Medical Image Analysis 93 (2024) 103090

Contents lists available at ScienceDirect

Medical Image Analysis

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

Fighting the scanner effect in brain MRI segmentation with a progressive level-of-detail network trained on multi-site data

Michele Svanera [a,*], Mattia Savardi [b], Alberto Signoroni [b], Sergio Benini [c,1], Lars Muckli [a,1]

[a] Center for Cognitive Neuroimaging at the School of Psychology & Neuroscience, University of Glasgow, UK
[b] Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health, University of Brescia, Italy
[c] Department of Information Engineering, University of Brescia, Italy

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
3D segmentation
Brain MRI
Progressive level-of-detail architecture
Multi-site learning

ABSTRACT

Many clinical and research studies of the human brain require accurate structural MRI segmentation. While traditional atlas-based methods can be applied to volumes from any acquisition site, recent deep learning algorithms ensure high accuracy only when tested on data from the same sites exploited in training (*i.e.*, internal data). Performance degradation experienced on external data (*i.e.*, unseen volumes from unseen sites) is due to the inter-site variability in intensity distributions, and to unique artefacts caused by different MR scanner models and acquisition parameters. To mitigate this site-dependency, often referred to as the *scanner effect*, we propose `LOD-Brain`, a 3D convolutional neural network with progressive levels-of-detail (LOD), able to segment brain data from any site. Coarser network levels are responsible for learning a robust anatomical prior helpful in identifying brain structures and their locations, while finer levels refine the model to handle site-specific intensity distributions and anatomical variations. We ensure robustness across sites by training the model on an unprecedentedly rich dataset aggregating data from open repositories: almost 27,000 T1w volumes from around 160 acquisition sites, at 1.5 - 3T, from a population spanning from 8 to 90 years old. Extensive tests demonstrate that `LOD-Brain` produces state-of-the-art results, with no significant difference in performance between internal and external sites, and robust to challenging anatomical variations. Its portability paves the way for large-scale applications across different healthcare institutions, patient populations, and imaging technology manufacturers. Code, model, and demo are available on the project website.

1. Introduction

Brain structure segmentation in magnetic resonance imaging (MRI) plays a pivotal role in both research and clinical routines for assessing and monitoring brain morphology, volumetry, and connectivity, in both normal and pathophysiological conditions. As more and more studies analyse data derived from thousands of MRI brain scans (Bethlehem et al., 2022), there is a growing need for tools able to perform automatic, fast, and reliable segmentation of brain structures, with benefits on downstream research and clinical studies in terms of accuracy, statistical power, and reproducibility of findings.

Well-established segmentation methods in neuroimaging, such as FreeSurfer (Fischl, 2012), FSL (Jenkinson et al., 2012), SPM (Friston et al., 1995), and CAT12 (Gaser et al., 2022), exploit one or more atlases *i.e.*, reference volumes and their manual trusted segmentation: first, the target is registered with the reference volume, then the anatomical prior knowledge from the manual segmentation is transferred to the target volume (Yaakub et al., 2020). Although computationally expensive and slow, these methods easily adapt to images from different scanners or acquired using different sequences.

Recently, achievements in deep learning (DL) methods applied to automatic brain MRI segmentation (Akkus et al., 2017) such as DeepNat Wachinger et al. (2018), QuickNat Roy et al. (2019), and CEREBRUM (Bontempi et al., 2020; Svanera et al., 2021), have made remarkable progress in competing with the reliability offered by atlas-based segmentation methods. However, most DL methods usually include, for both training and testing, only MRI volumes collected from a single or few centres with almost homogeneous characteristics in terms of image statistics, acquisition parameters, and artefacts. Consequently, when challenged on external data *i.e.*, unseen volumes from unseen sites, DL methods face the so-called *scanner effect*, a drop in performance on handling the data variability originated by different MRI site acquisitions. This mismatch between the distributions of internal and external data, which is common in MRI (see *e.g.*, the competitions in Sun et al. (2021), Campello et al. (2021)) is a problem more broadly known as *distribution*

* Correspondence to: 62 Hillhead Street, G12 8QB, Glasgow, UK.
E-mail address: Michele.Svanera@glasgow.ac.uk (M. Svanera).
[1] Shared authorship.

https://doi.org/10.1016/j.media.2024.103090
Received 16 February 2023; Received in revised form 30 October 2023; Accepted 12 January 2024
Available online 17 January 2024

shift (Wiles et al., 2021). Some researchers in brain segmentation propose to tackle it by applying aggressive data augmentation (Zhao et al., 2019) or harmonisation (Beer et al., 2020), by using domain adaptation or randomisation (Billot et al., 2021), or by generating synthetic data with the needed variations (Shin et al., 2018). Despite achieving good robustness on a wide range of MRI contrasts and resolutions, these approaches keep showing limitations in matching statistics of real data distributions, struggling with morphological variabilities and atypical scanner artefacts.

Given the current availability of open datasets, a concrete opportunity for handling inter-site diversity and improving the model portability is training a model directly on out-of-the-scanner data coming from multiple sites, to cover a broad range of vendors, resolutions, slice thickness, participant demographics, and pathological conditions. By integrating anatomical information acquired by a large number of volumes, such approach builds on the idea of generating the equivalent of an anatomical *brain prior*. Some studies on infant brains demonstrated the effectiveness of using a site-independent scanner-independent prior in helping the tissue segmentation (Wang et al., 2018a,b). However, in these works, prior knowledge of human brain is learned on a limited quantity of individuals and anatomical labels, adopting traditional machine learning classifiers. Also the work in Cerri et al. (2023) makes use of a subject-based segmentation prior based on a deformable probabilistic atlas. Specifically, the method is used for whole-brain and white matter lesion segmentation of longitudinal MRI scans (over 4500 volumes), and it is adaptive to different scanners, field strengths and acquisition protocols.

In general, previous approaches to multi-site learning for segmentation in different medical imaging domains show, on the one hand, that these methods help generalising on external data. On the other hand, they often perform worse on internal ones (*i.e.*, unseen volumes from sites included in the training set) (Styner et al., 2002). This apparently contradictory situation has been observed also in other medical image analysis tasks (Zech et al., 2018), reinforcing the concept that effective learning from multiple sources is highly challenging.

1.1. Main contributions

Driven by the goal of obtaining a brain prior that exploits the information contained in multi-source data, we design a dedicated neural network solution capable of integrating the anatomical knowledge acquired from a large number of volumes. This solution overcomes the memory requirements needed for multiple 3D whole-brain segmentations, and is additionally able to handle the high degree of variability that characterises data from different sites and scanner vendors. To date, we are not aware of existing effective solutions which exploit extensive multi-site data and deep neural networks to automatically generate the equivalent of a robust anatomical brain prior on multiple labels. Our contributions revolve around the following points:

Unprecedented dataset: In our work, we leverage an extensive dataset consisting of nearly 27,000 T1w brain MRI volumes from approximately 160 acquisition sites. This dataset represents on of the largest ever utilised for segmentation, surpassing (Pati et al., 2022), which reported to be the largest dataset in the literature for brain MRI (data from 71 sites, amounting to 6314 volumes).

Data insight: We shed light on a critical issue never tackled before in the field of multi-site segmentation: determine the required number of training datasets and the optimal number of volumes per site to learn a robust model.

Innovative model: We introduce a Level-of-Detail DL network architecture tailored to the unique characteristics of brain imaging data. This innovation eliminates the need for complex registration steps, significantly enhancing segmentation efficiency and speed. The model takes raw MRI data, bypassing computation-intensive pre-processing, and works seamlessly without the need for fine-tuning, making it easier to use. Furthermore, fully 3D segmentation masks are returned within

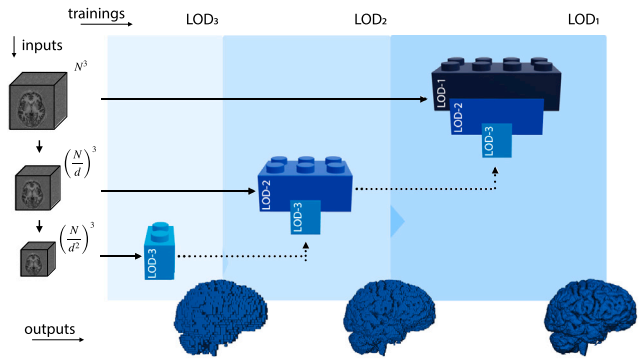


Fig. 1. LOD-Brain is a level-of-detail (LOD) network, where each LOD is a U-net which processes 3D brain multi-data at a different scale. Lower levels learn a coarse and site-independent brain representation, while superior ones incorporate the learnt spatial context, and refine segmentation masks at finer scales. Examples of outputs (grey matter renderings) at different LODs are shown in blue at the bottom.

few seconds, instead of many hours as is the case with the most commonly used tools in the field. In the specific, `LOD-Brain` is a progressive level-of-detail network able to train a robust brain MRI segmentation model from a huge variety of multi-site and multi-vendor data. `LOD-Brain` architecture is organised on multiple levels of detail (LOD), as shown in Fig. 1. Each level is a convolutional neural network (CNN) which processes 3D brain data at a different scale obtained via progressively down-sampling the input volume. Thanks to the rich variability of brain samples from different MRI acquisition sites, the proposed architecture learns, at lower levels, a robust brain anatomical prior. Concurrently, higher levels handle site-specific intensity distributions and scanner artefacts. Through inter-level connections between networks and a bottom-up training procedure, such architecture integrates contributions from all levels to produce an accurate and fast segmentation.

Efficiency and performance: Despite having significantly fewer model parameters compared to competing methods, our approach achieves state-of-the-art segmentation performance. `LOD-Brain` performs better than other solutions on almost every novel site, with no need for retraining nor fine-tuning, and with no relevant performance offset in segmenting either internal or external sites. Furthermore, it proves to be general and robust across sites against different population demographics, anatomical challenges, clinical conditions, and technical specifications (*e.g.*, field strength, manufacturer).

Usage and reproducibility: To maximise research reproducibility and state-of-the-art comparisons, we adopt for testing the MICCAI anatomical structure labels proposed in Mendrik et al. (2015), using FreeSurfer (Fischl, 2012) segmentation masks as silver ground-truth (*i.e.*, a ground-truth with errors). Moreover, as we release both the model and the code at the project website, `LOD-Brain` can be re-trained from scratch to deal with any set of structures and labels obtained by any manual or automatic software. A working demo is also available here.

2. Related work

Atlas- or multi-atlas based methods, such as FreeSurfer (Fischl, 2012) or FLS (Jenkinson et al., 2012), SPM (Friston et al., 1995), and CAT12 (Gaser et al., 2022), are still largely adopted for brain MRI segmentation (Cabezas et al., 2011). Despite the needed registration procedure usually provides a good alignment between volumes, these methods typically require hours of processing for each scan (Klein et al., 2017), thus imposing barriers to groups with limited computational capabilities in case of large-scale studies (Bethlehem et al., 2022). Furthermore, atlas-based strategies are hardly effective on data with

abnormalities, either in terms of anatomy or intensity distributions, requiring manual intervention for error fixing.

