IJCARS manuscript No.
(will be inserted by the editor)

An Unsupervised Learning Strategy for Postoperative Brain Cavity Segmentation Simulating Resections During Training

Fernando Pérez-García[1,2] · Michele Rizzi[3] · Francesco Cardinale[3] · Valerio Frazzini[4,5,6] · Katia Lehongre[4] · Vincent Navarro[4,5,6] · Caroline Essert[7] · Irène Ollivier[8] · Rachel Sparks[2] · John S. Duncan[9,10] · Sébastien Ourselin[2]

Received: date / Accepted: date

Abstract *Purpose* Accurate segmentation of brain resection cavities aids in postoperative analysis or determining appropriate follow-up treatment. Convolutional neural networks (CNNs) are the state-of-the-art image segmentation technique, but require large annotated datasets for training. Annotation of 3D medical images is time-consuming, requires highly-trained raters, and may suffer from high inter-rater variability. Self- and semi-supervised learning can be used to leverage large amounts of unlabeled data for training.

Methods We developed an algorithm to simulate resections on preoperative magnetic resonance images (MRIs). We curated a new dataset, EPISURG, comprising 430 postoperative and 269 preoperative MRIs from 430 patients who underwent resective surgery. We performed unsupervised training of a 3D CNN for resection cavity segmentation, using our resection cavity simulation. We finetuned our model on four small annotated datasets from different institutions comprising 20, 33, 19 and 133 subjects, respectively. Finally, we

Fernando Pérez-García – E-mail: fernando.perezgarcia.17@ucl.ac.uk
[1]Department of Medical Physics and Biomedical Engineering, University College London (UCL), London, United Kingdom
[2]School of Biomedical Engineering & Imaging Sciences (BMEIS), King's College London, London, United Kingdom
[3]"C. Munari" Epilepsy Surgery Centre ASST GOM Niguarda, Milan, Italy
[4]Paris Brain Institute, ICM, INSERM, CNRS, F-75013, Paris, France
[5]Sorbonne Université, F-75013, Paris, France
[6]AP-HP, Pitié-Salpêtrière Hospital, Epilepsy Unit, Reference Center for Rare Epilepsies, and Departement of Clinical Neurophysiology, F-75013, Paris, France
[7]Université de Strasbourg, CNRS, ICube, Strasbourg, France
[8]Department of Neurosurgery, Strasbourg University Hospital, Strasbourg, France
[9]Department of Clinical and Experimental Epilepsy, UCL Queen Square Institute of Neurology, London, United Kingdom
[10]National Hospital for Neurology and Neurosurgery, Queen Square, London, United Kingdom

qualitatively evaluated model performance on segmenting resection cavities on one intraoperative MRI and 13 postoperative brain tumor MRIs.

Results The model trained on unlabeled data obtained median (interquartile range) Dice score coefficients (DSCs) of 81.7 (16.4), 82.4 (36.4), 74.9 (24.2) and 80.5 (18.7) for each of the four datasets. After finetuning, the DSCs were 89.2 (13.3), 84.1 (19.8), 80.2 (20.1) and 85.2 (10.8). For comparison, inter-rater agreement between human annotators from our previous study was 84.0 (9.9). Qualitative evaluation on intraoperative MRI and postoperative tumor resection MRI was promising.

Conclusion We present an unsupervised learning strategy for CNNs using simulated resection cavities, that can accurately segment real resection cavities on postoperative MRI. Our method generalizes well to data from different institutions, pathologies and modalities. Source code, segmentation models and the EPISURG dataset are available at `https://github.com/fepegar/resseg-ijcars`.

Keywords Unsupervised learning · Deep learning · Segmentation · Brain resection · Simulation · Neuroimaging

1 Introduction

1.1 Motivation

Approximately one third of epilepsy patients are drug-resistant. If the epileptogenic zone (EZ), i.e., 'the area of cortex indispensable for the generation of clinical seizures' [51], can be precisely localized, resective surgery may be used to cure the patient by removing the EZ. Currently, only 40% to 70% of patients with refractory focal epilepsy are seizure-free after resective surgery [28]. This is, in part, due to limitations identifying the EZ. Retrospective studies relating presurgical clinical features and resected brain structures (such as the hippocampus or the piriform cortex) to surgical outcome may provide useful insight to localize and guide resection of the EZ [17]. To quantify the resected structures, first, the resection cavity must be segmented on the postoperative magnetic resonance image (MRI). Then, a preoperative image with a corresponding brain parcellation can be registered to the postoperative MRI to identify resected structures.

Segmentation of the resection cavity is also necessary in other types of neurosurgery. For example, in the context of neuro-oncology, the gross tumor volume, which is defined as the sum of the volumes of the resection cavity and the residual tumor, must be estimated to plan postoperative radiotherapy [13].

After surgery, the resection cavities fill with cerebrospinal fluid (CSF) [63]. This causes an inherent uncertainty in delineating the resection cavity when adjacent to structures such as sulci, ventricles or edemas, due to a lack of separating intensity gradient. Moreover, brain shift can occur during surgery, causing either CSF filling in regions outside of the resection cavity or changes to the shape and volume of brain structures.

Convolutional neural networks (CNNs) have become the state of the art for medical image segmentation [55,56]. They have repeatedly shown superhuman accuracy in fully-supervised learning settings using massive annotated datasets [22]. However, the performance of neural networks trained with fully-supervised learning using small datasets is often poor. Annotated medical imaging datasets are often small due to the financial and time burden annotating the (often three-dimensional) data, and the need for highly-trained raters. In self-supervised learning, training instances are generated automatically using unlabeled data from a source domain to learn features that can be transferred to a target domain [27]. Synthetic data can be generated cheaply to perform self-supervised training [41]. A good compromise between supervised and unsupervised learning is training the neural network using a mix of labeled and unlabeled data, referred to as semisupervised learning [43]. These techniques can be used to leverage unlabeled medical imaging data to improve training in settings where acquiring annotations is time-consuming or costly.

Despite recent efforts to segment resection cavities in the context of brain cancer [36,23,13], little research has been published in the context of epilepsy surgery. Furthermore, previous work is limited by the lack of benchmark datasets, released code or trained models, and evaluation typically being restricted to single-institution datasets used for both training and testing.

1.2 Related works

Changes in brain position, caused by brain shift, remove the possibility of using symmetry measurements around the sagittal plane to locate the resection cavity. Nonlinear registration has been presented to segment the resection cavity for epilepsy [8] and brain cancer [5] surgeries by detecting non-corresponding regions between pre- and post-resection images. However, evaluation of these methods was limited to only six 3D T_1-weighted (T1w) MRIs from a private dataset and two 2D slices, respectively. Furthermore, in cases where large brain shift or edemas occur, non-corresponding voxels detected in the image may be either in the resection cavity or in regions with changes after resection.

Traditional machine learning methods such as decision trees have been used for brain cavity segmentation from T_2-weighted (T2w), fluid-attenuated inversion recovery (FLAIR), and pre- and post-contrast T1w MRI in the context of glioblastoma surgery [36,23]. These methods aggregate information across hand-crafted features extracted from different MRI modalities to train classifiers. These approaches can be sensitive to signal inhomogeneity and have difficulty distinguishing brain regions with intensity patterns similar to CSF from resection cavities. Recently, a 2D CNN was trained to segment the resection cavity on 2D MRI slices, comprising images from the aforementioned four modalities, in a cohort of 30 glioblastoma patients [13]. The final 3D segmentation is determined from averaging predictions across the three anatomical axes. They obtained a median (interquartile range) Dice score coefficient (DSC) of 84 (10) with respect to a ground-truth label obtained by majority

voting from three independent experts. These methods require four different modalities to segment the resection cavity. However, some of the modalities are often unavailable in a clinical setting [10], especially during presurgical evaluation of epilepsy surgery [11]. Furthermore, code and datasets have not been made publicly available, which hinders the possibility of performing a fair comparison across methods. Applying these methods requires curating a dataset with manually obtained annotations on which to trained the models, which is expensive.

Self-supervised learning methods have been presented to leverage large, unlabeled medical image datasets during model training. Unlabeled data are used to automatically generate training instances which are input to train a model for the pretext task. The model is then finetuned on a smaller labeled dataset to perform the desired downstream task [6]. For example, a CNN was trained to reconstruct a skull bone flap from simulated craniectomy images [35]. Realistic lesions have been simulated in chest CT of healthy subjects to train models for nodule detection, improving accuracy compared to training on smaller datasets containing only real lesions [46].

Semisupervised learning may be used when a large amount of unlabeled data from the target domain is available. A model is first trained on labeled data (which might have been self-labeled in a self-supervised setting). Then, the model can generate pseudolabels for the unlabeled data. Methods for uncertainty estimation have been presented to select data instances with pseudolabels having a low uncertainty for medical image segmentation tasks [59].

1.3 Contributions

We present a combined self- and semi-supervised learning approach to train a 3D CNN to segment brain resection cavities from T1w MRI without the need of annotated data. We ensure our work is easily reproducible by publishing model training code, the trained CNNs, the installable Python package to simulate resections and the dataset used for evaluation. To the best of our knowledge, we introduce the first open annotated dataset of postoperative MRI of epilepsy surgery patients.

We have substantially extended our conference paper [48] as follows:

1. we formalized our transfer learning strategies and generalize our resection simulation
2. For semisupervised learning, we used uncertainty estimation as a selection criteria for pseudolabeled instances
3. We performed a more extensive evaluation, including detailed assessment of the resection simulation components and evaluation of the trained model for data from different institutions and pathologies.

The rest of this paper is organized as follows: Section 2 describes our proposed framework to simulate brain resections and the unsupervised training paradigm used for resection cavity segmentation. Section 3 presents experiments to evaluate our proposed method on simulated and real resection data.

Finally, Section 4 discusses the results and concludes with future directions and potential applications.

2 Methods

Our learning strategy is based on a self-supervised learning approach, using our resection simulation method to generate training instances from publicly available MRI datasets during training (Section 2.1.2). Once trained, this model may be used to generate pseudolabels on available unlabeled postoperative images, which can be used to train a new model (Section 2.1.3). The self-supervised model can also be finetuned to improve performance on small labeled datasets from different institutions. A general diagram of these learning strategies is shown in Fig. 1.

We first define machine learning paradigms used by our learning strategy in Section 2.1. We introduce our approach to simulate resection cavities on preoperative MRI in Section 2.2.

2.1 Definitions

A domain $\mathcal{D}$ is defined by a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \in \mathcal{X}$ [43]. Given a specific domain, $\mathcal{D} = \{\mathcal{X}, P(X)\}$, a task comprises a label space $\mathcal{Y}$ and a predictive function $f(\cdot)$ (denoted by $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$). $f(\cdot)$ is not observed but learned from training data composed of pairs $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, where $\boldsymbol{X}_i \in \mathcal{X}$ and $\boldsymbol{Y}_i \in \mathcal{Y}$. At test time, $f(\cdot)$ can be used to predict the corresponding label, $\boldsymbol{Y} = f(\boldsymbol{X})$, for a new instance $\boldsymbol{X}$.

We denote a source domain dataset as $D_\mathrm{S} = \{(\boldsymbol{X}_{\mathrm{S}_i}, \boldsymbol{Y}_{\mathrm{S}_i})\}_{i=1}^{n_\mathrm{S}}$, where $\boldsymbol{X}_{\mathrm{S}_i} \in \mathcal{X}_\mathrm{S}$ is a data instance and $\boldsymbol{Y}_{\mathrm{S}_i} \in \mathcal{Y}_\mathrm{S}$ is the corresponding label. Similarly, a target domain dataset is $D_\mathrm{T} = \{(\boldsymbol{X}_{\mathrm{T}_i}, \boldsymbol{Y}_{\mathrm{T}_i})\}_{i=1}^{n_\mathrm{T}}$, where $\boldsymbol{X}_{\mathrm{T}_i} \in \mathcal{X}_\mathrm{T}$ and $\boldsymbol{Y}_{\mathrm{T}_i} \in \mathcal{Y}_\mathrm{T}$. In most cases, $n_\mathrm{S} \gg n_T \geq 0$, i.e., the source domain dataset is much larger than the target domain dataset.

