing methods in FLAIR.

Figure 10 shows a qualitative comparison on a FLAIR scan from the Singapore dataset, both in the high-resolution axial plane, and in a lower resolution orthogonal view (sagittal). LST produces crisp segmentations of the WMH at native resolution, but with many false positives around the septum pellucidum (between the ventricles). SAMSEG, which also operates at native resolution, struggles with partial voluming (e.g., for the cortex) and often undersegments WMH. Our method, on the other hand, produces *isotropic* segmentations that are accurate for both anatomy and WMH.

Method	FP volume (mm^3)
LST	N/A
SAMSEG	877
MindGland	1076
WMH-SynthSeg (NoWMH-noCE-noMTL)	1850
WMH-SynthSeg (NoWMH)	1150
WMH-SynthSeg (full)	950

Table 2: False Positives rate by different methods, in mm^3, obtained by each model on the HF, T1 images of the the ADHD dataset.

Table 2 shows the results for the FPR experiment. In it, the HF, T1 scans from the young controls in the ADHD dataset should result in no prediction load. Since there are no WMH, a higher volumetric (in mm^3) prediction load would be associated with a higher FPR and lower specificity from that algorithm. This control is done to prevent an algorithm from continuously segmenting WMH on the same space, regardless of the representative value intensity of the WM lesion. A model that segments every periventricular area might show a high accuracy in WMH segmentation, since that is a very recurrent anatomical space for WMH to appear. It however might have no value if it is merely a spacial decision (the model overfits based on the anatomical area) but is not able to distinguish cases in which no WMH is apparent (which involves a pixel intensity-based criteria, too).

Our method produces on average 950 mm^3. This is a low value comparable to that produced by SAMSEG (877 mm^3). It is actually lower than most of the small lesions, and can therefore be readily solved by means of a threshold. We note that LST is not compatible with the ADHD dataset as it has T1w contrast. The ablated versions show a slight increase when no WM prior is introduced, and a much more significant rise in FP volume when both the prior and the multi-task learning feature are removed, highlighting the contribution of these components to the accuracy of the algorithm. MindGland reports an average of 1050 mm^3, somewhat higher values to our final algorithm, but that still falls behind the ablated versions.

Finally, Table 3 shows the correlations between the volumetric measurements derived from the high-field scans (ground truth) and the pMRI, for the WMH and for a rep-

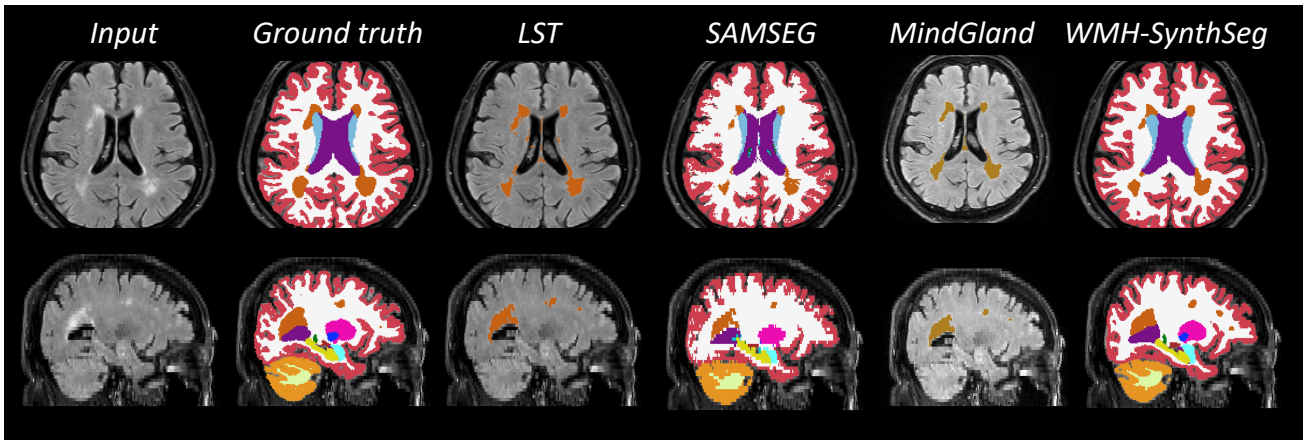

Fig. 10: Input, ground truth, and automated segmentations of a sample high-field scan from the Singapore dataset. The top row shows the high-resolution axial view; the bottom row shows a lower resolution orthogonal view (in sagittal orientation). Note that MindGland cannot predict so many brain structures as SynthSeg, so we only show the predicted lesion layer.

