FastSurfer - A fast and accurate deep learning based neuroimaging pipeline

Leonie Henschel [a], Sailesh Conjeti [a], Santiago Estrada [a], Kersten Diers [a], Bruce Fischl [b,c,d], Martin Reuter [a,b,c,*,1]

[a] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[b] A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston MA, USA
[c] Department of Radiology, Harvard Medical School, Boston, MA, USA
[d] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA

ABSTRACT

Traditional neuroimage analysis pipelines involve computationally intensive, time-consuming optimization steps, and thus, do not scale well to large cohort studies with thousands or tens of thousands of individuals. In this work we propose a fast and accurate deep learning based neuroimaging pipeline for the automated processing of structural human brain MRI scans, replicating FreeSurfer's anatomical segmentation including surface reconstruction and cortical parcellation. To this end, we introduce an advanced deep learning architecture capable of whole-brain segmentation into 95 classes. The network architecture incorporates local and global competition via competitive dense blocks and competitive skip pathways, as well as multi-slice information aggregation that specifically tailor network performance towards accurate segmentation of both cortical and subcortical structures. Further, we perform fast cortical surface reconstruction and thickness analysis by introducing a spectral spherical embedding and by directly mapping the cortical labels from the image to the surface. This approach provides a full FreeSurfer alternative for volumetric analysis (in under 1 min) and surface-based thickness analysis (within only around 1 h runtime). For sustainability of this approach we perform extensive validation: we assert high segmentation accuracy on several unseen datasets, measure generalizability and demonstrate increased test-retest reliability, and high sensitivity to group differences in dementia.

1. Introduction

The rapid emergence of standardized robust non-invasive imaging methods and infrastructure for big data analysis over the years has promoted the advent of a variety of large-scale neuroimaging studies. Different initiatives aim to understand the variability, development and anatomical layout of the human brain in e.g. neurodegeneration (ADNI (Mueller et al., 2005), OASIS (Marcus et al., 2007, 2010)), psychiatric diseases (LA5c (Poldrack et al., 2016)), neurodevelopmental disorders (ABIDE (Di Martino et al., 2017), MIRIAD (Malone et al., 2013)) or within populations (Rotterdam Study (Ikram et al., 2017), Human Connectome Project (Van Essen et al., 2012), UKBiobank (Sudlow et al., 2015), Rhineland Study (Breteler et al., 2014)). A core challenge within all neuroimaging studies is the need to process and analyze the continuing stream of data in a timely manner. As Magnetic Resonance

Imaging (MRI) is one versatile imaging modality and integral part of all these studies, developing efficient tools to identify clinically-relevant imaging biomarkers with MRI are in high demand. In this work we, therefore, develop a fast method for volumetric segmentation, reconstruction of cortical geometry, and morphometric estimation of brain structures including cortical thickness. It is the first work that aims at integrating an advanced deep learning method for image segmentation into a complete processing pipeline that includes cortical surface reconstruction and segmentation.

1.1. Neuroimage analysis

To date, a few well maintained neuroimage processing pipelines such as FreeSurfer (Fischl et al., 2002), BrainSuite (Shattuck and Leahy, 2002), SPM (Friston et al., 2007), ANTs (Avants et al., 2009), or FSL (Jenkinson

---

et al., 2012) are the only means available to process and evaluate the incoming flow of data. These pipelines usually employ multiple image transformation steps, some of which require careful fine-tuning of parameters such as convergence thresholds, smoothing levels, or iteration numbers. Furthermore, due to extensive numerical optimization, e.g. non-linear registration or Bayesian segmentation, these approaches are computationally expensive and suffer from long runtimes. Hence, several hours are required to process a single volume, significantly limiting scalability to large cohort studies with thousands of cases or to clinical workflows where immediate results are essential.

Supervised deep learning approaches are an attractive alternative to replace time-intensive steps within these pipelines such as whole-brain segmentation because of their faster runtimes. Fully convolutional neural networks (F–CNN), for example, are able to learn the correct feature representations in an end-to-end fashion from the image itself without requiring lengthy pre-processing steps. These methods can be effectively parallelized on graphical processing units (GPU) resulting in an enormous speed-up (seconds instead of hours). Additionally, these networks often outperform traditional approaches with respect to accuracy and have become increasingly popular for pixel or voxel-wise semantic segmentation tasks in computer vision and biomedical imaging (Long et al., 2015; Ronneberger et al., 2015; Noh et al., 2015; Badrinarayanan et al., 2017; Milletari et al., 2016; Jégou et al., 2017). In this work we propose a neuroimaging pipeline based on an advanced neural network architecture for whole-brain segmentation that induces local and global competition in the dense block and skip-connections.

1.2. Deep learning for whole-brain segmentation

The task of whole-brain segmentation in particular is challenging due to the complex 3D architecture and spatial dependency between slices, the large number of labels, the size of the scanning volumes (memory requirements), and variability across scanners and subjects. While several deep learning based approaches have been proposed for specific tasks, such as tumor segmentation (Rani and Vashisth; Dong et al., 2017; Arunachalam and Savarimuthu, 2017; Havaei et al., 2017; Amin et al., 2018; Brainless Glioma , 2018), brain lesion segmentation (Kamnitsas et al., 2017; Varghese et al., 2017; Rezaei et al., 2017; Roa-Barco et al., 2017; Chen and Konukoglu, 2018), MR image reconstruction (Jin et al., 2017; Mardani et al., 2019; Schlemper et al., 2018; Yang et al., 2018; Dedmari et al., 2018), prediction of brain related diseases and their progression (Payan and Montana, 2015; Qi and Tejedor, 2016; Hosseini-Asl et al., 2016; Lee et al. Kim) or segmentation of a smaller number of brain (sub-)structures (Zhang et al., 2015; Akkus et al., 2017; Milletari et al., 2017; Fedorov et al., 2017; Dolz et al., 2018; Thyreau et al., 2018; Chen et al., 2018; Nogovitsyn et al., 2019; Li et al., 2019; Sun et al., 2019; Ito et al., 2019) full brain segmentation into more than 25 classes has - so far - only been achieved by a few groups (de Brêbisson and Montana, 2015; Moeskops et al., 2016; Mehta et al., 2017; Wachinger et al., 2018; Roy et al., 2017, 2019; Jog et al., 2019; Huo et al., 2019; Coupé et al., 2019) - yet with the exception of (Roy et al., 2019) only with direct comparison of segmentation accuracy on a test-set, lacking extensive validations, e.g., of reliability and sensitivity to real neuroanatomical effects.

Most of these brain segmentation networks were trained on extracted 3D patches (de Brêbisson and Montana, 2015; Wachinger et al., 2018; Mehta et al., 2017) or 2D slices (Moeskops et al., 2016; Roy et al., 2017, 2019; Jog et al., 2019). Both approaches loose spatial information critical for correct classification of a given structure. Due to memory constraints of existing GPUs it is, however, currently not feasible to train a full 3D deep neural segmentation network with a large number of labels for a whole-brain MRI volume at full resolution. Even for a lower number of labels, batch size can usually not be larger than 1, which would be another detrimental factor. Segmenting downsampled images is also not an option, e.g. due to small cortical structures, that may not be sufficiently resolved at lower resolutions. Additionally, full 3D networks have

a much higher demand for training data as one MRI volume constitutes a single training case. Recently, approaches to overcome memory demands for whole-brain segmentation have been proposed, such as the combination of multiple (up to 250) patch-based 3D CNNs that process overlapping sub-volumes of the divided whole-brain input (Huo et al., 2019; Coupé et al., 2019). These segmentation tools, however, require a number of additional and potentially error-prone processing steps including image registration to a standard space and bias field removal, as well as inference for a large number of networks. Overall, most of the architectures proposed for brain segmentation so far have runtimes of around 15–60 min on the GPU (de Brêbisson and Montana, 2015; Moeskops et al., 2016; Mehta et al., 2017; Wachinger et al., 2018; Huo et al., 2019; Coupé et al., 2019) which makes them unattractive for integration into a maximally fast neuroimaging pipeline. The networks presented in (Roy et al., 2017) and (Roy et al., 2019) can produce a whole-brain segmentation in less than 1 min and are thus most relevant to our work.

The SkipDeconv-Net (SD-Net (Roy et al., 2017)) was the first whole-brain segmentation F–CNN and is based on a classic encoder-decoder architecture reminiscent of the U-net (Ronneberger et al., 2015). The SD-Net introduces a novel loss-function that addressed the inherent class imbalance problem and alleviated segmentation errors along anatomical boundaries. Subsequently, the network architecture was extended into an F–CNN called Quick segmentation of Neuroanatomy (QuickNAT (Roy et al., 2019)), which allows segmentation of a whole 3D brain volume into 27 structures. In this architecture, short-range skip connections were employed within each encoder-decoder block - these dense blocks were introduced in (Huang et al., 2017) for classification tasks. Further, QuickNAT uses three 2D F-CNNs to produce predictions for axial, coronal and sagittal slices which are combined in a final multi-view aggregation step and thus allows partial recapture of spatial information in the third dimension (e.g. 2.5D approach). Here, we compare to the different 2D architectures (with multi-view aggregation switched on for all of them for better comparability) and additionally include a modified patch-based 3D version of the UNet (Çiçek et al., 2016) which allows inference within a similar time frame. To this end, we use the self-adapting framework "no-new-Net" (nnU-Net) (Isensee et al., 2019) to fine-tune a 3D UNet architecture to the given image geometry and number of classes in our application.

Here, we propose FastSurferCNN a deep learning architecture capable of segmenting a whole brain into 95 classes in merely 1 min on the GPU (and 14 min sequential processing on the CPU). The basic architecture is inspired by QuickNAT (Roy et al., 2019). Each F–CNN has the same encoder/decoder-based architecture with skip connections (Ronneberger et al., 2015), enhanced with unpooling layers (Noh et al., 2015) and dense connections (Huang et al., 2017) within each block. The main methodological improvements of FastSurferCNN compared to QuickNAT are the introduction of competition within each block (competitive dense blocks) by replacing concatenation with maxout operations (Goodfellow et al., 2013; Estrada et al., 2018, 2019), as well as the inclusion of a wider image context within each 2D F–CNN (spatial information aggregation) in order to retain enough spatial information for the accurate segmentation of neuroanatomical structures, such as cortical gray matter regions.

Note, that voxel-based image segmentation, on its own, is limited with regard to neuroimage analysis and biomarker extraction. Especially surface-based analysis has proven pivotal for e.g. correct estimation of thickness - an issue which has so far not been addressed in comparative publications on deep learning. Existing traditional pipelines go far beyond image segmentation and provide utilities such as creation of cortical surface models, estimation of thickness, construction of fiber tracts or functional connectivity graphs, and tools for group comparison, such as registration and statistical frameworks. A major focus of this work is to fill this gap by integrating the developed deep learning framework (FastSurferCNN) into a complete, self-contained imaging pipeline called FastSurfer (Fig. 1).

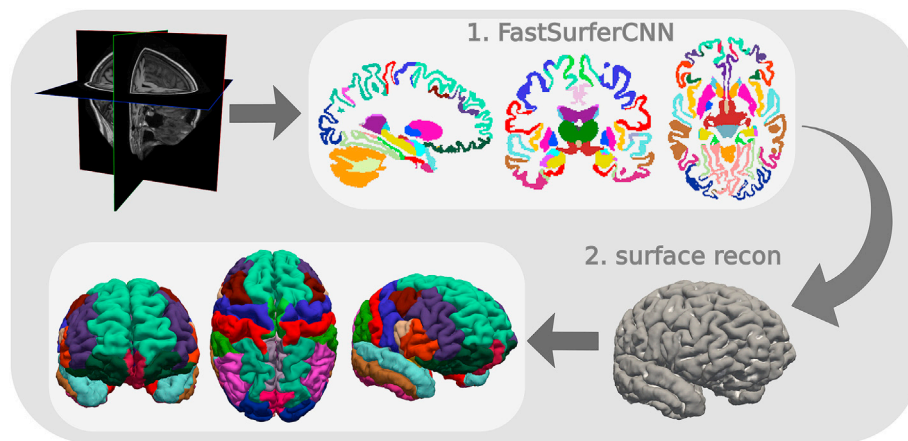Starting from the accurate 3D whole-brain segmentation, provided by

Fig. 1. FastSurfer - A fast and accurate deep learning based neuroimaging pipeline.

our deep learning framework, we perform cortical surface reconstruction and fast spherical mapping via a novel spectral approach that quickly maps the cortex using Laplace Eigenfunctions. Furthermore, we map cortical labels and include traditional point-wise and ROI thickness analysis, resulting in a full FreeSurfer alternative with approximately 60 min runtime (depending on image quality and process parallelization) of which only 1 min is attributed to the whole-brain segmentation. Hence, FastSurfer combines the speed of supervised deep learning approaches and the convenience of the broad spectrum of surfaced-based features and analysis methodologies provided by traditional neuroimaging pipelines.

We extensively validate the quality of our deep learning based neuroimaging pipeline through assessment of segmentation accuracy, generalizability to unseen datasets and acquisition parameters, test-retest reliability, and sensitivity to group level differences in imaging cohorts in a number of publicly available datasets. In fact, this is the first work within deep learning approaches with such an exhaustive validation. We demonstrate that despite being orders of magnitude *faster* than traditional approaches, FastSurfer *increases reliability and sensitivity* making it a dependable tool for future large-scale population analysis tasks. The source code of FastSurfer is available on Github: https://github.com/Deep-MI/FastSurfer.

2. Methodology

2.1. Datasets

MRI volumes from eight publicly available datasets were selected for training, testing, and for extensive validation of the FastSurfer pipeline. In brief, selected subjects from the Autism Brain Imaging Data Exchange II (ABIDE II) (Di Martino et al., 2017), the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005), the UCLA Consortium for Neuropsychiatric Phenomics LA5c Study (LA5c) (Poldrack et al., 2016), and the Open Access Series of Imaging Studies 1 and 2 (OASIS-1 (Marcus et al., 2007) and OASIS-2 (Marcus et al., 2010)) were used for network training. Further, subjects never encountered during training from ADNI and OASIS-1, the Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) (Malone et al., 2013), the Human Connectome Project (HCP) (Van Essen et al., 2012), a multi-subject, multi-model neuroimaging dataset (MMND) (Wakeman and Henson, 2015), the Traveling Human Phantom (THP) (Magnotta et al., 2012) as well as the Mindboggle-101 dataset (Klein and Tourville, 2012) were used for validation and testing of the FastSurfer pipeline. For a detailed description of these datasets as well as a usage summary see Appendix Section 5.1 and Table A.1.

All datasets were processed using FreeSurfer v6.0. FreeSurfer is an open source neuroimage analysis suite (Fischl et al., 2002; Fischl, 2012)

(http://surfer.nmr.mgh.harvard.edu/). Freesurfer morphometric procedures have been demonstrated to show good test-retest reliability across scanner manufacturers and across field strengths (Han et al., 2006; Reuter et al., 2012). In this work FreeSurfer parcellation following the "Desikan–Killiany–Tourville'' (DKT) protocol atlas (Klein and Tourville, 2012; Desikan et al., 2006) is used for training and evaluation. In order to limit the number of segmentation labels, cortical regions touching each other across the hemispheres, are lateralized while all others are combined thus reducing the total number of labels from 95 (DKT without corpus callosum segmentations which are added later) to 78 during network training. Association to the left or right hemisphere is restored in the final prediction by estimating the closest white matter centroid (left or right hemisphere) to each label cluster. A list of all segmentation labels is provided in the appendix (see Table A.2). In accordance with FreeSurfer, all MRI brain volumes are conformed to standard slice orientation and resolution (1 mm isotropic) before feeding them to the different deep learning networks. No further image processing is required afterwards (e.g. no skull stripping or intensity normalization).