2.1.1 Transfer learning and domain adaptation

Transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ to perform a target task $\mathcal{T}_\mathrm{T}$ in a target domain $\mathcal{D}_\mathrm{T}$ using a source domain $\mathcal{D}_\mathrm{S}$ and learning task $\mathcal{T}_\mathrm{S}$, where $\mathcal{D}_\mathrm{S} \neq \mathcal{D}_\mathrm{T}$, or $\mathcal{T}_\mathrm{S} \neq \mathcal{T}_\mathrm{T}$. If the target and source domains and tasks are the same, i.e., $\mathcal{D}_\mathrm{S} = \mathcal{D}_\mathrm{T}$ and $\mathcal{T}_\mathrm{S} = \mathcal{T}_\mathrm{T}$, the learning problem is a traditional fully-supervised machine learning problem. In transductive transfer learning, the target and source domains are different (but related) and the tasks are the same, i.e., $\mathcal{D}_\mathrm{S} \neq \mathcal{D}_\mathrm{T}$ and $\mathcal{T}_\mathrm{S} = \mathcal{T}_\mathrm{T}$. In *domain adaptation*, the gap between source and target domains is reduced in the raw feature space or in a latent space. A function $\psi : \mathcal{D}_\mathrm{S} \rightarrow \mathcal{D}_{\mathrm{S}'} \approx \mathcal{D}_\mathrm{T}$ may reduce the domain gap.

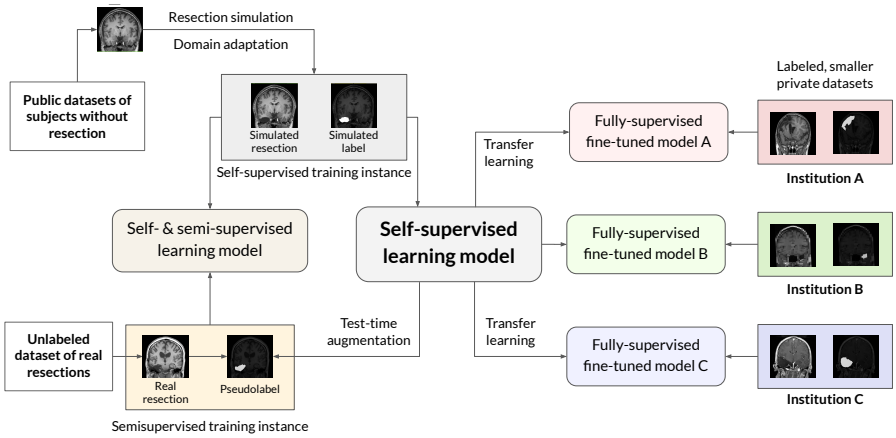


Fig. 1: Learning strategy. 3D images without resections (top left) are used with our resection simulation method to generate training instances. These instances are used to train a baseline model in a self-supervised manner (middle). The baseline model is used to generate pseudolabels from an unlabeled dataset of images from patients who underwent resective surgery (bottom left). Instances from the self-supervised training and pseudolabeled data are used to train new model in a self- and semi-supervised learning setting (left). The baseline model may be finetuned to improve its performance on small labeled datasets of real resections from a single institution, using a standard fully-supervised learning approach (right).

2.1.2 Self-supervised learning

Unsupervised learning refers to any learning method without human-annotated labels. In self-supervised learning, there is a large source dataset without labels and a smaller target dataset with labels, i.e., $D_{\mathrm{S}} = \{\boldsymbol{X}_{\mathrm{S}_1}\}_{i=1}^{n_{\mathrm{S}}}$ and $D_{\mathrm{T}} = \{(\boldsymbol{X}_{\mathrm{T}_i}, \boldsymbol{Y}_{\mathrm{T}_i})\}_{i=1}^{n_{\mathrm{T}}}$. The goal is to generate a model $f(\cdot)$ leveraging knowledge from $\mathcal{D}_{\mathrm{S}}$ to perform the downstream task $\mathcal{T}_{\mathrm{T}}$ by generating an intermediate dataset $D_{\mathrm{S}'} = \{(\boldsymbol{X}_{\mathrm{S}'_i}, \boldsymbol{Y}_{\mathrm{S}'_i})\}_{i=1}^{n'_{\mathrm{S}}}$ using a function $\phi : \mathcal{X}_{\mathrm{S}} \rightarrow \{\mathcal{X}_{\mathrm{S}'}, \mathcal{Y}_{\mathrm{S}'}\}$. Typically, some information from $\boldsymbol{X}_{\mathrm{S}}$ is withheld to generate a training instance $(\boldsymbol{X}_{\mathrm{S}'}, \boldsymbol{Y}_{\mathrm{S}'})$. $\mathcal{Y}_{\mathrm{S}'}$ can be used to define a pretext task $\mathcal{T}_{\mathrm{S}'}$ to train the model $f_{\mathrm{S}'}(\cdot)$. Finally, $f_{\mathrm{S}'}(\cdot)$ may be finetuned on D_{T} to perform the downstream task $\mathcal{T}_{\mathrm{T}}$ if labeled data is available. We denote $\tilde{f}_{\mathrm{AB}}(\cdot)$ as the model initially trained on a source domain $\mathcal{D}_{\mathrm{A}}$ and finetuned on a target domain $\mathcal{D}_{\mathrm{B}}$.

2.1.3 Semisupervised learning

In a semisupervised learning setting, we assume the presence of a labeled dataset $D_{\mathrm{A}} = \{(\boldsymbol{X}_{\mathrm{A}_i}, \boldsymbol{Y}_{\mathrm{A}_i})\}_{i=1}^{n_{\mathrm{A}}}$ and an unlabeled dataset $D_{\mathrm{B}} = \{\boldsymbol{X}_{\mathrm{B}_i}\}_{i=1}^{n_{\mathrm{B}}}$ from two similar domains $\mathcal{D}_{\mathrm{A}}$ and $\mathcal{D}_{\mathrm{B}}$. As manual annotations are expensive,

typically $n_{\mathrm{A}} \ll n_{\mathrm{B}}$. A predictive model $f_{\mathrm{A}}(\cdot)$ is trained using D_{A}. This model can then be used to generate pseudolabels from D_{B}, creating a new dataset

$$D'_{\mathrm{B}} = \{(\boldsymbol{X}_{\mathrm{B}_i}, f_{\mathrm{A}}(\boldsymbol{X}_{\mathrm{B}_i}))\}_{i=1}^{n'_{\mathrm{B}}} = \{(\boldsymbol{X}_{\mathrm{B}_i}, \widetilde{\boldsymbol{Y}}_{\mathrm{B}_i})\}_{i=1}^{n'_{\mathrm{B}}} \tag{1}$$

The reliability of the pseudolabeled instances may be assessed by estimating the prediction uncertainty $u(f'_{\mathrm{A}}, \boldsymbol{X}_{\mathrm{B}}, N)$, and only instances considered reliable enough will be included in D'_{B}, therefore $n'_{\mathrm{B}} \leq n_{\mathrm{B}}$. f'_{A} is a predictive function with stochastic behaviour modified. The stochastic behaviour can be derived from test-time augmentation (TTA) [61], where a transformation is applied to $\boldsymbol{X}_{\mathrm{B}}$ before (and sometimes after) prediction, or test-time dropout (TTD) [12], where weights in f_{A} are randomly zeroed [57]. N is the number of iterations used to estimate the uncertainty.

Finally, a new model f_{AB} is trained on $D_{\mathrm{A}} \cup D'_{\mathrm{B}}$ and evaluated on a test dataset.

2.2 Resection simulation for self-supervised learning

Let $\mathcal{D}_{\mathrm{C}}$ be the domain corresponding to brain MRI of subjects without a resection cavity (controls) and $\mathcal{D}_{\mathrm{P}}$ the domain corresponding to real postoperative images (with a resection cavity). To reduce the domain gap between $\mathcal{D}_{\mathrm{C}}$ and $\mathcal{D}_{\mathrm{P}}$, we present a function $\phi_{\mathrm{R}} : \mathcal{X}_{\mathrm{C}} \rightarrow \{\mathcal{X}_{\mathrm{R}}, \mathcal{Y}_R\} \simeq \{\mathcal{X}_{\mathrm{P}}, \mathcal{Y}_P\}$ that takes $\boldsymbol{X}_{\mathrm{C}}$, the MRI of a control subject, and generates a training instance $(\boldsymbol{X}_{\mathrm{R}}, \boldsymbol{Y}_{\mathrm{R}})$ composed of an image with a simulated resection cavity and a corresponding label map representing the cavity segmentation. ϕ_{R} generates the training instance using a shape model to determine the location of the resection cavity and a texture model to simulate realistic patterns inside and around the cavity. Instances from D_R are generated to train a resection cavity segmentation model $f_R(\cdot)$. Within our framework, we expect $\mathcal{D}_{\mathrm{R}}$ to approximately model $\mathcal{D}_{\mathrm{P}}$ and therefore $f_{\mathrm{R}} = f_{\mathrm{P}}$.

In the following sections, we will use definitions and notation from [48] to describe the image processing steps used for ϕ_{R}. We present the tools related to shape and texture generation in Sections 2.2.1, 2.2.2 and 2.2.4. We describe how ϕ_{R} is applied to control subjects in Sections 2.2.3 and 2.2.5 to 2.2.7.

2.2.1 Initial cavity shape perturbed with simplex noise

To simulate a realistic resection cavity, we considered the properties of resections: the cavity is a single, continuous volume whose shape is generally not be smooth. We first generate a geodesic polyhedron with frequency f by subdividing the edges of an icosahedron f times and projecting each vertex onto a parametric sphere with a unit radius. This polyhedron models a spherical surface $S = \{V, F\}$ with vertices $V = \left\{\boldsymbol{v}_i \in \mathbb{R}^3\right\}_{i=1}^{n_V}$ and faces $F = \{\boldsymbol{f}_k\}_{k=1}^{n_F}$. Each face $\boldsymbol{f}_k = \{i_1^k, i_2^k, i_3^k\}$ is defined as a sequence of three non-repeated vertex indices. S is centered at the origin.

To create a non-smooth surface, S is perturbed with simplex noise [45], a smooth procedural noise generated by interpolating pseudorandom gradients defined on a multidimensional simplicial grid. We chose this type of noise as it simulates natural-looking textures or terrains and is computationally efficient for multiple dimensions.

The noise at point $\boldsymbol{p} \in \mathbb{R}^3$ is a weighted sum of the noise contribution at ω different octaves, with weights $\gamma^{n-1} : n \in \{1, 2, \ldots, \omega\}$ controlled by the persistence parameter γ. The displacement $\delta : \mathbb{R}^3 \to [-1, 1]$ in mm is proportional to the noise function $\xi : \mathbb{R}^3 \to [-\tau, \tau]$:

$$\delta(\boldsymbol{p}) = \tau\xi\left(\frac{\boldsymbol{p}+\boldsymbol{\mu}}{\zeta}, \omega, \gamma\right) \tag{2}$$

where τ controls the noise amplitude, ζ is a scaling parameter to control smoothness and $\boldsymbol{\mu}$ is a shifting parameter that adds stochasticity (equivalent to a random number generator seed).

Each vertex $\boldsymbol{v}_i \in V$ is displaced radially:

$$\boldsymbol{v}_{\delta i} = \boldsymbol{v}_i + \delta(\boldsymbol{v}_i)\frac{\boldsymbol{v}_i}{\|\boldsymbol{v}_i\|}, \qquad \forall i \in \{1, 2, \ldots, n_V\} \tag{3}$$

to create a perturbed sphere $S_\delta = \{V_\delta, F\}$ with vertices $V_\delta = \{\boldsymbol{v}_{\delta i}\}_{i=1}^{n_V}$.