Method	T1w		FLAIR	
	Hippo	WMH	Hippo	WMH
LST (LPA)	N/A	N/A	N/A	-0.33
SAMSEG	0.71	0.63	0.69	0.64
MindGland	N/A	0.71	N/A	0.78
WMH-SynthSeg (full)	**0.89**	**0.75**	**0.86**	**0.85**

Table 3: Correlation between ground truth volumetric measurements obtained from high-field (FreeSurfer from T1w for anatomy, LST from FLAIR for WMH) and from automated segmentations of the pMRI (MGH dataset). The hippocampal volumes ("Hippo") are left-right averaged.

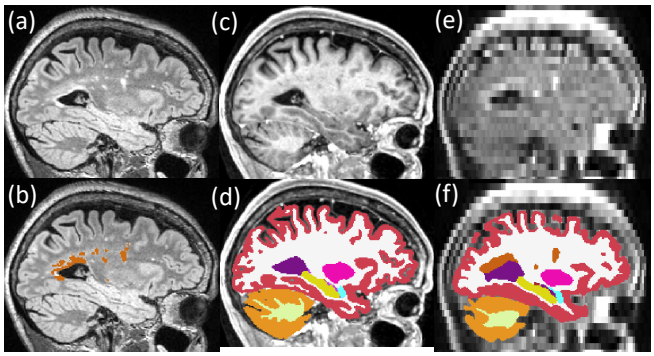

Fig. 11: (a) High-field 1mm isotropic FLAIR from MGH dataset. (b) LST segmentation, used as ground truth for WMH. (c) High-field 1mm T1w. (d) FreeSurfer segmentation of (c), used for ground truth for anatomy. (e) pMRI of the same subject at 2x2x5.8mm axial resolution. (f) WMH-SynthSeg segmentation. We note that, despite affine alignment of the high-field images to the pMRI, the anatomy on the slices is slightly different due to nonlinear distortion.

resentative brain ROI (the hippocampus). LST completely fails at low field, as it was not designed for it. Being contrast agnostic, SAMSEG yields fairly strong correlations (between .63 and .71). MindGland is still able to perform better than the aforementioned methods in the LF domain. WMH-SynthSeg produces very strong correlations (12-21 points higher than SAMSEG and 4-7 higher than MindGland). This is attributed to its excellent ability to adapt to low-field images, which is qualitatively exemplified in Figure 11.

6. CONCLUSION

Deep Learning algorithms have proved more successful in overall brain segmentation, including WM lesions. The use of synthetic images poses a major advantage in fighting the domain gap challenged, accentuated by different MRI contrasts and field strengths, leading to higher generalizability and domain adaptation.

SynthSeg showed the best overall performance in brain segmentation up to now, and MindGland in WMH prediction. The latter leverages a new loss function with CE and an immense dataset (20k images) for great performance in both HF and LF. Joining SynthSeg with some of the innovations by MindGland gave rise to WMH-SynthSeg, which which shows high performance in both brain and WMH segmentation. In the absence of a large dataset, WMH-SynthSeg proves that adding a prior can significantly improve model performance.

WMH-SynthSeg is the first published method that can simultaneously segment brain ROIs and WMH in scans of any resolution and contrast, including low-field pMRI. Future work will include more realistic modeling of WMH in images, in order to bridge the so-called "reality gap" between synthetic and real data. Our approach is publicly available and has huge potential in analyzing pMRI acquired in medically underserved areas.

7. ACKNOWLEDGMENTS

Supported by a grant from the Jack Satter Foundation and by NIH grants RF1MH123195, R01AG070988, R01EB031114, UM1MH130981, RF1AG080371, and R01NS112161.

8. REFERENCES

[1] A de Havenon, NR Parasuram, , et al., "Identification of white matter hyperintensities in routine emergency department visits using portable bedside magnetic resonance imaging," *J of the American Heart Association*, vol. 12, no. 11, pp. e029242, 2023.

[2] R Dobson and G Giovannoni, "Multiple sclerosis–a review," *European J Neurology*, vol. 26, pp. 27–40, 2019.

[3] E Fisher, J Lee, et al., "Gray matter atrophy in multiple sclerosis: a longitudinal study," *Annals of Neurology*, vol. 64, no. 3, pp. 255–265, 2008.