2.2. FastSurfer CNN

Here, we introduce the network architecture (FastSurferCNN) for whole-brain segmentation into 95 classes (excluding background) in under 1 min on the GPU (and approximately 14 min on the CPU). FastSurferCNN is composed of three F-CNNs operating on coronal, axial and sagittal 2D slices and a final view aggregation stage. The basic architecture of all three F-CNNs follows that of (Roy et al., 2019), namely a sequence of 4 dense encoder and decoder blocks separated by a bottleneck layer as illustrated in Fig. 2. Within FastSurferCNN, we integrate two improvements - competitive dense blocks and spatial information aggregation - targeted to promote information recovery and increase network connectivity. In the following sections, each of these elements will be explained in detail.

2.2.1. Competitive dense block

With the exception of our recent work in (Estrada et al., 2018, 2019), dense connections within convolutional blocks have been implemented via concatenation of feature maps (see QuickNAT (Roy et al., 2019)) - effectively doubling the numbers of learnable parameters (Fig. 3 top) within each encoder and decoder block and thus considerably increasing memory requirements. Here, we employ competitive dense blocks in which concatenations are replaced with maxout activations (Goodfellow et al., 2013; Estrada et al., 2018, 2019). The maxout activations induce competition between feature maps and significantly reduce the number of parameters compared to the classical dense blocks, thus creating a lightweight model (Fig. 3 bottom). Instead of stacking the output of previous layers on top of each other, only the maximum value at a given
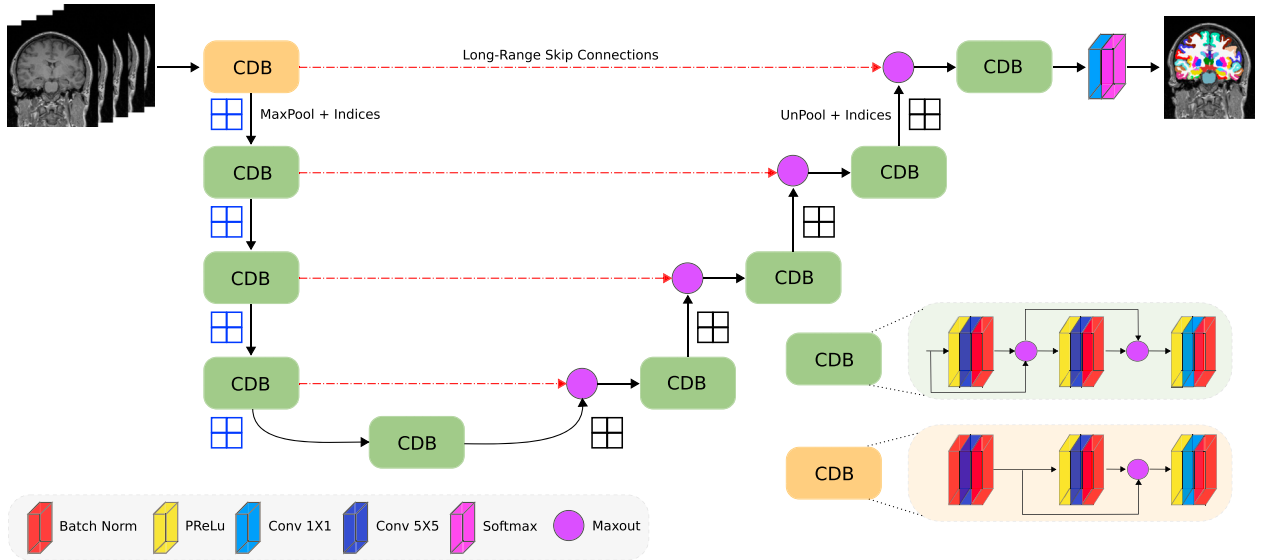
Fig. 2. FastSurfer Network Architecture. The network consists of four competitive dense blocks (CDB) in the encoder and decoder part, separated by a bottleneck layer. Each block is composed of three sequences of parametric rectified linear unit (PReLU), convolution (Conv) and batch normalization (BN) with exception of the very first encoder block. In the first block, the PReLU is replaced with a BN to normalize the raw inputs.
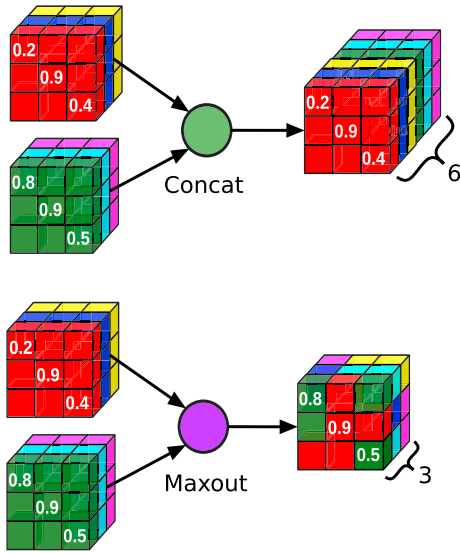

Fig. 3. The standard "concat" operation (top) appends the two incoming blocks doubling memory requirements in each step. The efficient "maxout" operation (bottom) retains the max value at each position, hence inducing local competition between the blocks while keeping memory requirements constant.

position is retained, keeping the number of input channels as well as parameters constant in each convolution layer. Assuming L inputs, denoted as $\mathbf{X} = \{\mathbf{x}^l\}_{l=1}^{L}$, with each $\mathbf{x}^l = [x_{ijk}^l]_{i,j,k=1}^{H,W,C}$, where H is height, W is width and C are number of channels for a particular feature map ($\mathbf{x}^l$), the maxout($\mathbf{X}$) output is given by:

$$\text{maxout}\left(\mathbf{X}\right) = \left[y_{ijk}\right]_{i,j,k=1}^{H,W,C} \text{ where} y_{ijk} = \max\left\{x_{ijk}^1, \cdots, x_{ijk}^L\right\}. \quad (1)$$

The difference between dense blocks and competitive dense blocks can thus be described as:

$$\mathbf{X}_l = H_3^l(\mathbf{y}_2) \quad (2)$$

$$\mathbf{y}_2 = \text{concat}\left(H_2^l(\mathbf{y}_1), \mathbf{y}_1, \mathbf{X}_{l-1}\right) \quad (3)$$

$$\underbrace{\mathbf{y}_1 = \text{concat}\left(H_1^l(\mathbf{X}_{l-1}), \mathbf{X}_{l-1}\right)}_{\text{Densely Connected Block}} \quad (4)$$

$$\mathbf{X}_l = H_3^l(\mathbf{y}_2) \quad (5)$$

$$\mathbf{y}_2 = \text{maxout}\left(H_2^l(\mathbf{y}_1), \mathbf{y}_1\right) \quad (6)$$

$$\underbrace{\mathbf{y}_1 = \text{maxout}\left(H_1^l(\mathbf{X}_{l-1}), \mathbf{X}_{l-1}\right)}_{\text{Competitive Dense Block}} \quad (7)$$

Here, H_j^l represents a composite function of three consecutive operations: parametric rectified linear unit (PReLU) followed by convolution and batch normalization (BN) with exception of the very first encoder block. In the first block, the raw inputs are passed through BN, convolution and another BN before following the previously described architecture (see Fig. 2, CDBi vs. CDB block). The sequence of operations in H_j^l guarantees normalized inputs to the maxout activation thereby improving convergence (Liao and Carneiro, 2016) and increasing the exploratory span of the created sub-networks (Liao and Carneiro, 2017) simultaneously. Furthermore, filter co-adaptation is implicitly discouraged by the short-range skip-connections within the dense blocks (Liao and Carneiro, 2017).

In addition to the competitive dense blocks we also implement competition across long-range skip connections. Instead of concatenating the unpooled information from the decoder arm with the corresponding feature maps from the encoder arm, we perform a maxout operation before further feeding the inputs to the competitive dense decoder blocks. All our competitive dense blocks are designed such that the inputs to this operation are already normalized (unpooling and skip transfer after BN; see Fig. 2).

2.2.2. Spatial information aggregation

As mentioned in the introduction, training a full 3D deep neural network for whole brain segmentation is currently not feasible for a large number of classes. However, 2D networks with single slice inputs loose information on the 3D spatial dependency between the inputs which can be crucial for correct segmentation of neuroanatomical structures. In order to retain as much information as possible, we provide the network with a larger volumetric context by passing a multi-slice input similar to

(Ding et al., 2017; Yu et al., 2018; Ghavami et al., 2018) where a sequence of neighbouring slices instead of a single 2D slice are used for network training.

Our spatial information aggregation approach consists of passing a 7-channel image by stacking the three preceding, the current, and the three succeeding slices for segmenting only the middle slice. Fundamentally, this spatial information aggregation combines the advantages of 3D patches (local neighbourhood) and 2D slices (global view). We, furthermore, analyze the impact of multi-slice images by comparing directly to single slice inputs.

2.2.3. View aggregation

In order to account for the inherent 3D geometry of the brain, one F–CNN per anatomical plane is trained and their outputs combined in a final view aggregation step. Depending on the orientation of the 2D slices, each network therefore learns the anatomical representation of the brain structures within the coronal, axial or sagittal view. The final segmentation is generated by aggregating the probability maps of each model through a weighted average. Combination of the three principal views can boost accuracy for cortical folds and subcortical structures, some of which are better represented in one of the individual planes. In addition, view aggregation acts as a regularizer to reduce erroneous predictions. As it is not possible to differentiate between the left and right hemispheres in the sagittal view, we merge lateral labels, effectively reducing the number of classes from 78 to 50 in the sagittal network. The probability maps of these lateralized classes are finally restored by copying the softmax output of the combined label to both left and right hemispheres. To account for this remapping step, the weight with which the sagittal predictions influence the final segmentation is reduced by one half compared to the other two views.

2.2.4. Model learning

Training Dataset: 140 representative subjects from ABIDE-II, ADNI, LA5C, and OASIS (see Section 5.1) were selected for training the F–CNN models and 20 subjects from MIRIAD were used for validation. Empty slices were filtered from the volumes, leaving on average 145 single view planes per subject and a total training size of above 20 k images per network. In addition, we use data augmentation (random translation of maximally 16 mm) to artificially increase the training set size further.

The training set is balanced with regard to gender, age, diagnosis, and spans various other parameters (i.e. scanners, field strength, and acquisition parameters); the distribution of the subset is presented in Table 1. Sufficient anatomical and acquisition variety in training images can be expected to improve network robustness, generalizability, and ultimately segmentation accuracy on most unseen scans without the need to fine-tune model weights. We will analyze generalizability to unseen datasets below.

Table 1

Characteristics of the participants (n = 160) showing mean ($\pm$ standard deviation) for continuous and counts (PCT) for categorical variables.

Dataset	Subjects	Age ($\pm$ SD)	Women, n(%)	Controls, n(%)
LA5c (Poldrack et al., 2016)	20	35 ($\pm$ 6.25)	10 (50)	10 (50)
MIRIAD (Malone et al., 2013)	20	68.85 ($\pm$ 5.60)	10 (50)	10 (50)
OASIS-1 (Marcus et al., 2007)	40	37.48 ($\pm$ 13.54)	21 (52.5)	40 (100)
OASIS-2 (Marcus et al., 2010)	20	77.70 ($\pm$ 7.23)	11 (55)	10 (50)
ABIDE II (Di Martino et al., 2017)	20	25.94 ($\pm$ 4.89)	10 (50)	10 (50)
ADNI (Mueller et al., 2005)	40	74.65 ($\pm$ 6.71)	20 (50)	20 (50)
Total	160	53.97($\pm$ 22.19)	82 (51.25)	100 (62.5)

F–CNN Implementation: Independent models for coronal, axial, and sagittal plane are implemented in PyTorch (Paszke et al., 2017) and trained for 30 epochs using two NVIDIA Titan Xp GPU with 12 GB RAM and the following parameters: batch size of 16, constant weight decay of 10^{-04}, and an initial learning rate of 0.01 decreased by 95% every 5 epochs. The networks are trained with Adam optimizer (Kingma and Ba, 2015) and a composite loss function of median frequency balanced logistic loss and Dice loss (Roy et al., 2017). This loss function encourages correct segmentation of tissue boundaries and counters class imbalances by up-weighting less frequent classes.

2.3. FastSurfer pipeline

Based on FreeSurfer methods and novel contributions, we also introduce a surface processing pipeline, that integrates our neural network architecture at its core to provide FreeSurfer volume and surface results, including cortical surfaces, thickness maps, and summary statistics in cortical regions following the DKT protocol atlas (Klein and Tourville, 2012; Desikan et al., 2006).

Traditionally, surfaces are generated via a pipeline of several time-consuming processing steps: First, based on a white matter segmentation, which is patched to remove holes and ensure connectivity, initial surface triangle meshes are created for each hemisphere (Dale et al., 1999, Fischl et al., 2001). Meshes are smoothed, mapped to a sphere to localize topological defects (i.e. holes or handles should not exist as each hemisphere should be topologically equivalent to the sphere) (Fischl et al., 1999a; Ségonne et al., 2007). Once all defects are fixed, surface placement along the white matter is fine-tuned and a second expanded surface (pial surface) is placed at the outer gray matter (GM) boundary, also providing thickness estimates at every point on the cortex (Dale et al., 1999; Fischl and Dale, 2000). Then, surfaces are carefully mapped to the sphere a second time (minimizing metric distortions), registered to a spherical atlas (Fischl et al., 1999b), and segmented into cortical parcellations (DKT atlas) (Desikan et al., 2006; Fischl et al., 2004; Klein and Tourville, 2012).

Based on the DKT volume segmentation available from the FastSurferCNN we modify the above FreeSurfer pipeline to yield surface results of FreeSurfer (including thickness and cortical ROI measures). A significant speed-up compared to FreeSurfer can be achieved by omitting several steps that have become obsolete, such as skull stripping and non-linear atlas registration, given that a high-quality full brain segmentation has already been achieved. Furthermore, we innovate some traditional approaches by developing novel modules based on spectral mesh processing. Specifically:

1. We use the full DKT brain volume segmentation to create a brainmask by closure, i.e. dilation and erosion, of the labels (including the ventricle label). This mask covers all labeled areas. Cortical regions are padded by one voxel layer to allow the pial surface to find its final position in some partial volume voxel between GM and CSF. Exceptions are the lateral orbital frontal and pars orbitalis to avoid capture of the optic nerve.
2. We retrospectively construct a quick bias field corrected brain image and a linear Talairach registration as these results are needed later for some relevant statistics (e.g. intracranial vault volume for head size estimation (Buckner et al., 2004)). Here we follow FreeSurfer, except that we can initialize the NU correct (Sled et al., 1998) with the already existing brainmask.
3. We generate initial surfaces by using a marching cube (Lorensen and Cline, 1987) algorithm rather than the traditional approach (Dale et al., 1999) aiming at higher mesh quality at a slightly reduced number of vertices.
4. We develop a fast mapping to the sphere using the Eigenfunctions of the Laplace operator to perform a spectral embedding of the original white matter surfaces quickly (for the topology fixer). Precisely, we solve the Laplace-Beltrami Eigenvalue problem $\Delta f = -\lambda f$ (Reuter

et al., 2006; Reuter, 2010) on the original cortical surface mesh to obtain the first three non-constant Eigenfunctions with smallest Eigenvalues. After correcting sign flips and swaps, these functions parametrize the surface smoothly in anterior-posterior, superior-inferior and lateral-medial directions. The spherical map can then be quickly obtained by projecting the 3D spectral embedding to the sphere, i.e. by scaling the 3D Eigenfunction vector to unit length for each vertex.

