CEREBRUM-7T: FAST AND FULLY-VOLUMETRIC BRAIN SEGMENTATION OF OUT-OF-THE-SCANNER 7T MR VOLUMES

A PREPRINT

Michele Svanera*
Institute of Neuroscience and Psychology
University of Glasgow
UK

Sergio Benini
Department of Information Engineering
University of Brescia
Italy

Dennis Bontempi
Department of Information Engineering
University of Brescia
Italy

Lars Muckli
Institute of Neuroscience and Psychology
University of Glasgow
UK

July 7, 2020

ABSTRACT

Ultra high field MRI enables sub-millimetre resolution imaging of human brain, allowing to disentangle complex functional circuits across different cortical depths. The capability of using these innovative scanners at 7 Tesla (7T) poses new challenges, as for example those related to the current lack of standardised acquisition protocols, and of automatised pipelines for image analysis. Segmentation, meant as the partition of MR brain images in multiple anatomical classes, is an essential step in many functional and structural neuroimaging studies. In this work, we design and test `CEREBRUM-7T`, an optimised end-to-end CNN architecture, that allows to segment a whole 7T $T1_w$ MRI brain volume at once, without the need of partitioning it into 2D or 3D tiles. Despite deep learning (DL) methods are recently starting to emerge in 3T literature, to the best of our knowledge, `CEREBRUM-7T` is the first example of DL architecture directly applied on 7T data with the purpose of segmentation. Training is performed in a weakly supervised fashion, since it exploits a ground-truth (GT) not exempt from errors. The generated model is able to produce accurate multi-structure segmentation masks on six different classes, in only few seconds. In the experimental part, a combination of objective numerical evaluations, and subjective analysis carried out by experienced neuroimaging users, confirms that the proposed solution outperforms the GT it was trained on in segmentation accuracy, and it is suitable for many neuroimaging studies. Furthermore, to allow replicability and encourage extensions, we release the code, 7T data (145 volumes), and other materials, including the GT and the survey.

Keywords Brain MRI Segmentation · Convolutional Neural Networks · Weakly Supervised Learning · 3D Image Analysis

*Corresponding author: Michele.Svanera at glasgow.ac.uk

1 Introduction

Magnetic Resonance Imaging (MRI) is a method for non-invasively measuring brain anatomy and function, widespread in research and clinical environments. Most MRI scanners in clinical practice have a field strength of 3 Tesla or less. However, ultra-high field magnetic resonance with a field strength of 7T are now CE labeled and FDA approved, and becoming more common in clinical settings. The introduction of 7T scanners, together with improvements in acquisition methods, has led to substantial advances in both signal and contrast to noise ratio, and increased the imaging resolution to a sub-millimetre level (Duyn, 2012). This technology enables imaging of different cortical depths and columns in order to investigate functional circuits of human cortex with high spatial specificity, and the visualisation of structures not as clearly seen in 1.5T or 3T.

However, the capability of using these innovative systems also poses challenges, including standardisation across 7T sites: there have been recent efforts showing improved consistency of findings by synchronising sequence protocols across centres, for example by the UK7T Network (Clarke et al., 2020). Further, standardised brain atlases and intensity-based pipelines specific for the analysis of 7T data will help to minimise variability in analyses across sites (Botvinik-Nezer et al., 2020). Another issue is magnetic field inhomogeneity and increased susceptibility artefacts during acquisition (Uğurbil et al., 2003). Inhomogeneities in the magnetic field can lead to intensity variations across the image, presenting a particular challenge for the analysis of structural data. For example, non-uniform voxel intensities within a given region can lead to low quality brain segmentation outcomes.

Image segmentation is an essential quantitative analysis step when assessing both healthy brain anatomy and patho-physiological conditions. The segmentation of brain structures is of paramount importance for monitoring anatomical variations during the development of neuro-degenerative processes, psychiatric disorders, and neurological diseases. Additionally, segmentation is an essential step in functional MRI (fMRI) studies, to isolate specific brain regions where to investigate patterns of brain activity. In this work, we propose a deep-learning based algorithm for automated brain segmentation, optimised for performance and speed compared to current approaches.

1.1 Existing methods for 7T brain segmentation

Popular fully-automatic segmentation tools available for 3T data - such as FreeSurfer (Fischl, 2012) - which usually apply atlas-based or multi-atlas-based segmentation strategies, are not maximally effective on 7T volumes, due to increased data complexity, voxel intensity inhomogeneity, and different intensity distributions.

As a consequence, manual segmentation protocols (Wenger et al., 2014; Berron et al., 2017), although time consuming (Zhan et al., 2018; Koizumi et al., 2019), are still a common practise for 7T data. To partially reduce the laboursome process of manual segmentation, the solution proposed by Gulban et al. (2018) combines manual and semi-automatic segmentation, by adopting a multi-dimensional transfer function to single out non-brain tissue voxels in 7T MRI data of nine volunteers. Other semi-automated methods developed in the past for generic MRI data, such as `ITK-SNAP` (Yushkevich et al., 2006), have been adapted by (Archila-Meléndez et al., 2018) for tackling also ultra-high field brain imaging.

Often, given the lack of harmonised neuroimaging analysis protocols, multiple 7T sites created in-house pipelines to perform MRI segmentation on site-data specifically. Fracasso et al. (2016) developed a custom workflow (used for example in Bergmann et al., 2019) which we also adopt as the ground-truth segmentation of gray and white matter (GM and WM). The placement of the GM/CSF (cerebrospinal fluid) boundary is based on the location of the 75% quantile of the variability of $T1_w$ partial volume estimates across cortical depth, while GM/WM boundary from a combination of `AFNI-3dSeg` (Cox, 1996) and a clustering procedure (see Section 2.2 for more details).

As an attempt to develop a site-independent approach, Bazin et al. (2014) presented a computational framework for whole brain segmentation at 7T, specifically optimiSed for MP2RAGE sequences. The authors develop a rich atlas of brain structures, on which they combine a statistical and a geometrical model. The method, which includes a non-trivial preprocessing chain for skull stripping and dura estimation, achieves whole brain segmentation and cortical extraction, all within a computation time below six hours. Despite these efforts, most existing solutions, including Bazin et al.

(2014), still generate a variety of segmentation errors that needs to be manually addressed, as reported in Gulban et al. (2018).

1.2 Deep learning methods for MRI segmentation

In recent years, the advanced classification and segmentation capabilities ensured by deep learning (DL) methods have impacted several medical imaging domains (Litjens et al., 2017; Hamidinekoo et al., 2018). Various segmentation algorithms, which exploit the generalization capabilities of DL and Convolutional Neural Networks (CNN) on unseen data, made possible a drastic improvement in the performance of other traditional, mostly atlas-based, segmentation tools.

To the best of our knowledge, no DL architectures have been directly applied on 7T data for segmentation purposes. The only attempt made by Bahrami et al. (2016) to use CNN in this field, aimed at reconstructing 7T-like images from 3T MRI data. Specifically, from the 3T image intensities and the segmentation labels of 3T patches, the CNN learns a mapping function so as to generate the corresponding 7T-like image, with quality similar to the ground-truth 7T MRI.

Restricting the scope to 3T data only, recent DL-based methods such as QuickNat (Roy et al., 2019), MeshNet (Fedorov et al., 2017; McClure et al., 2019), NeuroNet (Rajchl et al., 2018), DeepNAT (Wachinger et al., 2018), and FastSurfer (Henschel et al., 2020) have been the most effective solutions among those which proposed to obtain a whole brain segmentation. However, a common trait of all the aforementioned methods is that none of them fully exploit the 3D spatial nature of MRI data, thus making segmentation accuracy sub-optimal. In fact such solutions partition the brain into 3D sub-volumes (DeepNAT, MeshNet, and NeuroNet) or 2D patches (QuickNAT, FastSurfer), which are processed independently and only eventually reassembled; as recently shown in Reina et al. (2020), the use of "tiling" introduces small but relevant differences during inference that can negatively affect the overall quality of the segmentation result. For example in MRI segmentation, tiling entails a loss of global contextual information, such as the absolute and relative positions of different brain structures, which negatively impacts the segmentation outcome.

The DL model `CEREBRUM` (Bontempi et al., 2020) is the first attempt to fully exploit the 3D nature of MRI 3T data, taking advantage of both global and local spatial information. This 3D approach adopts an end-to-end encoding/decoding fully convolutional structure. Trained in a weakly supervised fashion with 900 whole brain volumes segmented with Freesurfer (Fischl, 2012), `CEREBRUM` learns to segment out-of-the-scanner brain volumes in just $\sim$5-10 seconds on a desktop GPU, with neither atlas-registration, pre-processing, nor filtering.