Next, a series of transforms is applied to S_δ to modify its volume, shape and position, as follows. Let $T_{\mathrm{T}}(\boldsymbol{p})$, $T_{\mathrm{S}}(\boldsymbol{s})$ and $T_{\mathrm{R}}(\boldsymbol{\theta})$ be translation, scaling and rotation transforms. Random rotations around each axis are applied to S_δ with the rotation transform $T_{\mathrm{R}}(\boldsymbol{\theta}_{\mathrm{r}}) = R_x(\theta_x) \circ R_y(\theta_y) \circ R_z(\theta_z)$, where $\circ$ indicates a transform composition, $R_i(\theta_i)$ is a rotation of θ_i radians around axis i, and $\theta_i \sim \mathcal{U}(0, 2\pi)$. A scaling transform $T_{\mathrm{S}}(\boldsymbol{r})$ is applied to S_δ, where $(r_1, r_2, r_3) = \boldsymbol{r}$ are the semiaxes of an ellipsoid with volume v modeling the cavity shape. The semiaxes are computed as $r_1 = r$, $r_2 = \lambda r$ and $r_3 = r/\lambda$, where $r = (3v/4)^{1/3}$ and λ controls the semiaxes length ratios[1]. The transforms are composed as $T_{\mathrm{E}} = T_{\mathrm{S}}(\boldsymbol{r}) \circ T_{\mathrm{R}}(\boldsymbol{\theta}_{\mathrm{r}})$ and applied to S_δ to obtain the resection surface $S_{\mathrm{E}} = T_{\mathrm{E}} \circ S_\delta$. This will define the volume and extent of the initial resection cavity surface S_{E}.

2.2.2 Shape restrictions

Let $\boldsymbol{M}_{\mathrm{A}} : \Omega \to \{0, 1\}$ be a binary image where positive voxels are candidates for the center of a mesh surface S_0, centered at the origin. Let $\boldsymbol{M}_{\mathrm{B}} : \Omega \to \{0, 1\}$ be a binary image that will restrict the final shape, where $\Omega \in \mathbb{R}^3$.

We define a function $\xi(\boldsymbol{M}_{\mathrm{A}}, \boldsymbol{M}_{\mathrm{B}}, S_0)$ that returns an image $\boldsymbol{M}_{\mathrm{R}}$ representing the shape S_0 centered on a positive voxel from $\boldsymbol{M}_{\mathrm{A}}$, restricted by positive voxels in $\boldsymbol{M}_{\mathrm{B}}$.

$\xi(\cdot)$ is computed by first selecting a random voxel $\boldsymbol{a}$ such that $\boldsymbol{M}_{\mathrm{A}}(\boldsymbol{a}) = 1$. Next, a translation transformation $T_{\mathrm{T}}(\boldsymbol{p})$ is defined to translate a point by $\boldsymbol{p}$.

[1] Note the volume of an ellipsoid with semiaxes (a, b, c) is $v = \frac{4}{3}\pi abc$.

(a) S_A on $\boldsymbol{M}_{\mathrm{A}}$ (b) S_A on $\boldsymbol{M}_{\mathrm{B}}$ (c) $\boldsymbol{M}_{\mathrm{R}} = \boldsymbol{M}_{S_A} \odot \boldsymbol{M}_{\mathrm{B}}$

Fig. 2: Example of processing steps in $\xi(\boldsymbol{M}_{\mathrm{A}}, \boldsymbol{M}_{\mathrm{B}}, S_0)$. S_A is derived from S_0 and is centered on a random positive voxel of $\boldsymbol{M}_{\mathrm{A}}$ (a). $\boldsymbol{M}_{S_A}$ is a binary mask derived from S_A. Then, the intersection of $\boldsymbol{M}_{S_A}$ and $\boldsymbol{M}_{\mathrm{B}}$ (b) is computed to get $\boldsymbol{M}_{\mathrm{R}}$ (c)

This transformation is applied to S so that $S_A = T_{\mathrm{T}}(\boldsymbol{a}) \circ S$ is centered on the randomly chosen voxel.

A binary image $\boldsymbol{M}_{S_A} : \Omega \rightarrow \{0, 1\}$ is generated from S_A such that $\boldsymbol{M}_{S_A}(\boldsymbol{p}) = 1$ for all $\boldsymbol{p}$ within S_A and $\boldsymbol{M}_{S_A}(\boldsymbol{p}) = 0$ outside. Finally, $\boldsymbol{M}_{S_A}$ is restricted by $\boldsymbol{M}_{\mathrm{B}}$ so that $\boldsymbol{M}_{\mathrm{R}} = \boldsymbol{M}_{S_A} \odot \boldsymbol{M}_{\mathrm{B}}$, where $\odot$ represents the Hadamard product. Fig. 2 shows an example of the process.

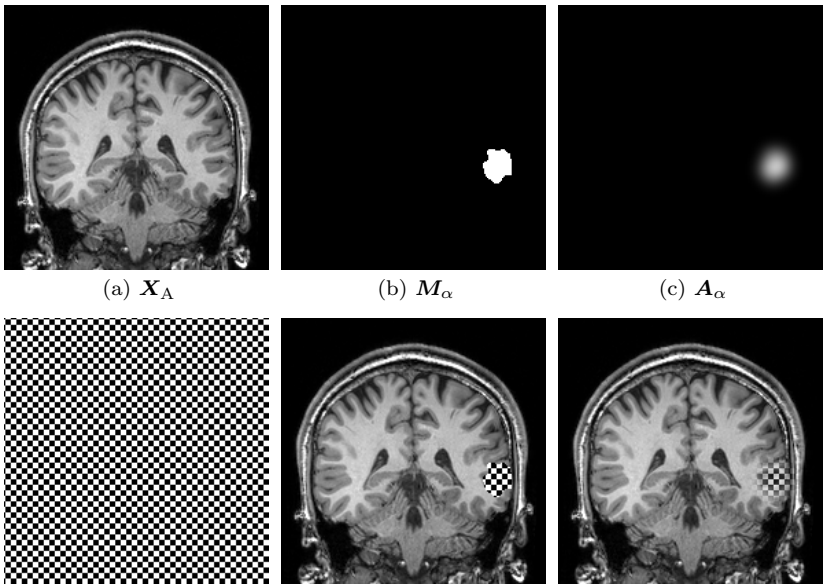
2.2.3 Ensuring the cavity stays within the brain

A T1w MRI is defined as $\boldsymbol{X}_{\mathrm{C}} : \Omega \rightarrow \mathbb{R}$. A full brain parcellation $\boldsymbol{P} : \Omega \rightarrow Z$ is generated [4] for $\boldsymbol{X}_{\mathrm{C}}$, where Z is the set of segmented brain structures. A cortical gray matter mask $\boldsymbol{M}^h_{\mathrm{GM}} : \Omega \rightarrow \{0, 1\}$ of hemisphere h is extracted from $\boldsymbol{P}$, where h is randomly chosen from $H = \{\mathrm{left}, \mathrm{right}\}$ with equal probability.

The simulated resection cavity should not span both hemispheres or include extracerebral tissues such as bone or scalp. To eliminate unrealistic regions, a 'resectable hemisphere mask' is generated from $\boldsymbol{P}$ and h as $\boldsymbol{M}^h_{\mathrm{R}}(\boldsymbol{p}) = 1$ if $\boldsymbol{P}(\boldsymbol{p}) \neq \{M_{\mathrm{BG}}, M_{\mathrm{BT}}, M_{\mathrm{CB}}, M_{\hat{h}}\}$ and 0 otherwise, where M_{BG}, M_{BT}, M_{CB} and $M_{\hat{h}}$ are the sets of labels in Z corresponding to the background, brainstem, cerebellum and contralateral hemisphere, respectively. $\boldsymbol{M}^h_{\mathrm{R}}$ is smoothed using a series of binary morphological operations. The cavity label is then computed by $\boldsymbol{Y}_{\mathrm{cavity}} = \xi(\boldsymbol{M}^h_{\mathrm{GM}}, \boldsymbol{M}^h_{\mathrm{R}}, S_{\mathrm{E}})$ (see Section 2.2.2 and Fig. 2).

2.2.4 Image blending for realistic texture simulation

Let $\boldsymbol{X}_{\mathrm{A}} : \Omega \rightarrow \mathbb{R}$ and $\boldsymbol{X}_{\mathrm{B}} : \Omega \rightarrow \mathbb{R}$ be two scalar-valued images. We use a binary image $\boldsymbol{M}_{\alpha} : \Omega \rightarrow \{0, 1\}$ to determine how to blend $\boldsymbol{X}_{\mathrm{B}}$ into $\boldsymbol{X}_{\mathrm{A}}$ as follows. A Gaussian filter is applied to $\boldsymbol{M}_{\alpha}$ to obtain a smooth alpha channel $\boldsymbol{A}_{\alpha} : \Omega \rightarrow [0, 1]$ defined as $\boldsymbol{A}_{\alpha} = \boldsymbol{M}_{\alpha} * \boldsymbol{G}_{\mathcal{N}}(\boldsymbol{\sigma})$, where $*$ is the convolution operator and $\boldsymbol{G}_{\mathcal{N}}(\boldsymbol{\sigma})$ is a 3D Gaussian kernel with standard deviations $\boldsymbol{\sigma} =$



(a) $\boldsymbol{X}_{\mathrm{A}}$ (b) $\boldsymbol{M}_{\alpha}$ (c) $\boldsymbol{A}_{\alpha}$

(d) $\boldsymbol{X}_{\mathrm{B}}$ (e) $\alpha(\boldsymbol{X}_{\mathrm{A}}, \boldsymbol{X}_{\mathrm{B}}, \boldsymbol{M}_{\alpha}, 0)$ (f) $\alpha(\boldsymbol{X}_{\mathrm{A}}, \boldsymbol{X}_{\mathrm{B}}, \boldsymbol{M}_{\alpha}, 5)$

Fig. 3: Example of processing steps in $\alpha(\boldsymbol{X}_{\mathrm{A}}, \boldsymbol{X}_{\mathrm{B}}, \boldsymbol{M}_{\alpha}, \boldsymbol{\sigma})$. Two scalar-valued images $\boldsymbol{X}_{\mathrm{A}}$ (a) and $\boldsymbol{X}_{\mathrm{B}}$ (d) are blended using $\boldsymbol{M}_{\alpha}$ (b) and $\sigma_i = 0$ mm to create a mix with hard boundaries (e) and $\sigma_i = 5$ mm (c) to create a mix with soft boundaries (f), mimicking partial-volume effects