[4] B Fischl, D Salat, et al., "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, pp. 341–355, 2002.

[5] B Patenaude, SM Smith, et al., "A bayesian model of shape and appearance for subcortical brain segmentation," *Neuroimage*, vol. 56, no. 3, pp. 907–922, 2011.

[6] P Schmidt, C Gaser, et al., "The LST toolbox for lesion segmentation and quantification," in *Computer Methods in Biomechanics and Biomedical Engineering*, 2012, vol. 16, pp. 196–200.

[7] L Griffanti, G Zamboni, et al., "Bianca (brain intensity abnormality classification algorithm): A new tool for automated segmentation of white matter hyperintensities," *NeuroImage*, vol. 141, pp. 191–205, 2016.

[8] O Puonti, JE Iglesias, et al., "Fast and sequence-adaptive whole-brain segmentation using parametric bayesian modeling," *NeuroImage*, vol. 143, pp. 235–249, 2016.

[9] S Cerri, O Puonti, et al., "A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in MS," *NeuroImage*, vol. 225, pp. 117471, 2021.

[10] A Roy, S Conjeti, et al., "QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy," *NeuroIm*, vol. 186, pp. 713–727, 2019.

[11] L Henschel, S Conjeti, et al., "Fastsurfer – a fast and accurate deep learning based neuroimaging pipeline," *NeuroImage*, vol. 219, pp. 117012, 2020.

[12] T Brosch, LYW Tang, et al., "Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to MS lesion segmentation," *IEEE Trans Med Im*, vol. 35, no. 5, pp. 1229–1239, 2016.

[13] M Ghafoorian, N Karssemeijer, et al., "Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities," *Scientific Reports*, vol. 7, no. 1, pp. 5110, 2017.

[14] M Wang and W Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–53, 2018.

[15] B Billot, DN Greve, et al., "SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining," *Med Im Anal*, vol. 86, pp. 102789, 2023.

[16] Iglesias JE, B Billot, et al., "SynthSR: A public AI tool to turn heterogeneous clinical brain scans into high-resolution T1-weighted images for 3D morphometry," *Science Advances*, vol. 9, no. 5, pp. eadd3607, 2023.

[17] B Billot, S Cerri, et al., "Joint segmentation of multiple sclerosis lesions and brain anatomy in MRI scans of any contrast and resolution with CNNs," in *ISBI*. IEEE, 2021, pp. 1971–1974.

[18] Adam de Havenon, Nethra R. Parasuram, Anna L. Crawford, Mercy H. Mazurek, Isha R. Chavva, Vineetha Yadlapalli, Juan E. Iglesias, Matthew S. Rosen, Guido J. Falcone, Seyedmehdi Payabvash, Gordon Sze, Richa Sharma, Steven J. Schiff, Basmah Safdar, Charles Wira, William T. Kimberly, and Kevin N. Sheth, "Identification of white matter hyperintensities in routine emergency department visits using portable bedside magnetic resonance imaging," *Journal of Magnetic Resonance Imaging*, vol. 54, no. 4, pp. 1105–1115, 2021.

[19] T. Campbell Arnold, Danni Tu, Serhat V. Okar, Govind Nair, Samantha By, Karan D. Kawatra, Timothy E. Robert-Fitzgerald, Lisa M. Desiderio, Matthew K. Schindler, Russell T. Shinohara, Daniel S. Reich, and Joel M. Stein, "Sensitivity of portable low-field magnetic resonance imaging for multiple sclerosis lesions," *NeuroImage: Clinical*, vol. 35, pp. 103101, 2022.

[20] Alessandra M. Valcarcel, Kristin A. Linn, Simon N. Vandekar, Theodore D. Satterthwaite, John Muschelli, Peter A. Calabresi, and Russell T. Shinohara, "Mimosa: An automated method for intermodal segmentation analysis of multiple sclerosis brain lesions," *Journal of Neuroimaging*, vol. 28, no. 4, pp. 389–398, 2018.

[21] Elizabeth M. Sweeney, Russell T. Shinohara, Navid Shiee, Farrah J. Mateen, Avni A. Chudgar, Jennifer L. Cuzzocreo, Peter A. Calabresi, Dzung L. Pham, Daniel S. Reich, and Ciprian M. Crainiceanu, "Oasis

is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in mri," *NeuroImage: Clinical*, vol. 2, pp. 402–413, 2013.