1.3 Aims and contributions

In this work we reshape the framework presented in Bontempi et al. (2020) to handle UHF high-resolution data, delivering `CEREBRUM-7T`, the first DL fully automatic solution for brain MRI segmentation on out-of-the-scanner 7T data[2].

Similarly to Bontempi et al. (2020), `CEREBRUM-7T` acts in a fully 3D fashion on brain volumes, as shown in Figure 1. This is possible simplifying the model architecture with respect to other DL-based segmentation methods, and also to `CEREBRUM`, since the increase in the data size and the GPU memory constraints. Such design choice enables to achieve a full brain segmentation in few seconds, as in Bontempi et al. (2020), compared to several hours of other currently used methods for 7T data. Furthermore, exploiting the ability of DL methods to efficiently learn internal representations of data, segmentation happens in native space, with neither the support of ad-hoc preprocessing, nor the alignment to reference atlases.

To compensate for the shortage of hand-labelled 7T data, the model is trained, in a weakly-supervised mode on a GT obtained by an in-house procedure, using the anatomical labels of the MICCAI challenge (Mendrik et al., 2015): gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), ventricles, cerebellum, brainstem, and basal ganglia. The GT segmentation masks for GM/WM are obtained by combining geometric and clustering approaches (Cox, 1996;

[2]With the term "out-of-scanner" we refer to the reconstructed data saved in DICOM 2D images.

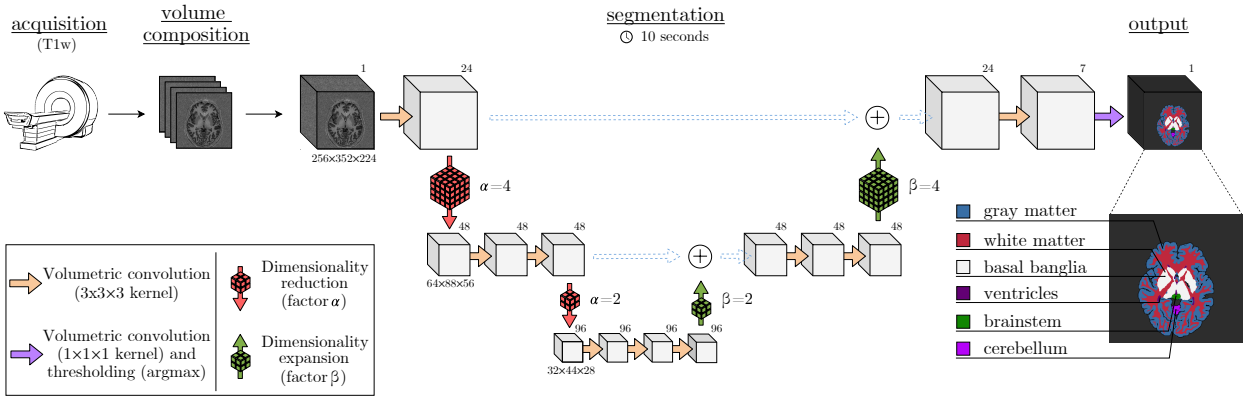


Figure 1. Method overview. The database is composed of 145 $T1_w$ volumes (MP2RAGE at 7T, 0.63 mm^3-iso) and the ground-truth obtained using a combination of `AFNI - 3dSeg` (Cox, 1996) and methods as in Fracasso et al. (2016). Segmentation results are provided in only few seconds ($\sim$ 10 on a desktop CPU). Data dimensions are provided below each volume, while the numbers of filters are in the top-right corners.

Fracasso et al., 2016), while other labels by using atlas-based techniques (Fischl, 2012) adequately processed, corrected, and adapted for 7T data. Although available anatomical labelling exhibits several errors, we show how the adopted learning procedure, exploiting "inaccurate supervision" (Zhou, 2017), ensures good generalisation capabilities and improved labelling performance with respect to the GT it was trained on.

Training and testing are carried out on the largest brain MRI 7T database for segmentation openly available (145 anatomical scans). Differently from Bontempi et al. (2020), where the dimension of the GT was well enough for training (900 volumes), `CEREBRUM-7T` model allows learning also on a reduced number of volumes, thanks to a data augmentation strategy.

Given the lack of other DL methods working on 7T data, `CEREBRUM-7T` is first quantitatively compared against the reference GT using the same metrics used in the MICCAI MR Brain Segmentation challenge, i.e., the Dice Similarity Coefficient, the 95th Hausdorff Distance, and the Volumetric Similarity Coefficient (Mendrik et al., 2015). We then present the results of a survey carried out on three volumes left out from previous analysis, in which seven expert neuroscientists choose, during a blind survey, the most accurate segmentation among the one produced by `CEREBRUM-7T`, the reference GT, and the result of a manual segmentation process, designed to add a further verification step on the obtained results. The comparison of the survey is also numerically examined using the above mentioned MICCAI metrics on $2.7M$ annotated voxels.

In the experimental part, a combination of objective numerical evaluations, and subjective analysis carried out by experienced neuroscientists, assesses the superiority of `CEREBRUM-7T` segmentation masks with respect to the GT. As a last contribution, we make publicly available the set of 145 7T scans[3], the related GT, and all the code[4] necessary to train `CEREBRUM-7T` and to perform the survey[5].

2 Material and methods

2.1 Data acquisition and split

The database consists of 145 out-of-the-scanner volumes obtained with a MP2RAGE sequence at 0.63 mm^3 isotropic resolution, using a 7-Tesla MRI scanner with 32-channel head coil. All volumes were collected, as reconstructed

[3]Data will be openly available under EBRAINS knowledge graph after publication.
[4]`https://github.com/rockNroll87q/cerebrum7t`: available after publication.
[5]`https://github.com/rockNroll87q/segmentation_survey`: available after publication.

DICOM images, at the Imaging Centre of Excellence (ICE) at the Queen Elizabeth University Hospital, Glasgow (UK). The columns of Figure 2 show some selected slices of the out-of-the-scanner $T1_w$, the segmentation resulting from FreeSurfer, the one from Fracasso et al. (2016), the corresponding automatic GT, the `CEREBRUM-7T` mask, and the manual segmentation, respectively.

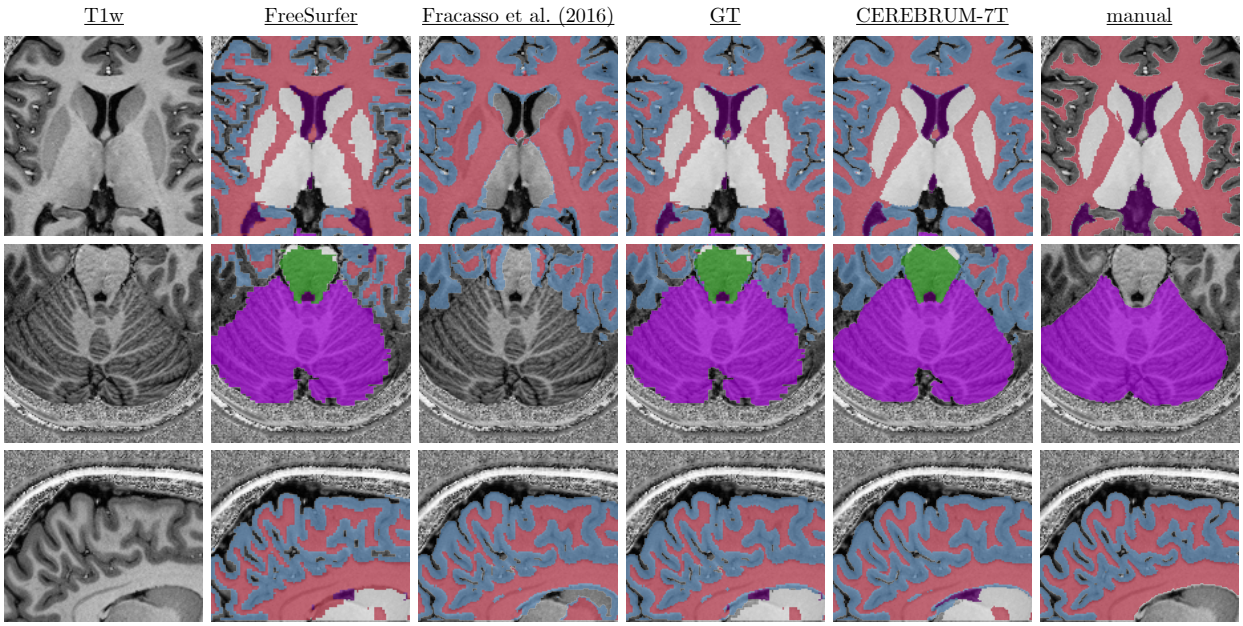


Figure 2. Visual examples of the dataset and results. Columns, from left to right, show: $T1_w$ scan (left), FreeSurfer segmentation, Fracasso et al. (2016), GT (obtained as described in Section 2.2), `CEREBRUM-7T`, and manual segmentation. During manual segmentation, only specific areas useful for the survey were annotated.