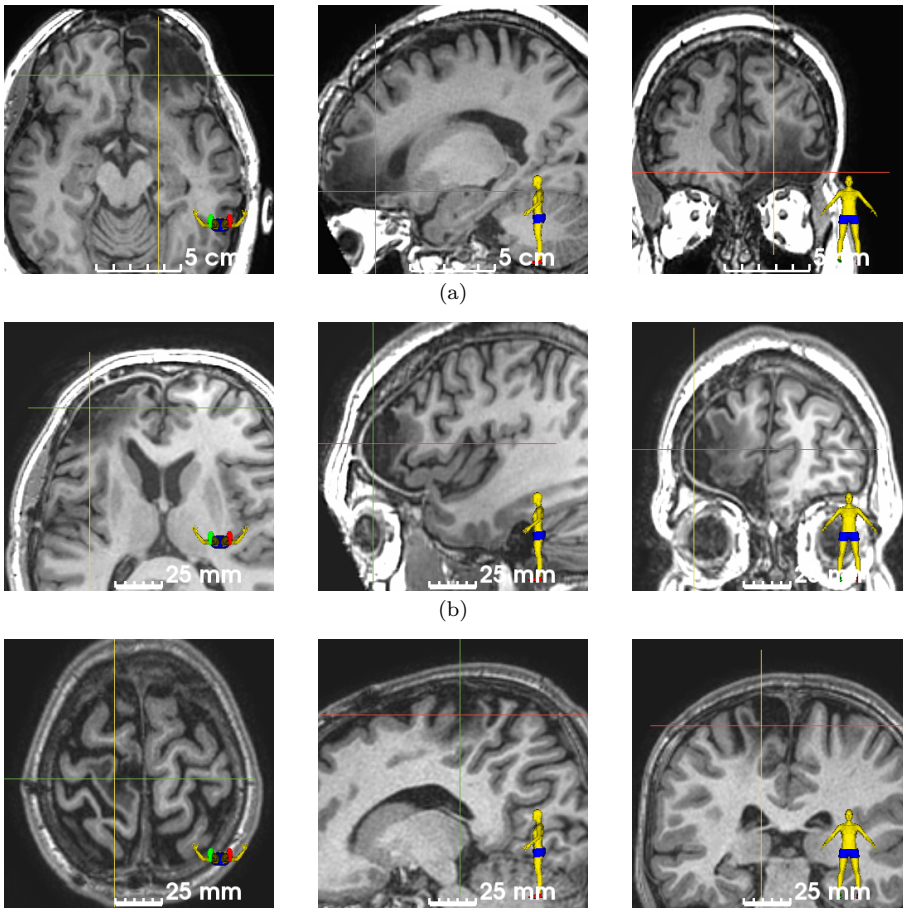
$(\sigma_x, \sigma_y, \sigma_z)$ mm. Then, the two images are blended by the convex combination

$$\boldsymbol{X}_{AB} = \boldsymbol{A}_{\alpha} \odot \boldsymbol{X}_{\mathrm{B}} + (1 - \boldsymbol{A}_{\alpha}) \odot \boldsymbol{X}_{\mathrm{A}} \tag{4}$$

We indicate this texture blending process by $\alpha(\boldsymbol{X}_{\mathrm{A}}, \boldsymbol{X}_{\mathrm{B}}, \boldsymbol{M}_{\alpha}, \boldsymbol{\sigma})$ (Fig. 3).

2.2.5 Simulating cavities filled with CSF

Brain resection cavities are normally filled with CSF. To generate a realistic CSF texture, we create a ventricle mask $\boldsymbol{M}_{\mathrm{V}} : \Omega \rightarrow \{0, 1\}$ from $\boldsymbol{P}$, such that $\boldsymbol{M}_{\mathrm{V}}(\boldsymbol{p}) = 1$ for all $\boldsymbol{p}$ within the ventricles and $\boldsymbol{M}_{\mathrm{V}}(\boldsymbol{p}) = 0$ outside. Intensity values within ventricles are assumed to have a normal distribution [20] with a mean μ_{CSF} and standard deviation σ_{CSF} calculated from voxel intensity values in $\boldsymbol{X}_{\mathrm{C}}(\boldsymbol{p}) : \forall \boldsymbol{p} \in \Omega$ where $\boldsymbol{M}_{\mathrm{V}}(\boldsymbol{p}) = 1$. A CSF-like image is then generated as $\boldsymbol{X}_{\mathrm{CSF}}(\boldsymbol{p}) \sim \mathcal{N}(\mu_{\mathrm{CSF}}, \sigma_{\mathrm{CSF}}), \forall \boldsymbol{p} \in \Omega$, and the resected image is $\boldsymbol{X}_{\mathrm{cavity}} = \alpha(\boldsymbol{X}_{\mathrm{C}}, \boldsymbol{X}_{\mathrm{CSF}}, \boldsymbol{Y}_{\mathrm{cavity}}, \boldsymbol{\sigma}_{\mathrm{cavity}})$ (see Section 2.2.4). We use $\boldsymbol{\sigma}_{\mathrm{cavity}} > 0$ to mimic partial-volume effects at the resection boundaries.



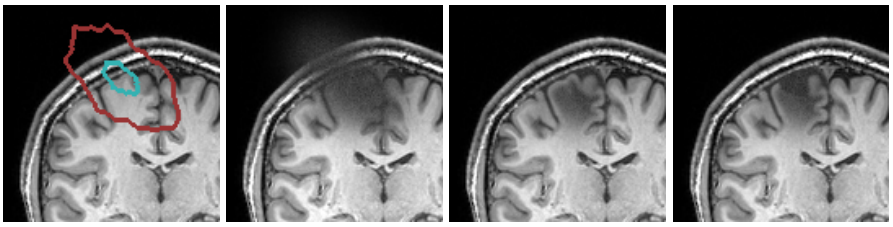
(a) (b) (c)

Fig. 4: Three subjects with white matter hypointensity around the resection cavity. Representative axial (left), sagittal (middle) and coronal (right) slices

2.2.6 Hypointense white matter lesion

Retraction injury from resective surgery may cause ischemic changes and degeneration of white matter. These lesions typically cause white matter hypointensity around the resection cavity (Fig. 4).

To simulate a white matter lesion, we first resize the binary image corresponding to the resection cavity (Section 2.2.1): $\boldsymbol{M}_{\mathrm{WM}} = T_{\mathrm{S}}(s_{\mathrm{WM}}) \circ \boldsymbol{Y}_{\mathrm{cavity}}$, where $T_{\mathrm{S}}(s_{\mathrm{WM}})$ is an isotropic scaling transform that will increase the size of the binary component in $\boldsymbol{Y}_{\mathrm{cavity}}$ corresponding to the cavity by a factor of s_{WM}. Then, a Gaussian kernel with a large standard deviation $\boldsymbol{\sigma}_{\mathrm{WM}}$ is used to generate a simulated white matter lesion: $\boldsymbol{X}'_{\mathrm{WM}} = \alpha(\boldsymbol{X}_{\mathrm{C}}, \boldsymbol{X}_{\mathrm{CSF}}, \boldsymbol{M}_{\mathrm{WM}}, \boldsymbol{\sigma}_{\mathrm{WM}})$.



(a) $\boldsymbol{X}_{\mathrm{C}}$ (b) $\boldsymbol{X}'_{\mathrm{WM}}$ (c) $\boldsymbol{X}_{\mathrm{WM}}$ (d) $\boldsymbol{X}_{\mathrm{WM,cavity}}$

Fig. 5: Simulation of postoperative white matter lesions. The binary image representing the cavity (blue) is scaled up (red) (a). A CSF-like image is blended with the MRI using a large standard deviation for the Gaussian kernel (b). The lesion is restricted so that it affects only the white matter (c). The cavity is added as explained in Section 2.2.1 (d)

For realism, we restrict the simulated lesion to the white matter voxels only. We use $\boldsymbol{P}$ to generate a white matter mask $\boldsymbol{M}^h_{\mathrm{WM}}$ and blend $\boldsymbol{X}'_{\mathrm{WM}}$ with the original image: $\boldsymbol{X}_{\mathrm{WM}} = \alpha(\boldsymbol{X}_{\mathrm{C}}, \boldsymbol{X}'_{\mathrm{WM}}, \boldsymbol{M}^h_{\mathrm{WM}}, \boldsymbol{\sigma}_{\mathrm{GM}})$. Finally, we simulate the resection cavity on $\boldsymbol{X}_{\mathrm{WM}}$, similar to Section 2.2.1: $\boldsymbol{X}_{\mathrm{WM,cavity}} = \alpha(\boldsymbol{X}_{\mathrm{WM}}, \boldsymbol{X}_{\mathrm{CSF}}, \boldsymbol{Y}_{\mathrm{cavity}}, \boldsymbol{\sigma}_{\mathrm{cavity}})$. The process is illustrated in Fig. 5.

2.2.7 Hyperintense blood products

Hyperintense postoperative blood products are often found in brain resection cavities (Fig. 6).

To simulate these hyperintensities, we first generate a perturbed sphere S_{BP} (Section 2.2.1), resize it using a random scaling transform $T_{\mathrm{S}}(\boldsymbol{s}_{\mathrm{BP}})$ and apply a random rotation $T_{\mathrm{R}}(\boldsymbol{\theta}_{\mathrm{BP}})$ to obtain the final shape S_{BP}. To ensure that blood products are not in contact with tissue outside the resection cavity, we erode $\boldsymbol{Y}_{\mathrm{cavity}}$ to obtain $\boldsymbol{Y}'_{\mathrm{cavity}}$, which is used to restrict S_{BP}. The blood product binary image is $\boldsymbol{M}_{\mathrm{BP}} = \xi(\boldsymbol{Y}'_{\mathrm{cavity}}, \boldsymbol{Y}'_{\mathrm{cavity}}, S_{\mathrm{BP}})$ (see Section 2.2.2).

To simulate the hyperintense texture, we generate a new image $\boldsymbol{X}_{\mathrm{BP}} \sim \mathcal{N}(\mu_{\mathrm{BM}}, \sigma_{\mathrm{CSF}})$. We typically choose μ_{BM} to be a high percentile of the intensity values in $\boldsymbol{X}_{\mathrm{C}}$. The simulated resection cavity including a blood product is $\boldsymbol{X}_{\mathrm{cavity,BP}} = \alpha(\boldsymbol{X}_{\mathrm{cavity}}, \boldsymbol{X}_{\mathrm{BP}}, \boldsymbol{M}_{\mathrm{BP}}, \boldsymbol{\sigma}_{\mathrm{BP}})$.

2.3 Leveraging unlabeled images for semisupervised learning

2.3.1 Data distillation

Data distillation is a method that ensembles predictions from multiple transformations applied to data, using a single model, to generate pseudolabels [50]. We perform Monte Carlo simulation to generate each pseudolabel using TTA, as this method is known to improve the performance of segmentation models [40]. Let N represent the total number of simulation runs. In the n-th

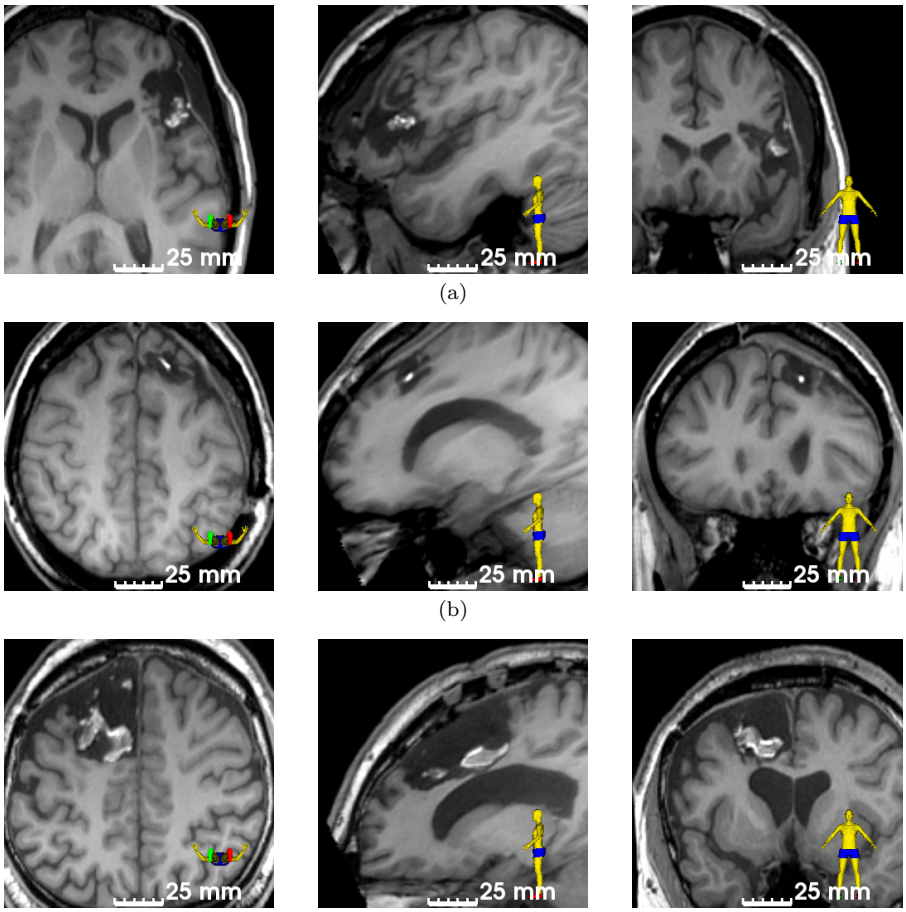


Fig. 6: Three subjects with postoperative blood products inside the resection cavity. Representative axial (left), sagittal (middle) and coronal (right) slices

simulation run, a trained model $f(\cdot)$ is used to compute the probability that each voxel in an MRI $\boldsymbol{X}$ represents a resection cavity:

$$\boldsymbol{Y}'_n = T_\beta^{-1}\left[f\left(T_\beta(T_\alpha(\boldsymbol{X}))\right)\right] = f_{\alpha\beta}(\boldsymbol{X}) \tag{5}$$

where the TTA transforms T_α and T_β represent the intensity and spatial transforms used for data augmentation during training, respectively (see Section 3.3.2), and $f_{\alpha\beta}$ represents the composition of the transforms and the trained model. We ensure that T_β is invertible by using diffeomorphic spatial transformations. To preserve image quality and ensure that probabilities stay within $[0, 1]$, we use tricubic and trilinear interpolation for T_β and T_β^{-1}, respectively.