5. After topology fixing and GM surface creation, we map the DKT GM segmentations from the image onto the surface and compute surface ROI statistics, such as mean thickness and curvature averages per region - mimicking FreeSurfer's surface segmentation pipeline without requiring the non-linear spherical atlas registration and segmentation. Spherical atlas registration can, however, be included if cross-subject correspondence is required, e.g. for local surface-based thickness analysis.

Overall, this yields a fast alternative to the FreeSurfer pipeline. We will evaluate the speed-up, reliability, and sensitivity of the full Fast-Surfer pipeline below.

2.4. Statistics

We will thoroughly validate the novel FastSurfer pipeline in terms of accuracy, generalizability, reliability and sensitivity using, Dice overlap, intraclass correlation and group analyses on volume and thickness ROI's, as well as thickness maps. In the following sections we explain these statistical methods in detail.

2.4.1. Dice similarity coefficient

The Dice similarity coefficient (DSC) is a metric to evaluate the segmentation performance of the deep learning networks and can mathematically be expressed as

$$DSC(\mathbf{G}, \mathbf{P}) = \frac{2 \times \mathbf{G} \cap \mathbf{P}}{\mathbf{G} + \mathbf{P}} \tag{8}$$

with binary label maps of ground truth $\mathbf{G}$ and prediction $\mathbf{P}$ (pixels of the given class indicated with 1, all others with 0). Here, the DSC is used two fold: first, to directly compare the performance of different network architectures against each other, and second, to estimate similarity of the predictions achieved with FastSurferCNN and FreeSurfer v6.0 for a number of previously unseen datasets (generalizability). The DSC will be calculated separately for each cortical and subcortical structure. Note, that neural networks tend to smooth results, e.g. removing segmentation noise such as incorrect protrusions that are encountered in only a few training images and usually in random locations. While this kind of smoothing can improve segmentation accuracy, it can decrease the DSC which is partially affected by noise in the ground truth data. This is one reason why it is essential to perform additional validations (such as reliability or sensitivity analysis).

2.4.2. Average Hausdorff distance

The average Hausdorff distance (AVG HD) is a metric for measuring the similarity between two sets of points and is often used to evaluate the quality of segmentation boundaries. It is defined as

$$AVG\,HD(\mathbf{G}, \mathbf{P}) = \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} min_{p \in \mathbf{P}} d(g,p) + \frac{1}{|\mathbf{P}|} \sum_{p \in \mathbf{P}} min_{g \in \mathbf{G}} d(p,g) \tag{9}$$

with $|\mathbf{G}|$ and $|\mathbf{P}|$ representing the number of voxels in the binary label maps of ground truth $\mathbf{G}$ and prediction $\mathbf{P}$, respectively. In contrast to the DSC, a smaller AVG HD indicates a better capture of the segmentation boundaries with a value of zero being the minimum (perfect match).

Here, we use the AVG HD to evaluate the segmentation performance of the different network architectures against each other.

2.4.3. Intraclass correlation coefficient

The intraclass correlation coefficient (ICC) is a widely used metric to assess both, the degree of correlation and agreement between measurements. Thus, it is an ideal metric to judge the reliability of a given method. The ICC ranges from 0 to 1, with values close to 1 representing high reliability. Here, we use the degree of absolute agreement among measurements also known as criterion-referenced reliability (McGraw and Wong, 1996) to compare the FastSurfer pipeline (deep learning segmentation + post-processing) to FreeSurfer. To this end, we calculate the agreement between cortical thickness and subcortical volumes in consecutive scans using the OASIS1 test-retest set. Prior to ICC calculations, volume and thickness estimates of the subcortical and cortical structures are extracted from the segmented brains. After averaging across hemispheres, the ICC as well as the upper and lower bound with $\alpha = 0.05$ level of significance are calculated for each region using the method described in (McGraw and Wong, 1996). Additionally, cortical thickness maps are mapped to a common space (*fsaverage*) and smoothed at 15 FWHM before calculating the ICC separately for each hemisphere. The resulting overlay maps are visualized on the semi-inflated *fsaverage* surfaces.

2.4.4. Group analysis

A segmentation method can potentially reach high ICC while being insensitive to actual effects in the data. Therefore, it is important to validate the sensitivity of a given method with respect to its capability to detect known significant variations in brain morphology between diagnostic groups (group separability). We, therefore, fit identical linear models to FreeSurfer and to FastSurfer results each, explaining thickness or volume (dependent variable) by diagnosis controlling for age, sex, and (only for volume analysis) head size, using FreeSurfer's *mri_glmfit*. The p-values of the diagnosis effect are monotonically connected to the absolute value of the t-statistic (effect size divided by standard error) which in turn is a scaled version of Cohen's *d*, where the scaling factor depends on sample size. Given that here both methods operate on the same input images (and thus same sample sizes), a direct comparison of p-values across processing methods is meaningful, where a smaller p-value indicates a better group separability.

These analyses are performed for volume of subcortical structures and average thickness of cortical regions of interest (ROIs), as well as for vertex-wise cortical thickness maps on both hemispheres. ROI measures are reported for structures that reach $p < 10^{-5}$ and are averaged across hemispheres for the sake of brevity. Cortical thickness is calculated as the minimal distance between the white matter and pial surface (Fischl and Dale, 2000). Prior to statistical analysis, thickness estimates of each subject were mapped to a common space (*fsaverage*) and smoothed at 15 FWHM. Significance cut-off is set to p < 0.05 without correction for multiple comparison as we are only interested in the relative differences between the two methods (FastSurfer and FreeSurfer). The p-value maps for each hemisphere are displayed on the semi-inflated *fsaverage* surfaces.

3. Results

3.1. Accuracy

In this section we evaluate segmentation performance with respect to FreeSurfer as a reference and with respect to manual labels. The comparison with FreeSurfer is relevant, given that the main goal of this work is to replicate FreeSurfer at faster processing speed without sacrificing quality. Furthermore, the comparison with manual labels is important to validate accuracy and rank the methods with respect to a manual reference standard.

3.1.1. Comparison to FreeSurfer

To evaluate segmentation accuracy, we first compute the DSC and AVG HD of each candidate method with FreeSurfer labels on five testsets (ADNI, OASIS1, HCP, MIRIAD, and THP). Here, scans from subjects used for network training and validation (i.e. 40 subjects each from OASIS and ADNI, 20 from MIRIAD) are excluded. No subjects from HCP and THP have been included in the training set at any point. Five subjects from OASIS1 were further excluded from the testset due to heavy white matter lesion load and resulting downstream topological surface defects in FreeSurfer and FastSurfer.

We benchmark our proposed network against traditional whole-brain segmentation F-CNNs namely SDNet (Roy et al., 2017) and QuickNAT (Roy et al., 2019), as well as a patch-based 3D UNet. Additionally, we incrementally test the importance of our network modifications. First, we evaluate the effects of competition within the dense blocks and across the long-range skip connections (CDB, see Section 2.2.1). For this we compare to QuickNAT which inherently uses only vanilla dense blocks within its architecture. Second, we increase the information input to the network by passing the stacked 7-channel image to the network (spatial information aggregation (SPI), see Section 2.2.2). Both architecture changes together comprise our final proposed FastSurferCNN. To permit a fair comparison, all benchmark networks were trained on the same data and - with exception of the 3D UNet - follow the same architectural design of 4 encoder and decoder blocks separated by a bottleneck block. Each block contains the same convolutional layer architecture as

illustrated in Fig. 2. The baseline architectures were further suitably adopted by modifying the final classification layer to predict 78 classes as the original implementations do not target cortical parcellations and hence comprise a much lower number of output labels (27 for QuickNAT (Roy et al., 2019) and 26 for SDNet (Roy et al., 2017)). Furthermore, all comparative 2D models were implemented with the above-mentioned view aggregation (see Section 2.2.3). Care was taken to confirm that the adaptations are acceptable to the first author of the original papers. The 3D UNet was optimized such that it allows segmentation at 1 mm isotropic resolution with a patch size of $128 \times 128 \times 112$, a batch size of 2, and 32 feature maps in the highest layer using a self-adapting framework (Isensee et al., 2019).

In Fig. 4, we report the DSC (left side) and AVG HD (right side) against 649 previously unseen scans from ADNI (180 subjects), OASIS1 (370 subjects), HCP (45 subjects), MIRIAD (49 subjects) and THP (5 subjects). We calculate the average DSC and AVG HD on 33 subcortical structures and 62 cortical regions (31 per hemisphere after remapping). For a complete list of structures the reader is referred to Appendix Table A.2. The AVG HD should be as small as possible (optimal value: 0). An optimal segmentation in terms of DSC should be close to 100.

In comparison to the 2.5D network approaches, the 3D UNet has the overall lowest DSC on the subcortical structures (average DSC of 87.27 - see Fig. 4 left, gray bars), and is comparable only to the SDNet on the cortical structures (average DSC of 83.90, a 0.4% increase compared to SDNet, light green bars). However, it has by far the highest AVG HD (see
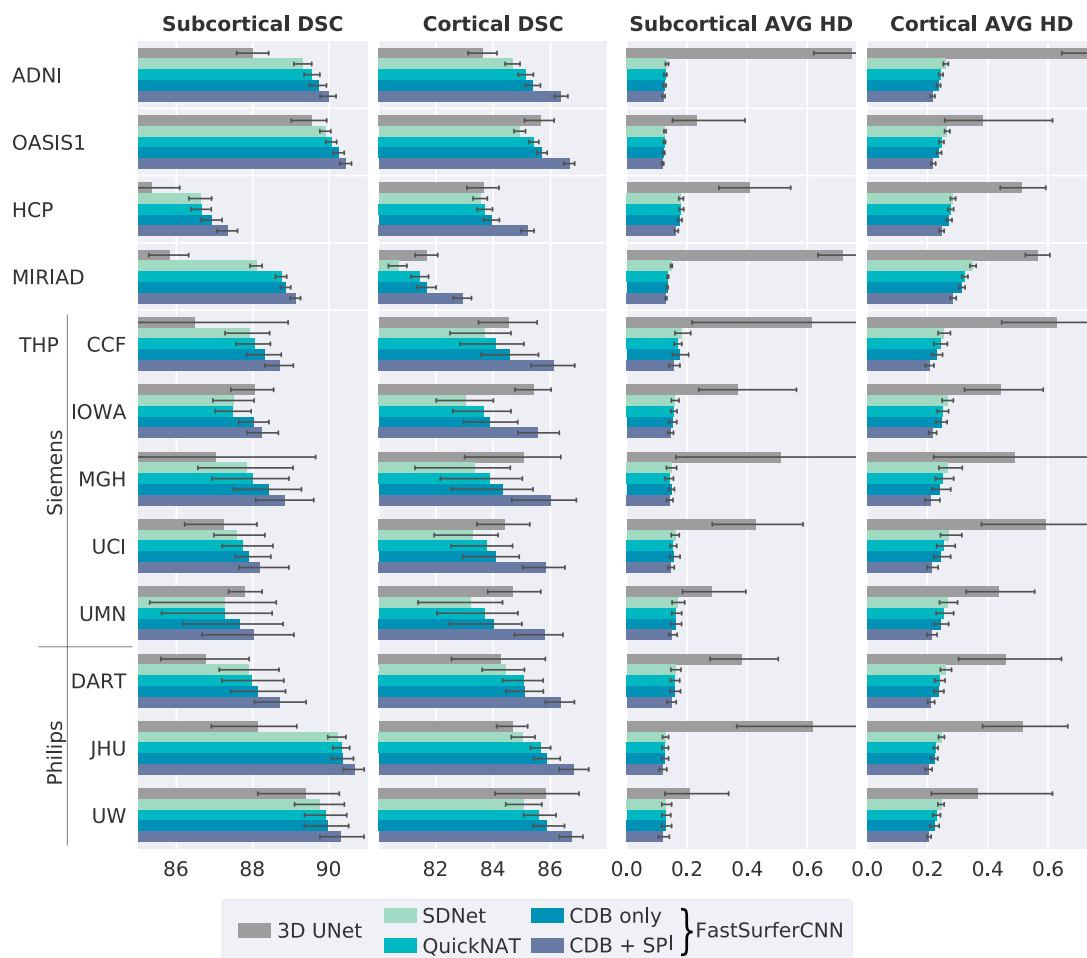

Fig. 4. Dice similarity coefficient (DSC, left, larger = better) and average Hausdorff distance (AVG HD, right, smaller = better) comparison of baselines and the proposed FastSurferCNN on four different datasets (mean ± standard deviation). Network modifications (i) competitive dense blocks (CDB) and (ii) spatial information aggregation (SPI) are incrementally tested. The final FastSurferCNN (dark blue, CDB + SPI) outperforms all other models on both, subcortical and cortical structures.

Fig. 4 right, gray bars) for both subcortical and cortical structures (0.5063 and 0.5357, respectively). The errorbars for the 3D UNet were cropped at 0.76 in the AVG HD plot.

Each successive network modification within the 2.5D network results in an increase of the DSC (see left part of Fig. 4) and a decrease of the AVG HD (see right part of Fig. 4) for all five datasets. Introduction of competition within the network (blue, CDB) already outperforms QuickNAT (green) and SDNet (light green) with an up to 0.3% improvement of the DSC and 5% of the AVG HD for both, subcortical and cortical structures. On average, competition increases the DSC to 88.74 and 84.55 for the subcortical and cortical structures, respectively. The AVG HD is decreased to 0.247 on the cortex and 0.1486 on the subcortex. Note, that this improvement is achieved while simultaneously reducing the number of trainable parameters by one half (from approx. $3.6*10^6$ to $1.8*10^6$)! The final FastSurferCNN (CDB + SPI, dark blue) further increases segmentation accuracy on average by 0.6% (DSC) and 5.7% (AVG HD) on the subcortical and 1.9% (DSC) or 12.7% (AVG HD) on the cortical structures compared to QuickNAT (final DSC of 89.08 and 85.88; AVG HD of 0.1400 and 0.2222). Therefore, increasing the local information content provided to the network via the spatial information aggregation (SPI) is particularly useful for recognizing cortical folding patterns. The same trend can be observed when analyzing the worst instead of the average DSC (data not shown). Statistical testing further confirmed a highly significant increase in DSC and decrease in AVG HD for both improvements (competition and information aggregation) compared to QuickNAT (Wilcoxon signed-rank test, $p < 10^{-20}$ after Bonferroni correction for multiple testing).

FastSurferCNN also outperforms all other models on the challenging THP dataset. This data source contains scans from eight different sites and scanning conditions with strong variations in data quality (e.g. motion artifacts). FastSurferCNN, however, maintains high accuracy for all eight sites (average DSC 89.00, AVG HD: 0.1426 (QuickNAT: 88.37 and 0.1514, respectively) for subcortical and DSC: 86.16, AVG HD: 0.2136 (QuickNAT: 84.43 and 0.2465, respectively) for cortical structures). Interestingly, MRI data acquired on Philips Scanner (DART, JHU and UW) are segmented quite accurately even though the training set predominantly included Siemens scans. Additionally, the AVG HD and DSC are more stable across imaging sites for FastSurferCNN compared to QuickNAT on the cortical structures (40% less difference in DSC and 25% less difference in AVG HD between best (JHU) and worst (IOWA) site). Notably, the algorithm also generalizes well to defaced and downsampled input images (HCP dataset) even though such examples are absent from the training set (DSC of 85.21 for the cortical and 87.36 for the subcortical structures, AVG HD of 0.2474 and 0.1643, respectively).