Out of the total 145 volumes, 120 are used for training, 4 for validation, and 21 for testing (3 of which for the survey only). The only preprocessing applied is the neck cropping using the INV2 scan obtained during acquisition. Dataset details are shown in Table 1.

Table 1
Dataset details. MR Images acquired at the Imaging Centre of Excellence (ICE) at the Queen Elizabeth University Hospital, Glasgow (UK).

Parameter	**Value**
Sequence used	$T1_w$ MP2RAGE
Field strenght	7 Tesla
Voxel size	$0.63 \times 0.625 \times 0.625$
Original volume sizes	$256 \times 360 \times 384$
Training volume sizes	$256 \times 352 \times 224^{\dagger}$
Training	120 volumes
Validation	4 volumes
Testing	21 volumes
Testing (survey)	3 vol. $\times 8$ areas

† neck cropping

2.2 Automatic ground-truth generation

Similarly to most approaches employing DL frameworks for brain MRI segmentation (Roy et al., 2019; McClure et al., 2019; Fedorov et al., 2017; Rajchl et al., 2018; Bontempi et al., 2020), we also adopt an almost fully automatic procedure for GT creation, since the prohibitive time cost required to produce a manual annotation on such large dataset. Such a decision is also driven by the consideration that, despite we use a GT with label errors, in some cases the trained models already proved to perform the same (Rajchl et al., 2018), or even better (Bontempi et al., 2020; Roy et al., 2019), than the automated GT.

Differently from Bontempi et al. (2020), we cannot use FreeSurfer (Fischl, 2012) as the only tool for segmenting all tissue and sub-cortical brain structures. As also hinted in Figure 2 (column b) the quality of WM and GM segmentation masks obtained with such tool is not acceptable for our purposes, even considering inaccurate supervision for learning.

An overview of the GT generation process is shown in Figure 3. The pipeline accounts for two main branches: the upper

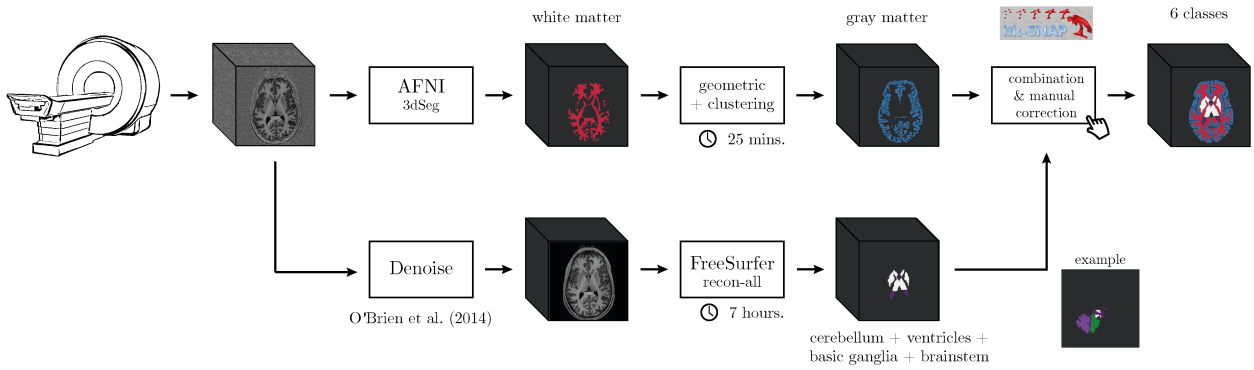


Figure 3. Processing pipeline used to generate the GT starting from the reconstructed $T1_w$.

one deals with WM and GM segmentations, while the lower one isolates other brain structures such as cerebellum, ventricles, brainstem, and basal ganglia. The two processing branches are combined afterwards, when a manual correction step is also carried out to reduce major errors.

In the upper branch the white matter (WM) mask is obtained using a combination of `AFNI - 3dSeg` (Cox, 1996) followed by geometric and clustering methods as in Fracasso et al. (2016). Specifically $T1_w$ images are co-registered to an atlas (Desikan et al., 2006) and a brain mask is overlaid to the $T1_w$ images to remove the cerebellum and subcortical structures. The $T1_w$ images are then separated in six different parts along the posterior to anterior direction to improve intensity homogeneity. Each part is afterwards separately processed by the `3dSeg` function in `AFNI`, to isolate WM. The WM masks obtained from each part are summed together resulting in whole brain WM mask (see Fracasso et al. (2016) for further details).

The GM segmentation exploits such whole brain WM segmentation and an atlas co-registered to the $T1_w$ images (Desikan et al., 2006). Next, a distance map from the WM/GM boundary to the pial surface is built computing the Euclidean distance of each voxel from the WM/GM border. Negative distances are assigned inside WM and positive distances are assigned from WM borders onwards. Each region of interest (ROI) in the atlas by Desikan et al. (2006) is selected iteratively. For each ROI the coordinates are divided into four separate subparts using k-means clustering. For each subpart, voxels within $-2mm$ and $7mm$ (Euclidian distance) from the WM/GM border are selected and their $T1_w$ intensity stored for further analysis. For each cluster 10 bins between $-2mm$ and $7mm$ are obtained - with each bin containing 10% of the data. For each of them, a partial volume estimate, defined as the standard deviation of $T1_w$ intensity as well as the average Euclidean distance for the same bin, is computed. A linear model is then fit between the average Euclidean distance of each bin and the corresponding partial volume estimate. The slope of the linear model can be either positive or negative: if there is a positive slope, the 75% quantile of the standard deviation values is computed among the 10 bins; if, on the other hand, the slope is negative, the 25% quantile is computed. The

Euclidean distance of the 25%/75% quantile corresponds to a drop or rise in $T1_w$ variability and is considered as the transition between GM and CSF. To improve the obtained gray matter segmentation, the WM and GM masks are fed to the Cortical Reconstruction using Implicit Surface Evolution (CRUISE) algorithm in the `nighres` software package (Huntenburg et al., 2018).

Despite the method ensures a robust result in segmenting GM/WM boundary, no cerebellum, ventricles, and basic ganglia areas are computed. To address such lack of GT structures, in the lower branch of the pipeline, we use FreeSurfer (Fischl, 2012), which first requires to denoise the $T1_w$ volume (O'Brien et al., 2014), to add the following labels: cerebellum, ventricles, brainstem, and basic ganglia. When necessary, a final step of manual correction is carried out to reduce major errors (especially on cerebellum classified as GM) using `ITK-SNAP` (Yushkevich et al., 2006).

The z-scoring procedure, applied to normalise the data before `CEREBRUM-7T` training, is obtained using the mean and standard deviation volumes computed on the entire dataset (shown in Figure 4).

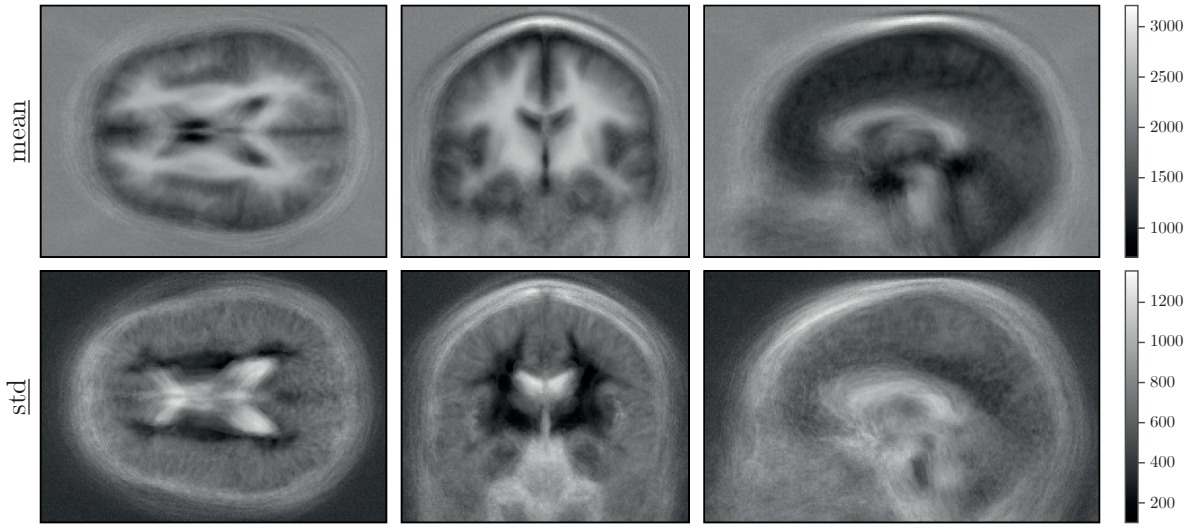


Figure 4. Mean and standard deviation volumes of the database used to z-score the data. Denoised mean/std volumes are found in Supplementary Material.