The predictions are then averaged to obtain $\boldsymbol{Y}' : \Omega \to [0, 1]$

$$\boldsymbol{Y}' = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{Y}_n \tag{6}$$

and the corresponding binary pseudolabel $\widetilde{\boldsymbol{Y}} : \Omega \to \{0, 1\}$ is obtained applying a threshold of 0.5 to $\boldsymbol{Y}'$.

2.3.2 Uncertainty estimation

To ensure that only pseudolabels with high reliability are used for training, we estimate the subject-level prediction uncertainty $u(f_{\alpha\beta}, \boldsymbol{X}, N)$ (Section 2.1.3). We use the N TTA predictions to estimate aleatoric uncertainty, which is expected to capture noise inherent in the observation [30]. Aleatoric uncertainty has been shown to be a good indicator of segmentation quality and used successfully as a selection criterion for pseudolabels in semisupervised learning settings for medical image segmentation [61,59]. Moreover, it can be used to inform users about the reliability of the inferred segmentations.

For N samples from the Monte Carlo simulation, let $L = \{l_n\}_{n=1}^{N}$ denote the set of (soft) volumes of the segmented cavity, where l_n is the sum of all probabilities in the result of the n-th sample $\boldsymbol{Y}'_n$. We use the coefficient of quartile variation (CQV) of the volumes [65,61] to estimate the image-level uncertainty $u : L \to [0, 1]$:

$$u = \frac{q_3 - q_1}{q_3 + q_1} \tag{7}$$

where q_1 and q_3 are the first and third quartiles of L, respectively. The CQV is agnostic to the volume of the segmented resection cavity and therefore circumvents the bias introduced by naturally-occurring uncertainty along the resection boundaries [29]. A threshold t_{u} is used to select images with low uncertainty that will be added to the self-supervised dataset for the semisupervised setting.

3 Experiments and results

3.1 Data

The datasets we used for experiments are summarized in Table 1.

3.1.1 Public data for simulation

T1w magnetic resonance (MR) images were collected from publicly available datasets Information eXtraction from Images (IXI)[2], Alzheimer's Disease Neuroimaging Initiative (ADNI) [26], and Open Access Series of Imaging Studies (OASIS) [32], for a total of 1813 images.

[2] `https://brain-development.org/ixi-dataset/`

Table 1: Datasets used in this study. If multiple resolutions are present, the minimum, mean and maximum along each dimension are shown. 'Gad' indicates that gadolinium, a contrast enhancement agent, was used

Dataset	Modality	Resolution (mm)	Subjects	Surgery	Annotated
IXI	T1w	$0.94 \times 0.94 \times 1.20$ $0.94 \times 0.94 \times 1.20$ $0.98 \times 0.98 \times 1.20$	566	-	-
ADNI	T1w	$1.00 \times 1.00 \times 1.00$	467	-	-
OASIS	T1w	$1.00 \times 1.00 \times 1.00$ $1.05 \times 1.01 \times 1.02$ $1.20 \times 1.05 \times 3.00$	780	-	-
EPISURG	T1w	$0.75 \times 0.75 \times 0.75$ $0.96 \times 0.96 \times 1.08$ $1.09 \times 1.09 \times 1.60$	430	Epilepsy	133
Milan	T1w	$0.46 \times 0.46 \times 0.90$	20	Epilepsy	20
Strasbourg	T1w & T1w Gad	$0.23 \times 0.23 \times 0.50$ $0.61 \times 0.61 \times 2.79$ $1.00 \times 1.00 \times 5.00$	33	Epilepsy	33
Paris	T1w	$0.47 \times 0.47 \times 0.49$ $0.82 \times 0.76 \times 1.06$ $1.20 \times 0.98 \times 1.20$	19	Epilepsy	19
BITE	T1w Gad	$1.00 \times 0.47 \times 0.47$ $2.31 \times 0.53 \times 0.53$ $5.50 \times 0.55 \times 0.55$	13	Tumor	0

For self-supervised learning, publicly available data is used to build a dataset $D_{\mathrm{C}} = \{\boldsymbol{X}_{\mathrm{C}_i}\}_{i=1}^{n_{\mathrm{C}}}$ corresponding to control subjects (Section 2.1). Note that we use the term 'control' to refer to subjects that have not undergone resective surgery, but they may have other neurological conditions. For example, subjects in ADNI suffer from Alzheimer's disease.

3.1.2 EPISURG dataset

We curated the EPISURG dataset using images from patients with refractory focal epilepsy who underwent resective surgery between 1990 and 2018 at the National Hospital for Neurology and Neurosurgery (NHNN), London, United Kingdom. This was an analysis of anonymized data that had been previously acquired as a part of clinical care, so individual patient consent was not required. All images in EPISURG were defaced using a predefined face mask in the Montreal Neurological Institute (MNI) space to preserve patient identity. In total, there were 430 patients with postoperative T1w MRI, 269 of which had a corresponding preoperative MRI. The distribution of resection types is shown in Table 2.

Annotations used for evaluation in this study were performed semiautomatically using a fast grow-cut algorithm implemented in 3D Slicer 4.10 [64, 14]. EPISURG is available online and can be freely downloaded for future research [47].

Table 2: Distribution of resection types in EPISURG

Lobe	**Type**	**Subjects**
Temporal	lobectomy	317
Temporal	lesionectomy	30
Temporal-frontal	lobectomy	2
Temporal-parietal	lobectomy	1
Frontal	lobectomy	47
Frontal	lesionectomy	10
Parietal	lesionectomy	11
Parietal	lobectomy	4
Occipital-parietal	lobectomy	2
Occipital	lobectomy	2
-	multiple subpial	2
-	hemispherectomy	2
Total		430

3.1.3 Multicentric epilepsy data

We evaluate the generalizability of our learning strategy to data from several institutions (*Milan*, *Strasbourg*, *Paris*) that may use different acquisition protocols and surgical approaches. The same human rater (F.P.G.) annotated all images shared by these institutions using the same protocol as used for EPISURG.

3.1.4 Brain tumor datasets

The Brain Images of Tumors for Evaluation (BITE) dataset [37] consists of T1w MRI with gadolinium contrast enhancement (T1wCE) of 13 patients with brain tumors[3]. We use the postoperative images in BITE to perform a qualitative assessment of the potential of our models to generalize to images from a substantially different domain (as images in BITE are contrast-enhanced) and different pathology and, therefore, potentially different surgical techniques that may effect the resection cavity appearance.

3.1.5 Preprocessing

For all images, the brain was segmented using ROBEX [24]. Voxels within the brain were used to register the images to the nonlinear symmetric ICBM152 MNI template [15,16] using a pyramidal approach to compute the affine transformation [39]. All images were resampled into the MNI space using sinc interpolation to preserve image quality. After resampling, images had a 1-mm isotropic resolution and size $193 \times 229 \times 193$.

[3] We only use the postoperative images from group 3.

3.2 Network architecture and implementation details

We used the PyTorch deep learning framework [44], training with automatic mixed precision (AMP) on two 32-GB TESLA V100 GPUs.

We implemented a variant of 3D U-Net [9] using two downsampling and upsampling blocks, upsampling with trilinear interpolation for the synthesis path, and 1/4 of the filters for each convolutional layer. We used dilated convolutions [7], starting with a dilation factor of one, then increased or decreased in steps of one after each downsampling or upsampling block, respectively. This results in a model with the same receptive field (a cube of length 88 mm) but $\approx 77\times$ fewer parameters (246 156) than the original 3D U-Net, reducing overfitting and computational burden.

He initialization was used for all convolutional layers, followed by batch normalization and nonlinear PReLU activation functions [25, 22] A dropout layer with probability 0.5 [57] was added before the last convolutional layer to reduce overfitting and estimate epistemic uncertainty. We used adaptive moment estimation (AdamW) [31, 33] to adjust the learning rate during training, with weight decay of 10^{-2}, and a learning scheduler that divides the learning rate by 10 every 20 epochs. We optimized our network to minimize the mean soft Dice loss [38] of each minibatch, for all the experiments. A minibatch size of 10 images (5 per GPU) was used for training. Unsupervised training took about 27 hours. Finetuning on a small annotated dataset took about 7 hours.

We used Sacred [19] and TensorBoard [1] to configure, log and visualize our experiments.

3.3 Processing during training

3.3.1 Resection simulation

We perform the resection simulation on the fly, i.e., during training. Simulation requires 0.6 to 2.2 s for a image of size $193 \times 229 \times 193$, depending on the addition of white matter lesions and blood products (Sections 2.2.6 and 2.2.7). In practice, we perform expensive operations such as convolutions on subvolumes to reduce computational burden. The simulation is implemented using SimpleITK [34], VTK [52] and NumPy [60]. To generate the noisy sphere, we used `pyDome`[4] and `noise`[5].

3.3.2 Preprocessing and augmentation

We use TorchIO transforms to load, preprocess and augment our data during training [49]. Instead of preprocessing the images with denoising or bias removal, we simulate different artifacts so that our models are robust to them.

[4] `https://github.com/badassdatascience/pyDome`

[5] `https://github.com/caseman/noise`

Our preprocessing and augmentation transforms are described below. For transforms that are not applied to all images, we show the probability p of the transform being applied.

1. Random resection simulation (for self-supervised training only)
2. Histogram standardization [42]
3. Simulation of low resolution artifacts ($p = 0.75$). Sampled uniformly from
   (a) Random simulation of anisotropic spacing [2] and
   (b) Gaussian blurring with random variance
4. Random simulation of MRI ghosting artifacts [54] ($p = 0.2$)
5. Random simulation of MRI spike artifacts [54] ($p = 0.2$)
6. Random simulation of MRI motion artifacts [53] ($p = 0.2$)
7. Random simulation of bias field inhomogeneity [58] ($p = 0.5$)
8. Standardization to zero-mean and unit variance using only voxels with intensity above the mean to compute the statistics
9. Gaussian noise with random variance ($p = 0.75$)
10. Diffeomorphic spatial transform, sampled with probabilities 0.9 and 0.1 from
    (a) Random rotation and anisotropic scaling and
    (b) Random elastic deformation
11. Random flip around the sagittal plane ($p = 0.5$)
12. Crop images using a tight bounding box around the brain, to a size of $176 \times 216 \times 160$ voxels.

We refer the reader to the GitHub repository for a detailed inspection of the transforms parameters used for our experiments.

3.4 Experiments

All overlap measurements are expressed as 'median (interquartile range)' DSC. No postprocessing is performed for evaluation. We analyzed differences in model performance using a one-tailed Mann-Whitney U test (as DSCs were not normally distributed) with a significance threshold of $\alpha = 0.05$, and a Bonferroni correction for each set of N experiments: $\alpha_{\text{Bonf}} = \frac{\alpha}{N \times (N-1)}$.