3.1.2. Comparison to manual reference

Here we compare segmentation performance of the networks to a manual standard. We compute the DSC and AVG HD on the Mindboggle-101 dataset (78 subjects) with manually labeled cortical as well as manual subcortical segmentations on 20 of these subjects.

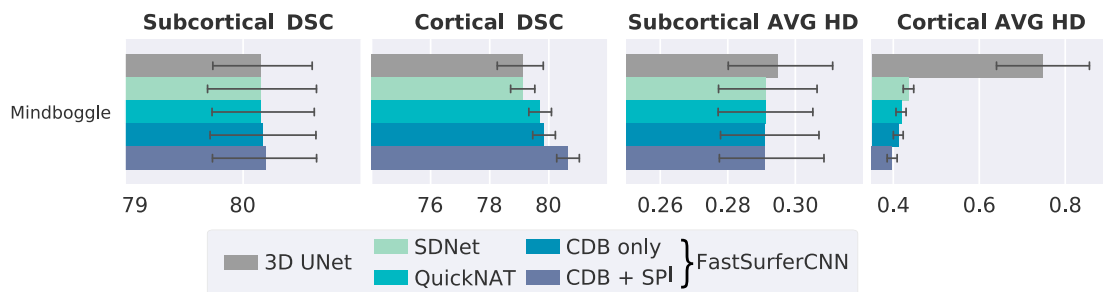In Fig. 5, we report the DSC (left side) and AVG HD (right side) for the

subcortical and cortical structures separately. The DSC on the subcortical structures are very similar for all five networks: between a DSC of 80.15 for 3D UNet, SDNet and QuickNAT and 80.19 for the final FastSurferCNN (CDB + SPI). On the cortical structures the difference is more pronounced, with the final FastSurferCNN (CDB + SPI) reaching a value of 80.65 (1% increase compared to CDB (blue) and QuickNAT (green), 2% compared to SDNet (light green) and 3D UNet (gray)). The same trend is observed for the AVG HD (right plot). Here, introduction of competition (CDB) decreases the distance between segmentation and the manual labels by 0.025% on the subcortical and 1.56% on the cortical structures (CDB (blue) versus QuickNAT (green)). The final FastSurferCNN (CDB + SPI, dark blue) reaches a value of 0.2909 for the subcortical (0.085% improvement compared to SDNet (light green) and QuickNAT (green)) and 0.3973 on the cortical structures (10% and 5% compared to SDNet and QuickNAT, respectively). The 3D UNet is again the lowest performing out of all the five networks with an AVG HD of 0.3443 for the subcortical and 0.7472 for the cortical structures, indicating large inaccuracies in especially the cortical regions.

The improved segmentation performance of FastSurferCNN on the manual cortical labels with respect to DSC and AVG HD was further confirmed by statistical testing (Wilcoxon signed-rank test, $p < 10^{-10}$ after Bonferroni correction for multiple testing).

Inherently, cortical regions are usually defined on the cortical surface based on folding patterns and curvature, e.g. to determine boundaries inside the sulcii. Thus, assessment of cortical segmentation quality may not be optimal in a volume based comparison. Therefore, we also calculate the surface-based DSC for Mindboggle-101 (manual surface labels). To this end, the volumetric segmentation of FastSurferCNN are
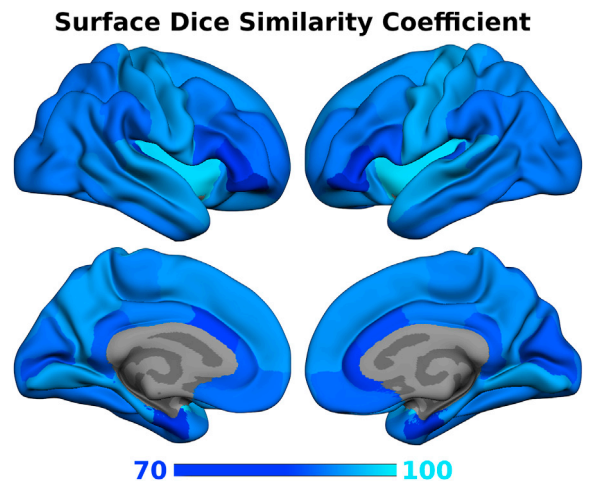
Surface Dice Similarity Coefficient

Fig. 6. Surface Dice similarity coefficient per region of the proposed FastSurferCNN for the 31 cortical parcels on manual surface labels from Mindboggle-101. FastSurferCNN achieves accurate results across structures with an average DSC of above 86.

Fig. 5. Dice similarity coefficient (DSC, left, larger = better) and average Hausdorff distance (AVG HD, right, smaller = better) comparison across networks with respect to a manual reference (Mindboggle-101). FastSurferCNN outperforms all other models on both subcortical and cortical structures.

projected onto the Mindboggle-101 surfaces where the area-related DSC of the mapped regions are then directly calculated. Overall, a high average surface DSC of 86.35 on the right and 86.97 on the left hemisphere is reached. Further, no structure has a DSC below 72.3 (see Fig. 6). The good performance of the surface analysis corroborates the volume based DSC comparisons and underlines the high segmentation quality achieved with FastSurferCNN.

3.2. Generalizability

High generalizability will ensure that the proposed method can be applied across different sites, vendors, field strengths, and for large multi-center studies. Fig. 4 indicates that networks generalize well across these parameters and respective image qualities, as the DSC remains quite stable. For example, the HCP dataset consists of 0.7 mm isotropic images, downsampled to 1 mm and de-faced, which were never encountered during training. MIRIAD is a IR-FSPGR sequence on a GE scanner. Furthermore, in the THP dataset DSC scores vary only around 1 or 2% across the 8 sites spanning Siemens and Philips scanners. The five subjects of THP, however, might not be representative, which is why in this section we quantify generalizability by computing the agreement of FastSurferCNN with FreeSurfer across different scanner types (Siemens, Philips and GE) as well as disease states (CN, MCI and AD patients) in a larger dataset. For this purpose we employ an independent testset consisting of 180 scans from ADNI balanced with regard to vendor, disease group, gender and age.

In the upper part of Fig. 7 the DSC across different disease states is shown. Scans from subjects with later-stage dementia are more difficult to segment, potentially due to increased motion and lower gray-white contrast. Specifically, the increase in ventricle volume, subsequent shrinkage of GM and/or increased white matter lesion load can have a

profound effect and are frequently difficult to segment with traditional neuroimaging pipelines such as FreeSurfer. A small decrease in segmentation DSC (i.e. increased deviation from FreeSurfer) can be observed with disease progression (CN to AD) in our proposed method (cortical structures: DSC decreases from 86.8 to 85.8, subcortical structures: DSC decreases from 90.3 to 89.6). However, with FastSurferCNN DSC scores do not decrease more than 1.11% between diseased (MCI, AD) and control state (CN) for cortical and 0.77% for subcortical structures. A similar trend is observed with regard to generalizability across vendors (lower part of Fig. 7). Even though FastSurferCNN was predominantly trained on MRI scans acquired on Siemens scanners, the segmentations of Philips and GE scans are comparable to FreeSurfer with only a minor decrease in DSC of 1.6% for the cortical structures (see Fig. 7). Here, GE and Philips scans can be segmented similarly well (DSC of 85.50 and 85.57, respectively). Segmentation of subcortical regions is equally consistent on Siemens and GE scans (DSC of 90.30 and 90.23, respectively). Philips scans show more variation with a difference similar to the one observed for the cortical structures (1.6% difference compared to Siemens; DSC of 88.85). These findings in combination with the high DSC scores achieved on MIRIAD (1.5 T GE Scanner) and HCP (downsampled and defaced 0.7 mm 3 T Siemens), see Fig. 4, show that FastSurferCNN is capable of generalizing stably across disease stages, field strength, vendors and pre-processing.

While very few GE scans were used in training (only 3 cases of ADNI), still the network has at least seen some images spanning the above parameters which might explain its good generalizability. All datasets so far acquire some kind of MPRAGE sequence, except for MIRIAD with IR-FSPGR (note, MIRIAD was not used in training, only for validation). We now test generalizability across sequences to an unseen MEF sequence (MMND dataset, see Section 5.1). This datasets provides 16 subjects each with MEF and MPRAGE scans. We first confirm, see Fig. 8 (orange column), that FastSurferCNN obtains a high DSC in comparison to FreeSurfer on the MPRAGE images, again corroborating good generalizability to Philips scanners. A 6.7% point drop in DSC can be observed when comparing both segmentation methods on the MEF images (red
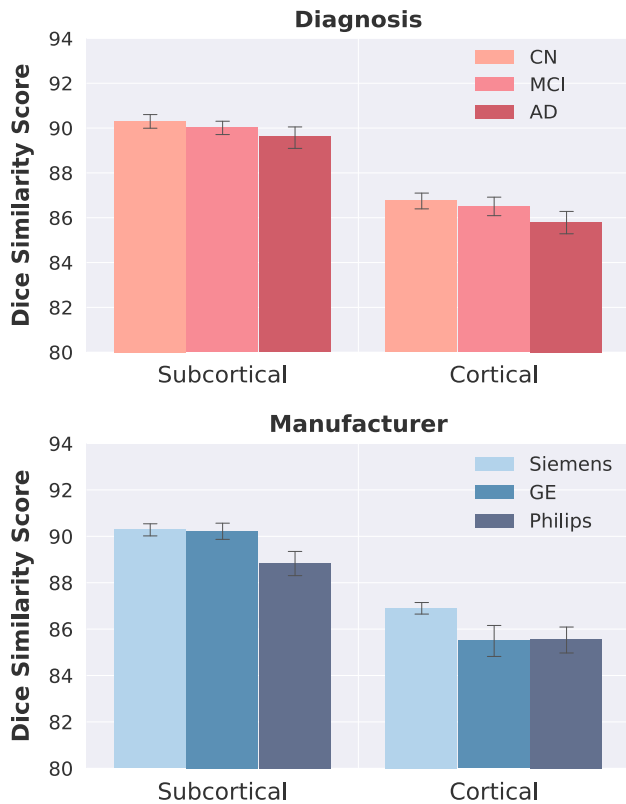

Fig. 7. Comparison of the DSC (mean ± standard deviation) across neurodegenerative states (cognitive normal (CN), mild cognitive impaired (MCI), demented (AD); top) and vendor (GE, Philips, Siemens; bottom). FastSurfer achieves high accuracy and low variability across all of them.
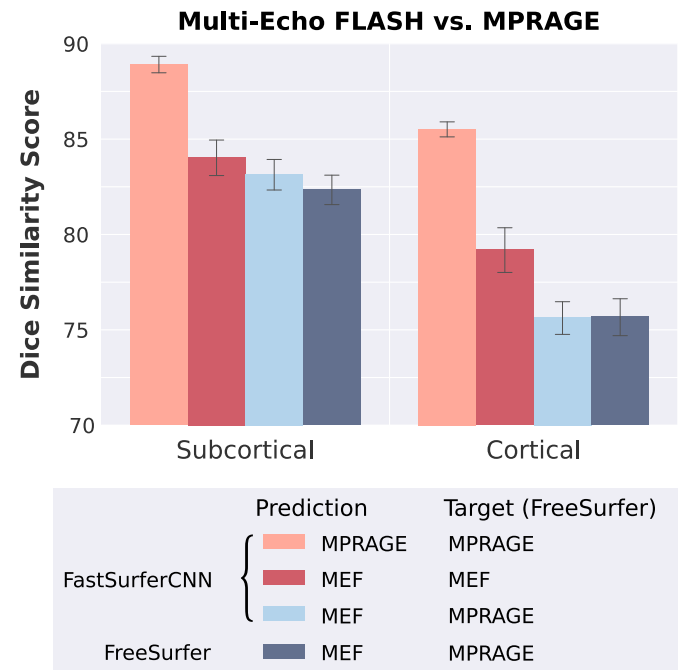

Fig. 8. Comparison of segmentation accuracy between different image acquisition sequences (MPRAGE versus Multi-echo FLASH (MEF)). FreeSurfer's MPRAGE segmentations are considered ground truth. FastSurferCNN generalizes well to MEF acquired images and achieves closer results to FreeSurfers MPRAGE than FreeSurfer itself on the subcortical structures.

L. Henschel et al.

NeuroImage 219 (2020) 117012

column). Reduced agreement on MEF images can, of course, be explained by a deviation of either (or both) methods from ground truth. In the absence of ground truth, we set FreeSurfer's MPRAGE segmentation as the standard. To test how well FreeSurfer generalizes to MEF, we compare FreeSurfer's own segmentations across MEF and MPRAGE after robust rigid registration (Reuter et al., 2010) and observe a significantly reduced DSC (82.37 on subcortical and 75.70 on cortical structures) (Fig. 8 dark blue). In comparison FastSurferCNN's MEF segmentation (using the same registrations) slightly outperforms FreeSurfer's generalizability as it is actually closer to ground truth (FreeSurfer MPRAGE) than FreeSurfer itself for subcortical structures (DSC of 83.17) and similar for the cortex (DSC of 75.67) (Fig. 8 bright blue). These results highlight an excellent generalizability of FastSurfer to the unseen T1-weighted MEF sequence.

The above DSC comparisons assume accurate labels provided by FreeSurfer or a human rater respectively. This is of course not granted, as both automated and manual segmentation quality can degrade across scanners, sequences, or advanced neurodegeneration. Furthermore, highly folded 3D structures such as cortical gray matter are difficult to segment manually on a voxel grid when viewing 2D slices. Whenever "ground truth" cannot be trusted, it is difficult to quantify performance with direct comparisons, as a small DSC can also indicate noisy or

erroneous ground truth. Therefore, we also perform validations in the next sections that are independent of ground truth labels but rather rely on the assumptions, that (i) anatomy remains stable in small time frames (test-retest reliability) and that (ii) established disease effects should be detectable with high statistical significance (sensitivity). For these comparisons we run the full FastSurfer pipeline, extending the FastSurferCNN with subsequent surface processing.

3.3. Reliability

Test-retest reliability is assessed as the agreement between the evaluations of two scans in a short time frame. We calculate the intraclass correlation coefficient (McGraw and Wong, 1996) on the OASIS1 test-retest dataset with 20 participants. Note, that the acquisition source of variation (motion, noise etc.) will be identical for different image processing methods. Higher agreement can therefore be taken as an index of method stability and consistency of results.