2.2.1 Data augmentation

The corpus of data we use for our experiments is one of the biggest 7T MRI publicly available datasets. Yet, given the complexity of the DL architecture, i.e., the number of learnable parameters, its size makes such dataset not off-the-shelf suitable for a DL application. Therefore, we decide to implement data augmentation procedures.

Offline augmentation, too computationally demanding to be performed in training-time, applies small random shifts ($max_shift = [10, 15, 10]$ voxels) or rotations ($max_rotation = 5°$, on all three axes), and elastic deformations. This ensures an augmentation factor of 10 of the training set. Details on data augmentation with ranges of applied transformations are presented in Figure 5.

Online data augmentation, performed during training, comprises variations on voxel intensities only: gaussian, salt & pepper, and inhomogeneous-field noise. In MRI, and especially in UHF, the inhomogeneity in the magnetic field produces an almost linear shift in the voxel intensity distributions for different areas in the 3D space (Sled et al., 1998). In other words: the same anatomical structure, for example GM, has different voxel intensities in the frontal and in the occipital regions. One of the main limitations of segmentation methods that heavily rely on intensity values is the inability to correctly classify the same class having different local distributions. To increase our model invariance, we introduce, as an additional data augmentation strategy, a synthetic inhomogeneous field noise. We start by pre-computing a 3D multivariate normal distribution, with zero-means and twice the dimension (for each axis, i.e.,

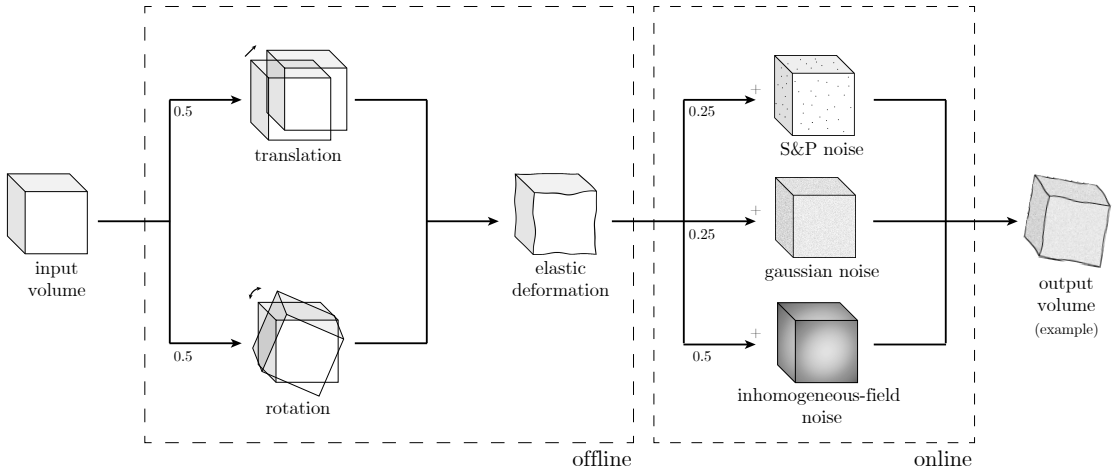


Figure 5. Data augmentation procedure. Offline (with respect to the training procedure), geometric augmentation is performed with rotation or translation ($p = \{0.5, 0.5\}$) and elastic deformation (Çiçek et al., 2016). Online, voxel intensity changes are applied: salt and pepper noise ($p = 0.25$), gaussian noise ($p = 0.25$), or inhomogeneous-field noise ($p = 0.5$). An example of data augmentation is shown in the output volume, where one rotation and one elastic deformation are followed by an additional gaussian noise.

$8\times$ the volume) of the original volume. For each training volume, we randomly sample, from the 3D multivariate normal distribution, a noise volume as big as the volume to augment. The so-generated noise volume is then added to the anatomical MRI, adding further variability to the volume intensities and simulating distortions along different directions. In Figure 6, a sketch of the method, when applied on a 2D slice example, is shown (see Supplementary Material for further examples on 1D and 3D cases).

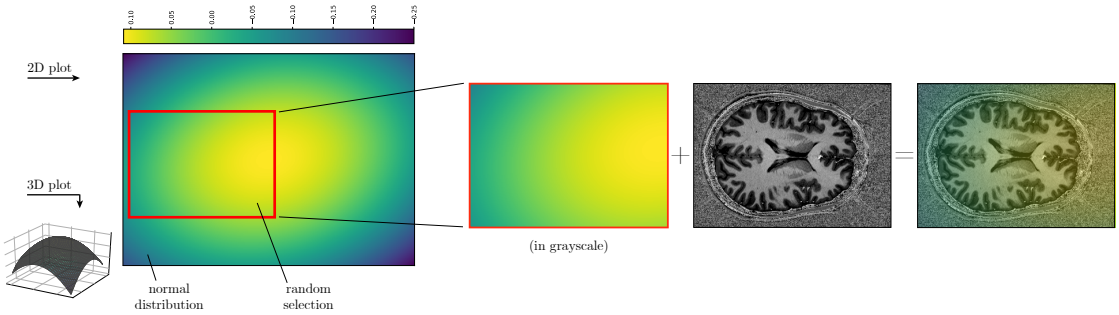


Figure 6. Cartoon to describe the inhomogeneous-field noise for the 2D case. After pre-computing a multivariate normal distribution, we randomly extract a noise sample as big as the sample to augment, for every training batch. The extracted patch is then added to the original sample.

2.3 Manual segmentation

A portion of the considered dataset has been manually annotated by one of the author from the University of Glasgow, who accumulated several years of experience in neuroscience, and reviewed by other co-authors and expert colleagues in the field. In particular, three subject volumes have been randomly selected from the 7T MRI dataset: for each of them, eight regions of interest in neuroscience research - i.e., early visual cortex (EVC), high-level visual areas (HVC), motor cortex (MCX), cerebellum (CER), hippocampus (HIP), early auditory cortex (EAC), brainstem (BST), and basal ganglia (BGA) - have been selected and labelled. Since each sub-volume of interest includes 5 adjacent slices of dimension 150×150, the manually labelled dataset accounts for a total number of $2.7M$ voxels.

2.4 Deep learning model

Similarly to Bontempi et al. (2020), the model architecture is designed so as to deal with the dimensionality of the training data (i.e., $256 \times 352 \times 224$ voxels) at once. As shown in Figure 1, the model is a deep encoder/decoder network with three layers, with one, two, and three 3D convolutional blocks in the first, second, and third level respectively ($n_filters = 24, 48, 96$).

Since the network is fed with a whole volume as an input, each convolutional block (kernel size $3 \times 3 \times 3$), processes the whole brain structure. The full volume helps the model to learn both local and global structures and spatial features (e.g., the absolute and relative positions of different brain components), which are then propagated to subsequent blocks. The receptive field for each convolutional block of `CEREBRUM-7T` is provided in the Supplementary Material. Dimensionality reduction is achieved using strided convolutions instead of max-pooling, which contributes to learning the best down-sampling strategy. A dimensionality reduction (of factor 4 on each dimension) is computed after the first layer, to explore more abstract spatial features. Eventually, the adoption of both tensor sum and skip connections, instead of concatenation, helps in containing the dimension of the parameter space to $\sim 1.2M$.

Training is performed on a multi-GPU machine equipped with 4 GeForce® GTX 1080 Ti, on which different parts of the model are distributed. During training, we optimise the categorical cross-entropy function using Adam (Kingma and Ba, 2014) with a learning rate of $5 \cdot 10^{-4}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, using dropout ($p = 0.1$ on 2th and 3rd level), and without batch normalisation (Ioffe and Szegedy, 2015), achieving convergence after ~ 23 epochs. The code is written in TensorFlow and Keras.

3 Results

To check the validity of `CEREBRUM-7T` as a segmentation tool for brain 7T data, in Section 3.1, we first provide a quantitative assessment of the obtained segmentation, with and without data augmentation, with respect to the automatically obtained GT labelling. Then, in Section 3.2, we present the outcome of the survey carried out on a data portion by experienced neuroscientists who were asked to subjectively evaluate the best segmentations among `CEREBRUM-7T`, the automatic GT, and the manual segmentation. Finally, in Section 3.3 the manual segmentation masks are used as gold standard for the numerical comparison among `CEREBRUM-7T` outcome, the automatic GT, and the manual segmentation.

3.1 Numerical comparison against GT

`CEREBRUM-7T` architecture is compared against the automatically obtained GT in two variants: with and without data augmentation. The two models are trained by minimising the same loss and using the same learning rate.

