Automatic hemisphere segmentation in rodent MRI with lesions

Juan Miguel Valverde\textsuperscript{a}, Artem Shatillo\textsuperscript{b}, Riccardo de Feo\textsuperscript{a}, Jussi Tohka\textsuperscript{a}

\textsuperscript{a}A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, 70150 Kuopio, Finland
\textsuperscript{b}Charles River Discovery Services, Kuopio 70210, Finland.

Abstract

We present MedicDeepLabv3+, a convolutional neural network that is the first completely automatic method to segment brain hemispheres in magnetic resonance (MR) images of rodents with lesions. MedicDeepLabv3+ improves the state-of-the-art DeepLabv3+ with an advanced decoder, incorporating spatial attention layers and additional skip connections that, as we show in our experiments, lead to more precise segmentations. MedicDeepLabv3+ requires no MR image preprocessing, such as bias-field correction or registration to a template, produces segmentations in less than a second, and its GPU memory requirements can be adjusted based on the available resources. Using a large dataset of 723 MR rat brain images, we evaluated our MedicDeepLabv3+, two state-of-the-art convolutional neural networks (DeepLabv3+, UNet) and three approaches that were specifically designed for skull-stripping rodent MR images (Demon, RATS and RBET). In our experiments, MedicDeepLabv3+ outperformed the other methods, yielding an average Dice coefficient of 0.952 and 0.944 in the brain and contralateral hemisphere regions. Additionally, we show that despite limiting the GPU memory and the training data to only three images, our MedicDeepLabv3+ also provided satisfactory segmentations. In conclusion, our method, publicly available at https://github.com/jmlipman/MedicDeepLabv3Plus, yielded excellent results in multiple scenarios, demonstrating its capability to reduce human workload in rodent neuroimaging studies.

Keywords: Hemisphere segmentation, MRI, Convolutional Neural Networks, Rodent imaging

1. Introduction

Rodents are widely used in preclinical research to investigate brain diseases \cite{1}. These studies often utilize in-vivo imaging technologies, such as magnetic resonance imaging (MRI), to visualize brain tissue at different time-points, which is necessary for studying disease progression. MRI permits the acquisition of brain images with different contrasts in a non-invasive manner, making MRI a particularly advantageous in-vivo imaging technology. However, these images typically need to be segmented before conducting quantitative analysis. As an example, the size of the hemispheric brain edema relative to the brain size is an important biomarker for acute stroke that requires accurate hemisphere segmentation \cite{2,3}.

With brain edema biomarkers in mind, our work focuses on hemisphere segmentation in rodent MR images with lesions. Segmenting these images is particularly challenging since lesions’ size, shape, location, and contrast can vary even within images from the same study, hampering, as we show in our experiments, traditional segmentation methods. In addition, rodents’ small size makes image acquisition sensitive to misalignments, potentially producing slices with asymmetric hemispheres and particularly affecting anisotropic data. These difficulties have led researchers and technicians to annotate rodent brain hemispheres manually \cite{4,5}, which is laborious and time-consuming, and motivates this work.

In recent years, convolutional neural networks (ConvNets) have been widely used to segment medical images...
due to their outstanding performance [6, 7, 8]. ConvNets can be optimized end-to-end, require no preprocessing, such as bias-field correction and costly registration, and can produce segmentation masks in real time [9]. ConvNets can also be tailored to specific segmentation problems by incorporating domain constraints and shape priors [10]. In particular, DeepLabv3+ architecture with its efficient computation of large image regions via dilated convolutions has demonstrated excellent results on various segmentation tasks [11], leading researchers to its application on medical image segmentation. Xie et al. [12] utilized DeepLabv3+ to estimate the segmentation maps and subsequently refined such estimation with a second ConvNet. Ma et al. [13] modified DeepLabv3+ for applying style transfer to homogenize MR images with different properties. Khan et al. [14] showed that DeepLabv3+ outperforms other ConvNets on prostate segmentation of T2-weighted MR scans. However, in our preliminary experiments, DeepLabv3+ provided unsatisfactory results, especially in the masks borders.

We present and make publicly available MedicDeepLabv3+, the first method for segmenting brain hemispheres in MR images of rodents with lesions. MedicDeepLabv3+ improves DeepLabv3+ architecture with a new decoder with spatial attention layers [15, 16] and more long and short skip connections that facilitate the optimization. We evaluated our method on a large and challenging dataset of 723 MR rat brain scans from 11 studies acquired at multiple lesion stages. Our experiments show that MedicDeepLabv3+ outperformed the baseline state-of-the-art DeepLabv3+ [11], UNet [17], and, particularly for skull-stripping, it also outperformed Demon [18], RATS [19], and RBET [20]. Additionally, we evaluated MedicDeepLabv3+ with very limited GPU memory and training data, and our experiments demonstrate that, despite such restrictions, MedicDeepLabv3+ yields satisfactory segmentations, showcasing its usability in multiple real-life situations and environments.

1.1. Related work

Anatomical segmentation of rodent brain MRI without lesions. The vast majority of anatomical segmentation methods for rodent brain images have been exclusively developed for brains without lesions. These methods can be classified into three categories. First, atlas-based segmentation approaches, which apply registration to one or more brain atlases [23] and, afterwards, label candidates are refined or combined with, for instance, Markov random fields [24]. As these approaches heavily rely on registration, they underperform in the presence of anatomical deformations. Second, methods that group nearby voxels with similar properties. These approaches typically start by proposing one or several candidate regions, and later adjust such regions with an energy function and, optionally, shape priors. Examples of these methods include surface deformation models [20], graph-based segmentation algorithms [19], and a more recent approach that combines blobs into a single region [25]. These approaches can handle different MRI contrasts and require no registration. However, they also rely on local features, such as nearby image gradients and intensities. Thus, these methods can be very sensitive to intensity inhomogeneities, and small brain deformities. Third, machine learning algorithms that classify brain features. These features can be handcrafted, such as in [26, 27] where authors employed support vector machines to classify voxels into different neuroanatomical regions based on their intensity, location, neighbor labels, and probability maps. On the contrary, deep neural networks, a subclass of machine learning algorithms, can automatically find relevant features and learn meaningful non-linear relationships between such features. Methods based on neural networks, such as pulse-coupled neural networks [28, 29] and ConvNets [18, 30, 9], have been used in the context of rodent MRI segmentation.