Fig. 9 shows the ICC value for each structure separately including the upper and lower bound at significance level $\alpha = 0.05$ (black error bar) for FastSurfer (dark blue bars) and FreeSurfer (light blue bars). A higher ICC indicates a better reliability with the maximum being 1. The ICC values for the volume of subcortical structures are higher for FastSurfer on all 13
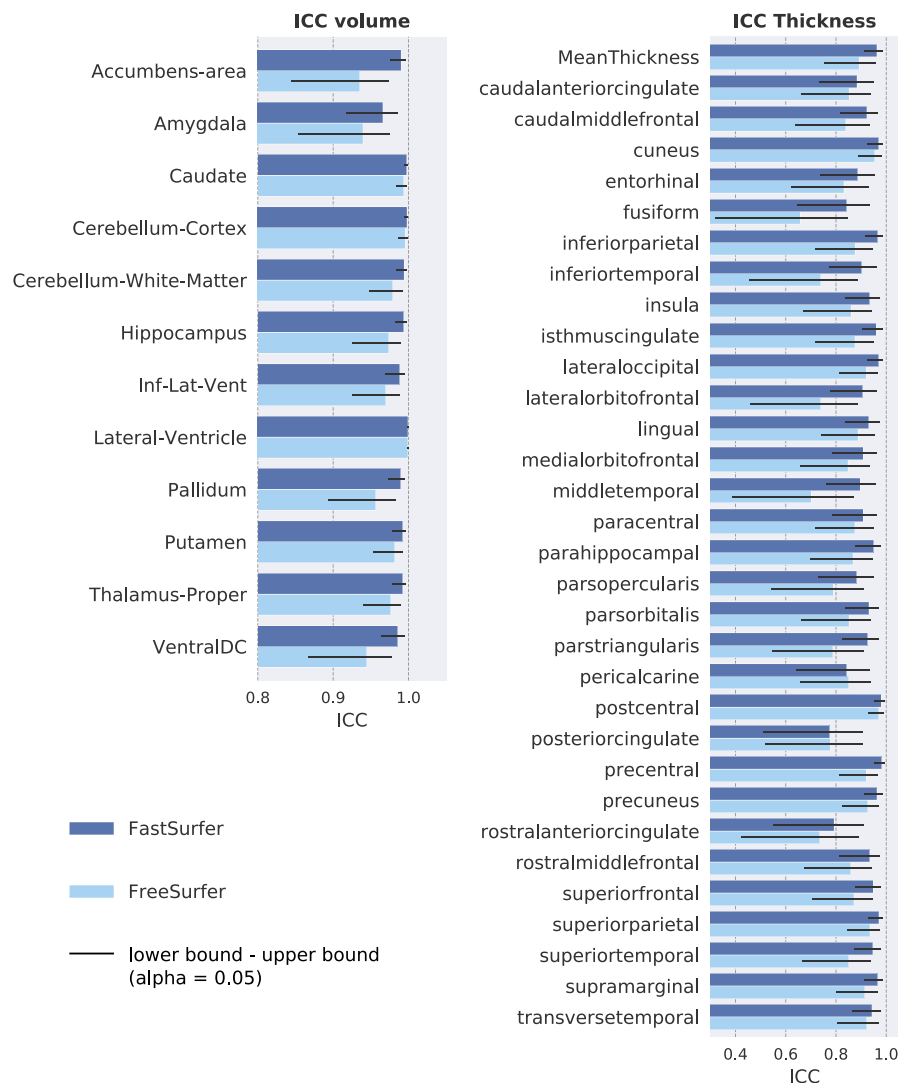

Fig. 9. Intraclass correlation coefficient on the Test-Retest OASIS1 dataset for FastSurfer (dark blue) and FreeSurfer (light blue). Error bars indicate upper and lower bound of the calculated ICC (significance level $\alpha = 0.05$).

10

structures (0.99 on average versus 0.97 with FreeSurfer). On six structures (including Hippocampus, Putamen, Caudate and Thalamus) the agreement between scans is above 0.99. Furthermore, all confidence intervals (lower to upper bound) are smaller for FastSurfer indicating better segmentation consistency across the 20 participants. The ICC values for the thickness of cortical regions show a similar pattern, with FastSurfer achieving higher ICC on both hemispheres in 30 out of 31 regions. On average, an ICC of 0.92 is achieved with FastSurfer (0.87 with FreeSurfer) and 24 structures show a correlation of above 0.9. No cortical structure has an ICC below 0.78 (0.73 for FreeSurfer). Correspondingly, visualization of the ICC directly on the surface (see Fig. 10) demonstrates that FastSurfer segmentations yield larger regions on the cortex with high ICC values (light blue) compared to FreeSurfer. Here, it is further apparent that the majority of the cortex reaches ICC values of more than 0.8 (blue areas).

3.4. Sensitivity to group differences in dementia

Sensitivity to group differences in dementia is an essential component of our evaluation. While the segmentation accuracy of our approach (as quantified by the DSC to a manual standard) has already been established, we now determine to what extent our results are relevant and useful in an applied research setting. To this end, we analyze whether our proposed method is capable of reproducing well-known group differences in dementia: The sensitivity of both FreeSurfer and FastSurfer is determined by evaluating their ability to separate diagnostic groups in OASIS1 (AD versus CN), as indicated by the p-value.

In Fig. 11 structures with p-values below 10^{-5} are shown (based on both FreeSurfer and FastSurfer). The signed p-values presented in the figure indicate the direction of the effect (below zero represent atrophy and above zero volume increase). Consequently, one can directly observe that the ventricle volume (lateral ventricles, inferior lateral ventricles) increases while all other structures atrophy for both, FreeSurfer and

FastSurfer. In the subcortical domain, a volume reduction is specifically detected for the hippocampus, the amygdala and the thalamus, which is congruent with other research results on AD (de Jong et al., 2008; Henneman et al., 2009; Schuff et al., 2009; Poulin et al., 2011; Aggleton et al., 2016; Pini et al., 2016). FastSurfer reaches lower p-values for all three structures indicating a higher sensitivity to differences between the groups. FastSurfer and FreeSurfer are further capable of detecting significant differences in areas related to disease progression in the cortex (e.g. bilateral frontal, temporal and parietal lobe (Braak and Braak, 1995; Baron et al., 2001; Wenk, 2003; Lerch et al., 2004; Poulin et al., 2011)). Specifically, parts of the temporal (superiortemporal, middletemporal, inferiortemporal, entorhinal) and parietal lobes (inferiorparietal, supramarginal) are significantly thinner (p < 10^{-11} detected with FastSurfer for all areas). The overall thickness also correlates with disease progression (MeanThickness p < 10^{-12} for FastSurfer). These results remain stable when analyzing lateralized ROI measures instead of the mean (not shown).

Fig. 12 depicts the detected differences of cortical thickness in patients with AD compared to CN subjects with the original FreeSurfer stream (left) and our proposed FastSurfer pipeline (right) directly on the surface (vertex-wise analysis). The visualization complements the results of Fig. 11, clearly indicating the ability of FreeSurfer and FastSurfer to detect thinning effects across hemispheres. Again, the differences between groups are more pronounced with the FastSurfer pipeline (smaller uncorrected p-value, as indicated by larger yellow regions on both hemisphere). The proposed pipeline is thus able to separate groups in this dataset very clearly.

3.5. Pipeline innovations

As described in Section 2.3, FastSurfer is a full surface reconstruction pipeline based on FreeSurfer. Additionally to skipping/replacing some steps the two main modifications are (i) reconstructing surfaces with the marching cube algorithm and (ii) implementing a novel, fast spectral mapping to the sphere. Here we compare these two changes with a pipeline that uses the original FreeSurfer modules (*mri_tessellate* and *mris_sphere*) in these two steps and is otherwise identical to FastSurfer. We quantify a) the number of topological surface defects, b) the overall processing time for the surfaces, and c) the average quality of the produced surface triangle meshes. Surface meshes should be topologically equivalent to the sphere, which is why topological defects introduced during surface construction, i.e. handles and tunnels, as well as self-intersections have to be corrected in time-intensive downstream processing (Fischl et al., 1999a; Ségonne et al., 2007). This is why a smaller number of topological defects on the original meshes is favourable. The mesh quality is computed by averaging triangle qualities $Q = 4\sqrt{3}\,A/(e_1^2 + e_2^2 + e_3^2)$, where A is the triangle area and e_i the edges (Bank, 1990). Q is 1 for the equilateral triangle and close to zero for degenerated triangles.

To evaluate the two new modules, the following statistics are based on the OASIS1 dataset and summarized in Table 2. The average number of defects in the pipeline with FreeSurfer modules (27.2 defects per hemisphere) decreased by 12% (24.0 defects) when constructing the surfaces with marching cube, and 15.3% (23.1 defects) when the proposed pipeline (FastSurfer: marching cube + spectral spherical projection) was used. Also the average surface processing time decreased significantly by 15 min per hemisphere with FastSurfer. Finally the triangle mesh quality index of FastSurfer (Q = 0.902) was significantly higher than for the pipeline with the traditional modules (Q = 0.888), due to the marching cube algorithm. Additionally, construction of surfaces with marching cube instead of *mri_tessellate* slightly reduces the number of vertices on average by 120 per hemisphere (133,154 versus 133,034 for tessellate and marching cube, respectively). All three tests (defects, time, quality) of the final FastSurfer pipeline with respect to the one with original modules are highly significant: the Wilcoxon signed-
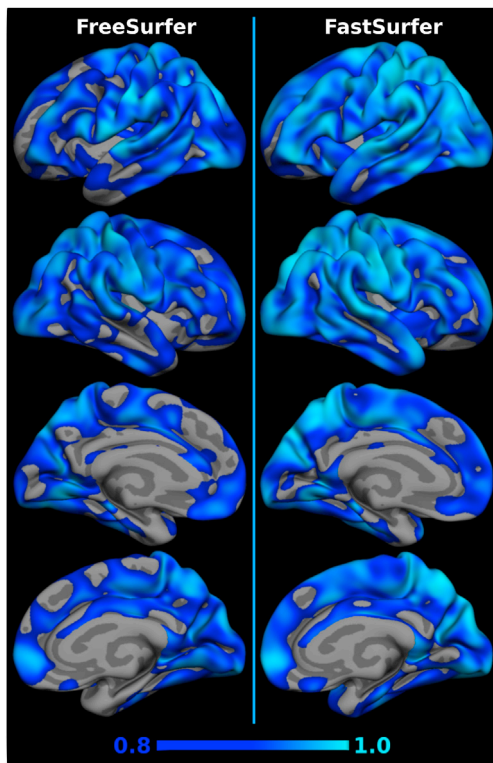

Fig. 10. Visualization of the intraclass correlation coefficient on the Test-Retest OASIS1 dataset for FreeSurfer (left) and FastSurfer (right). ICC ranges from 0.8 (dark blue) to 1.0 (light blue) are shown.
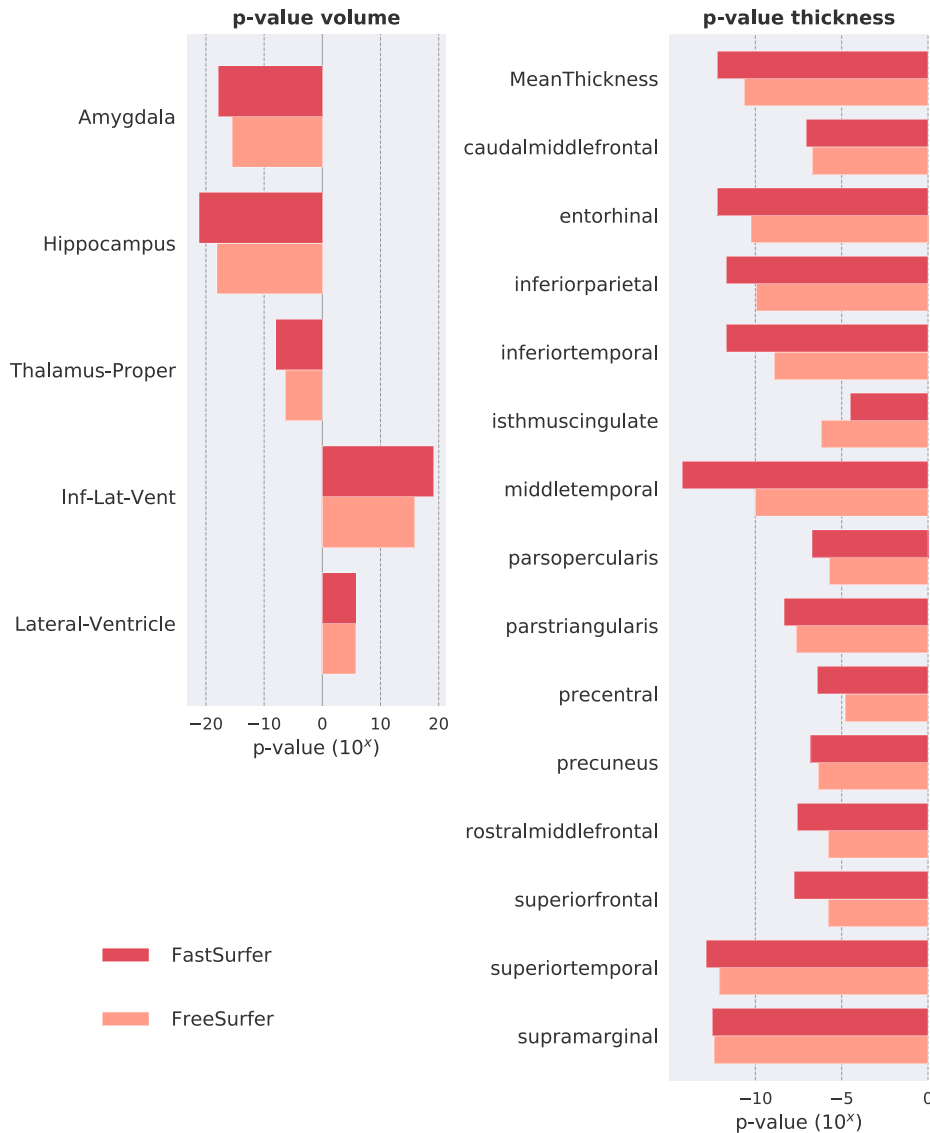
Fig. 11. Significance (p-value) of cortical thickness changes in disease groups on the OASIS1 dataset for FastSurfer (red) and FreeSurfer (light red). Structures with p-values below 10^{-5} are shown. Effect directions are indicated by the sign (atrophy: negative values, enlargement: positive values). Volume of ventricles increases while all other structures show atrophy. FastSurfer and FreeSurfer detect significant changes in areas related to disease progression.

rank test reports a p-value below machine precision ($p < 10^{-16}$).

Overall, in addition to these methodological innovations, our pipeline saves time by replacing many FreeSurfer steps, such as skull stripping, spherical segmentation etc., since we can directly build on the high-quality image segmentations provided by the deep neural network. Here, we compare runtimes of three approaches: (i) complete regular FreeSurfer processing, (ii) FastSurfer pipeline without spherical registration, (iii) and complete FastSurfer with spherical registration (Table 3). Note, while spherical registration is not needed to obtain surface segmentations and ROI thickness measures in FastSurfer, it is required to construct cross-subject correspondence, e.g. when performing vertex-wise surface thickness analyses. All pipelines are evaluated on identical subjects (10 representative subjects from OASIS1, balanced with regard to gender (5 male, 5 female), age range (21–86 years), and diagnosis (4 AD, 6 CN)) and identical hardware (CPU: Intel Xeon Gold 6154 @ 3 Ghz) using both, sequential processing and parallel processing (4 threads and simultaneous processing of the two hemispheres).