Performance is assessed by three metrics adopted by the MICCAI MRBrainS18 challenge, which are among the most popular ones used in the context of segmentation (Taha and Hanbury, 2015): the Dice Coefficient (DC), a similarity measure which accounts for the overlap between segmentation masks; the Hausdorff Distance computed on its 95th percentile (HD95), which evaluates the mutual proximity between segmentation contours (Huttenlocher et al., 1993); the Volumetric Similarity (VS) as in (Cárdenes et al., 2009), a non-overlap-based metric which considers the similarity between volumes.

The quantitative comparison is outlined in Figure 7 where average results for DC, HD95, and VS) obtained on the 18 test volumes are shown class-wise (i.e., on GM, WM, ventricles, cerebellum, brainstem, and basal ganglia).

Overall, the architecture with data augmentation outperforms the baseline solution on every class, independently from the observed metric. This is especially noticeable for HD95, where the difference in the average score between the two configurations (computed on all test volumes) is proportionally more prominent than for other metrics. This might be due either to a larger variability (which might affect the reliability of the measure), or to the fact that, since HD95 accounts for differences in segmentation contours, the beneficial effects given by offline data augmentation (i.e, shifts, rotations, and morphing) reflects on an increased accuracy of the segmentation borders. Such interpretation is

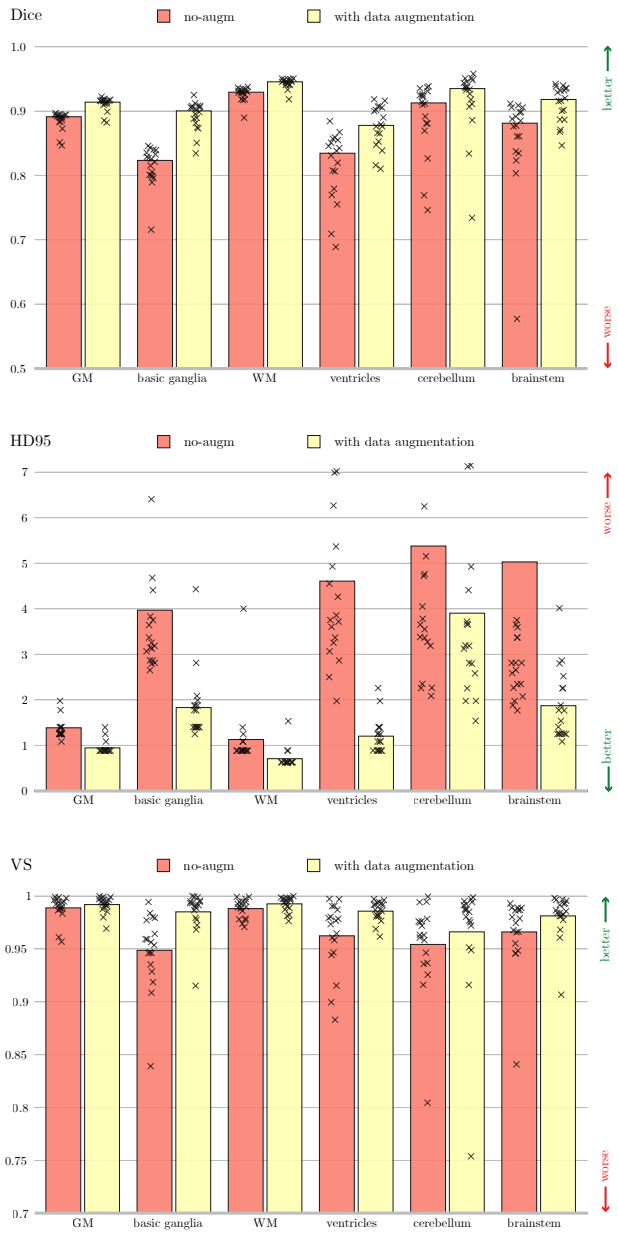


Figure 7. Dice Coefficient, 95th percentile Hausdorff Distance, and Volumetric Similarity computed using the automatic GT segmentation as a reference. The data augmented model (yellow), and the model trained without data augmentation (red) are compared. The height of the bar indicates the mean across all the test subjects, while every mark is a tested volume.

supported by the observation that smaller brain structures, such as ventricles, brainstem, etc., where the identification of segmentation boundaries is most critical, are the ones that benefit the most from such augmentation.

3.2 Survey results

From the quantitative assessment presented in Section 3.1 emerges that there is a relative difference in performance between `CEREBRUM-7T` architectures and the automatic GT used for training. Nevertheless, since performance are measured with respect to the automatic GT, `CEREBRUM-7T` segmentations might be even superior to those provided in

the automatic GT, as in Bontempi et al. (2020) and Roy et al. (2019), as we also suggest in Figure 2 where CEREBRUM-7T masks appear more accurate than GT masks.

To test this hypothesis, we design a survey in which seven expert neuroscientists (different from those who generated the manual segmentation) are asked to choose the most accurate results among three provided ones: the mask produced by CEREBRUM-7T, the automatic GT, and the manual segmentation (intended as gold standard).

If systematically proven, the superiority of CEREBRUM-7T against the automatic GT would confirm the validity of the weakly supervised learning approach, resulting in a learnt model with generalisation capability over its training set. Furthermore, a human expert evaluation, compared to a purely numerical measure, has the advantage to account for the grade of severity of every single segmentation error, giving important feedbacks on the suitability of the segmentation for the application (Taha and Hanbury, 2015).

The survey participants are presented with a set of randomly arranged slices taken from the manually annotated dataset: they are either axial, sagittal, or coronal views from the eight selected areas of interests (see Section 2.3 for details) segmented with the three compared methods (CEREBRUM-7T, GT, and manual segmentation). For each presented couple of segmentation results, the expert is asked to choose the best one between the two, or, in case he/she is unsure, to skip to the next slice set. Each participant inspects all eight areas of interest, for each of the three test volumes. To better compare results in a volumetric fashion, it is also possible for the participant to browse among neighbouring slices (two slices before, and two after), and interactively change the mask opacity to more easily check for an exact anatomical overlap. A snapshot of the survey interface, coded with PsychoPy (Peirce et al., 2019), is provided in the Supplementary Material.

As evident when inspecting the survey aggregated results shown in Figure 8(a), the direct comparison between CEREBRUM-7T and the GT shows that survey participants clearly favoured our proposed solution. Moreover, when both CEREBRUM-7T and the GT are compared against manual segmentation, CEREBRUM-7T obtains more favourable results than GT. This is confirmed also when results are split per different brain areas: in the comparison against manual, GT is almost never chosen, while in selected areas (i.e., EVC, MCX, HIP) CEREBRUM-7T becomes competitive also against the gold standard offered by manual segmentation.

3.3 Numerical comparison against manual segmentation

For the sake of completeness, we also present a purely numerical evaluation based on Dice coefficient of the same dataset used in the survey (three volumes, eight selected areas per volume). Considering manual annotations as a reference, Figure 9 presents a quantitative comparison among CEREBRUM-7T, the automatic GT, the results obtained by FreeSurfer, and those by the method presented in Fracasso et al. (2016) on applicable classes only (GM and WM). The results, separately presented for each of the six brain categories, confirm that CEREBRUM-7T returns the most accurate segmentation on all brain structures.

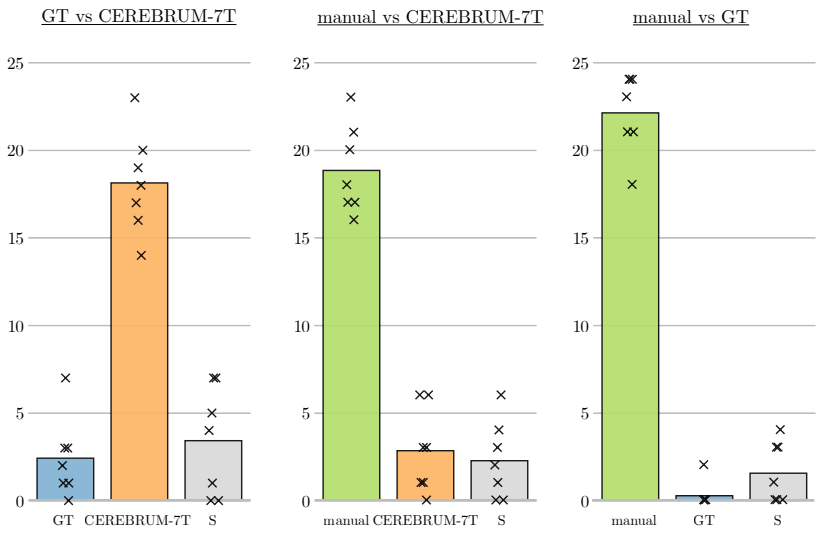
4 Discussion

In this work, we design and test CEREBRUM-7T, an optimised end-to-end CNN architecture that allows for the segmentation of a whole MRI brain volume acquired at 7T. The speed of computation, i.e., few seconds, and the quality of obtained results, make CEREBRUM-7T an advantageous solution for 7T MRI brain segmentation among the few currently available. Furthermore, in order to allow other researchers to replicate and build upon CEREBRUM-7T findings, we make code, 7T data and other materials (including GT and survey) available to readers[6].