In recent years, deep learning (DL) techniques deeply impacted medical imaging (Litjens et al., 2017) and image segmentation tools (Isensee et al., 2021). Regarding the brain, first DL-based methods were limited in handling the 3D nature of MRI data, as they processed single 2D slices only. QuickNAT (Roy et al., 2019) tries to overcome the drawbacks imposed by 2D segmentation by aggregating the predictions of three different 2D slice-based encoder–decoder models, one per canonical slicing plane (longitudinal, sagittal, and coronal), and combining the three results for obtaining the segmentation. FastSurferCNN (Henschel et al., 2020) applies the same 2D approach training the network on a sequence of 2D neighbouring slices, instead of a single slice. To reduce the loss of 3D context and minimise inter-slice artefacts, methods processing 3D-patches and aggregating the resulting sub-volumes are proposed in Dolz et al. (2019), Wachinger et al. (2018). However, all these tools exploit only local 3D spatial information, while global spatial clues, such as the absolute and relative positions of different brain structures, are disregarded, hindering any possible learning of anatomical priors. Other ensemble approaches based on multiple CNNs processing different overlapping brain sub-volumes, such as AssemblyNet (Coupé et al., 2020) or SLANT (Huo et al., 2019), achieve whole brain segmentation, at the cost of an explosion of parameter cardinality. To avoid these drawbacks typical of the tiling process on 2D or 3D patches (Reina et al., 2020), CEREBRUM tools represent a fully 3D solution to brain MRI segmentation for 3T (Bontempi et al., 2020), and 7T scans (Svanera et al., 2021). However, similarly to DL methods which are trained on single-site MRIs, they also do not perform well on volumes from unseen sites, as they require training from scratch, or fine-tuning for each new target distribution (Svanera et al., 2021).

Drawing inspiration from the recent achievements of transformers in the field of Natural Language Processing (NLP), early research works (Hatamizadeh et al., 2022; Zhou et al., 2023) reformulate the challenge of volumetric segmentation as a sequence-to-sequence prediction problem. Yet, these initial investigations primarily focus on multi-organ segmentation, automatic cardiac diagnosis, and the segmentation of brain tumours, without confronting the issues arising in large-scale analyses of independently collected neuroimaging data.

Data harmonisation strategies, when oriented to an explicit removal of site-related effects in multi-site data (Pomponio et al., 2020), constitute a valid strategy to partially alleviate the unwanted performance drop due to the scanner effect. To mitigate inter-site differences, Beer et al. (2020) propose a longitudinal version of the ComBat method: an empirical Bayesian approach which applies a multivariate linear mixed-effects regression to account for both the biological variables and the scanner. The model is able to adjust for additive and multiplicative effects by calculating a site-specific scaling factor. A joint normalising function across multiple datasets is instead learnt by Delisle et al. (2021) by means of two fully-convolutional 3D CNNs: the first normalises image intensities across multiple datasets, while the second optimises images for a downstream segmentation task. The study discussed in Robinson et al. (2020) suggests instead the application of image-and-spatial transformer networks (ISTNs) to address domain shift harmonisation at the image-feature level within multi-site imaging data. Despite harmonisation algorithms mitigate scanner-specific effects, they not always preserve the inter-subject biological variability from each site, and are sometimes sensitive to changes in pre-processing steps (Cetin-Karayumak et al., 2020).

Closely related to harmonisation, domain adaptation methods try to adapt the segmentation networks trained on a source domain to produce correct outputs also on samples from a target domain. As an example, DeepHarmony (Dewey et al., 2019) exploits a fully-convolutional CNN architecture to map brain scans of a subject from one source acquisition protocol to a target one. However, DeepHarmony cannot be extended to more than two sites since it relies on learning a protocol-to-protocol mapping.

SynthSeg (Billot et al., 2021) is an effective adaptation method which, starting from a full domain randomisation of the training set, segment brain MRI scans of any contrast and resolution, without re-training nor fine-tuning. As traditional data augmentation has limited ability to emulate real variations, SynthSeg is trained with synthetic scans obtained by leveraging a generative model with fully randomised parameters (intensity, shape, etc.). Despite its high accuracy, peculiar scanner artefacts and the absence of alignment parameters in the image header can determine the presence of errors in the segmentation.

Far from applying full domain randomisation, Zhao et al. (2019) propose an alternative but still aggressive augmentation solution. This approach first learns independent spatial and appearance transform models to capture the variations in a dataset of brain scans. Then, it uses these transform models to synthesise a dataset of labelled examples starting from only a single selected scan. The synthesised dataset is eventually used to train a supervised network, which significantly improves over previous methods for one-shot biomedical image segmentation, but with less clear advantages in the presence of larger labelled training sets. Other synthetic approaches adopt generative adversarial networks (GAN) to create synthetic abnormal MRI images with brain tumours, so as to improve tumour and brain segmentation (Shin et al., 2018). A DL framework based on GAN is used by Liu et al. (2021) to consider cross-site MRI image harmonisation as a style transfer problem rather than a domain transfer problem; in particular, authors show that MR images can be harmonised by inserting the style information encoded from a reference image directly, without knowing their site/scanner labels a priori. Also the work in Chen et al. (2021) presents a generative framework for improving cross-site segmentation on cardiac imaging datasets. It includes a cooperative training approach with fast and slow-thinking networks and a method for generating challenging training examples that enhance generalisation and robustness to unforeseen data shifts. While synthetic methods can enhance generalisation, aggressive augmentations do not always improve model performance. In these cases, they do not represent a valid solution for coping with distinct scanners and protocols.

One of the first multi-site learning-based attempts at making a model that is robust to the scanner effect is described in Liu et al. (2020) in the domain of prostate segmentation. Authors first perform feature normalisation for each site separately, and then extract more generalisable representations from multi-site data by a novel learning paradigm. Other seminal works that adopt deep learning techniques to cope with the multi-site variability can be found in Rundo et al. (2019) again for prostate segmentation, and in Dou et al. (2020) for multi-organ segmentation from unpaired CT and MRI. However, most of these approaches confirm to perform well on internal subjects, whereas require additional external images for the adaptation step (*e.g.,* see Karani et al. (2018)) to adequately cope with testing data obtained using different imaging protocols or scanners. An accurate review presenting other retrospective techniques to compensate site effects in multi-site neuroimaging analyses, with a thorough discussion on the benefits and drawbacks for each of different use cases, can be found in Bayer et al. (2022). What emerges is that efficiently handling multi-site data is still an open challenge and how the development of models able to jointly handle structure segmentation and site adaptation is highly needed. Learning directly from out-of-the-scanner MRI brain volumes (*i.e.*, with no atlas-based pre-alignment) from multiple-sites, with no fine-tuning nor adaptation steps, is an option that has remained unexplored until now, despite the recent availability of a large amount of brain open data repositories.

3. Brain MRI multi-site data

To address the huge brain MRI variability in intensity statistics and scanning artefacts, we collect almost 27,000 brain T1-weighted volumes of both healthy and clinical subjects, mainly scanned with `mprage/mp2rage` sequences, and released in 80 databases covering

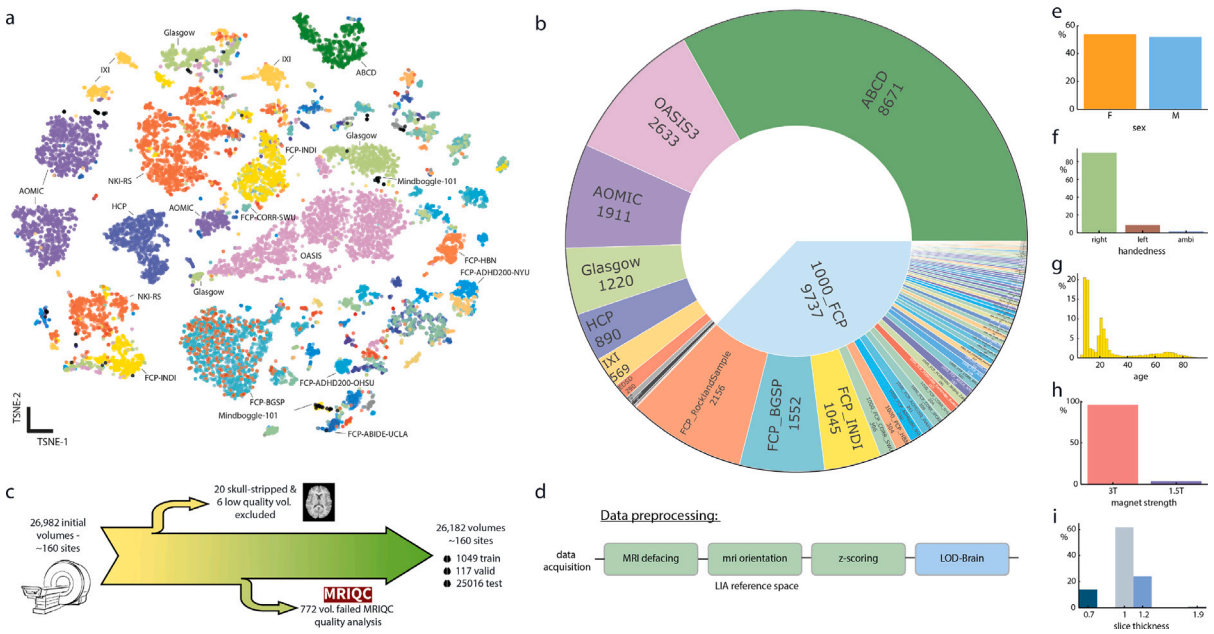


Fig. 2. Multi-site dataset: we collect and analyse with MRIQC (Esteban et al., 2017) almost 27,000 volumes originating from around 160 different sites (26,182 volumes after the quality check). (a) A visualisation by t-SNE (Van der Maaten and Hinton, 2008) of the 68 MRIQC features (different colour for each dataset). Note that one dataset (*e.g.*, IXI in yellow colour) may contain volumes from more than one site or acquired with different scanners, and thus separate in clusters in the t-SNE space. (b) Dataset cardinalities. (c) Details on data quality assessment and (d) pre-processing. From (e) to (i), different demographic features and scanner properties are reported.

approximately 160 world sites.[2] We first aggregated data from well-known open repositories, such as HCP, ABCD, OASIS, and datasets contained in the INDI project including NKI-RS, IXI, ABIDE, and ADHD. Then we added datasets from open platforms as OpenNeuro, OSF, neuGRID2 and NIMH, avoiding paid repositories such as UKBiobank. Other public datasets included are Mindboggle101, AOMIC, BrainWeb, and IBSR. Apart from Glasgow data,[3] all repositories are available without fees, to maximise the reproducibility of this work. A full data table is provided on the project website.

In Fig. 2, we present the composition of the dataset, its cardinalities and features, the quality assessment process done by MRIQC (Esteban et al., 2017), and details on training and testing splits. The 26,182 volumes that passed the MRIQC quality control analysis undergo defacing first, and then simple pre-processing steps before neural network feeding, including FreeSurfer's `mri_convert` to reorient volumes to LIA (left, inferior, anterior) reference space, padding to 256^3, and z-scoring.