In our test a complete FreeSurfer run takes approximately 7 h (4 h parallel) on the CPU, which can vary depending on image quality, disease severity etc. Obtaining only the *aseg.mgz* (36 labels including subcortical

structures) takes approximately 2–3 h for FreeSurfer. However, improved GM and WM segmentations as well as cortical segmentation (DKT Atlas) is only available after surface reconstruction and remapping to the volume as one of the final steps. Thus, obtaining the final segmentation takes the full processing time (7 h, 4 h parallel). The proposed FastSurfer pipeline achieves the volumetric segmentation (both subcortical and cortical regions) in only 1 min (on the GPU, 14 min on the CPU), surface processing including cortical ROI thickness measures in 1.7 h (0.9 h parallel), or complete processing including the spherical registration for potential subsequent group analysis of surface maps in 3.7 h (1.6 h parallel) on the CPU.

4. Discussion

In this work we introduce an advanced semantic segmentation neural network architecture and a fast pipeline for the processing of neuroanatomical surfaces that outperforms FreeSurfer with respect to runtime, reliability and sensitivity. We contribute a deep learning architecture (FastSurferCNN) by incorporating competition into the network (within dense blocks and long-range skip connections) and increasing the initial
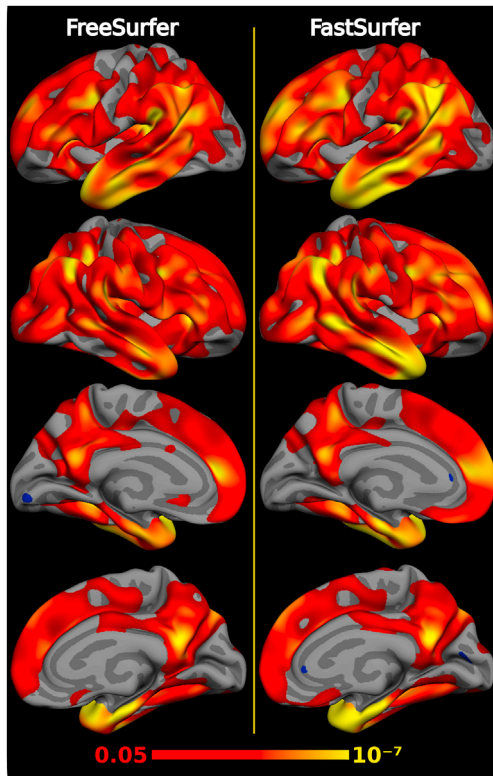
Fig. 12. Group Analysis of cortical thickness variations in Alzheimer's disease compared to controls based on the OASIS1 dataset for FreeSurfer (left) and FastSurfer (right). The color-coded uncorrected p-value map ranges from 0.05 (red) to 10^{-7} (yellow). Differences in cortical thickness are more pronounced in the FastSurfer analysis stream.

Table 2
Average number of topological defects, mesh quality and processing time per hemisphere when using original FreeSurfer modules, marching cube, and spectral spherical projection in the FastSurfer pipeline.

	Defects	Quality	Time/hemi
orig	27.2	0.888	41.6 min
+ mc	24.0	0.902	25.7 min
+ qspec	23.1	0.902	25.4 min

Table 3
Runtime comparisons of the different pipelines (recon-all, recon-surf and recon-surf including registration (reg) of the orig surface to the spherical atlas for cross-subject correspondence). Average processing times and standard deviation in minutes. Parallel = parallelization of hemispheres, 4 threads.

	Sequential	Parallel
FreeSurfer	424 min ($\pm$ 37)	244 min ($\pm$ 20)
FastSurfer	104 min ($\pm$ 20)	54 min ($\pm$ 27)
FastSurfer + sphere.reg	223 min ($\pm$ 53)	97 min ($\pm$ 31)

information content provided to the network via spatial information aggregation. Competition significantly reduces the number of network weights resulting in a slimmer architecture with lower memory requirements. We demonstrate its superior performance for the fast and detailed (close to 100 structures) segmentation of whole-brain MRI compared to existing deep learning approaches. FastSurferCNN outperforms 3D UNet, SDNet and QuickNAT in terms of accuracy by a significant margin both with respect to FreeSurfer and a manual standard. Across five different datasets our network achieves the highest DSC compared to FreeSurfer as a reference on the subcortical and cortical

structures (89.08 and 85.87 on average), as well as the lowest AVG HD (0.1400 and 0.2222 on average). In addition, FastSurferCNN is the best performing network on the manually labeled Mindboggle-101 dataset with a DSC of 80.19 and 80.65 and an AVG HD of 0.2909 and 0.3973 on the subcortical and cortical structures, respectively.

Of note, the optimized 3D UNet was not able to outperform the other view-aggregating (2.5D) architectures. Specifically in terms of the AVG HD, the 3D UNet demonstrated underwhelming performance. In general, it can be expected that full 3D approaches eventually outperform 2.5D networks for whole-brain segmentation due to their potential to better capture the inherent 3D geometry of the cortex. The better performance of the 2.5D approaches here might in part be explainable by their advanced architectures which are difficult to directly translate to a full 3D approach due to the increase in parameters and hence memory requirements. Furthermore, adding spatial information (SPI) as well as view-aggregating the three 2D networks provides spatial context at a much lower memory prize and training effort. In depth comparison of 2.5D versus 3D networks as well as improvements of 3D architectures is however out of scope of this work and an important topic for future exploration.

Further, FastSurferCNN's segmentation results for a 3D 1 mm isotropic MRI brain scan are achieved at a processing time below 1 min. Fast MRI segmentation opens up multiple avenues of potential applications, ranging from direct feedback or field-of-view localization during image acquisition, fast clinical decision support by quantitative personalized measurements, and scalability to very large cohort data sets. Many such applications require no surface models and can terminate after the 1 min image segmentation step, allowing rapid processing of the incoming data.

One frequently quoted limitation of learning-based approaches is the uncertain generalizability beyond image types encountered during training. This limitation is valid and it remains unclear how far networks generalize, e.g. to different sequences, disease or age groups. Various domain-adaptation approaches have been introduced to accommodate fine-tuning a network to a new type of data, and should be considered when network performance degrades. In this work, we put an emphasis on evaluating generalizability of our method. We first demonstrate good generalizability to different sites, vendors, field strength, scanner types, and across disease groups. Analysis on HCP further highlights generalizability to de-faced and down-sampled high-resolutional images, which were never encountered during training. Furthermore, we were able to demonstrate good generalizability to an unseen multi-echo FLASH sequence, even outperforming FreeSurfer. These results are very promising, yet, as with any automated software, we recommend users to visually inspect images to ensure good quality for their acquisition setting as stable generalizability to any T1-weighted sequence can certainly not be guaranteed.

Extending the image segmentation network, our full FastSurfer pipeline permits the fast analysis of cortical thickness (vertex-wise and region-wise) following the DKT atlas. This is achieved by both optimizing and replacing multiple steps of the FreeSurfer pipeline, e.g. by mapping segmentation results from the image to the surfaces. Processing of a single MRI volume with parallelization can thus be achieved in below 1 h including thickness ROI analysis, and 1.6 h including surface registration for cross-subject correspondence - a fraction of the time a whole FreeSurfer run needs to complete (4 h with parallelization). Some of this speed-up can also be attributed to the reduced number of detected topological defects. Marching cube seems to be reducing the number of defects on the initial surfaces already, but also the new spectral spherical mapping helps further reduce detected defects, potentially due to the smooth embedding of the Eigenfunctions and resulting reduction of self-folds. Future work will focus on increasing processing speed further, e.g. by including deep learning based registration procedures. Also note, that ongoing activities to parallelize and speed-up traditional FreeSurfer code will directly impact multiple components of the FastSurfer pipeline (such as the topology fixer, surface reconstruction, cross subject registration

etc.).

Our extensive validation of FastSurfer further includes test-retest reliability and sensitivity studies. FastSurfer exhibits improved test-retest reliability relative to FreeSurfer. This is reflected in higher ICC values across both hemispheres for FastSurfer (on average 0.92 for all cortical and 0.99 for all subcortical structures). Given that increased reliability can be bought by extensive smoothing, potentially at the cost of sensitivity, we evaluate FastSurfer's capability to separate groups in dementia. Here, we can replicate group differences between CN and AD patients with high sensitivity. Specifically, AD-related significant volume reductions in amygdala and hippocampus, increased ventricle volume, as well as cortical thinning in the temporal and parietal lobes were detected with both FastSurfer and FreeSurfer.

While these group differences may not purely reflect neuro-degeneration, but can include indirect factors such as head motion (Reuter et al., 2015; Tisdall et al., 2016) or hydration levels (Biller et al., 2015) the good agreement of FastSurfer with FreeSurfer and with established findings (Braak and Braak, 1995; Baron et al., 2001; Wenk, 2003; Lerch et al., 2004; de Jong et al., 2008; Henneman et al., 2009; Schuff et al., 2009; Poulin et al., 2011; Aggleton et al., 2016; Pini et al., 2016) indicate the validity of our method. FastSurfer's smaller p-values in the majority of these discrimination tasks could potentially be explained by its implicit noise reduction: While consistent boundaries in FreeSurfer segmentations will be learned, random segmentation noise such as local inaccuracies or protrusions are averaged out and might allow the network to achieve superior results, as also corroborated by the improved test-retest reliability results.

The inherent training paradigm of FastSurferCNN can be considered another contributing factor. During training, the network has been exposed to various pathological scans with high anatomical and acquisition variability in contrast to the limited number of cases (40) within the FreeSurfer atlas. The larger corpus likely improves the resulting segmentations and derived volume and thickness estimates. In fact, it is remarkable that the 140 training cases (plus augmentation) are sufficient to provide these excellent results. This is put into perspective by considering that in fact 20 k 2D images (plus augmentation, some of these highly correlated - of course) are used for training each view. Still it can be expected that training with more cases will improve accuracy and generalizability further, leaving space for future exploration.

Finally, one of the major advantages of supervised learning over traditional pipelines is that consistent errors can be removed by manually fixing existing or adding new training cases. This is in stark contrast to model based pipelines, where updates or fixes to the algorithm can only be introduced by a handful of experts and often have unintended consequences. Future work can thus explore training on very large and heterogeneous datasets, as well as the inclusion of manual labels or manually corrected automated labels to improve segmentation quality even further.

It should, however, be noted that availability of manually labeled or corrected training data is often inherently limited due to time intensive production, restricting the applicability of deep learning approaches to domains were enough training data exists. This is especially true for full 3D segmentation networks were one whole-brain MRI is equivalent to a single training case. 2D networks operating on slices or 3D patch approaches are more forgiving due to the much larger number of training inputs. In this work, training on FreeSurfer generated outputs allows us to generate a wide range of training cases with the disadvantage that consistent FreeSurfer segmentation errors may be learned by the networks. Our successful comparison to a manual reference standard (Mindboggle-101) is, therefore, helpful to ascertain high segmentation quality. Yet, it does not replace quality control when applying FastSurfer to new images, as is the case with any automated processing method.

Overall we introduce a fast, stable, reliable and sensitive pipeline for automated neuroimage analysis that scales well to large datasets and enables various new applications where segmentation speed is essential, for example: to localize structures during image acquisition, to provide quantitative measures in clinical workflows, or to process large cohort studies efficiently.

CRediT authorship contribution statement

Leonie Henschel: Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Sailesh Conjeti:** Methodology, Software, Investigation, Writing - original draft. **Santiago Estrada:** Methodology, Software, Writing - review & editing. **Kersten Diers:** Validation, Formal analysis, Writing - review & editing. **Bruce Fischl:** Methodology, Software, Validation, Writing - review & editing, Supervision. **Martin Reuter:** Conceptualization, Methodology, Software, Validation, Resources, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Appendix A

Datasets

ABIDE II: The Autism Brain Imaging Data Exchange II (Di Martino et al., 2017) contains cross-sectional data and focuses on autism spectrum disorders covering a wide age rage (5–64 years of age). The dataset contains 1044 MRI scans from 19 different institutions. The 3D magnetization prepared rapid gradient echo (MP-RAGE) sequence, or a vendor specific variant, was used to acquire all data. The corresponding sequence parameters vary depending on the site (see Table 2 of the corresponding paper for details (Di Martino et al., 2017)). With the exception of a single collection (IP_1, 1.5 T, Philips Achieva), all MRI data were acquired using 3 T scanners (1 Ingenia and 4 Achieva (Philips), 2 MR750 (GE), 7 TriTim (Siemens), 3 Allegra (Siemens) and 1 Skyra (Siemens)) at voxel resolutions varying from 1.30 mm to 0.7 mm (majority at 1.0 mm) and is available: http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html. 20 cases from the ABIDE-II were used for training.

ADNI: The Alzheimer's Disease Neuroimaging Initiative (Mueller et al., 2005) was launched in 2003 as a public-private partnership, led by principal investigator Michael W. Weiner, MD. The dataset contains 1.5 T and 3 T-MRI scans acquired at a resolution of $1.0 \times 1.0 \times 1.2$ mm with scanners from the three largest MRI vendors (GE, Philips and Siemens) and includes Alzheimer's disease patients, mild cognitive impaired subjects, and elderly controls. Data were acquired with a MP-RAGE sequence whose parameters are optimized for the different vendors (see (Jack et al., 2008) for details). The ADNI database has > 2000 participants and is available at: http://adni.loni.usc.edu. 40 cases from ADNI where used for training. 180 different cases were used for assessing accuracy and generalizability across disease groups and scanners.

HCP: The Human Connectome Projects Young Adult (Van Essen et al., 2012) is a cross-sectional dataset and contains 1200 healthy participants from ages 22 to 35. The 3 T MR images were acquired with a single scanner (customized Connectome Skyra, Siemens) using a MP-RAGE sequence with a repetition time (TR) of 2.4 s, echo time (TE) of 2.14 ms, an inversion time (TI) of 1 s, and a flip angle of 8°. Images are 0.7 mm isotropic with a field of view of 224×224, and are de-faced. Data is available at: https://www.humanconnectome.org/study/hcp-young-adult. 45 cases from HCP were used for assessing accuracy and generalizability to inputs with pre-processing (de-facing, downsampling).

LA5c: The UCLA Consortium for Neuropsychiatric Phenomics LA5c Study (Poldrack et al., 2016) is a cross-sectional study and includes 142 participants diagnosed with a neuropsychiatric or neurodevelopmental disorder (schizophrenia, bipolar disorder, ADHD) and 130 normal controls (ages 21–50). T1-weighted MP-RAGE images were acquired on a 3 T Siemens Trio at a single-center with field of view of 250, 256×256 matrix, and 176 1.0 mm sagittal partitions. An TI of 1.1 s, TE of 3.5–3.3 ms, TR of 2.53 s and flip angle of 7° was used for all scans. This data was obtained from the OpenfMRI database (https://openfmri.org/dataset/ds000030/). Its accession number is ds000030.20 cases from LA5c were used for training.

Mindboggle-101: a manually corrected set of 101 labeled brain images based on a consistent human cortical labeling protocol (DKTatlas) (Klein and Tourville, 2012). It is the largest and most complete set of free, publicly accessible, manually labeled human brain images. The dataset consists of anatomically labeled brain surfaces and volumes derived from T1-weighted brain MRIs of healthy individuals and is available at: https://osf.io/nhtur/. Five subjects were scanned specifically for this dataset (MMRR-3T7T-2, Twins-2, and Afterthought-1), all others are from publicly available datasets (i.e. Test-Retest OASIS1 (Marcus et al., 2007), the Multi-Modal Reproducibility Resource (Landman et al., 2011), Nathan Kline Institute Test–Retest and Nathan Kline Institute/Rockland Sample, Human Language Network subjects (Morgan et al., 2009), and Colin Holmes 27 template (Holmes et al., 1998), see (Klein and Tourville, 2012) for details). The OASIS1 Test-Retest portion (20 subjects) also includes manually labeled subcortical segmentations (labelings by Neuromorphometrics, Inc. (http://Neuromorphometrics.com/) released under Creative Commons License BY-NC-ND 4.0 (http://creativecommons.org/licenses/by-nc-nd/4.0) within Mindboggle-101). Here, we use all volumes with isotropic voxel sizes (78 in total) from Mindboggle-101 to evaluate network performance with respect to a manual reference.