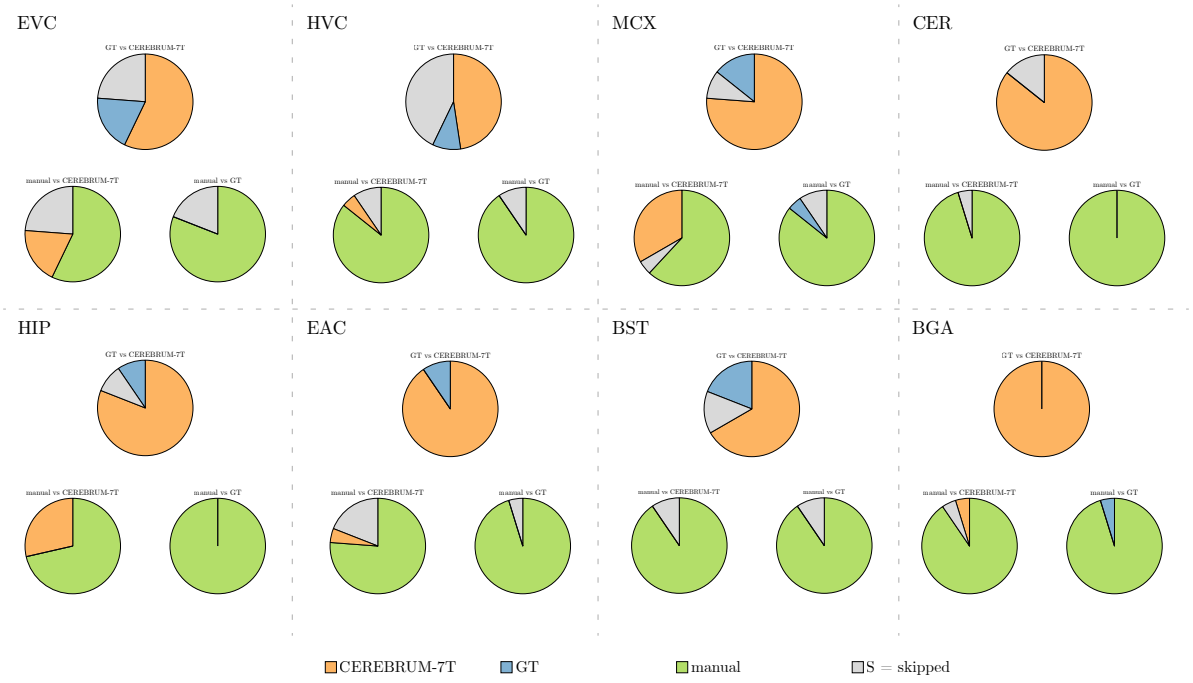
Full volumetric segmentation

Similarly to CEREBRUM, also CEREBRUM-7T processes the whole brain volume as one, avoiding the drawbacks of the tiling process (Reina et al., 2020), thus preserving both global and local contexts. This partially resembles what happens

[6] After publication



(a) Aggregated survey results.



(b) Results per area.

Figure 8. Results of the survey. (a) The three subplots show the three comparisons questioned during the survey (manual vs `CEREBRUM-7T`, manual vs GT, GT vs `CEREBRUM-7T`), since segmentations masks were presented in couples. GT votes are displayed in blue, `CEREBRUM-7T` in orange, while skipped responses (`S`), meaning participants could not choose between the two segmentations, are displayed in gray. The height of bars indicate the means across subjects (i.e, how many times a selection was made, where max. is 3 volumes $\times$ 8 areas $=$ 24); every mark `x` is a participant. (b) Results are split per area of interest: early visual cortex (EVC), high-level visual areas (HVC), motor cortex (MCX), cerebellum (CER), hippocampus (HIP), early auditory cortex (EAC), brainstem (BST), and basal ganglia (BGA).

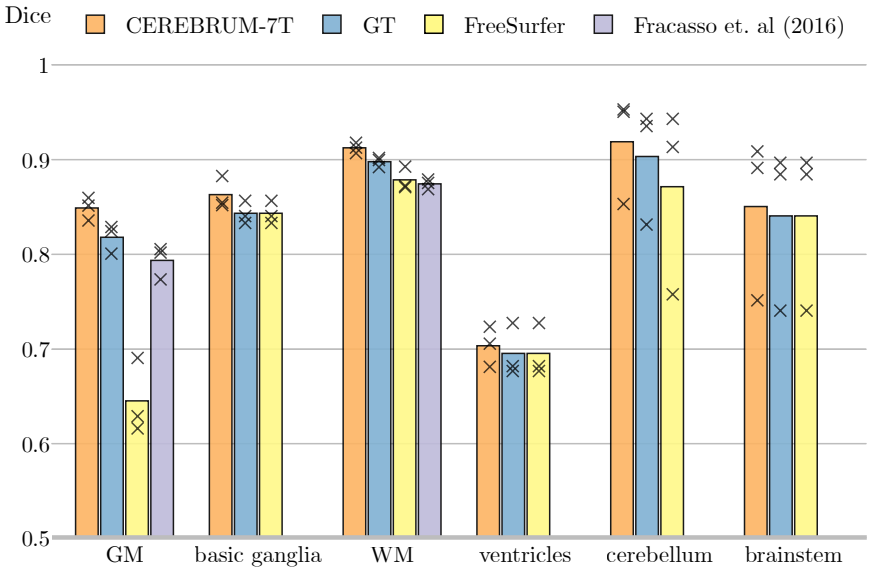


Figure 9. Using manual annotation as a reference, comparison of Dice coefficient between our method (CEREBRUM-7T), the automatic GT used for training, FreeSurfer, and the segmentation tool in Fracasso et al. (2016) (only GM/WM). Every mark is a tested volume from the manually annotated testing set.

during manual segmentation: first, the expert looks at the brain volume from afar to identify where different brain structures are located (global clues). Once a coarse segmentation is apparent, the expert begins to segment voxel by voxel at the pixel level, focusing only on a specific area (local processing). For a human, both of these two levels (or scales) of information are needed to perform the segmentation.

CEREBRUM preserves such two-scale analysis: global features are obtained by analysing the volume at once, without partitioning. The full-resolution processing of the first layer enables to perform a maximum resolution analysis. A table reporting the receptive fields for each convolutional block of CEREBRUM-7T can be found in the Supplementary Material.

Classical automatic (pre-DL) segmentation tools, instead, emulate these two steps using atlases to gain global clues and, for most of them, gradient methods for the local processing. For what concerns DL segmentation methods based on tiling, they conceptually lack in the gain of global clues. Furthermore, limitations in memory size of accelerator cards, prevented so far large medical volumes from being processed as a whole: thanks to the reduction of network layers we applied on the model architecture, it was possible to make the exploitation of global spatial information computationally tractable. In fact increasing the depth of a CNN does not always allows the model to capture richer structures and yielding better performance. On the contrary, as highlighted by works such as Perone et al. (2018), in some cases the low-level features extracted by the network prove to be the most important ones - even if the task is complex. Maintaining a small number of layers allow us to analyse the volume at full resolution and at once, gaining both global and local scale: this brings in a sense our DL model closer to an atlas, with respect to any other previous approach, since it finally learns a-priori probabilities for every voxel.

Mesh reconstruction

The CEREBRUM-7T approach is able to efficiently develop 3D high-quality models useful for the neuroscientific and biomedical communities. For example, in functional MRI (fMRI) studies, researchers need first to isolate specific brain structures, for example the GM, in order to analyse the spatio-temporal patterns of activity happening within it. As such, we show in Figure 10 a view on a reconstructed mesh obtained from a testing volume processed by CEREBRUM-7T. The left mesh is reconstructed by using Brainvoyager (Goebel, 2012) directly from the obtained GM/WM boundary, without

any manual correction; the right one was subjected to a light smoothing operation performed on 50 iterations to enhance the visual quality of the 3D model.