3.4.1 Self-supervised learning: training with simulated resections only

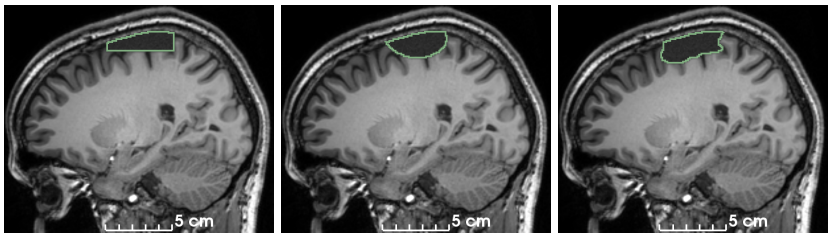
In our first experiment, we assess the relation between the complexity of the resection simulation and segmentation performance. We train using simulated resections on the publicly available dataset $D_{\text{C}} = \{\boldsymbol{X}_{\text{C}_i}\}_{i=1}^{n_{\text{C}}}$, where $n_{\text{C}} = 1813$ (Section 3.1). We use 90% of the images in D_{C} for the training set $D_{\text{Ctrain}} = \{\boldsymbol{X}_{\text{C}_i}\}_{i=1}^{n_{\text{Ctrain}}}$ and 10% for the validation set $D_{\text{Cval}} = \{\boldsymbol{X}_{\text{C}_i}\}_{i=1}^{n_{\text{Cval}}}$ ($n_{\text{Ctrain}} = 1632$ and $n_{\text{Cval}} = 181$). The 133 annotated postoperative images in EPISURG are used for evaluation.

Before training, we precompute and cache a validation set

$$D_{\text{Rval}} = \{T_{\text{Aug}} \circ \phi_{\text{R}}(\boldsymbol{X}_{\text{C}_i})\}_{i=1}^{n_{\text{Cval}}} = \{(\boldsymbol{X}_{\text{R}_i}, \boldsymbol{Y}_{\text{R}_i})\}_{i=1}^{n_{\text{Cval}}} \tag{8}$$

Table 3: Quantitative evaluation on the annotated images in EPISURG of models trained with simulated resections only. DSCs are expressed as 'median (interquartile range)'

White matter lesion	Blood products	Cavity shape	DSC
No	No	Cuboid	57.9 (73.1)
No	No	Ellipsoid	79.0 (20.0)
No	No	Noisy	**80.5 (18.7)**
No	Yes	Noisy	79.6 (16.5)
Yes	No	Noisy	78.2 (20.3)
Yes	Yes	Noisy	78.0 (18.0)



(a) (b) (c)

Fig. 7: Simulation of resection cavities with increasing shape complexity: cuboid (a), ellipsoid (b) and ellipsoid perturbed with simplex noise (c)

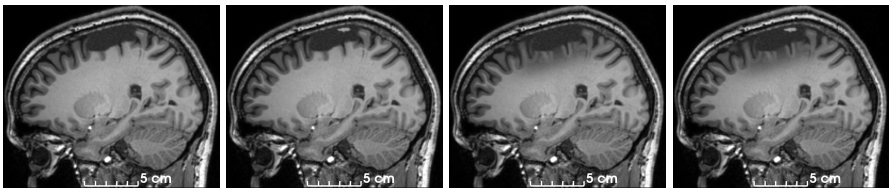
where ϕ_{R} is the resection simulation described in Section 2.2 and T_{Aug} represents the preprocessing and augmentation described in Section 3.3.2.

At each training iteration, b images from D_{Ctrain} are loaded, resected, preprocessed and augmented to obtain a mini-batch of b training instances $\{(\boldsymbol{X}_{\mathrm{R}_i}, \boldsymbol{Y}_{\mathrm{R}_i})\}_{i=1}^{b}$. Note that the resection simulation is performed on the fly, which allows us to ensure that the network never sees the same resection twice.

All models were trained for 60 epochs, using an initial learning rate of 10^{-3}. We use for evaluation the model with the lowest mean validation loss obtained during training.

Effect of resection shape To investigate the effect of the simulated cavity shape on the model performance on real data, we modify ϕ_{R} to generate cuboid- (Fig. 7a) or ellipsoid-shaped (Fig. 7b) resections, and compare the performance with the baseline simulation of a 'noisy' ellipsoid (Fig. 7c). The cuboids and ellipsoid meshes are not perturbed using simplex noise, and cuboids are not rotated. The performance of the baseline model is only marginally better than the model trained with rotated ellipsoids ($p = 0.123$) (Table 3). The model trained with cuboid-shaped resection cavities performed significantly worse than the baseline model ($p < 10^{-8}$).

Effect of resection texture We investigate the effect of simulating additional postoperative phenomena such as white matter lesions around the cavity and



(a) (b) (c) (d)

Fig. 8: Simulation of resection cavities with increasing texture complexity: baseline (a), blood products (b), white matter lesion (c) and both (d)

blood products inside (Fig. 8). Adding either of these effects did not improve results, but the superiority of the baseline model with respect to the models trained with white matter lesions ($p = 0.163$), blood products ($p = 0.323$) or both ($p = 0.054$) was not statistically significant.

3.4.2 Finetuning on small clinical datasets

We assess the generalizability of our baseline model by finetuning it using small datasets from different institutions, with different scanners, voxel resolution and acquisition protocols. Additionally, we finetune the model on 20 cases from EPISURG with the lowest DSC (Section 3.4.1).

For each dataset, we load the pretrained baseline model, initialize the optimizer with an initial learning rate of $5 \cdot 10^{-4}$, initialize the learning rate scheduler and finetune all layers simultaneously for 40 epochs using 5-fold cross-validation. We use model weights from the epoch with the lowest mean validation loss for evaluation. To minimize data leakage, we chose the above hyperparameters using the validation set of one fold in the *Milan* dataset.

We observed a consistent increase in DSC for all finetuned models, up to a maximum of 89.2 (13.3) for the *Milan* dataset. For comparison, inter-rater agreement between human annotators in our previous study was 84.0 (9.9) [48].

Quantitative and qualitative evaluations are illustrated in Figs. 9 and 10, respectively.

3.4.3 Semisupervised learning: leveraging real unlabeled resections

We assess the ability of semisupervised learning to improve the performance of our baseline model. We first computed uncertainty $u(f_{\alpha\beta}, \boldsymbol{X}, N)$ (Section 2.3.2) for all unlabeled images in EPISURG $D_{\mathrm{R}} = \{\boldsymbol{X}_{\mathrm{R}_i}\}_{i=1}^{n_{\mathrm{R}}}$, where $n_{\mathrm{R}} = 297$ (Fig. 11), using the baseline model and the transforms used for preprocessing and augmentation Section 3.3.2. We generated pseudolabels using data distillation (Section 2.3.1) for all images in D_{R} with $u(\boldsymbol{X}_{\mathrm{R}_i}, \cdot) < 0.2$ (Fig. 11a) to obtain $D_{\mathrm{P}} = \{(\boldsymbol{X}_{\mathrm{P}_i}, \widetilde{\boldsymbol{Y}}_{\mathrm{P}_i})\}_{i=1}^{n_{\mathrm{P}}}$, where $n_{\mathrm{P}} = 256$. We computed uncertainty and generated the pseudolabels from 50 Monte Carlo TTA iterations (Fig. 11e).

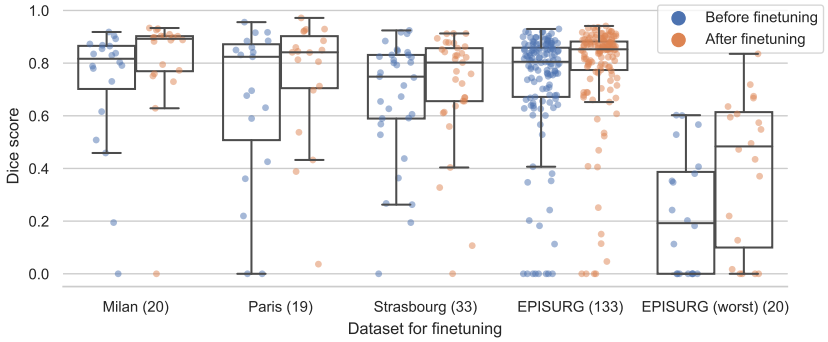


Fig. 9: DSC before (blue) and after (orange) finetuning the self-supervised model on datasets from different institutions. Horizontal lines in the boxes represent the first, second (median) and third quartiles. Numbers in parentheses represent the number of subjects in each dataset

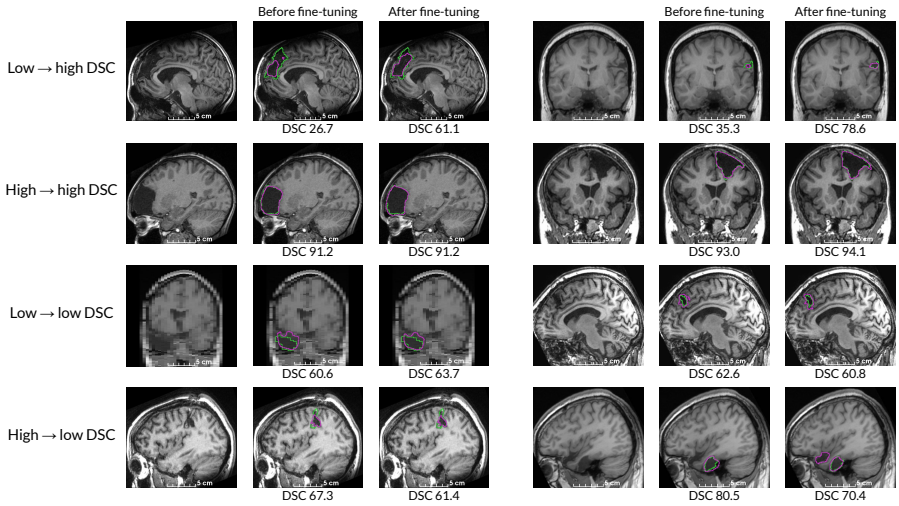


Fig. 10: Qualitative evaluation of finetuning for the *Strasbourg* (left) and EPISURG (right) datasets. Rows correspond, from top to bottom, to cases for which the DSC 1) became much higher, 2) remained high, 3) remained low and 4) became much lower after finetuning the self-supervised model. Manual annotations (green) and thresholded model predictions (magenta) are overlaid. For interpretation of this figure, the reader is referred to the web version of this article

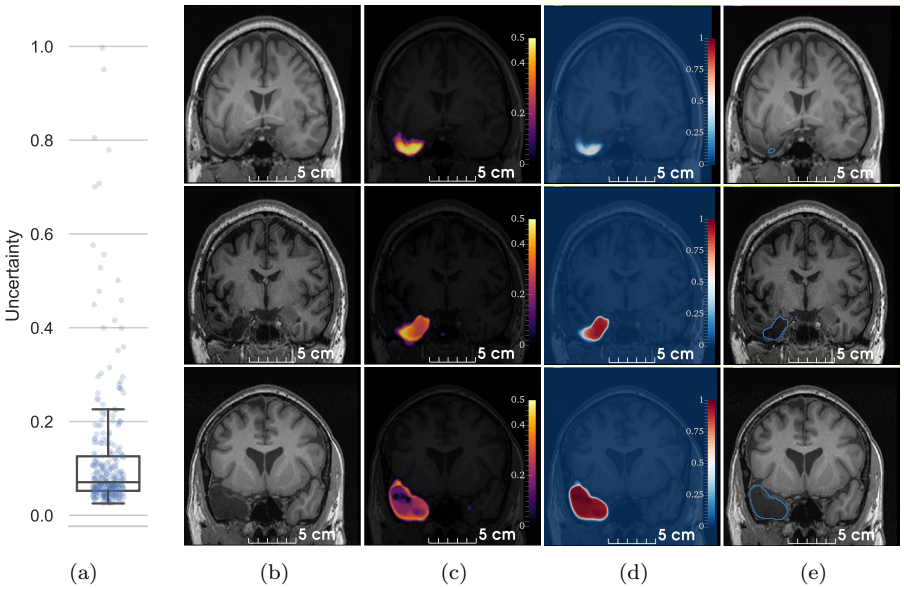


Fig. 11: Generating reliable pseudolabels for semisupervised learning. Image-level uncertainty for the 297 unlabeled postoperative images in EPISURG (a). For each subject, probabilities for the resection cavity are generated from an MRI (b) using 50 TTA Monte Carlo iterations. The voxel-wise uncertainty (c) is estimated as the standard deviation of the probabilities across all iterations. The mean prediction (d) is thresholded at 0.5 to generate the pseudolabel (e) used for semisupervised learning. Image-level uncertainties for the three cases are 0.805 (top), 0.195 (middle) and 0.025 (bottom)

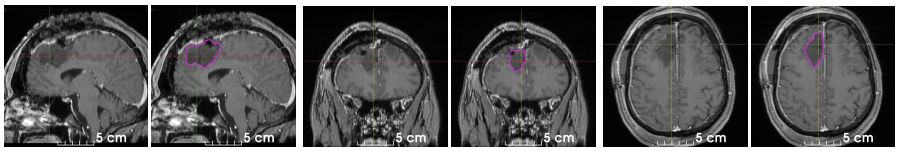
We used the self-supervised training dataset D_{Ctrain} in addition to D_{P} to train a new model $f_{\text{P}}(\cdot)$, using the same hyperparameters as in the self-supervised setting (Section 3.4.1). To ensure that all batches contain real resections, we use $b-b_{\text{P}}$ images from D_{Ctrain} and b_{P} images from D_{P} to compose each minibatch of size b. We chose $b_{\text{P}} = 2$ for our experiments.