MIRIAD: The Minimal Interval Resonance Imaging in Alzheimer's Disease (Malone et al., 2013) dataset includes longitudinal scans from 46 elderly individuals (ages 55+) diagnosed with Alzheimer's disease and 23 elderly normal controls. The 3D MR images were acquired at a single center with a 1.5 T Signa MRI scanner (GE Medical systems), using an IR-FSPGR (inversion recovery prepared fast spoiled gradient recalled) sequence, field of view of 24 cm, 256×256 matrix, 124 1.5 mm coronal partitions (voxel size $0.9 \times 0.9 \times 1.5$), TR 15 ms, TE 5.4 ms, flip angle 15°, and TI 650 ms. The data is available at: https://www.ucl.ac.uk/drc/research/methods/minimal-interval-resonance-imaging-alzheimers-disease-miriad. 20 cases from MIRIAD were used for validation. 49 different cases were used to assess accuracy and generalizability to a different T1-weighted acquisition sequences (IR-FSGPR) and scanner (GE).

MMND: This multi-subject, multi-model neuroimaging dataset was acquired with multiple functional and structural neuroimaging modalities (structural MRI + functional MRI + MEG + EEG) on the same 16 healthy volunteers (Wakeman and Henson, 2015). Here, we only use the structural data which include T1-weighted MPRAGE and Multi-Echo Fast Low Angle Shot (MEF) sequences. The data was collected from a Siemens 3 T TIM TRIO with a standard 1 mm isotropic resolution. For each participant, a T1-weighted image was acquired using an MPRAGE sequence (TR 2,250 ms, TE 2.98 ms, TI 900 ms, 190 Hz/pixel; flip angle 9°) as well as two bandwidth-matched MEF sequences (651 Hz/pixel; TR 20 ms) at both 5° and 30° flip angles for each of 7 echo-times (TE 1.85 ms; 4.15 ms; 6.45 ms; 8.75 ms; 11.05 ms; 13.35 ms; 15.65 ms). This data was obtained from the OpenNeuro database. Its accession number is ds000117. All MMND cases were used to assess generalizability to MEF sequences.

Oasis-1 (Marcus et al., 2007) and **Oasis-2** (Marcus et al., 2010): The Open Access Series of Imaging Studies 1 and 2, both contain scans from a 1.5-T Vision scanner and a TIM Trio 3 T MRI scanner (both Siemens) acquired in a single-center from non-demented and demented individuals diagnosed with very mild to moderate Alzheimer's disease. All subjects were scanned in sagittal orientation with a voxel resolution of $1.0 \times 1.0 \times 1.25$ mm and a MP-RAGE sequence with the following parameters: TR 9.7 ms, TE 4.0 ms, flip angle 10°, and TI 20 ms. The Oasis-1 set is cross-sectional and contains 416

subject aged 18 to 96. In addition, it contains a test-retest component consisting of 20 subjects that were scanned no more than 90 days apart (all except 5 less than 30 days). The Oasis-2 set focuses on older adults (age 60+) and contains longitudinal scans from 150 subjects. Both dataset are available at: https://www.oasis-brains.org/. 40 cases from Oasis-1 and 20 from Oasis-2 were used for training. 20 different cases from Oasis-1 where used to assess test-retest reliability and 370 cases for quantifying sensitivity to group effects.

THP: The Traveling Human Phantom (Magnotta et al., 2012) is a dataset collected for assessing multi-site neuroimaging reliability. The THP includes 3D MP-RAGE MRI scans at 1.0 mm isotropic voxel resolution from 5 healthy subjects acquired at 8 different imaging centers. The sites involved in this study had either a Siemens 3 T TIM Trio scanner (five sites: IOWA, UMN, UCL, MGH, CCF) or a Philips 3 T Achieva scanner (three sites: JHU, DART, UW). The data is available at: https://openneuro.org/datasets/ds000206. All THP cases were used to quantify accuracy and generalizability across sites and scanners.

Participants of the individual studies gave informed consent in accordance with the Institutional Review Board at each of the participating sites. Complete ethic statements are available at the respective study webpages.

Table A.1

Summary of training, validation and testing sets. Table lists the usage, scanner, field strength, (disease) state, age range and number of used subjects for each dataset.

Usage	Dataset	Scanner	1.5 T/3 T	State	Age	Subjects
Training	ABIDE-II	Phillips	3 T	Autism/Normal	20–39	20
	ADNI	Philips, GE, Siemens	1.5 T/3 T	AD/MCI/Normal	56–90	40
	LA5C	Siemens	3 T	Neuropsych/Normal	23–44	20
	OASIS1	Siemens	1.5 T	Normal	18–60	40
	OASIS2	Siemens	1.5 T	AD/Normal	66–90	20
Validation	MIRIAD	GE	1.5 T	AD/Normal	60–77	20
Accuracy	ADNI	Philips, GE, Siemens	1.5 T/3 T	AD/MCI/Normal	58–85	180
	HCP	Siemens	3 T	Normal	22–35	45
	OASIS1	Siemens	1.5 T	AD/Normal	18–96	370
	MIRIAD	GE	1.5 T	AD/Normal	55–80	49
	THP	Philips, Siemens	3 T	Normal	–	5
	Mindboggle-101	Philips, Siemens	1.5 T/3 T	Normal	19–61	78
Generalizability	ADNI	Philips, GE, Siemens	3 T	AD/MCI/Normal	58–85	180
	HCP	Siemens	3 T	Normal	22–35	45
	MIRIAD	GE	1.5 T	AD/Normal	55–80	49
	MMND	Siemens	3 T	Normal	23–31	16
	THP	Philips, Siemens	3 T	Normal	–	5
	Mindboggle-101	Philips, Siemens	1.5 T/3 T	Normal	19–61	78
Reliability	OASIS1	Siemens	1.5 T	Normal	19–34	20
Sensitivity	OASIS1	Siemens	1.5 T	AD/Normal	18–96	370

Labels

Table A.2

FastSurferCNN (proposed) segmentation labels and mapping to FreeSurfer (FS) for subcortical (left) and cortical (right) structures.

Subcortical Structures	Proposed	FS	Cortical Structures	Proposed	FS
Cortical white matter (lh)	1	2	caudalanteriorcingulate (lh)	34	1002
Lateral Ventricle (lh)	2	4	caudalmiddlefrontal (lh, rh)	35	1003, 2003
Inferior Lateral Ventricle (lh)	3	5	cuneus (lh)	36	1005
Cerebellar White Matter (lh)	4	7	entorhinal (lh, rh)	37	1006, 2006
Cerebellar Cortex (lh)	5	8	fusiform (lh, rh)	38	1007, 2007
Thalamus (lh)	6	10	inferiorparietal (lh, rh)	39	1008, 2008
Caudate (lh)	7	11	inferiortemporal (lh, rh)	40	1009, 2009
Putamen (lh)	8	12	isthmuscingulate (lh)	41	1010
Pallidum (lh)	9	13	lateraloccipital (lh, rh)	42	1011, 2011
3rd-Ventricle	10	14	lateralorbitofrontal (lh)	43	1012
4th-Ventricle	11	15	lingual (lh)	44	1013
Brain Stem	12	16	medialorbitofrontal (lh)	45	1014
Hippocampus (lh)	13	17	middletemporal (lh, rh)	46	1015, 2015
Amygdala (lh)	14	18	parahippocampal (lh)	47	1016
CSF	15	24	paracentral (lh)	48	1017
Accumbens (lh)	16	26	parsopercularis (lh, rh)	49	1018, 2018
Ventral DC (lh)	17	28	parsorbitalis (lh, rh)	50	1019, 2019
Choroid Plexus (lh)	18	31	parstriangularis (lh, rh)	51	1020, 2020
Cortical white matter (rh)	19	41	pericalcarine (lh)	52	1021
Lateral Ventricle (rh)	20	43	postcentral (lh)	53	1022
Inferior Lateral Ventricle (rh)	21	44	posteriorcingulate (lh)	54	1023
Cerebellar White Matter (rh)	22	46	precentral (lh)	55	1024
Cerebellar Cortex (rh)	23	47	precuneus (lh)	56	1025
Thalamus (rh)	24	49	rostralanteriorcingulate (lh, rh)	57	1026, 2026
Caudate (rh)	25	50	rostralmiddlefrontal (lh, rh)	58	1027, 2027
Putamen (rh)	26	51	superiorfrontal (lh)	59	1028
Pallidum (rh)	27	52	superiorparietal (lh, rh)	60	1029, 2029
Hippocampus (rh)	28	53	superiortemporal (lh, rh)	61	1030, 2030
Amygdala (rh)	29	54	supramarginal (lh, rh)	62	1031, 2031
Accumbens (rh)	30	58	transversetemporal (lh, rh)	63	1034, 2034

Table A.2 (*continued*)

Subcortical Structures	Proposed	FS	Cortical Structures	Proposed	FS
Ventral DC (rh)	31	60	insula (lh, rh)	64	1035, 2035
Choroid Plexus (rh)	32	63	caudalanteriorcingulate (rh)	65	2002
WM-hypointensities	33	77	cuneus (rh)	66	2005
			isthmuscingulate (rh)	67	2010
			lateralorbitofrontal (rh)	68	2012
			lingual (rh)	69	2013
			medialorbitofrontal (rh)	70	2014
			parahippocampal (rh)	71	2016
			paracentral (rh)	72	2017
			pericalcarine (rh)	73	2021
			postcentral (rh)	74	2022
			posteriorcingulate (rh)	75	2023
			precentral (rh)	76	2024
			precuneus (rh)	77	2025
			superiorfrontal (rh)	78	2028

References

Aggleton, J.P., Pralus, A., Nelson, A.J.D., Hornberger, M., 2016. Thalamic pathology and memory loss in early Alzheimer's disease: moving the focus from the medial temporal lobe to Papez circuit. Brain 139 (7), 1877–1890.

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. Digit. Imag. 30 (4), 449–459.

Amin, J., Sharif, M., Yasmin, M., Fernandes, S.L., 2018. Big data analysis for brain tumor detection: deep convolutional neural networks. Future Generat. Comput. Syst. 87, 290–297.

Arunachalam, M., Savarimuthu, S.R., 2017. An efficient and automatic glioblastoma brain tumor detection using shift-invariant shearlet transform and neural networks. Int. J. Imag. Syst. Technol. 27 (3), 216–226.

Avants, B.B., Tustison, N., Song, G., 2009. Advanced normalization tools (ants). Insight Journal 2, 1–35.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Bank, R.E., 1990. PLTMG: a software package for solving elliptic partial differential equations: users' guide 6.0. In: Frontiers in Applied Mathematics, vol. 7. SIAM.

Baron, J., Chêtelat, G., Desgranges, B., Perchey, G., Landeau, B., de la Sayette, V., Eustache, F., 2001. In vivo mapping of gray matter loss with voxel-based morphometry in mild alzheimer's disease. Neuroimage 14 (2), 298–309.

Biller, A., Reuter, M., Patenaude, B., Homola, G., Breuer, F., Bendszus, M., Bartsch, A., 2015. Responses of the human brain to mild dehydration and rehydration explored in vivo by 1h-mr imaging and spectroscopy. AJNR. American journal of neuroradiology 36 (12), 2277–2284.

Braak, H., Braak, E., 1995. Staging of alzheimer's disease-related neurofibrillary changes. Neurobiol. Aging 16 (3), 271–278 (the Schmitt Symposium: The Cytoskeleton and Alzheimer's Disease).

Brainlesion Glioma, 2018. Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop. BrainLes. Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I, Vol. 11383 of Lecture Notes in Computer Science, Springer, 2019.

Breteler, M.M., Stöcker, T., Pracht, E., Brenner, D., Stirnberg, R., 2014. Mri in the rhineland study: a novel protocol for population neuroimaging, Alzheimer's & Dementia. J Alzheimer's Assos. 10 (4), P92.

Buckner, R.L., Head, D., Parker, J., Fotenos, A.F., Marcus, D., Morris, J.C., Snyder, A.Z., 2004. A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. Neuroimage 23, 724–738.

Chen, X., Konukoglu, E., 2018. Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. In: Proceedings of the 1st International Conference on Medical Imaging with Deep Learning MIDL 2018. Amsterdam, Netherlands, 4-6 July 2018.

Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.-A., 2018. Voxresnet: deep voxelwise residual networks for brain segmentation from 3d mr images. Neuroimage 170, 446–455.

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 - 19th International Conference, pp. 424–432. Athens, Greece, October 17-21, 2016, Proceedings, Part II, Vol. 9901 of Lecture Notes in Computer Science.

Coupé, P., Mansencal, B., Clément, M., Giraud, R., de Senneville, B.D., Ta, V.-T., Lepetit, V., Manjon, J.V., 2019. Assemblynet: a novel deep decision-making process for whole brain mri segmentation. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, pp. 466–474.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis: I. segmentation and surface reconstruction. Neuroimage 9 (2), 179–194.

de Brébisson, A., Montana, G., 2015. Deep neural networks for anatomical brain segmentation. In: In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops. CVPRW, pp. 20–28.

de Jong, L.W., van der Hiele, K., Veer, I.M., Houwing, J.J., Westendorp, R.G.J., Bollen, E.L.E.M., de Bruin, P.W., Middelkoop, H.A.M., van Buchem, M.A., van der Grond, J., 2008. Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. Brain 131 (12), 3277–3285.

Dedmari, M.A., Conjeti, S., Estrada, S., Ehses, P., Stöcker, T., Reuter, M., 2018. Complex fully convolutional neural networks for mr image reconstruction. In: Knoll, F., Maier, A., Rueckert, D. (Eds.), Machine Learning for Medical Image Reconstruction. Springer International Publishing, pp. 30–38.

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., 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.

Di Martino, A., O'connor, D., Chen, B., Alaerts, K., Anderson, J.S., Assaf, M., Balsters, J.H., Baxter, L., Beggiato, A., Bernaerts, S., et al., 2017. Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. Scientific data 4, 170010.

Ding, J., Li, A., Hu, Z., Wang, L., 2017. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2017 - 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III, Vol. 10435 of Lecture Notes in Computer Science. Springer, pp. 559–567.

Dolz, J., Desrosiers, C., Ayed, I.B., 2018. 3d fully convolutional networks for subcortical segmentation in mri: a large-scale study. Neuroimage 170, 456–470.

Dong, H., Yang, G., Liu, F., Mo, Y., Guo, Y., July . Automatic brain tumor detection and segmentation using u-net based fully convolutional networks. In: Medical Image Understanding and Analysis - 21st Annual Conference. MIUA 2017, Edinburgh, UK, pp. 506–517. Proceedings, 2017.

Estrada, S., Conjeti, S., Ahmad, M., Navab, N., Reuter, M., 2018. Competition vs. concatenation in skip connections of fully convolutional networks. In: Machine Learning in Medical Imaging. Springer International Publishing, pp. 214–222.

Estrada, S., Lu, R., Conjeti, S., Orozco-Ruiz, X., Panos-Willuhn, J., Breteler, M.M.B., Reuter, M., 2019. Fatsegnet : a fully automated deep learning pipeline for adipose tissue segmentation on abdominal dixon MRI. Magn. Reson. Med. 83, 1471–1483.