By operating as a true 3D structure model, `CEREBRUM-7T` ensures globally smoother and more coherent surfaces across slices with respect to 2D methods, both manual and automatic. Commonly adopted editing tools, such as ITK-SNAP (Yushkevich et al., 2006) or `ilastik` (Berg et al., 2019), usually display 3 synchronised 2D orthogonal views onto which the operator draws the contour of the structures. The extraction of a continuous 3D surface from the collection of 2D contours, as well as from 3D tiles, is a nontrivial postprocessing task, where bumps in the reconstructed 3D surface are often inevitable due to inter-slice inconsistencies in segmentation.

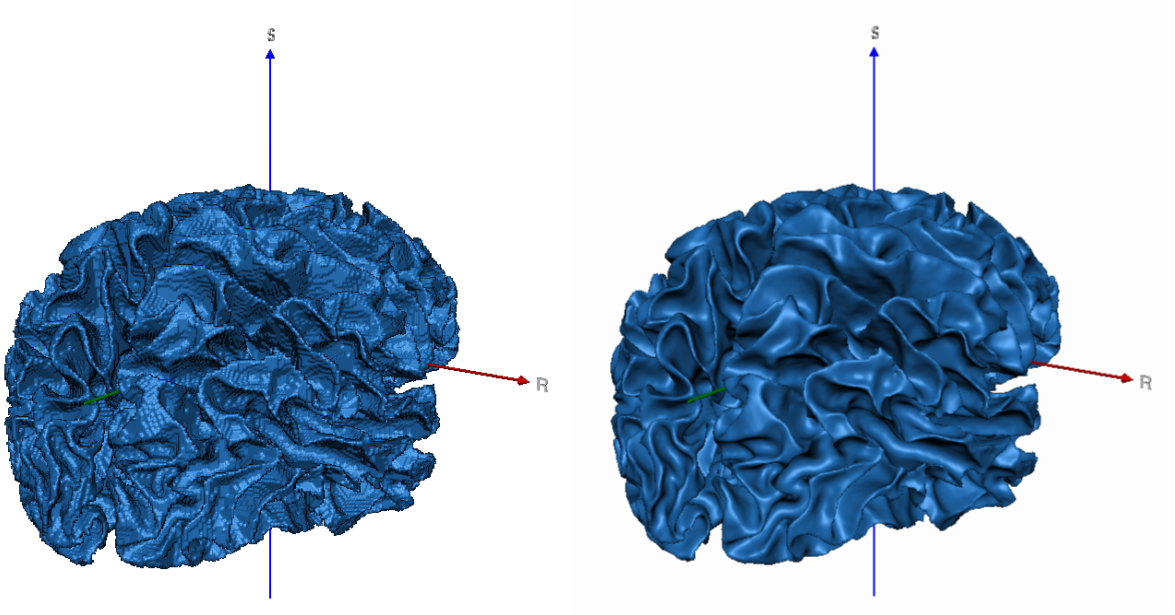


Figure 10. Reconstructed mesh from GM/WM boundary of a testing volume (BrainVoyager, Brain Innovation, Maastricht, The Netherlands). A light smoothing operation is performed on the right mesh (50 iterations) - no manual corrections are performed.

Probability maps

To further appreciate the quality of `CEREBRUM-7T` output, in Figure 11 we show the segmentation inferred by the model before thresholding (percent probability). In such probability maps, each voxel intensity is associated with the probability of belonging to the most likely class. Since most voxels are associated with significant probability of belonging to their correct brain class, such maps demonstrate the ability of the model to make use of both global and local spatial cues. Furthermore, the almost total absence of spurious activations, confirms the high level of certainty achieved by the model.

Limitations

Having decided to process the whole volume at once, which required to maintain a model with low level of complexity, it was not possible to include network elements which are very popular in recent architectures (e.g., dense layers). As another downside of the choice of processing the whole brain volume at once, it was not possible to increase the batch size to a value greater than one, due to the technological constraints of GPU memory. We chose to analyse the volume in its entirety, instead of exploiting the advantages that the increase of the batch size could carry.

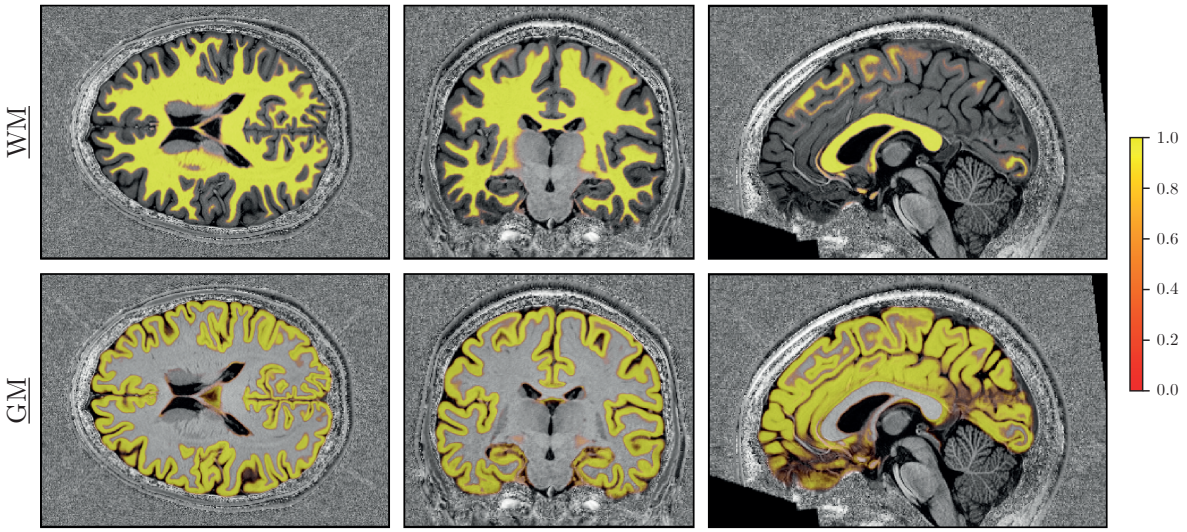


Figure 11. Soft segmentation maps (i.e., probability maps) of a testing volume for WM and GM classes. The model produces maps very consistent probabilities, giving additional flexibility than a hard thresholded map. Remaining classes are shown in Supplementary Material.

The presented experiments are performed only on one sequence (i.e., $T1_w$) in order to limit the scanning time, which is a constraint often imposed for reducing the patient discomfort. However, the code provided can be easily extended to multiple sequences without adding much more complexity to the model.

Being developed on proprietary data, and especially inserted in a study pipeline mostly focused on WM/GM analysis, we performed neck cropping on data, which led to an increase in the filter number.

Lastly, since the number of Ultra-High Field MRI scanners with field larger or equal to 7T for humans is still limited world-wide[7], imaging protocols for 7T are often site-dependent and hardware-specific; as a consequence, `CEREBRUM-7T` is currently highly optimised only for data collected in Glasgow (UK).

Future extensions and applications

In order to make `CEREBRUM-7T` scan-independent, we plan to apply to our model the necessary fine-tuning procedures in order to operate also on data from other 7T sites. This operation will be possible as public databases of 7T data become more available.

In some functional studies, it is important to have a model capable of segmenting only a specific area (for example the hippocampus), since, in order to save scanning time, the anatomical scan is sometimes acquired only on the area of interest. To provide an example of the flexibility of the method, we crop both $T1_w$ and GT on the visual cortex and we retrain the model on GM and WM classes only. Segmentation results are presented in the Supplementary Material. Starting from the code and the GT we provide, `CEREBRUM-7T` can be easily converted and effectively employed for these types of fMRI studies.

Last, the already mentioned extension of `CEREBRUM-7T` to other MRI sequence types, would make easier the differentiation between brain structures, or between normal structures and abnormalities.

[7]Less than hundred, see “Ultra-High Field MRI scanners”.

Acknowledgements

This project has received funding from the European Union Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 and 945539 (Human Brain Project SGA2 and SGA3).

We thank Dr. Alessio Fracasso for assistance with the ground truth creation, the Muckli lab for the survey and neuroimaging knowledge, Dr. Mattia Savardi for very useful technical comments, Dr. Alberto Signoroni for fruitful discussion on learning strategies, and Dr. Lucy Petro for comments improving the manuscript.

Author Contributions

Michele Svanera: Conceptualisation, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Review & Editing, Visualisation. **Sergio Benini**: Conceptualisation, Writing - Original Draft, Writing - Review & Editing, Supervision, Project Administration. **Dennis Bontempi**: Conceptualisation, Methodology, Software, Writing - Original Draft, Writing - Review & Editing. **Lars Muckli**: Resources, Funding Acquisition.

References

Archila-Meléndez, M. E., Valente, G., Correia, J. M., Rouhl, R. P. W., van Kranen-Mastenbroek, V. H., and Jansma, B. M. (2018). Sensorimotor representation of speech perception. cross-decoding of place of articulation features during selective attention to syllables in 7t fmri. *eNeuro*, 5(2).