Semisupervised learning improved the performance of the baseline model from 80.5 (18.7) to 81.5 (17.8) ($p = 0.474$).

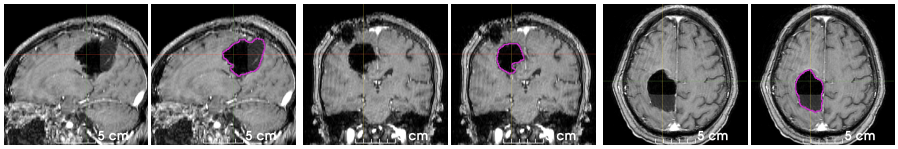
3.4.4 Qualitative evaluation on brain tumor resection dataset

We used the BITE dataset [37] to evaluate the ability of our self-supervised model to segment resection cavities on images from a different institution, modality and pathology with respect to the datasets used for quantitative validation. Probabilities were thresholded at 0.5 and all but the largest binary connected component were removed.

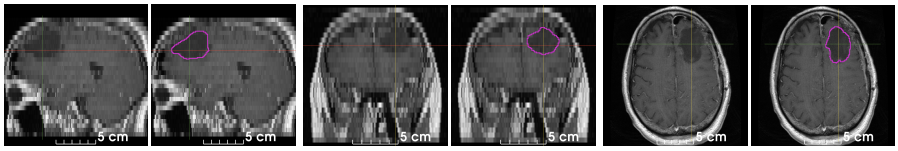
The model was able to successfully segment the cavity on 11/13 images, even though some presented challenging features (Fig. 12).



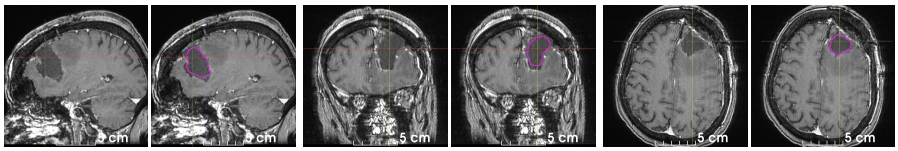
(a)



(b)



(c)



(d)

Fig. 12: Qualitative evaluation of the self-supervised model on a dataset of postoperative brain tumor T1wCE MRI. The model is robust to multiple challenging scenarios: low contrast between the cavity and the brain (a), air and CSF within the resection cavity (b), highly anisotropic resolution (c), motion artifacts and edema (d), and different modality to the one used for training (all)

3.4.5 Qualitative evaluation on intraoperative image

We used our baseline model to segment the resection cavity on a preoperative MRI. Despite the large domain shift between the training dataset and the intraoperative image, which includes a retracted skin flap and a missing bone flap, the model was able to correctly estimate the resection cavity segmentation, discarding other similar regions filled with CSF or air (Fig. 13).

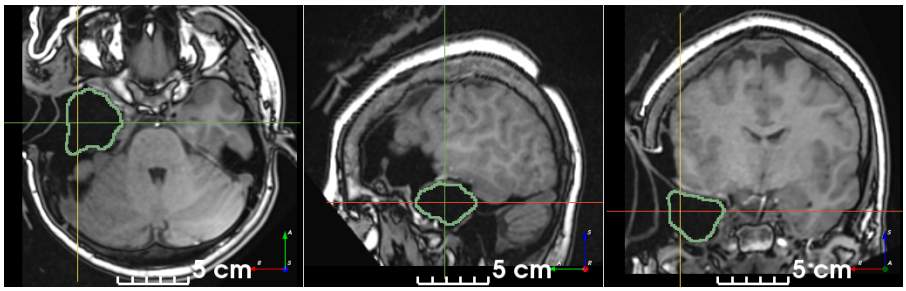


Fig. 13: Qualitative result on an intraoperative MRI. The baseline model correctly discarded regions filled with air or CSF that do not correspond to the resection cavity

4 Discussion and conclusion

This work addresses the challenge of segmenting postoperative brain resection cavities from T1w MRI without using annotated data. The main contributions are 1) a method to simulate resection cavities on normal MRI, 2) a learning strategy using the resection simulation to train without annotated data and 3) a large annotated dataset of pre- and post-operative MRI of refractory epilepsy patients. Our novel approach is conceptually simple, easy to implement and relies on assumptions based on clinical knowledge about postoperative phenomena. The resection simulation is fast enough that it can be executed during training. The trained models do not require any preprocessing such as denoising or bias field correction, thanks to our substantial data augmentation including simulation of multiple MRI artifacts. Moreover, in contrast with related works, we did not perform skull-stripping as it would be affected by the resection cavity. This is especially true in the context of epilepsy surgery, as the EZ is always located in the cortical gray matter and therefore all resections partially overlap with this brain region.

Model performance is poor when the simulated cavities are cuboids, and best results were obtained using ellipsoids perturbed with procedural noise. This indicates that modelling a realistic cavity shape is important, and there is room for improvement by optimizing the hyperparameters of the resection simulation. While the addition of simulated white matter lesions and blood products did not improve model performance, we believe our approach can be easily extended to simulate other objects such as air within the cavity or residual tumors, which would improve performance in the context of brain cancer surgery [13].

Our model generalizes well to clinical data from different institutions and pathologies such as epilepsy and glioma, and may be easily finetuned using small annotated clinical datasets to further improve performance. Moreover, our resection simulation and learning strategy may be trivially extended to

train with arbitrary modalities, or with synthetic modalities generated from brain parcellations [2].

The main causes of failure are very small cavities, where the cavity was not detected, and large brain shift or subdural edema, where this regions were incorrectly classified as resection cavities. The former issue may be overcome using curriculum learning [21], i.e., training using increasingly challenging instances, which can be performed by modifying the hyperparameters of the resection simulation and data augmentation during training. The latter can be addressed by extending our method to simulate these phenomena using biomechanical brain models and nonlinear deformation [18].

We demonstrated that uncertainty estimates may be used as a selection criterion for pseudolabeled images in a semisupervised learning pipeline, and to inform users of segmentation reliability. Our strategy can be adopted by institutions with a large amount of unlabeled data which can be for training, while finetuning and testing may be performed on a smaller labeled dataset.

We showed that our model correctly segmented an intraoperative image, respecting imaginary boundaries between brain and skull, suggesting a good inductive bias of human neuroanatomy. Qualitative results and execution time (< 1 s) suggest that our method could be used intraoperatively to improve registration with preoperative images by masking the cost function using the resection cavity segmentation [3,62].

We curated and release EPISURG, hoping that it will serve as a benchmark dataset for quantitative assessment of resective neurosurgery.

Declarations

Availability of data and material

EPISURG can be freely downloaded from the UCL Research Data Repository [47]: `https://doi.org/10.5522/04/9996158`.

Code availability

The code for resection simulation is available online[6] and can be installed from Python Package Index (PyPI) using Pip Installs Packages (PIP) with a single line of code on any Windows, macOS and Linux machine: `pip install resector`. The simulation integrates seamlessly with our preprocessing and augmentation framework, as it is implemented as a TorchIO transform [49].

The scripts used for training and evaluation and the trained models are also available on GitHub[7].

We wrote an open-source tool to segment the resection cavity using Python or the command line. It can be installed running `pip install resseg`.

[6] `https://github.com/fepegar/resector`

[7] `https://github.com/fepegar/ijcars-2020-resseg`

Authors' contributions

Conceptualization: F.P.G., R.S., J.S.D. and S.O.; *Methodology*: F.P.G. and R.S.; *Software, Formal Analysis, Investigation* and *Visualization*: F.P.G.; *Resources*: F.P.G., M.R., F.C., V.F., K.L., V.N., C.E., I.O. and J.S.D.; *Data Curation*: F.P.G. and J.S.D.; *Writing — Original Draft*: F.P.G.; *Writing – Review & Editing*: F.P.G., R.S., J.S.D. and S.O.; *Supervision*: R.S., J.S.D. and S.O.; *Project Administration*: J.S.D. and S.O.; *Funding Acquisition*: R.S., J.S.D. and S.O.

Funding

This publication represents, in part, independent research commissioned by the Wellcome Innovator Award (218380/Z/19/Z/). The views expressed in this publication are those of the authors and not necessarily those of the Wellcome Trust.

Computing infrastructure at the Wellcome / EPSRC Centre for Interventional and Surgical Sciences (WEISS) (UCL) was used for this study.

Conflicts of interest

The authors declare that they have no conflict of interest.

Research involving human participants

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. For this type of study, formal consent is not required.

Informed Consent

Informed consent was obtained from all individual participants included in the study.

Acknowledgements Some of the data used in preparation of this article was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at `http://adni.loni.usc.edu/`.

Data were also provided in part by the Open Access Series of Imaging Studies (OASIS) (`https://www.oasis-brains.org/`).