3.1. Data split and labelling

Out of the 80 datasets, 70 are considered as internal (INT), while 8 are left out for testing only (EXT). The 2 remaining sets are used for specific analyses: SIMON (Duchesne et al., 2019) contains scans from a single healthy individual who participated in a multi-centre study; the last is a dataset with five patients with only one brain hemisphere from Kliemann et al. (2019). As validated in Section 5.1.1, the model used for testing is trained on a randomised selection of 1049 volumes from internal data (15 volumes for each dataset, except one contributing with 14 volumes as it does not have enough data). This allows to obtain a balanced training set in terms of dataset representativeness and an appropriate total number of training volumes for the learning task. The 78 datasets used for testing (70 INT and 7 EXT) include a total of 25,016 volumes (15,841 INT and 9175 EXT). Since only 10% of the datasets include more than 80% of the testing volumes, we select up to 200 volumes per dataset to avoid biases and guarantee balanced results, ending up with a total of 5976 testing volumes (5360 INT and 616 EXT). The validation set, used for hyperparameter selection, includes 117 volumes from 72 datasets (91 INT and 26 EXT).

As no manual segmentations (gold standard) are available for most volumes, training adopts a weakly supervised learning strategy, exploiting segmentation labels obtained by FreeSurfer (Fischl, 2012) as a silver standard ground-truth (GT), similarly to what proposed in Bontempi et al. (2020). MindBoggle, a dataset with semi-manual labels (*i.e.*, FreeSurfer plus manual corrections), is exploited in validation and testing. The manual segmentations provided for IBSR and MALC2012 were discarded and replaced with FreeSurfer outputs, because of their low quality. The only dataset that provides gold standard segmentation masks is BrainWeb, which contains a set of synthetic MRI data volumes produced by an MRI simulator and is only used in testing.

As for the quality of the FreeSurfer's GT masks, these present high variability. In particular, out of the seven external datasets (testing only) labelled with FreeSurfer, four present an acceptable GT (covering a total of 32 sites), while the other three show low quality GT segmentations as they include clinical scans. Low quality GT masks are usually produced from low quality T1w volumes; while they are not used for training, since we do not want to compromise the model learning ability, they are still used for testing to explore model capabilities and limitations.

The labelling strategy follows the 7 classes adopted by MRBrainS challenge (Mendrik et al., 2015): grey matter, white matter, cerebrospinal fluid, ventricles, cerebellum, brainstem, and basal ganglia. Such labelling maximises the possibility of comparison with other state-of-the-art methods, and covers most of clinical and research studies and applications. However, there are no limitations in selecting different brain structures and related labels for retraining `LOD-Brain`.

4. Methods

`LOD-Brain` is a progressive level-of-detail 3D network designed for brain MRI segmentation. As shown in the general scheme in Fig. 1, each

[2] From open repositories, it is not always possible to retrieve the number and model of the scanners employed in acquisitions, nor the number of unique participants. This means that a database could contain more than one site. Henceforth, we try to distinguish between *dataset* and *site* wherever possible.

[3] Maintained by authors, with pending ethics permissions for sharing.

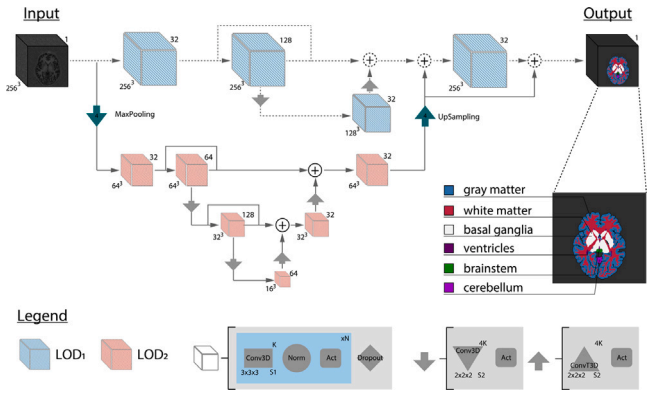


Fig. 3. LOD-Brain architecture selected for the experiments on the brain MRI segmentation task ($L = 2$, $d = 4$).

level of LOD-Brain is a U-net (Çiçek et al., 2016) which processes the input MRI volume (of initial dimensions 256^3) at a different scale obtained by successively down-sampling the volume by a factor d along each coordinate axis. The lowest network level, LOD_L, is in charge of learning a robust *anatomical prior*. Since down-sampling input volumes removes high frequency details and smoothes individual differences, LOD_L learns a coarse representation of brain structures, and their mutual locations, which is less dependent on the scan site. Training happens in a bottom-up approach: after convergence, LOD_L is frozen, and inter-level connections ensure that the 3D spatial context learnt at the lower level is embedded and propagated to LOD_{L-1} and, from there, to higher levels of the architecture. The process is iteratively repeated through superior levels until the upper one *i.e.*, LOD_1, which processes the input data at the full scale, refining the segmentation masks at the finest detail and accounting for site-specific intensity distributions.

The loss $\mathcal{L}$ adopted by LOD-Brain is a mixed per-channel dice function $\mathcal{L}_{dice}$ and cross entropy loss $\mathcal{L}_{CE}$:

$$\mathcal{L} = -\lambda \mathcal{L}_{C\mathcal{E}} - (1-\lambda)\mathcal{L}_{dice}$$

with λ balancing the two components. In particular, $\mathcal{L}_{CE}$ is:

$$\mathcal{L}_{CE} = \sum_{i=1}^{C}\sum_{k=1}^{V} y_{k,i} log(F_{k,i})$$

where V and C are the set of voxels and classes, respectively, y is the GT mask, and F is the output. Conversely, $\mathcal{L}_{dice}$ is:

$$\mathcal{L}_{dice} = 1 - \frac{2}{C}\sum_{i=1}^{C}\frac{\sum_{k=1}^{V} y_{k,i}F_{k,i}}{\sum_{k=1}^{V} y_{k,i}^2 + F_{k,i}^2}$$

Hyperparameter selection, network design, and the choice of parameters L, λ, d, etc. is described in Section 5.1.4.

The architecture which emerges as the best performing one during the experiments is presented in Fig. 3.

The network is made up of three basic 3D convolutional blocks. The first addresses feature learning: it is composed of a $3 \times 3 \times 3$ convolution layer followed by normalisation and non-linear activation, all repeated multiple times, ending with a dropout layer. The other two blocks perform down-sampling and up-sampling, with strided convolution and transposed convolutions, respectively, both followed by non-linear activations. These layers allow the network to learn optimal up/down-sampling strategies and process different extracted feature hierarchies. Moreover, skip connections and inter-level connections are implemented along with summation nodes, as it was proven to have a better trade-off between segmentation accuracy and parameter count compared to concatenation (Milletari et al., 2016).

Table 1
Details on selected augmentation methods.

Group	Augmentation	Prob.	Parameters
Geometrical	Flip	1/2	Sagittal plane only
	Grid distortion	1	Steps: 5; Distortion: .1
Noise	Salt and pepper	1/6	Amount: 0.01; Salt: 0.2
	Gaussian	1/6	Amount: 0.2
	Gamma	1/6	Clip: 0.025
	Contrast	1/6	Alpha: 0.5-3.0
	Blur	1/6	Limit: 3
	Downscale	1/6	Scale: 0.25-0.75
Artefacts	Ghosting	1/2	Max rep.: 4
	Inhomogeneity	1/2	See Svanera et al. (2021) for details

4.1. Data augmentation

Instead of performing a pre-selected set of common data augmentations, we perform an ad-hoc procedure to verify the usefulness of augmentations in advance. In the first step, we create a pool of realistic transformations belonging to three categories: geometrical transformations, noise distortions, and artefact introduction. In the first category, in addition to classical operations such as, flip, rotation, and translation, we also introduce grid distortion. The second category accounts for a comprehensive set of noises: salt and pepper, Gaussian, Gamma, and contrast noise. The last transformation family focuses on mimicking MRI artefacts like ghosting and MR field inhomogeneity, as described in Svanera et al. (2021). In the second step, we test which transformation is beneficial to increase the model performance. Validation is done by applying each transformation to the validation set volumes (by increasing transformation parameters), and then computing the performance of a model trained without any data augmentation. If the model is already robust to a specific transformation (*i.e.*, there is no performance gap in testing a volume with and without transformation), this is no further considered. Otherwise, in those situations in which the training set is not rich enough (*i.e.*, whenever transforming the input data introduces a performance drop), such transformation is considered suitable for augmentation, since it can introduce a realistic alteration to input volumes that the model is not able to handle yet. Table 1 reports details on selected augmentations only, showing probabilities of application and parameters justified by the experiments detailed in Section 5.1.3.

5. Results and discussion

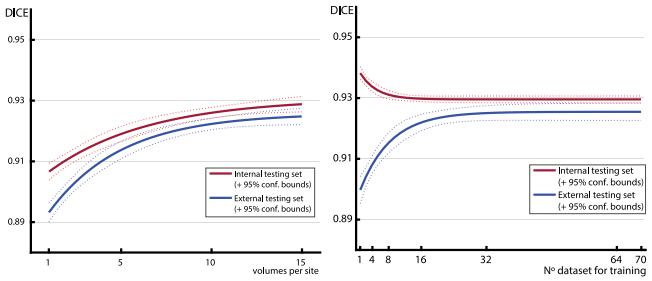
The experimental assessment of our multisite-based model is structured as follows. The first set of experiments aims to justify the choice of the adopted model. Next, we test the robustness and generality of LOD-Brain on different types of data (internal and external datasets, and data with marked anatomical variations), and the invariance of the model against different types of bias, including scanner vendors and models. Eventually, we quantitatively and qualitatively compare our method against the state-of-the-art. Unless differently stated, results are computed using Dice coefficient as performance metric, and FreeSurfer segmentation as silver ground-truth.

5.1. Model training and hyperparameter selection

5.1.1. Multi-site learning

Given the richness of the aggregated data, we first want to find suitable dimensions for the training set. This requires answering to the following questions: in order to reach good generalisation capability (*i.e.*, similar performance on both INT and EXT testing data), how many volumes per site should be considered in the training set? And, how many different datasets should be included?

In Fig. 4(a), we show the performance of models trained with 1, 5, 10, or 15 volumes per dataset, considering all 70 available training



(a) How many volumes? (b) How many datasets?

Fig. 4. (a) Using 70 available training datasets, we trained 4 models with $[1, 5, 10, 15]$ volumes per dataset. The model is tested on INT (red) and EXT data (blue). (b) Using 15 volumes per dataset, we train models with an increasing number of sites $[1, 4, 8, 16, 32, 64, 70]$. Testing is done on INT and EXT data (red and blue respectively). Both graphs fit exponential curves.