[22] Navid Shiee, Pierre-Louis Bazin, Arzu Ozturk, Daniel S. Reich, Peter A. Calabresi, and Dzung L. Pham, "A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions," *NeuroImage*, vol. 49, no. 2, pp. 1524–1535, 2010.

[23] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[24] Refaat E. Gabr, Ivan Coronado, Melvin Robinson, Sheeba J. Sujit, Sushmita Datta, Xiaojun Sun, William J. Allen, Fred D. Lublin, Jerry S. Wolinsky, and Ponnada A. Narayana, "Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: A large-scale study," *Multiple Sclerosis Journal*, vol. 26, no. 10, pp. 1217–1226, 2020.

[25] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.

[26] Konstantinos Kamnitsas, Christian Ledig, Virginia F Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.

[27] M. L. de Leeuw den Bouter, G. Ippolito, T. P. A. O'Reilly, R. F. Remis, M. B. van Gijzen, and A. G. Webb, "Deep learning-based single image super-resolution for low-field mr brain images," *Scientific Reports*, vol. 12, no. 1, pp. 6362, 2022.

[28] Adrian V Dalca, Eugene Yu, Polina Golland, Bruce Fischl, Mert R Sabuncu, and Juan Eugenio Iglesias, "Unsupervised deep learning for bayesian brain mri segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III*. Springer International Publishing, 2019, pp. 356–365.

[29] Min Chen Hubert Lee, Katja Petersen, Nick Pawlowski, Ben Glocker, and Michiel Schaap, "Tetris: Template transformer networks for image segmentation with shape priors," *IEEE Transactions on Medical Imaging*, vol. 38, no. 11, pp. 2596–2606, 2019.

[30] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Simon A Cook, Antonio De Marvao, Timothy Dawes, Declan P O'Regan, et al., "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 384–395, 2017.

[31] O Ronneberger, P Fischer, et al., "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, vol. 18, pp. 234–241.

[32] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox, "Discriminative unsupervised feature learning with convolutional neural networks," in *Advances in neural information processing systems*, 2014, vol. 27.

[33] Amy Zhao, Guha Balakrishnan, Frédo Durand, John V Guttag, and Adrian V Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8543–8553.

[34] Ö Çiçek, A Abdulkadir, et al., "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *Proc of MICCAI*, 2016, vol. 19, pp. 424–432.

[35] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan H. Heckers, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," 2019.

[36] Xin He, Kaiyong Zhao, and Xiaowen Chu, "Automl: A survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, pp. 106622, 2021.

[37] Yunzhe Xia and et al., "3d semi-supervised learning with uncertainty-aware multi-view co-training," in *Proc. IEEE Winter Conference on Applications of Computer Vision*. IEEE Computer Society, 2020, pp. 3646–3655.

[38] Fabian Isensee, Paul F Jaeger, Simon A Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[39] Jun Ma, "Cutting-edge 3d medical image segmentation methods in 2020: Are happy families all alike?," 2021.

[40] Nabil Ibtehaz and M Sohel Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural networks*, vol. 121, pp. 74–87, 2020.

[41] Ali Hatamizadeh, Yuyin Tang, Vinith Nath, Di Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Dong Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[42] Hanting Chen, Yunhe Wang, Chunjing Xu, Boxin Shi, Chao Xu, Qi Tian, and Chang Xu, "Addernet: Do we really need multiplications in deep learning?," 2021.

[43] Y Wu and K He, "Group normalization," in *ECCV*, 2018, pp. 3–19.

[44] DC Van Essen, SM Smith, et al., "The WU-Minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[45] CR Jack Jr, MA Bernstein, et al., "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J of MRI*, vol. 27, no. 4, pp. 685–691, 2008.

[46] HJ Kuijf, JM Biesbroek, et al., "Standardized assessment of automatic segmentation of white matter hyperintensities; results of the wmh segmentation challenge," *IEEE Trans Med Im*, 2019.

[47] A Carass, S Roy, et al., "Longitudinal MS lesion segmentation data," *Data in brief*, vol. 12, pp. 46–50, 2017.

[48] O Commowick, A Istace, et al., "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," *Scientific Reports*, vol. 8, no. 1, pp. 13650, 2018.

[49] P Bellec, C Chu, et al., "The neuro bureau ADHD-200 repository," *NeuroImage*, vol. 144, pp. 275–286, 2017.

[50] P Schmidt, *Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging*, Ph.D. thesis, LMU, 2017.

[51] "An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis," .