Fedorov, A., Johnson, J., Damaraju, E., Ozerin, A., Calhoun, V.D., Plis, S.M., May . End-to-end learning of brain tissue segmentation from imperfect labeling. In: 2017 International Joint Conference on Neural Networks. IJCNN 2017, Anchorage, AK, USA, pp. 3785–3792. IEEE, 2017.

Fischl, B., 2012. Freesurfer. Neuroimage 62 (2), 774–781.

Fischl, B., Dale, A., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc. Natl. Acad. Sci. U. S. A 97 (20), 11050–11055.

Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis: II. Inflation, flattening, and a surface-based coordinate system. Neuroimage 9 (2), 195–207.

Fischl, B., Sereno, M.I., Tootell, R., Dale, A., 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum. Brain Mapp. 8, 272–284.

Fischl, B., Liu, A.K., Dale, A.M., 2001. Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. IEEE Trans. Med. Imag. 20 (1), 70–80.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33 (3), 341–355.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D., Busa, E., Seidman, L., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., Dale, A., 2004. Automatically parcellating the human cerebral cortex. Cerebral cortex (New York, N.Y. : 1991 14, 11–22.

Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W., 2007. Statistical Parametric Mapping: the Analysis of Functional Brain Images. Academic Press.

Ghavami, N., Hu, Y., Bonmati, E., Rodell, R., Gibson, E., Moore, C., Barratt, D., 2018. Integration of spatial information in convolutional neural networks for automatic

segmentation of intraoperative transrectal ultrasound images. J. Med. Imag 6 (1), 1–6.

Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y., 2013. Maxout networks. In: Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume, vol. 28. JMLR. org, pp. III–1319.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., Fischl, B., 2006. Reliability of mri-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. Neuroimage 32 (1), 180–194.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2017. Brain tumor segmentation with deep neural networks. Med. Image Anal. 35, 18–31.

Henneman, W., Sluimer, J.D., Barnes, J., van der Flier, W.M., Sluimer, I.C., Fox, N.C., Scheltens, P., Vrenken, H., Barkhof, F., 2009. Hippocampal atrophy rates in alzheimer disease. Neurology 72 (11), 999–1007.

Holmes, C.J., Hoge, R., Collins, L., Woods, R., Toga, A.W., Evans, A.C., 1998. Enhancement of mr images using registration for signal averaging. J. Comput. Assist. Tomogr. 22, 324–333.

Hosseini-Asl, E., Keynton, R., El-Baz, A., 2016. Alzheimer's disease diagnostics by adaptation of 3d convolutional network. In: 2016 IEEE International Conference on Image Processing. ICIP, pp. 126–130.

Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, p. 3.

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.

Ikram, M.A., Brusselle, G.G., Murad, S.D., van Duijn, C.M., Franco, O.H., Goedegebure, A., Klaver, C.C., Nijsten, T.E., Peeters, R.P., Stricker, B.H., et al., 2017. The rotterdam study: 2018 update on objectives, design and main results. Eur. J. Epidemiol. 32 (9), 807–850.

Isensee, F., Petersen, J., Kohl, S.A.A., Jäger, P.F., Maier-Hein, K.H., 2019. nnU-Net: Breaking the Spell on Successful Medical Image Segmentation. CoRR abs/ 1904.08128.

Ito, R., Nakae, K., Hata, J., Okano, H., Ishii, S., 2019. Semi-supervised deep learning of brain tissue segmentation. Neural Network. 116, 25–34.

Jack, C., Bernstein, M.A., C Fox, N., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P., Whitwell, J.L., Ward, C., Dale, A., Felmlee, J., Gunter, J., Hill, D., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., Weiner, M., 2008. The alzheimer's disease neuroimaging initiative (adni): mri methods. J. Magn. Reson. Imag. : JMRI 27, 685–691.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., 2017. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. IEEE, pp. 1175–1183.

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

Jin, K.H., McCann, M.T., Froustey, E., Unser, M., 2017. Deep convolutional neural network for inverse problems in imaging. IEEE Trans. Image Process. 26 (9), 4509–4522.

Jog, A., Hoopes, A., Greve, D.N., Leemput, K.V., Fischl, B., 2019. Psacnn: pulse sequence adaptive fast whole brain segmentation. Neuroimage 199, 553–569.

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

Kingma, D.P., Ba, J., May . Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015. Conference Track Proceedings, 2015, San Diego, CA, USA.

Klein, A., Tourville, J., 2012. 101 labeled brain images and a consistent human cortical labeling protocol. Front. Neurosci. 6, 171.

Landman, B.A., Huang, A.J., Gifford, A., Vikram, D.S., Lim, I.A.L., Farrell, J.A., Bogovic, J.A., Hua, J., Chen, M., Jarso, S., Smith, S.A., Joel, S., Mori, S., Pekar, J.J., Barker, P.B., Prince, J.L., van Zijl, P.C., 2011. Multi-parametric neuroimaging reproducibility: a 3-t resource study. Neuroimage 54 (4), 2854–2866.

G. Lee, K. Nho, B. Kang, K.-A. Sohn, D. Kim, Predicting Alzheimer's disease progression using multi-modal deep learning approach, Sci. Rep. 9 (1).

Lerch, J.P., Pruessner, J.C., Zijdenbos, A., Hampel, H., Teipel, S.J., Evans, A.C., 2004. Focal decline of cortical thickness in alzheimer's disease Identified by computational neuroanatomy. Cerebr. Cortex 15 (7), 995–1001.

Li, J., Yu, Z.L., Gu, Z., Liu, H., Li, Y., 2019. Mman: multi-modality aggregation network for brain segmentation from mr images. Neurocomputing 358, 10–19. https://doi.org/ 10.1016/j.neucom.2019.05.025. URL. http://www.sciencedirect.com/science/artic le/pii/S0925231219306873.

Liao, Z., Carneiro, G., 2016. On the importance of normalisation layers in deep learning with piecewise linear activation units. In: Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, pp. 1–8.

Liao, Z., Carneiro, G., 2017. A deep convolutional neural network module that promotes competition of multiple-size filters. Pattern Recogn. 71, 94–105.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Lorensen, W.E., Cline, H.E., 1987. Marching cubes: a high resolution 3d surface construction algorithm. SIGGRAPH Comput. Graph. 21 (4), 163–169.

Magnotta, V.A., Matsui, J.T., Liu, D., Johnson, H.J., Long, J.D., Bolster Jr., B.D., Mueller, B.A., Lim, K., Mori, S., Helmer, K.G., et al., 2012. Multicenter reliability of diffusion tensor imaging. Brain Connect. 2 (6), 345–355.

Malone, I.B., Cash, D., Ridgway, G.R., MacManus, D.G., Ourselin, S., Fox, N.C., Schott, J.M., 2013. Miriad–public release of a multiple time point alzheimer's mr imaging dataset. Neuroimage 70, 33–36.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. J. Cognit. Neurosci. 19 (9), 1498–1507.

Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2010. Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. J. Cognit. Neurosci. 22 (12), 2677–2684.

Mardani, M., Gong, E., Cheng, J.Y., Vasanawala, S., Zaharchuk, G., Alley, M.T., Thakur, N., Han, S., Dally, W.J., Pauly, J.M., Xing, L., 2019. Deep generative adversarial networks for compressed sensing automates MRI. IEEE Trans. Med. Imag. 38, 167–179.

McGraw, K.O., Wong, S., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1 (1), 30–46.

Mehta, R., Majumdar, A., Sivaswamy, J., 2017. BrainSegNet: a convolutional neural network architecture for automated segmentation of human brain structures. J. Med. Imag 4 (2), 1–11.

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

Milletari, F., Ahmadi, S.-A., Kroll, C., Plate, A., Rozanski, V., Maiostre, J., Levin, J., Dietrich, O., Ertl-Wagner, B., Bötzel, K., Navab, N., 2017. Hough-cnn: deep learning for segmentation of deep brain regions in mri and ultrasound. Comput. Vis. Image Understand. 164, 92–102.

Moeskops, P., Viergever, M.A., Mendrik, A.M., de Vries, L.S., Benders, M.J.N.L., Išgum, I., 2016. Automatic segmentation of mr brain images with a convolutional neural network. IEEE Trans. Med. Imag. 35 (5), 1252–1261.

Morgan, V.L., Mishra, A., Newton, A.T., Gore, J.C., Ding, Z., 2009. Integrating functional and diffusion magnetic resonance imaging for analysis of structure-function relationship in the human language network. PloS One 4 (8), 1–8.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). Alzheimer's Dementia 1 (1), 55–66.

Nogovitsyn, N., Souza, R., Muller, M., Srajer, A., Hassel, S., Arnott, S.R., Davis, A.D., Hall, G.B., Harris, J.K., Zamyadi, M., Metzak, P.D., Ismail, Z., Bray, S.L., Lebel, C., Addington, J.M., Milev, R., Harkness, K.L., Frey, B.N., Lam, R.W., Strother, S.C., Goldstein, B.I., Rotzinger, S., Kennedy, S.H., MacQueen, G.M., 2019. Testing a deep convolutional neural network for automated hippocampus segmentation in a longitudinal sample of healthy participants. Neuroimage 197, 589–597. https:// doi.org/10.1016/j.neuroimage.2019.05.017.

Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1520–1528.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch. In: NIPS Workshop Autodiff.

Payan, A., Montana, G., 2015. Predicting Alzheimer's disease - a neuroimaging study with 3d convolutional neural networks. In: ICPRAM 2015 - Proceedings of the International Conference on Pattern Recognition Applications and Methods, ume 2. Portugal, Lisbon, pp. 355–362, 10-12 January, 2015.

Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., Galluzzi, S., Marizzoni, M., Frisoni, G.B., 2016. Brain atrophy in alzheimer's disease and aging. Ageing Res. Rev. 30, 25–48 brain Imaging and Aging.

Poldrack, R.A., Congdon, E., Triplett, W., Gorgolewski, K., Karlsgodt, K., Mumford, J., Sabb, F., Freimer, N., London, E., Cannon, T., et al., 2016. A phenome-wide examination of neural and cognitive function. Scientific data 3, 160110.

Poulin, S.P., Dautoff, R., Morris, J.C., Barrett, L.F., Dickerson, B.C., 2011. Amygdala atrophy is prominent in early alzheimer's disease and relates to symptom severity. Psychiatr. Res. Neuroimaging 194 (1), 7–13.

Qi, J., Tejedor, J., 2016. Deep multi-view representation learning for multi-modal features of the schizophrenia and schizo-affective disorder. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 952–956.

N. Rani, S. Vashisth, Brain tumor detection and classification with feed forward back-prop neural network, Int. J. Comp. Appl. 146.

Reuter, M., 2010. Hierarchical shape segmentation and registration via topological features of laplace-beltrami eigenfunctions. Int. J. Comput. Vis. 89 (2), 287–308.

Reuter, M., Wolter, F.-E., Peinecke, N., 2006. Laplace-beltrami spectra as "shape-dna" of surfaces and solids. Comput. Aided Des. 38 (4), 342–366.

Reuter, M., Rosas, H.D., Fischl, B., 2010. Highly accurate inverse consistent registration: a robust approach. Neuroimage 53 (4), 1181–1196.

Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61 (4), 1402–1418.

Reuter, M., Tisdall, M.D., Qureshi, A., Buckner, R.L., van der Kouwe, A.J.W., Fischl, B., 2015. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. Neuroimage 107, 107–115.

Rezaei, M., Yang, H., Meinel, C., 2017. Deep neural network with l2-norm unit for brain lesions detection. In: Neural Information Processing - 24th International Conference. ICONIP, Guangzhou, China, pp. 798–807. November 14-18, 2017, Proceedings, Part IV, 2017.

Roa-Barco, L., Serradilla-Casado, O., de Velasco-Vázquez, M., López-Zorrilla, A., Graña, M., Chyzhyk, D., Price, C., 22-24 May 2017. A 2d/3d convolutional neural network for brain white matter lesion detection in multimodal MRI. In: Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017. Polanica Zdroj, Poland, pp. 377–385, 2017.

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

Roy, A.G., Conjeti, S., Sheet, D., Katouzian, A., Navab, N., Wachinger, C., 2017. Error corrective boosting for learning fully convolutional networks with limited data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 231–239.

Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A.D.N., et al., 2019. Quicknat: a fully convolutional network for quick and accurate segmentation of neuroanatomy. Neuroimage 186, 713–727.

Schlemper, J., Caballero, J., Hajnal, J.V., Price, A.N., Rueckert, D., 2018. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. IEEE Trans. Med. Imag. 37 (2), 491–503.

Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L.M., Trojanowski, J.Q., Thompson, P.M., Jack, C.R., Weiner, M.W., 2009. The Alzheimer's; Disease Neuroimaging Initiative, MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. Brain 132 (4), 1067–1077.

Ségonne, F., Pacheco, J., Fischl, B., 2007. Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. IEEE Trans. Med. Imag. 26 (4), 518–529.

Shattuck, D.W., Leahy, R.M., 2002. Brainsuite: an automated cortical surface identification tool. Med. Image Anal. 6 (2), 129–142.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in mri data. IEEE Trans. Med. Imag. 17 (1), 87–97.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12 (3), 1–10.

Sun, L., Ma, W., Ding, X., Huang, Y., Liang, D., Paisley, J., 2019. A 3d spatially-weighted network for segmentation of brain tissue from mri. IEEE Trans. Med. Imag. 39 (4), 898–909.

Thyreau, B., Sato, K., Fukuda, H., Taki, Y., 2018. Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing. Med. Image Anal. 43, 214–228.

Tisdall, M.D., Reuter, M., Qureshi, A., Buckner, R.L., Fischl, B., van der Kouwe, A.J., 2016. Prospective motion correction with volumetric navigators (vnavs) reduces the bias and variance in brain morphometry induced by subject motion. Neuroimage 127, 11–22.

Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., et al., 2012. The human connectome project: a data acquisition perspective. Neuroimage 62 (4), 2222–2231.

Varghese, A., Vaidhya, K., Thirunavukkarasu, S., Kesavdas, C., Krishnamurthi, G., 2017. Semi-supervised learning using denoising autoencoders for brain lesion detection and segmentation. J. Med. Imag 4, 1–16.

Wachinger, C., Reuter, M., Klein, T., 2018. Deepnat: deep convolutional neural network for segmenting neuroanatomy. Neuroimage 170, 434–445.

Wakeman, D.G., Henson, R.N., 2015. A multi-subject, multi-modal human neuroimaging dataset. Scientific data, 150001, 02.

Wenk, G., 2003. Neuropathologic changes in alzheimer's disease. J. Clin. Psychiatr. 64 (Suppl. 9), 7–10.

Yang, G., Yu, S., Dong, H., Slabaugh, G.G., Dragotti, P.L., Ye, X., Liu, F., Arridge, S.R., Keegan, J., Guo, Y., Firmin, D.N., 2018. DAGAN: deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. IEEE Trans. Med. Imag. 37 (6), 1310–1321.

Yu, Q., Xie, L., Wang, Y., Zhou, Y., Fishman, E.K., Yuille, A.L., June . Recurrent saliency transformation network: incorporating multi-stage visual cues for small organ segmentation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018. IEEE Computer Society, 2018, Salt Lake City, UT, USA, pp. 8280–8289.

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., Shen, D., 2015. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. Neuroimage 108, 214–224.