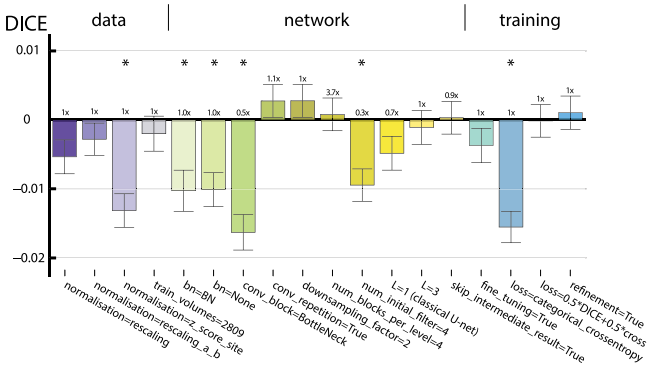


Fig. 5. Ablation study. Performance of models trained with different architectural options are shown with respect to the best model (on the zero x-axis). Results (Dice coefficient differences) are computed on the validation set (those marked with * are statistically significant according to t_{test} applying Bonferroni correction).

datasets (*i.e.*, number of training volumes = 70, 350, 700, 1049). Internal testing (INT) is performed on 110 unseen volumes equally distributed among the same datasets used for training at the considered step, while external testing includes again 110 volumes from 4 left out datasets (ABCD, MALC2012, 1000_FCP_CORR_NKI_TRT, and MindBoggle101). As expected, using more training volumes per dataset enhances the segmentation accuracy for both INT and EXT testing data. Since we reach a plateau of performance between 10 and 15 volumes per dataset, and to avoid the introduction of dataset-related biases, we decide to use 15 volumes per dataset, which is the maximum possible for maintaining balance across datasets.

In Fig. 4(b), we evaluate the accuracy of `LOD-Brain` as a function of the number of datasets included in training. Testing volumes, both INT and EXT, are the same used in Fig. 4(a) to allow comparisons. For each value $i \in [1, 4, 8, 16, 32, 64, 70]$, we retrain `LOD-Brain` with 1049 volumes selected from i datasets randomly chosen from those with enough samples. As shown in Fig. 4(b), as long as the number of sites increases, the gap of performance between internal and external testing data progressively decreases, until it fades. Therefore, unless otherwise specified, we set to 70 the number of datasets used to train `LOD-Brain`.

5.1.2. Parameter selection

In Fig. 5, we present the most relevant results of the ablation study carried out to select model parameters.

This includes investigations regarding data processing, network architecture, and training. All results are computed on the validation set by evaluating their statistical significance and, in case no significance is found, by preferring models with least parameters.

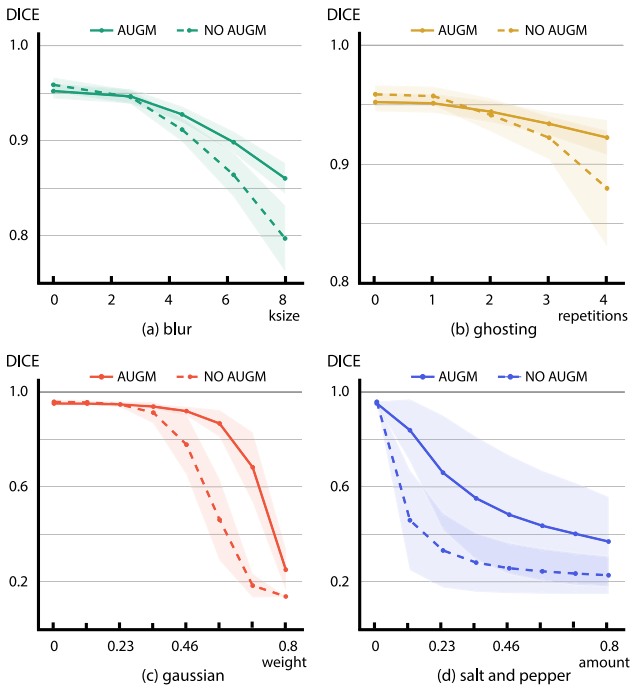


Fig. 6. Data augmentation: performance of models trained with versus without augmentation for four transformations (*i.e.*, blur, ghosting, gaussian, and salt and pepper noise).

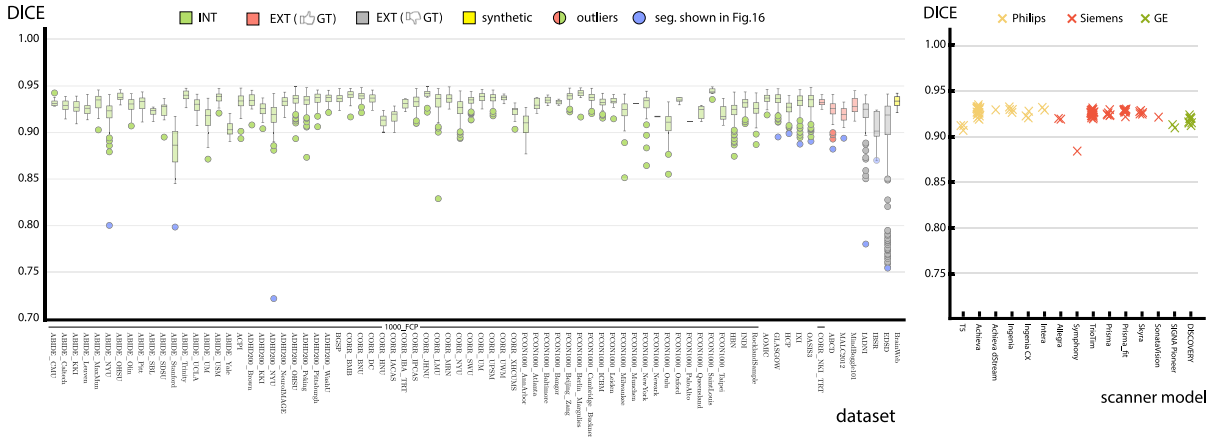
On data, we evaluate the most advantageous type of data normalisation, and we make an attempt to train with a larger training set (almost 3k samples). However, since this causes a unbalance in data, we observe a drop in performance.

Regarding the network, we test different design choices for its architecture *e.g.*, the number of levels L, the convolutional block (plain or residual), layer normalisation (batch or group), etc. As a result, the LOD network implemented for testing is configured on $L = 2$ levels and a down-sampling factor of $d = 4$, as shown in Fig. 3. It is relevant to note that working with two levels as resulting from the ablation study, somehow recalls the effectiveness of the approaches extensively used in classic reference methods for brain segmentation: atlas-based registration first, followed by voxel-level segmentation. Similarly here, the coarser level learns a robust brain prior which replaces the registration step in identifying brain structure locations, while, the finest level, handles site-specific intensity distributions and artefacts. The entire procedure also may resemble the steps of manual segmentation, in which the human expert first zooms out to identify major anatomical structures, and then zooms in refining structures until the task is complete at the finest level.

With respect to training choices, we compare, among others, different loss functions (best with $\lambda = 0$ *i.e.*, pure Dice loss) and evaluate as detrimental a refinement of the entire unfrozen network, thus confirming that the brain prior learnt at LOD_2 is robust, and that a joint fine-tuning with a higher level would negatively affect its site-independent brain representation.

5.1.3. Data augmentation

After selecting the useful transformations as in Section 4.1, we augment the validation set (117 volumes) and test the two models trained with and without augmentation. Fig. 6 reports the comparison as function of the augmentation parameters for four significant transformations.



(a) LOD-Brain results on all datasets (b) Results on SIMON dataset

Fig. 7. Testing results. (a) Dice coefficient on 5976 testing volumes (78 datasets), displayed per dataset: INT (green), EXT with good GT (red), with low-quality GT (grey), and synthetic GT (yellow). Segmentation masks of the worst numerical result (blue dots) are further displayed in Fig. 16. (b) Accuracy on SIMON (Single Individual volunteer for Multiple Observations across Networks) dataset (EXT) (Duchesne et al., 2019), comprising 94 volumes acquired with 15 different models of scanner by 3 major MR vendors (in different colours).

5.1.4. Implementation details

Training optimisation is done using Adam (Kingma and Ba, 2014) and training lasts 50 epochs for LOD_2 and 30 for LOD_1, with an initial learning rate of $5e-4$, reduced by $1/4$ on plateau. As non-linear activation, `relu` is applied for both encoder and decoder. For better regularisation, each convolutional block performs group normalisation, while the dropout rate is 0.05. Training lasts 3 days using a workstation with Nvidia©Quadro RTX 8000 GPUs and Weights & Biases for experiment tracking.

5.2. Robustness and generalisability

In the following series of experiments, we test `LOD-Brain` on a variety of scenarios to assess its robustness and capabilities.

5.2.1. Accuracy across datasets

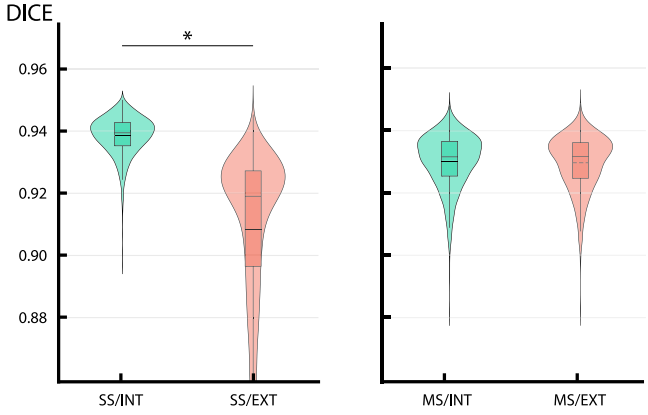
Fig. 7(a) reports segmentation performance for each of the 78 datasets used in this study. The overall accuracy (mean: 0.928, std: 0.017) proves the robustness of the method, showing similar results on both internal and external sites. The performance obtained on low-quality GT datasets (in grey in Fig. 7(a)) is justified by the presence of several scans with head movement artefacts due to participant populations (*e.g.*, elderly people with dementia in EDSD and children 7.5–12.9 y.o. in ABIDE_Stanford) which impair FreeSurfer segmentation.

5.2.2. Multi-site versus single-site models

To validate the need for multi-site data, we compare the generalisation abilities of multi-site (MS) training with those of single-site (SS) models, by testing both (MS vs. SS) on the same internal (INT) and external data (EXT). For enabling comparison, this experiment uses only the 4 single-site datasets with more than 1049 volumes (AOMIC: 1911 volumes, Glasgow: 1220, FCP_BGSP: 1552, FCP_RocklandSample: 2156).

For each of the 4 datasets, we train a SS model with 1049 volumes, and we test its segmentation accuracy on both internal data (*i.e.*, all remaining volumes from the same site) and external sites (*i.e.*, all left-out volumes from the other 3 datasets). A significant drop between INT and EXT performance, due to the scanner effect, is observed in Fig. 8(a).

As for multi-site training, we test our model trained with 1049 volumes from 70 datasets on the same INT dataset used in the single-site case. To test on external data, we train 4 additional MS models on 69 datasets, considering the left-out dataset (one among AOMIC, Glasgow, FCP_BGSP, and FCP_RocklandSample) as EXT data. As both experiments in Figs. 8(a) and 8(b) use the same testing sets, we observe



(a) Single-site model (b) Multi-site model

Fig. 8. (a) Performance obtained using single site (SS) models: in green, results on testing volumes from the same site (INT); in red, on testing volumes from different sites (EXT). Performed t_{test} shows a statistically significant difference (marked with *) between INT and EXT ($p_{value} < 1e^{-5}$).
(b) Performance of multi-sites models (MS): in green, results on testing volumes from the same sites (INT); in red, on testing volumes from different sites (EXT) (no statistically significant difference between INT and EXT $p_{value} = 0.11$).

that models trained on multiple sites almost reaches on EXT data the same performance of SS models tested on unseen volumes from their training sites, while exhibiting far superior generalisation ability (*i.e.*, a non-significant performance difference between INT and EXT data in Fig. 8(b)).