Bahrami, K., Shi, F., Rekik, I., and Shen, D. (2016). Convolutional neural network for reconstruction of 7t-like images from 3t mri using appearance and anatomical features. In Carneiro, G., Mateus, D., Peter, L., Bradley, A., Tavares, J. M. R. S., Belagiannis, V., Papa, J. P., Nascimento, J. C., Loog, M., Lu, Z., Cardoso, J. S., and Cornebise, J., editors, *Deep Learning and Data Labeling for Medical Applications*, pages 39–47, Cham. Springer International Publishing.

Bazin, P.-L., Weiss, M., Dinse, J., Schaefer, A., Trampel, R., and Turner, R. (2014). A computational framework for ultra-high resolution cortical segmentation at 7tesla. *NeuroImage*, 93:201 – 209. In-vivo Brodmann Mapping of the Human Brain.

Berg, S., Kutra, D., Kroeger, T., Straehle, C. N., Kausler, B. X., Haubold, C., Schiegg, M., Ales, J., Beier, T., Rudy, M., Eren, K., Cervantes, J. I., Xu, B., Beuttenmueller, F., Wolny, A., Zhang, C., Koethe, U., Hamprecht, F. A., and Kreshuk, A. (2019). ilastik: interactive machine learning for (bio)image analysis. *Nature Methods*.

Bergmann, J., Morgan, A. T., and Muckli, L. (2019). Two distinct feedback codes in v1 for 'real' and 'imaginary' internal experiences. *bioRxiv*.

Berron, D., Vieweg, P., Hochkeppler, A., Pluta, J., Ding, S.-L., Maass, A., Luther, A., Xie, L., Das, S., Wolk, D., Wolbers, T., Yushkevich, P., Düzel, E., and Wisse, L. (2017). A protocol for manual segmentation of medial temporal lobe subregions in 7tesla mri. *NeuroImage: Clinical*, 15:466 – 482.

Bontempi, D., Benini, S., Signoroni, A., Svanera, M., and Muckli, L. (2020). Cerebrum: a fast and fully-volumetric convolutional encoder-decoder for weakly-supervised segmentation of brain structures from out-of-the-scanner mri. *Medical Image Analysis*, 62:101688.

Botvinik-Nezer, R. et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Ourselin, S., Joskowicz, L., Sabuncu, M. R., Unal, G., and Wells, W., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432, Cham. Springer International Publishing.

Clarke, W. T., Mougin, O., Driver, I. D., Rua, C., Morgan, A. T., Asghar, M., Clare, S., Francis, S., Wise, R. G., Rodgers, C. T., Carpenter, A., Muir, K., and Bowtell, R. (2020). Multi-site harmonization of 7 tesla mri neuroimaging protocols. *NeuroImage*, 206:116335.

Cox, R. W. (1996). Afni: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29(3):162 – 173.

Cárdenes, R., de Luis-García, R., and Bach-Cuadra, M. (2009). A multidimensional segmentation evaluation for medical image data. *Computer Methods and Programs in Biomedicine*, 96(2):108 – 124.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31(3):968 – 980.

Duyn, J. H. (2012). The future of ultra-high field mri and fmri for study of the human brain. *NeuroImage*, 62(2):1241 – 1248. 20 YEARS OF fMRI.

Fedorov, A., Johnson, J., Damaraju, E., Ozerin, A., Calhoun, V., and Plis, S. (2017). End-to-end learning of brain tissue segmentation from imperfect labeling. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3785–3792.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, 62(2):774–781.

Fracasso, A., van Veluw, S. J., Visser, F., Luijten, P. R., Spliet, W., Zwanenburg, J. J., Dumoulin, S. O., and Petridou, N. (2016). Lines of baillarger in vivo and ex vivo: Myelin contrast across lamina at 7t mri and histology. *NeuroImage*, 133:163 – 175.

Goebel, R. (2012). BrainVoyager — Past, present, future. *NeuroImage*, 62(2):748–756.

Gulban, O. F., Schneider, M., Marquardt, I., Haast, R. A., and De Martino, F. (2018). A scalable method to improve gray matter segmentation at ultra high field mri. *PloS one*, 13(6):e0198335.

Hamidinekoo, A., Denton, E., Rampun, A., Honnor, K., and Zwiggelaar, R. (2018). Deep learning in mammography and breast histology, an overview and future trends. *Medical Image Analysis*, 47:45 – 67.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., and Reuter, M. (2020). Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219:117012.

Huntenburg, J. M., Steele, C. J., and Bazin, P.-L. (2018). Nighres: processing tools for high-resolution neuroimaging. *GigaScience*, 7(7). giy082.

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. (1993). Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koizumi, A., Zhan, M., Ban, H., Kida, I., De Martino, F., Vaessen, M. J., de Gelder, B., and Amano, K. (2019). Threat anticipation in pulvinar and in superficial layers of primary visual cortex (v1). evidence from layer-specific ultra-high field 7t fmri. *eNeuro*, 6(6).

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sanchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60 – 88.

McClure, P., Rho, N., Lee, J. A., Kaczmarzyk, J. R., Zheng, C. Y., Ghosh, S. S., Nielson, D. M., Thomas, A. G., Bandettini, P., and Pereira, F. (2019). Knowing what you know in brain segmentation using bayesian deep neural networks. *Frontiers in Neuroinformatics*, 13:67.

Mendrik, A. M., Vincken, K. L., Kuijf, H. J., Breeuwer, M., Bouvy, W. H., de Bresser, J., Alansary, A., de Bruijne, M., Carass, A., El-Baz, A., Jog, A., Katyal, R., Khan, A. R., van der Lijn, F., Mahmood, Q., Mukherjee, R., van Opbroek, A., Paneri, S., Pereira, S., Persson, M., Rajchl, M., Sarikaya, D., Smedby, O., Silva, C. A., Vrooman, H. A., Vyas, S., Wang, C., Zhao, L., Biessels, G. J., and Viergever, M. A. (2015). MRBrainS Challenge: Online Evaluation Framework for Brain Image Segmentation in 3T MRI Scans. *Computational Intelligence and Neuroscience*, 2015:1–16.

O'Brien, K. R., Kober, T., Hagmann, P., Maeder, P., Marques, J., Lazeyras, F., Krueger, G., and Roche, A. (2014). Robust t1-weighted structural brain imaging and morphometry at 7t using mp2rage. *PLOS ONE*, 9(6):1–7.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203.

Perone, C. S., Calabrese, E., and Cohen-Adad, J. (2018). Spinal cord gray matter segmentation using deep dilated convolutions. *Scientific Reports*, 8(1).

Rajchl, M., Pawlowski, N., Rueckert, D., Matthews, P. M., and Glocker, B. (2018). NeuroNet: Fast and Robust Reproduction of Multiple Brain Image Segmentation Pipelines. *arXiv:1806.04224*.

Reina, G. A., Panchumarthy, R., Thakur, S. P., Bastidas, A., and Bakas, S. (2020). Systematic evaluation of image tiling adverse effects on deep learning semantic segmentation. *Frontiers in Neuroscience*, 14:65.

Roy, A. G., Conjeti, S., Navab, N., Wachinger, C., and ADNI (2019). QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage*, 186:713–727.

Sled, J. G., Zijdenbos, A. P., and Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Transactions on Medical Imaging*, 17(1):87–97.

Taha, A. A. and Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):29.

Uğurbil, K., Adriany, G., Andersen, P., Chen, W., Garwood, M., Gruetter, R., Henry, P.-G., Kim, S.-G., Lieu, H., Tkac, I., Vaughan, T., Moortele], P.-F. V. D., Yacoub, E., and Zhu, X.-H. (2003). Ultrahigh field magnetic resonance imaging and spectroscopy. *Magnetic Resonance Imaging*, 21(10):1263 – 1281.

Wachinger, C., Reuter, M., and Klein, T. (2018). DeepNAT: Deep Convolutional Neural Network for Segmenting Neuroanatomy. *NeuroImage*, 170:434–445.

Wenger, E., Mårtensson, J., Noack, H., Bodammer, N. C., Kühn, S., Schaefer, S., Heinze, H.-J., Düzel, E., Bäckman, L., Lindenberger, U., and Lövdén, M. (2014). Comparing manual and automatic segmentation of hippocampal volumes: Reliability and validity issues in younger and older brains: Comparing Manual and Automatic Segmentation of Hc Volumes. *Human Brain Mapping*, 35(8):4236–4248.

Yushkevich, P. A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J. C., and Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128.

Zhan, M., Goebel, R., and de Gelder, B. (2018). Ventral and dorsal pathways relate differently to visual awareness of body postures under continuous flash suppression. *eNeuro*, 5(1).

Zhou, Z.-H. (2017). A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53.