We thank Pedro Borges for the fruitful discussions on uncertainty estimation.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous distributed systems
2. Billot, B., Greve, D.N., Leemput, K.V., Fischl, B., Iglesias, J.E., Dalca, A.: A learning strategy for contrast-agnostic MRI segmentation. In: Medical Imaging with Deep Learning, pp. 75–93. PMLR. ISSN: 2640-3498
3. Brett, M., Leff, A.P., Rorden, C., Ashburner, J.: Spatial normalization of brain images with focal lesions using cost function masking **14**(2), 486–500. DOI 10.1006/nimg.2001.0845
4. Cardoso, M.J., Modat, M., Wolz, R., Melbourne, A., Cash, D., Rueckert, D., Ourselin, S.: Geodesic information flows: Spatially-variant graphs and their application to segmentation and fusion **34**(9), 1976–1988. DOI 10.1109/TMI.2015.2418298
5. Chen, K., Derksen, A., Heldmann, S., Hallmann, M., Berkels, B.: Deformable image registration with automatic non-correspondence detection. In: J.F. Aujol, M. Nikolova, N. Papadakis (eds.) Scale Space and Variational Methods in Computer Vision, Lecture Notes in Computer Science, pp. 360–371. Springer International Publishing. DOI 10.1007/978-3-319-18461-6_29
6. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D.: Self-supervised learning for medical image analysis using image context restoration **58**, 101539. DOI 10.1016/j.media.2019.101539
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs
8. Chitphakdithai, N., Duncan, J.S.: Non-rigid registration with missing correspondences in preoperative and postresection brain images. In: T. Jiang, N. Navab, J.P.W. Pluim, M.A. Viergever (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010, Lecture Notes in Computer Science, pp. 367–374. Springer. DOI 10.1007/978-3-642-15705-9_45
9. Cicek, O., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: Learning dense volumetric segmentation from sparse annotation
10. Dorent, R., Booth, T., Li, W., Sudre, C.H., Kafiabadi, S., Cardoso, J., Ourselin, S., Vercauteren, T.: Learning joint segmentation of tissues and brain lesions from task-specific hetero-modal domain-shifted datasets **67**, 101862. DOI 10.1016/j.media.2020.101862
11. Duncan, J.S., Winston, G.P., Koepp, M.J., Ourselin, S.: Brain imaging in the assessment for epilepsy surgery **15**(4), 420–433. DOI 10.1016/S1474-4422(15)00383-X
12. Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., Cardoso, M.J.: Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions. In: A.F. Frangi, J.A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Lecture Notes in Computer Science, pp. 691–699. Springer International Publishing. DOI 10.1007/978-3-030-00928-1_78
13. Ermiş, E., Jungo, A., Poel, R., Blatti-Moreno, M., Meier, R., Knecht, U., Aebersold, D.M., Fix, M.K., Manser, P., Reyes, M., Herrmann, E.: Fully automated brain resection cavity delineation for radiation target volume definition in glioblastoma patients using deep learning **15**(1), 100. DOI 10.1186/s13014-020-01553-z
14. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J.V., Pieper, S., Kikinis, R.: 3d slicer as an image computing platform for the quantitative imaging network **30**(9), 1323–1341. DOI 10.1016/j.mri.2012.05.001
15. Fonov, V., Evans, A., McKinstry, R., Almli, C., Collins, D.: Unbiased nonlinear average age-appropriate brain templates from birth to adulthood **47**, S102. DOI 10.1016/S1053-8119(09)70884-5

16. Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L.: Unbiased average age-appropriate atlases for pediatric studies **54**(1), 313–327. DOI 10.1016/j.neuroimage.2010.07.033
17. Galovic, M., Baudracco, I., Wright-Goff, E., Pillajo, G., Nachev, P., Wandschneider, B., Woermann, F., Thompson, P., Baxendale, S., McEvoy, A.W., Nowell, M., Mancini, M., Vos, S.B., Winston, G.P., Sparks, R., Prados, F., Miserocchi, A., de Tisi, J., Van Graan, L.A., Rodionov, R., Wu, C., Alizadeh, M., Kozlowski, L., Sharan, A.D., Kini, L.G., Davis, K.A., Litt, B., Ourselin, S., Moshé, S.L., Sander, J.W.A., Löscher, W., Duncan, J.S., Koepp, M.J.: Association of piriform cortex resection with surgical outcomes in patients with temporal lobe epilepsy **76**(6), 690–700. DOI 10.1001/jamaneurol.2019.0204
18. Granados, A., Perez-Garcia, F., Schweiger, M., Vakharia, V., Vos, S.B., Miserocchi, A., McEvoy, A.W., Duncan, J.S., Sparks, R., Ourselin, S.: A generative model of hyperelastic strain energy density functions for multiple tissue brain deformation **16**(1), 141–150. DOI 10.1007/s11548-020-02284-y. URL `https://doi.org/10.1007/s11548-020-02284-y`
19. Greff, K., Klein, A., Chovanec, M., Hutter, F., Schmidhuber, J.: The sacred infrastructure for computational research pp. 49–56. DOI 10.25080/shinma-7f4c6e7-008. Conference Name: Proceedings of the 16th Python in Science Conference
20. Gudbjartsson, H., Patz, S.: The rician distribution of noisy MRI data **34**(6), 910–914
21. Hacohen, G., Weinshall, D.: On the power of curriculum learning in training deep networks. In: International Conference on Machine Learning, pp. 2535–2544. PMLR. URL `http://proceedings.mlr.press/v97/hacohen19a.html`. ISSN: 2640-3498
22. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification
23. Herrmann, E., Ermiş, E., Meier, R., Blatti-Moreno, M., Knecht, U.P., Aebersold, D.M., Manser, P., Reyes, M.: Fully automated segmentation of the brain resection cavity for radiation target volume definition in glioblastoma patients **102**(3), S194. DOI 10.1016/j.ijrobp.2018.07.087. Publisher: Elsevier
24. Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z.: Robust brain extraction across datasets and comparison with publicly available methods **30**(9), 1617–1634. DOI 10.1109/TMI.2011.2138152
25. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift
26. Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W.: The alzheimer's disease neuroimaging initiative (ADNI): MRI methods **27**(4), 685–691. DOI 10.1002/jmri.21049
27. Jing, L., Tian, Y.: Self-supervised visual feature learning with deep neural networks: A survey
28. Jobst, B.C., Cascino, G.D.: Resective epilepsy surgery for drug-resistant focal epilepsy: a review **313**(3), 285–293. DOI 10.1001/jama.2014.17426
29. Jungo, A., Balsiger, F., Reyes, M.: Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation **14**. DOI 10.3389/fnins.2020.00282. Publisher: Frontiers
30. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision?
31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization
32. LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., Raichle, M.E., Cruchaga, C., Marcus, D.: OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease p. 2019.12.13.19014902. DOI 10.1101/2019.12.13.19014902. Publisher: Cold Spring Harbor Laboratory Press
33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization
34. Lowekamp, B.C., Chen, D.T., Ibáñez, L., Blezek, D.: The design of SimpleITK **7**, 45. DOI 10.3389/fninf.2013.00045

35. Matzkin, F., Newcombe, V., Stevenson, S., Khetani, A., Newman, T., Digby, R., Stevens, A., Glocker, B., Ferrante, E.: Self-supervised skull reconstruction in brain CT images with decompressive craniectomy. In: A.L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M.A. Zuluaga, S.K. Zhou, D. Racoceanu, L. Joskowicz (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Lecture Notes in Computer Science, pp. 390–399. Springer International Publishing. DOI 10.1007/978-3-030-59713-9_38
36. Meier, R., Porz, N., Knecht, U., Loosli, T., Schucht, P., Beck, J., Slotboom, J., Wiest, R., Reyes, M.: Automatic estimation of extent of resection and residual tumor volume of patients with glioblastoma **127**(4), 798–806. DOI 10.3171/2016.9.JNS16146
37. Mercier, L., Del Maestro, R.F., Petrecca, K., Araujo, D., Haegelen, C., Collins, D.L.: Online database of clinical MR and ultrasound images of brain tumors **39**(6), 3253–3261. DOI 10.1118/1.4709600
38. Milletari, F., Navab, N., Ahmadi, S.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. DOI 10.1109/3DV.2016.79
39. Modat, M., Cash, D.M., Daga, P., Winston, G.P., Duncan, J.S., Ourselin, S.: Global image registration using a symmetric block-matching approach **1**(2). DOI 10.1117/1.JMI.1.2.024003
40. Moshkov, N., Mathe, B., Kertesz-Farkas, A., Hollandi, R., Horvath, P.: Test-time augmentation for deep learning-based cell segmentation on microscopy images **10**(1), 5068. DOI 10.1038/s41598-020-61808-3. Number: 1 Publisher: Nature Publishing Group
41. Nikolenko, S.I.: Synthetic data for deep learning
42. Nyúl, L.G., Udupa, J.K., Zhang, X.: New variants of a method of MRI scale standardization **19**(2), 143–150. DOI 10.1109/42.836373
43. Pan, S.J., Yang, Q.: A survey on transfer learning **22**(10), 1345–1359. DOI 10.1109/TKDE.2009.191
44. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. In: H. Wallach, H. Larochelle, A. Beygelzimer, F.d. Alché-Buc, E. Fox, R. Garnett (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc.
45. Perlin, K.: Improving noise **21**(3), 681–682. DOI 10.1145/566654.566636
46. Pezeshk, A., Petrick, N., Chen, W., Sahiner, B.: Seamless lesion insertion for data augmentation in CAD training **36**(4), 1005–1015. DOI 10.1109/TMI.2016.2640180
47. Pérez-García, F., Radionov, R., Alim-Marvasti, A., Sparks, R., Duncan, J., Ourselin, S.: EPISURG: Quantitative analysis of resection neurosurgery for refractory epilepsy. University College London. DOI 10.5522/04/9996158.v1
48. Pérez-García, F., Rodionov, R., Alim-Marvasti, A., Sparks, R., Duncan, J.S., Ourselin, S.: Simulation of brain resection for cavity segmentation using self-supervised and semi-supervised learning. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Lecture Notes in Computer Science, pp. 115–125. Springer International Publishing. DOI 10.1007/978-3-030-59716-0_12
49. Pérez-García, F., Sparks, R., Ourselin, S.: TorchIO: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning
50. Radosavovic, I., Dollár, P., Girshick, R., Gkioxari, G., He, K.: Data distillation: Towards omni-supervised learning
51. Rosenow, F., Lüders, H.: Presurgical evaluation of epilepsy **124**(9), 1683–1700. DOI 10.1093/brain/124.9.1683
52. Schroeder, W., Martin, K., Lorensen, B.: The Visualization Toolkit. Kitware
53. Shaw, R., Sudre, C., Ourselin, S., Cardoso, M.J.: MRI k-space motion artefact augmentation: Model robustness and task-specific uncertainty. In: International Conference on Medical Imaging with Deep Learning, pp. 427–436
54. Shaw, R., Sudre, C.H., Ourselin, S., Cardoso, M.J.: A heteroscedastic uncertainty model for decoupling sources of MRI image quality

55. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M., Golia-Pernicka, J., Heckers, S.H., Jarnagin, W.R., McHugo, M.K., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms
56. Singh, S.P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., Gulyás, B.: 3d deep learning on medical images: A review **20**(18), 5097. DOI 10.3390/s20185097. Number: 18 Publisher: Multidisciplinary Digital Publishing Institute
57. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting **15**, 1929–1958
58. Sudre, C.H., Cardoso, M.J., Ourselin, S.: Longitudinal segmentation of age-related white matter hyperintensities **38**, 50–64. DOI 10.1016/j.media.2017.02.007
59. Venturini, L., Papageorghiou, A.T., Noble, J.A., Namburete, A.I.L.: Uncertainty estimates as data selection criteria to boost omni-supervised learning. In: A.L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M.A. Zuluaga, S.K. Zhou, D. Racoceanu, L. Joskowicz (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Lecture Notes in Computer Science, pp. 689–698. Springer International Publishing. DOI 10.1007/978-3-030-59710-8_67
60. van der Walt, S., Colbert, S.C., Varoquaux, G.: The NumPy array: a structure for efficient numerical computation **13**(2), 22–30. DOI 10.1109/MCSE.2011.37
61. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks **338**, 34–45. DOI 10.1016/j.neucom.2019.01.103
62. Winston, G.P., Daga, P., Stretton, J., Modat, M., Symms, M.R., McEvoy, A.W., Ourselin, S., Duncan, J.S.: Optic radiation tractography and vision in anterior temporal lobe resection **71**(3), 334–341. DOI 10.1002/ana.22619
63. Winterstein, M., Münter, M.W., Burkholder, I., Essig, M., Kauczor, H.U., Weber, M.A.: Partially resected gliomas: diagnostic performance of fluid-attenuated inversion recovery MR imaging for detection of progression **254**(3), 907–916. DOI 10.1148/radiol09090893
64. Zhu, L., Kolesov, I., Gao, Y., Kikinis, R., Tannenbaum, A.: An effective interactive medical image segmentation method using fast GrowCut. Publication Title: Int Conf Med Image Comput Comput Assist Interv. Workshop on Interactive Methods.
65. Zwillinger, D., Kokoska, S.: CRC Standard Probability and Statistics Tables and Formulae. CRC Press. Google-Books-ID: tB3RVEZ0UIMC