5.2.3. SIMON dataset

As segmentation performance can vary for both scanner intensity distribution and variability in participants' anatomy, here we attempt to disentangle the two components. We therefore test our model on a left-out dataset where, in the context of a multi-centre study (Duchesne et al., 2019), a single subject has been scanned many times and sites during his lifetime from 29 to 46 years old. The dataset includes 73 sessions (94 volumes), 33 world locations, 15 different models of scanner, covering the 3 major MR vendors (GE, Philips, and Siemens). Fig. 7(b) results show an impressively high coherence among a large variety of scanner models by different vendors.

5.2.4. Robustness to challenging anatomical variations

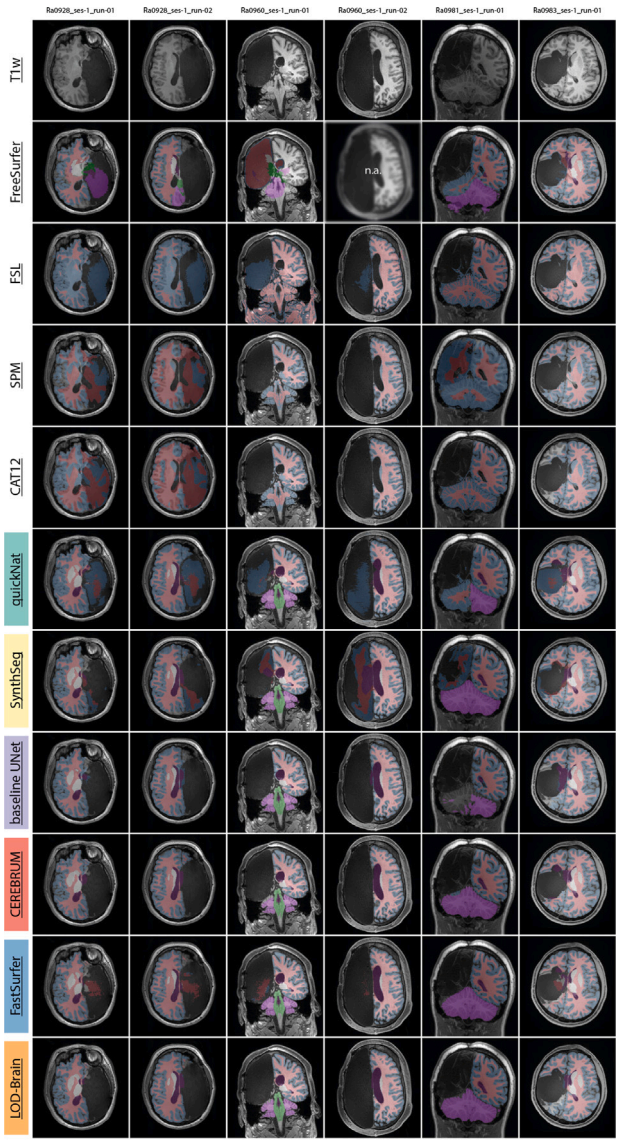
To test the robustness of the anatomical prior learnt, we test our model on a challenging scenario: a dataset with five individuals who had undergone surgical removal of one hemisphere (Kliemann et al., 2019). In Fig. 9(a), we show the visual results obtained by LOD-Brain compared with five competing deep learning methods and three atlas-based methods FSL, SPM, and CAT12, in addition to FreeSurfer. While FreeSurfer – and in general atlas-based methods – fails to generalise to such severe anatomical singularities, often inferencing non-existing structures, LOD-Brain reliably segments such cases, proving a high level of robustness to anatomical variations. In Fig. 9(b), we report different activation maps for three subjects of this dataset coming from different levels (i.e., layers) of the network. Skull stripping and cortex extraction is coarsely done already in LOD_2 (bottom layer) and then the information is propagated along the network upper levels. This result gives an intuition on how the coarse level acts as a prior, giving guidance to LOD_1 for finer segmentation.

5.2.5. Invariance to bias

To investigate the fairness of our segmentation model, we assess LOD-Brain for potential bias regarding demographic characteristics such as sex and age, or other technical characteristics of the scanner including scanner model, vendor, magnet strength, and slide thickness. Despite the training data imbalance for some of these characteristics (see those in Figs. 2g/h/i), on the test-set of 5976 volumes we observe no salient differences in Dice performance between different groups. Results are reported on the project website and in the Supplementary material (Fig. 2).

5.3. Quantitative method comparisons

A comparative assessment of our method against state-of-the-art techniques is proposed here in terms of both brain segmentation performance and model complexity. The selected methods are chosen based on stringent evaluation criteria, encompassing not only segmentation accuracy but also robustness to various challenges and practical applicability. Furthermore, our choice aims at ensuring a well-rounded assessment across various categories of segmentation techniques: for algorithms operating on 2D slices, we included QuickNat (Roy et al., 2019) as best method; for methods working with coronal, axial, and sagittal 2D slice stacks, we selected FastSurferCNN (Henschel et al., 2020); in the category of fully 3D methods, we evaluated CEREBRUM (Bontempi et al., 2020) and the baseline 3D-UNet (Çiçek et al., 2016); for synthetic data-driven methods, we incorporated SynthSeg (Billot et al., 2021). Fig. 10(a) shows the obtained results on the whole testing set grouped by segmented brain structure. Fig. 10(b) focuses instead on the comparative performance of different methods on external datasets only. Obtained results highlight LOD-Brain as one of the most competing methods on all brain labels, as it yields the best scores in almost all target structures and on the majority of external datasets with acceptable GT. The number of parameters for each model is also reported, highlighting LOD-Brain (337,719 parameters only) as the best overall model in terms of performance-to-complexity ratio. It is also relevant to note that LOD-Brain first outperforms all other methods on BrainWeb, a synthetic dataset that usually serves as a gold standard because of its correctly segmented ground truth. Furthermore, it achieves high performance on the ABCD dataset, despite it includes volumes from 32 diverse scanners that were previously skull-stripped and aligned to MNI152 reference space (a common situation in the field).



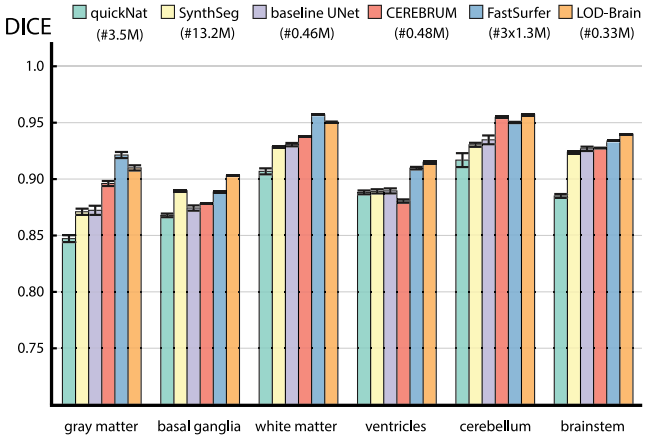
(a) Visual results

(b) Activation maps

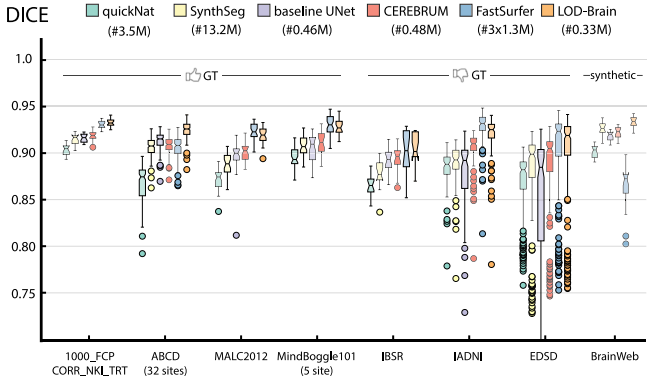
Fig. 9. Results on individuals who had undergone surgical removal of one hemisphere (Kliemann et al., 2019). (a) Inference for 5 subjects are shown for all methods, both DL- and atlas-based. (b) Activation maps for 3 subjects at different LODs *i.e.*, layers in the network (zoom in for better view).

5.3.1. Robustness to motion artefacts

During the image acquisition process, the quality of MRI images can be compromised by motion artefacts, which may adversely affect

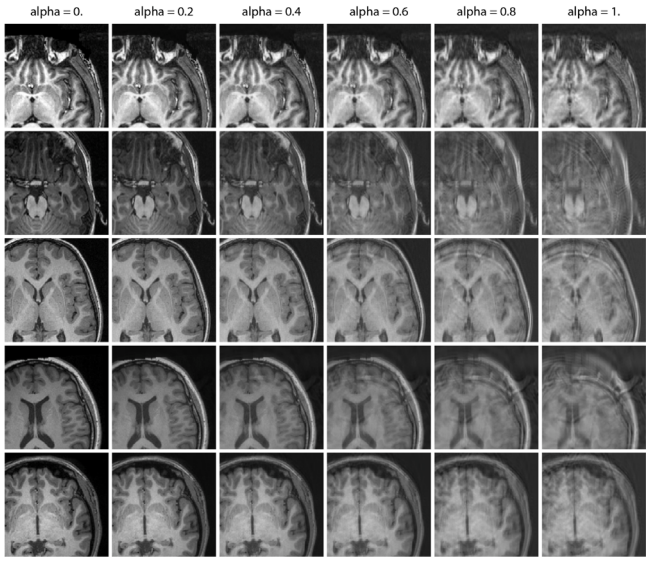


(a) Results grouped by brain structure

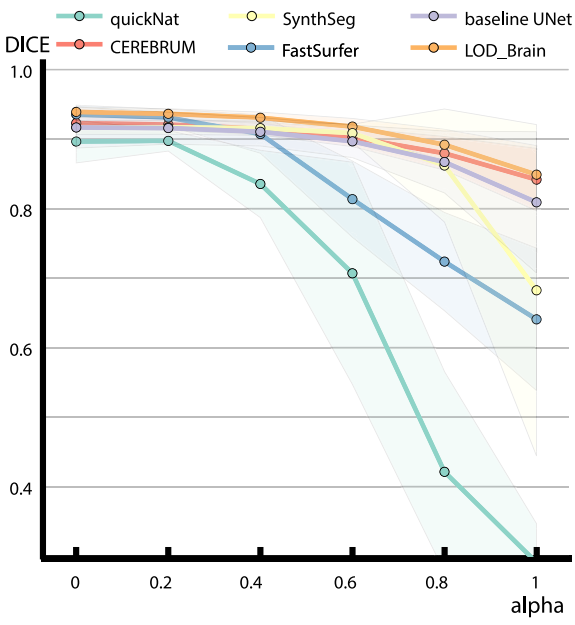


(b) External datasets only

Fig. 10. Performance comparison: QuickNat (Roy et al., 2019), SynthSeg (Billot et al., 2021), 3D-UNet (Çiçek et al., 2016), CEREBRUM (Bontempi et al., 2020), FastSurferCNN (Henschel et al., 2020), and our method. (a) Results are computed on the test set of 5976 volumes, using FreeSurfer as GT reference and grouped for brain structure. (b) Results on external sites only, divided in acceptable vs. low-quality vs. synthetic ground-truth. Numbers of parameters for each model are also reported.



(a) Simulated motion artefacts changing the alpha parameter



(b) Performance comparison

Fig. 11. Display of motion artefacts (a) changing parameters (zoom in for better view) and (b) performance comparison.

clinical diagnoses and automated image analysis. Therefore, ensuring the robustness of a segmentation method to motion artefacts is crucial. To assess the effectiveness of our approach in handling motion artefacts, we employ a method proposed by Shaw et al. (2020), which allows for the generation of realistic motion artefacts on existing MRI data. Specifically, we apply this technique to 754 randomly selected testing volumes. In Fig. 11(a), we provide examples of T1w images with increasing motion artefacts. Subsequently, in Fig. 11(b), we present a comparative analysis of `LOD-Brain`'s performance, along with competing methods. It is worth noting that while other existing methods struggle to deliver satisfactory results in the presence of motion artefacts, `LOD-Brain` demonstrates remarkable robustness even when subjected to larger distortion artefacts, as shown in the right columns of Fig. 11(a). Interestingly, other than `LOD-Brain`, the other methods that demonstrate significant robustness to motion artefacts are CEREBRUM and the 3D-UNet baseline. Both are trained with the same quantity of data from numerous and various different sites. Conversely, as we suggest in Fig. 4(b), methods that have not seen data from many different sites experience a significant drop in performance.

5.4. Qualitative comparisons

We here incorporate a comprehensive array of visual results obtained with different methods, which represents a clear insight into performance. In Fig. 16 we show a qualitative comparison performed on the 12 worst numerical results obtained with `LOD-Brain` (one for dataset — blue dots in Fig. 7(a)). We display FreeSurfer segmentation masks in the first row, and `LOD-Brain` in the second, with segmentation masks overlayed to the correspondent T1w. Despite the poor numerical results computed against FreeSurfer's masks, the segmentation boundaries returned by `LOD-Brain` show less errors and are much smoother than the silver GT produced by FreeSurfer.

Furthermore, in Figs. 12 and 13 we showcase the 30 MRI volumes with the highest variance in DICE scores computed on the segmentation masks. This approach aims to maximise the visual distinctions among different anatomical structures, ensuring that a visual comparison provides a clear and informative perspective on the effectiveness of our

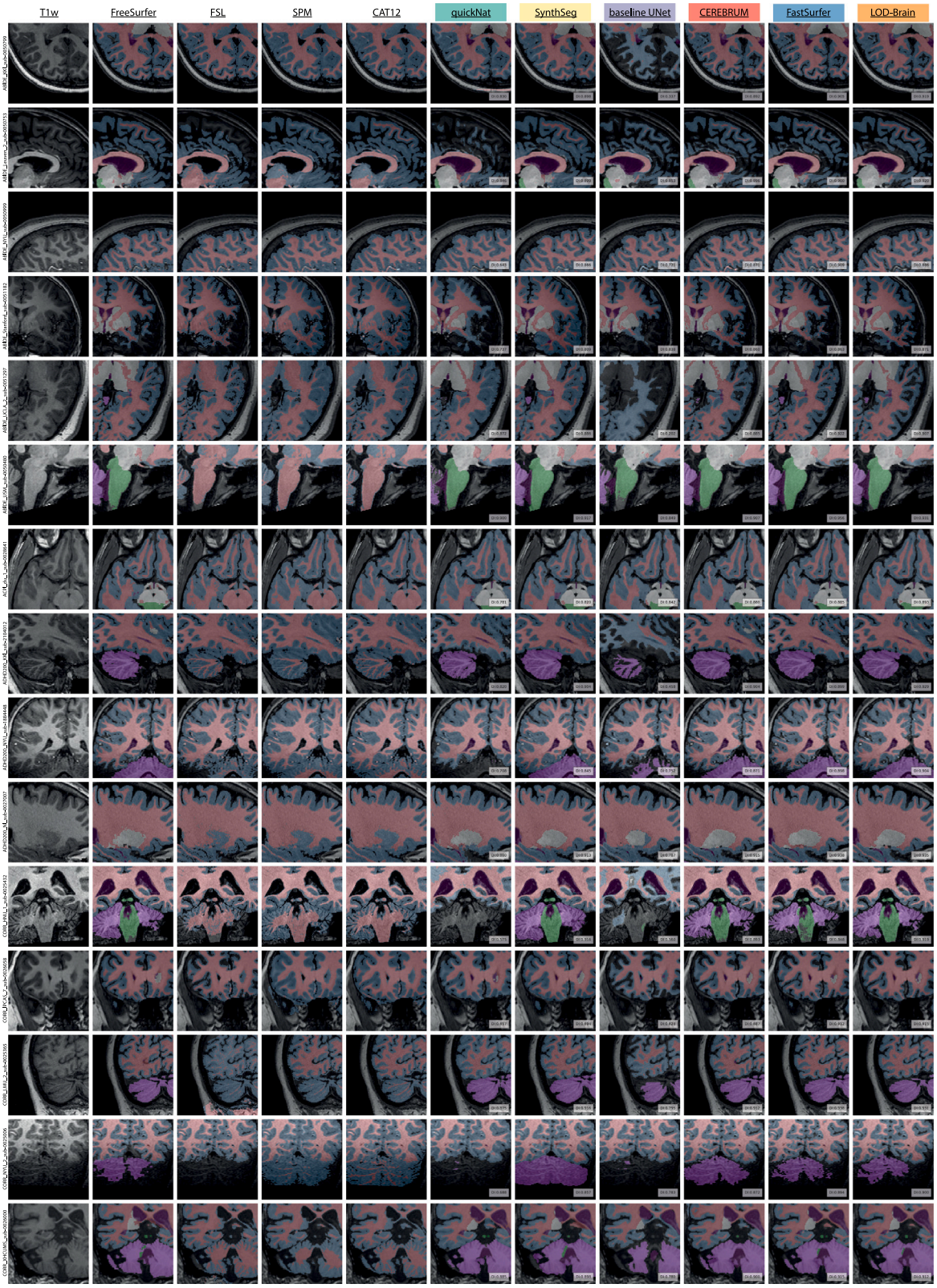


Fig. 12. Comparison on the first half of 30 volumes (vol. 1–15) with highest disagreement on segmentation masks among difference methods. Zoom in for better view.

method with respect to the selected atlas- and deep learning-based methods.

5.4.1. *Surface analysis*

We include a surface analysis obtained by all competing Deep Learning (DL) methods on both the inner grey matter surface (in Fig. 14(a))

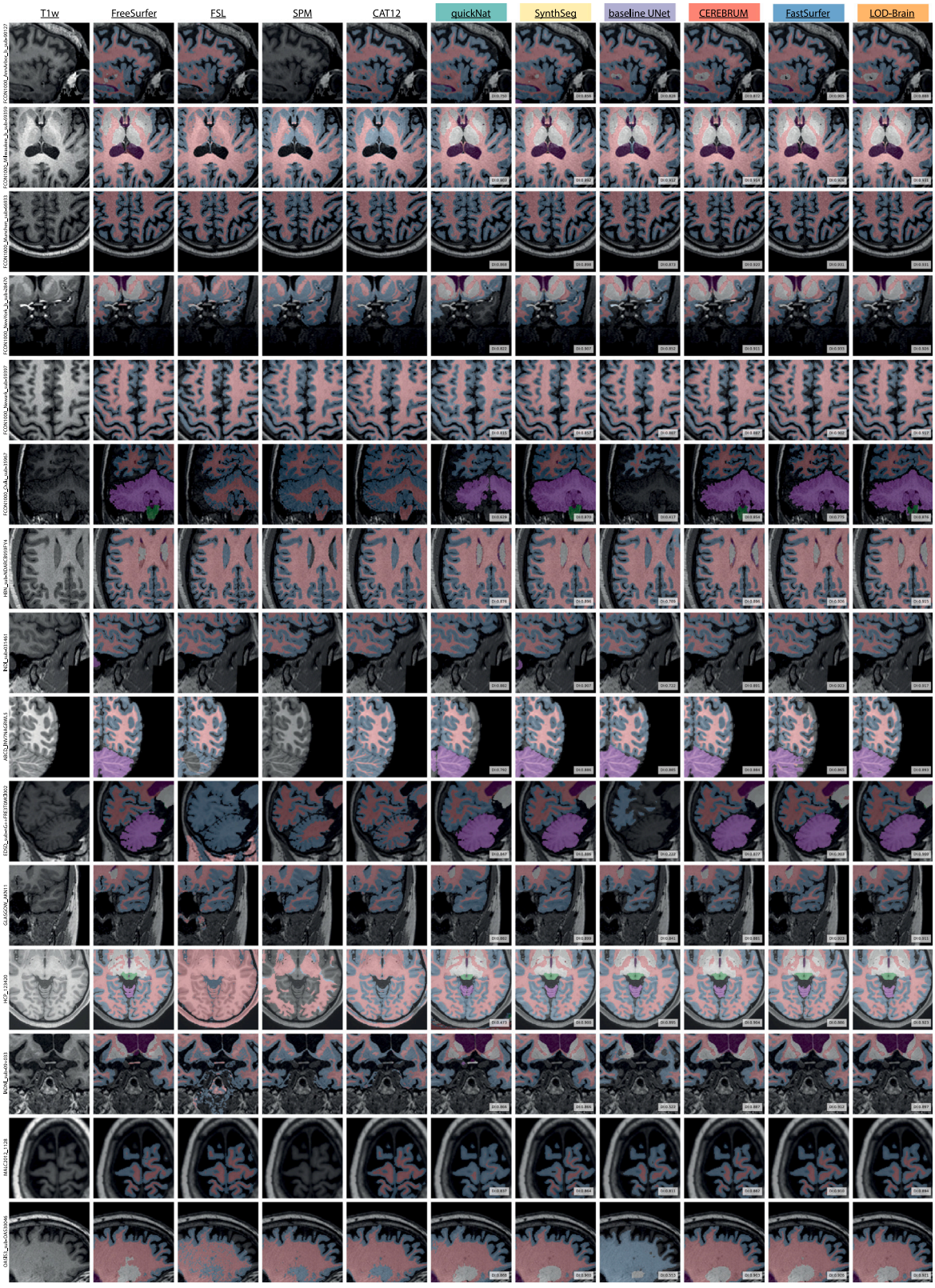
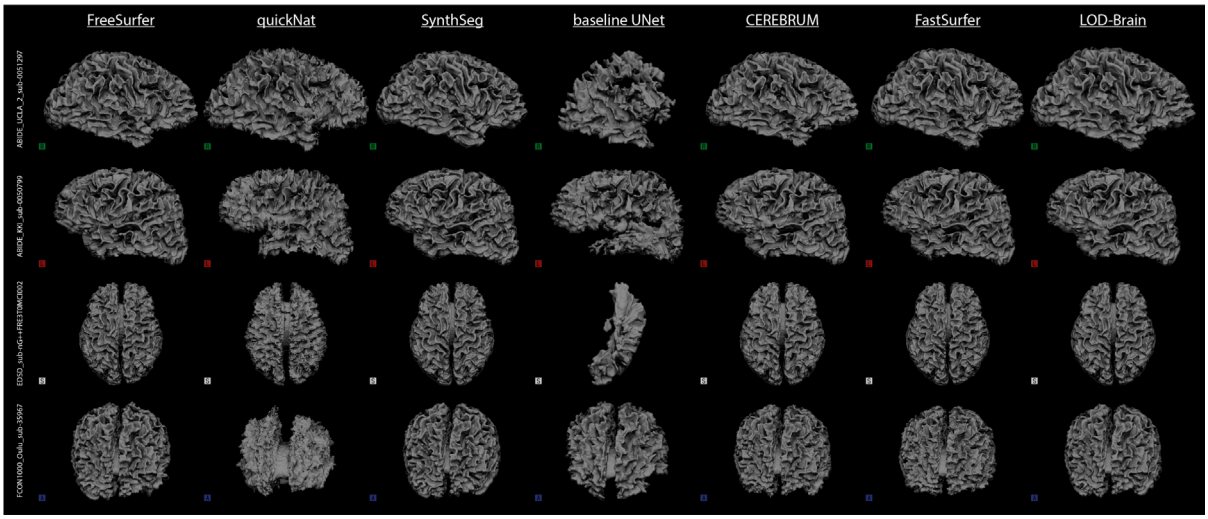
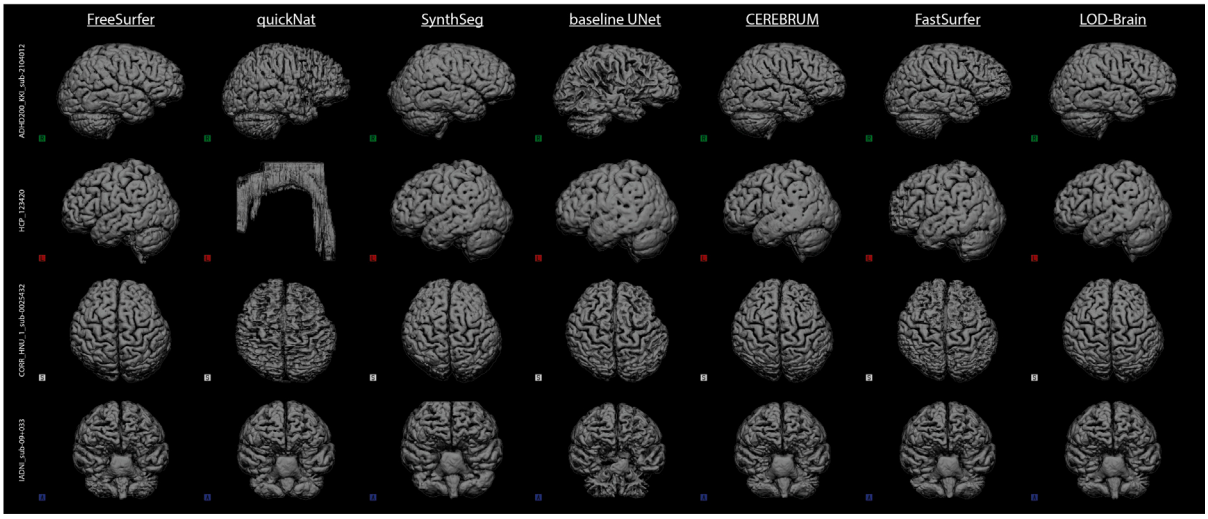


Fig. 13. Comparison on the second half of 30 volumes (vol. 16–30) with highest disagreement on segmentation masks among difference methods. Zoom in for better view.

and the outer grey matter surface (in Fig. 14(b)). The analysis is carried out on the MRI volumes with the maximum level of disagreement in segmentation masks as generated by all competing DL methods. By presenting these 3D surface representations generated with nii2mesh, we offer a more comprehensive assessment of our method's ability to capture anatomical brain structures, while allowing for a direct comparison of our approach with benchmark methods.



(a) Inner Gray Matter surfaces



(b) Outer Gray Matter surfaces

Fig. 14. (a) Inner and (b) Outer Grey Matter surface on the four volumes with the maximum DICE variance in the testing set (zoom in for better view).

5.5. Error analysis

To provide a more comprehensive analysis and investigate potential areas of underperformance, we generate brain maps for error assessment. To accomplish this, we compute the error (in native space) between the silver ground-truth generated by FreeSurfer and the results produced by `LOD-Brain` on 10 MRI volumes from each of the testing sites, resulting in a total of 719 volumes. Subsequently, we align the T1w and error volumes with the MNI atlas and plot the results in MNI space. These findings are presented in Fig. 15, where we can observe the excellent results obtained by `LOD-Brain` in the brain cortex region.

For a more detailed breakdown of the error assessment, we include additional brain maps in the Supplementary material (Fig. 1). These additional maps offer a label-specific analysis of misclassified voxels, providing a more granular understanding of `LOD-Brain`'s performance. They shed light on areas where misclassification occurs and where it excels. It is worth noting that `LOD-Brain`'s underperformance in specific brain regions, such as the basal ganglia and ventricles, is primarily attributed to the poor silver ground-truth segmentation

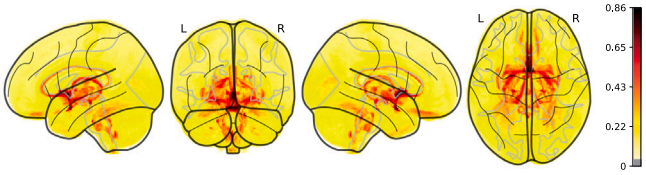


Fig. 15. Brain maps for error assessment.

in these areas, as demonstrated in another figure in the Supplementary material (Fig. 3), where we compare FreeSurfer and `LOD-Brain` segmentation masks.

6. Conclusion

In this work, we introduce `LOD-Brain`, a progressive level-of-detail network tailored for training a robust brain MRI segmentation model. Our multi-level approach is designed to achieve distinctive performance in capturing essential brain structures and accommodating site-specific and anatomical variations. The outcomes of our

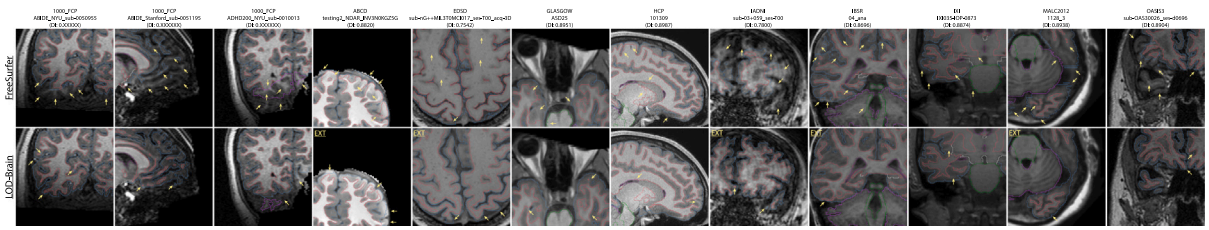


Fig. 16. We select 12 volumes with the worst numerical result (max one for dataset), and display FreeSurfer (first row) and `LOD-Brain` (second row) segmentation masks overlayed to the correspondent T1w image. Despite low numerical performance, the segmentation masks returned by `LOD-Brain` are more accurate than Freesurfer (yellow arrows indicate errors), which is indicative of low-quality ground-truth. Full 3D volumes are displayed on the project website.

research exhibit exceptional consistency across various scanners and sites, showcasing robust performance even in the face of challenging anatomical complexities. While its inherent resilience to artefacts and adaptability across different scanners and sites make it a powerful tool for researchers and clinicians, `LOD-Brain` focuses on adult brain MRI volumes only, and it may not be the most appropriate solution for segmenting the brains of children under the age of 2 years (see Supplementary material), due to the distinct anatomical differences in this population.

Conversely, one of the most valuable traits of `LOD-Brain` is its accessibility and ease of use. To ensure that our method is readily available to the broader research community, we have developed a user-friendly online demo where by simply drag and drop a nifti MRI brain T1w volume, it is possible to appreciate `LOD-Brain` capabilities without any software setup. In fact, `LOD-Brain` is designed to accept MRI volumes directly from the scanner, eliminating the need for labor-intensive pre-processing steps. To consider transparency and reproducibility, the complete python source code and instructions for `LOD-Brain` usage are available on our project page. Finally, `LOD-Brain` is a parameter-free method: it operates without the need for intricate parameter adjustments, ensuring that users can derive accurate segmentations effortless.

CRediT authorship contribution statement

Michele Svanera: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – review & editing, Visualisation. **Mattia Savardi:** Conceptualisation, Methodology, Software, Validation, Investigation, Writing – review & editing, Visualisation. **Alberto Signoroni:** Conceptualisation, Writing – review & editing, Supervision, Funding acquisition. **Sergio Benini:** Conceptualisation, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Lars Muckli:** Resources, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Lars Muckli reports financial support was provided by EU Horizon 2020 Framework Programme for Research and Innovation.

Data availability

All data used is publicly available and openly accessible.

Acknowledgements

This was supported from the EU Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3). We thank Ibrahim Hashim and Simona Fratus for assistance with the database creation and Mrs Frances Crabbe for useful feedback on manual segmentation.

We recognise the priceless contribution made by several openly available MRI datasets: OpenNeuro (https://openneuro.org/), ABCD (https://abcdstudy.org/), Open Science Framework (https://osf.io/), the Human Connectome Project (http://www.humanconnectomeproject.org/), the BrainWeb dataset (https://brainweb.bic.mni.mcgill.ca/brainweb/), the NIMH Data Archive (https://nda.nih.gov/), the Open Access Series of Imaging Studies (OASIS) (https://www.oasis-brains.org/), Mindboggle101 (https://mindboggle.info/data.html), the evaluation framework for MR Brain Image Segmentation (MRBrainS) (https://mrbrains18.isi.uu.nl/), the Amsterdam Open MRI Collection (AOMIC) (https://nilab-uva.github.io/AOMIC.github.io/), the Internet Brain Segmentation Repository (IBSR) (https://www.nitrc.org/projects/ibsr), and the great contribution provided by the International Neuroimaging Datasharing Initiative (INDI) (https://fcon_1000.projects.nitrc.org/), with many datasets, including but not limited to, the Nathan Kline Institute-Rockland Sample (NKI-RS) (http://fcon_1000.projects.nitrc.org/indi/enhanced/), the Information eXtraction from Images project (IXI) (https://brain-development.org/ixi-dataset/), the Autism Brain Imaging Data Exchange (ABIDE) (http://fcon_1000.projects.nitrc.org/indi/abide/), and the Attention Deficit Hyperactivity Disorder (ADHD) (https://fcon_1000.projects.nitrc.org/indi/adhd200/). The ABCD data repository grows and changes over time. The ABCD data used in this report came from DOI=10.15154/1528529. DOIs can be found at http://dx.doi.org/10.15154/1528529.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103090.

References

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D.L., Erickson, B.J., 2017. Deep learning for brain MRI segmentation: State of the art and future directions. J. Digital Imaging 30 (4), 449–459. http://dx.doi.org/10.1007/s10278-017-9983-4.

Bayer, J.M., Thompson, P.M., Ching, C.R., Liu, M., Chen, A., Panzenhagen, A.C., Jahanshad, N., Marquand, A., Schmaal, L., Sämann, P.G., 2022. Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. Front. Neurol. 13, 923988.

Beer, J.C., Tustison, N.J., Cook, P.A., Davatzikos, C., Sheline, Y.I., Shinohara, R.T., Linn, K.A., Initiative, A.D.N., et al., 2020. Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data. Neuroimage 220, 117129.

Bethlehem, R.A., Seidlitz, J., White, S.R., Vogel, J.W., Anderson, K.M., Adamson, C., Adler, S., Alexopoulos, G.S., Anagnostou, E., Areces-Gonzalez, A., et al., 2022. Brain charts for the human lifespan. Nature 604 (7906), 525–533.

Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E., 2021. SynthSeg: Domain randomisation for segmentation of brain MRI scans of any contrast and resolution. arXiv preprint arXiv:2107.09559.

Bontempi, D., Benini, S., Signoroni, A., Svanera, M., 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 Anal. 62.

Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., Bach Cuadra, M., 2011. A review of Atlas-based segmentation for magnetic resonance brain images. Comput. Methods Programs Biomed. 104 (3), e158–e177. http://dx.doi.org/10.1016/j.cmpb.2011.07.015.

Campello, V.M., Gkontra, P., Izquierdo, C., Martín-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al., 2021. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. IEEE Trans. Med. Imaging.

Cerri, S., Greve, D.N., Hoopes, A., Lundell, H., Siebner, H.R., Mühlau, M., Van Leemput, K., 2023. An open-source tool for longitudinal whole-brain and white matter lesion segmentation. NeuroImage: Clin. 38, 103354.

Cetin-Karayumak, S., Stegmayer, K., Walther, S., Szeszko, P.R., Crow, T., James, A., Keshavan, M., et al., 2020. Exploring the limits of ComBat method for multi-site diffusion MRI harmonization. bioRxiv.

Chen, C., Hammernik, K., Ouyang, C., Qin, C., Bai, W., Rueckert, D., 2021. Cooperative training and latent space data augmentation for robust medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. Springer, pp. 149–159.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention. MICCAI 2016, Springer, pp. 424–432.

Coupé, P., Mansencal, B., Clément, M., Giraud, R., de Senneville, B.D., Ta, V.-T., Lepetit, V., Manjon, J.V., 2020. AssemblyNet: A large ensemble of CNNs for 3D whole brain MRI segmentation. NeuroImage 219, 117026.

Delisle, P.-L., Anctil-Robitaille, B., Desrosiers, C., Lombaert, H., 2021. Realistic image normalization for multi-domain segmentation. Med. Image Anal. 74, 102191.

Dewey, B.E., Zhao, C., Reinhold, J.C., Carass, A., Fitzgerald, K.C., Sotirchos, E.S., Saidha, S., Oh, J., Pham, D.L., et al., 2019. DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. Magn. Reson. Imaging 64, 160–170.

Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ben Ayed, I., 2019. HyperDense-net: A hyper-densely connected cnn for multi-modal image segmentation. IEEE Trans. Med. Imaging 38 (5), 1116–1126. http://dx.doi.org/10.1109/TMI.2018.2878669.

Dou, Q., Liu, Q., Heng, P.A., Glocker, B., 2020. Unpaired multi-modal segmentation via knowledge distillation. IEEE Trans. Med. Imaging 39 (7), 2415–2425.

Duchesne, S., Chouinard, I., Potvin, O., Fonov, V.S., Khademi, A., Bartha, R., Bellec, P., Collins, D.L., Descoteaux, M., et al., 2019. The Canadian dementia imaging protocol: Harmonizing national cohorts. J. Magn. Reson. Imaging 49 (2), 456–465.

Esteban, O., Birman, D., Schaer, M., Koyejo, O.O., Poldrack, R.A., Gorgolewski, K.J., 2017. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. PLoS One 12 (9), e0184661.

Fischl, B., 2012. FreeSurfer. NeuroImage 62 (2), 774–781. http://dx.doi.org/10.1016/j.neuroimage.2012.01.021.

Friston, K.J., Ashburner, J., Frith, C.D., Poline, J.-B., Heather, J.D., Frackowiak, R.S., 1995. Spatial registration and normalization of images. Hum. Brain Map. 3 (3), 165–189.

Gaser, C., Dahnke, R., Thompson, P.M., Kurth, F., Luders, E., Initiative, A.D.N., 2022. CAT–A computational anatomy toolbox for the analysis of structural MRI data. pp. 2006–2022. http://dx.doi.org/10.1101/2022.06.11.495736, biorxiv.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D.U., 2022. Transformers for 3D medical image segmentation. 2022 IEEE/CVF winter conference on applications of computer vision (WACV). Waikoloa, HI, USA: IEEE.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M., 2020. FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. NeuroImage 219, 117012. http://dx.doi.org/10.1016/j.neuroimage.2020.117012.

Huo, Y., Xu, Z., Xiong, Y., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S.M., Cutting, L.E., Landman, B.A., 2019. 3D whole brain segmentation using spatially localized Atlas network tiles. NeuroImage 194, 105–119.

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

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. FSL. Neuroimage 62 (2), 782–790.

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

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

Klein, A., Ghosh, S.S., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Neto, E.C., Keshavan, A., 2017. Mindboggling morphometry of human brains. PLoS Comput. Biol. 13 (2), e1005350. http://dx.doi.org/10.1371/journal.pcbi.1005350.

Kliemann, D., Adolphs, R., Tyszka, J.M., Fischl, B., Yeo, B.T., Nair, R., Dubois, J., Paul, L.K., 2019. Intrinsic functional connectivity of the brain in adults with a single cerebral hemisphere. Cell Rep. 29 (8), 2398–2407.e4. http://dx.doi.org/10.1016/j.celrep.2019.10.067.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sanchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. http://dx.doi.org/10.1016/j.media.2017.07.005.

Liu, Q., Dou, Q., Yu, L., Heng, P.A., 2020. MS-Net: Multi-site network for improving prostate segmentation with heterogeneous MRI data. IEEE Trans. Med. Imaging 39 (9), 2713–2724. http://dx.doi.org/10.1109/TMI.2020.2974574.

Liu, M., Maiti, P., Thomopoulos, S., Zhu, A., Chai, Y., Kim, H., Jahanshad, N., 2021. Style transfer using generative adversarial networks for multi-site mri harmonization. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. Springer, pp. 313–322.

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., et al., 2015. MRBrainS challenge: Online evaluation framework for brain image segmentation in 3T MRI scans. Comput. Intell. Neurosci. 2015.

Milletari, F., Navab, N., Ahmadi, S., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision. 3DV, pp. 565–571. http://dx.doi.org/10.1109/3DV.2016.79.

Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S.-H., Reina, G.A., Foley, P., Gruzdev, A., Karkada, D., Davatzikos, C., et al., 2022. Federated learning enables big data for rare cancer boundary detection. Nat. Commun. 13 (1), 7346.

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I.M., Satterthwaite, T.D., et al., 2020. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage 208, 116450.

Reina, G.A., Panchumarthy, R., Thakur, S.P., Bastidas, A., Bakas, S., 2020. Systematic evaluation of image tiling adverse effects on deep learning semantic segmentation. Front. Neurosci. 14, 65. http://dx.doi.org/10.3389/fnins.2020.00065.

Robinson, R., Dou, Q., Coelho de Castro, D., Kamnitsas, K., de Groot, M., Summers, R.M., Rueckert, D., Glocker, B., 2020. Image-level harmonization of multi-site data using image-and-spatial transformer networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23. Springer, pp. 710–719.

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

Rundo, L., Han, C., Nagano, Y., Zhang, J., Hataya, R., Militello, C., Tangherloni, A., Nobile, M.S., Ferretti, C., Besozzi, D., et al., 2019. USE-net: Incorporating squeeze-and-excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. Neurocomputing 365, 31–43.

Shaw, R., Sudre, C.H., Varsavsky, T., Ourselin, S., Cardoso, M.J., 2020. A k-space model of movement artefacts: Application to segmentation augmentation and artefact removal. IEEE Trans. Med. Imaging 39 (9), 2881–2892.

Shin, H.-C., Tenenholtz, N.A., Rogers, J.K., Schwarz, C.G., Senjem, M.L., Gunter, J.L., Andriole, K.P., Michalski, M., 2018. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 1–11.

Styner, M.A., Charles, H.C., Park, J., Gerig, G., 2002. Multisite validation of image analysis methods: Assessing intra-and intersite variability. In: Medical Imaging 2002: Image Processing, Vol. 4684. International Society for Optics and Photonics, pp. 278–286.

Sun, Y., Gao, K., Wu, Z., Li, G., Zong, X., Lei, Z., Wei, Y., Ma, J., Yang, X., Feng, X., et al., 2021. Multi-site infant brain segmentation algorithms: The iSeg-2019 challenge. IEEE Trans. Med. Imaging 40 (5), 1363–1376.

Svanera, M., Benini, S., Bontempi, D., Muckli, L., 2021. CEREBRUM-7T: Fast and fully volumetric brain segmentation of 7 tesla MR volumes. Hum. Brain Map. 42 (17), 5563–5580.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).

Wachinger, C., Reuter, M., Klein, T., 2018. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. NeuroImage 170, 434–445. http://dx.doi.org/10.1016/j.neuroimage.2017.02.035, arXiv:1702.08192.

Wang, L., Li, G., Adeli, E., Liu, M., Wu, Z., Meng, Y., Lin, W., Shen, D., 2018a. Anatomy-guided joint tissue segmentation and topological correction for 6-month infant brain MRI with risk of autism. Hum. Brain Map. 39 (6), 2609–2623.

Wang, L., Li, G., Shi, F., Cao, X., Lian, C., Nie, D., Liu, M., Zhang, H., Li, G., Wu, Z., et al., 2018b. Volume-based analysis of 6-month-old infant brain MRI for autism biomarker identification and early diagnosis. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11. Springer, pp. 411–419.

Wiles, O., Gowal, S., Stimberg, F., Alvise-Rebuffi, S., Ktena, I., et al., 2021. A fine-grained analysis on distribution shift. arXiv preprint.

Yaakub, S.N., Heckemann, R.A., Keller, S.S., McGinnity, C.J., Weber, B., Hammers, A., 2020. On brain Atlas choice and automatic segmentation methods: A comparison of MAPER & FreeSurfer using three Atlas databases. Sci. Rep. 10 (1), 1–15.

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 15 (11), e1002683.

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

Zhou, H.-Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y., 2023. nnFormer: Volumetric medical image segmentation via a 3D transformer. IEEE Trans. Image Process..