Anatomical segmentation of rodent brain MRI with lesions. Anatomical segmentation in MR images of rodents with lesions is an underexplored area; Roy et al. [18] is the only study that investigated such problem. The authors showed that their Inception-based [21] skull-stripping ConvNet named Demon surpassed other methods on MR images of mice and humans with traumatic brain injury. Mulder et al. [22] developed a lesion segmentation pipeline that includes an atlas-based contralateral hemisphere segmentation step. However, these hemisphere segmentations were not compared to a ground truth, and this approach is sensitive to the lesion appearance because it relies on registration.
Lesion segmentation of rodent brain MRI. The high contrast between lesion and non-lesion voxels in certain rodent brain MR images motivated the development of thresholding-based methods \cite{31, 32}. However, these methods are not fully automatic, and they cannot be used in MR images with other contrasts, or lesions with different appearances. Mulder et al. \cite{22} introduced a fully-automated pipeline to segment lesions via level sets \cite{33, 34}. Images were first registered to a template, then skull-stripped, and their ventricles were segmented prior to the final lesion segmentation step. Arnaud et al. \cite{35} framed lesion segmentation as an anomaly-detection problem and developed a pipeline that detects voxels with unusual intensity values with respect to healthy rodent brains. Valverde et al. \cite{36} developed the first single-step method to segment rodent brain MRI lesions using ConvNets.

2. Materials and methods

2.1. MRI Data

The image data, provided by Charles River Laboratories Discovery site (Kuopio, Finland)\footnote{https://www.criver.com/products-services/discovery-services}, consisted of 723 MR T2-weighted brain scans of 481 adult male Wistar rats weighting between 250-300 g derived from 11 different studies. Rats were induced focal cerebral ischemia by middle cerebral artery occlusion for 120 minutes in the right hemisphere of the brain \cite{37}. MR data was acquired at multiple time-points after the occlusion; for each of the 11 studies time-points were different (see Fig. 1-A for details). In total, our dataset contained MR images from nine lesion stages: shams, 2h, 24h, D3, D7, D14, D21, D28, and D35. Figure 1-B shows representative images of these lesion stages in approximately the same brain area. All animal experiments were conducted according to the National Institute of Health (NIH) guidelines for the care and use of laboratory animals, and approved by the National Animal Experiment Board, Finland. Multi-slice multi-echo sequence was used with the following parameters; TR = 2.5 s, 12 echo times (10-120 ms in 10 ms steps) and 4 averages in a horizontal 7T magnet. T2-weighted images were calculated as the sum of the all echoes. Eighteen coronal slices of 1 mm thickness were acquired using a field-of-view of 30x30 mm\(^2\) producing 256x256 imaging matrices of resolution 117 µm × 117 µm.

2.2. Data preparation and train, validation and test sets

The T2-weighted images were not preprocessed (i.e., no registration, bias-field or artifact correction), and their intensity values were standardized to have zero mean and unit variance. Brain and contralateral hemisphere masks were annotated by several trained technicians employed by Charles River according to standard operating procedure. These annotations did not include the cerebellum and the olfactory bulb. Finally, we computed the ipsilateral hemisphere mask by subtracting the contralateral hemisphere from the brain mask, yielding non-overlapping regions (i.e., the background, ipsilateral and contralateral hemispheres) for optimizing the ConvNets.

MR images were grouped by their study and acquisition time-point (Fig. 1-A). Images from sham-operated...
animals were not included to the training and validation sets since our work focused on rodent brains with lesions. From the resulting 17 subgroups, our training and validation sets comprised 3 and 1 MR images, respectively, per subgroup (51 and 17 images in total). The remaining 655 MR images, including shams, formed the independent test set. This splitting strategy aimed to create a diverse training set, as brain lesions have notably different T2-weighted MRI intensities depending on the lesion stage, and annotations can differ slightly across studies due to the task subjectivity and the consequent low interrater agreement [22, 30].

2.3. MedicDeepLabv3+

MedicDeepLabv3+ is a 3D fully convolutional neural network (ConvNet) based on DeepLabv3+ [11] and UNet [17]. DeepLabv3+ employs Xception [38] as an encoder to reduce the data dimensionality, and a decoder, similar to UNet, that upsamples the transformed data, twice, by a factor of four. MedicDeepLabv3+ replaces this decoder with three stages of skip connections and convolutional layers, and it incorporates spatial attention layers into these blocks, enabling deep supervision (see Fig. 2).

2.3.1. Encoder

MedicDeepLabv3+ stacks several $3 \times 3 \times 3$ convolutional layers, normalizes the data with Batch Normalization [39], and incorporates residual connections to facilitate the parameters’ optimization [40]. In particular, the first layers of MedicDeepLabv3+ correspond to Xception [38], which uses depthwise-separable convolutions instead of regular convolutions. Depthwise-separable convolutions, popularized by MobileNets [41], are layers that contain depthwise and pointwise convolutions. The motivation behind depthwise-separable convolutions lies in the capability to decouple channel and spatial information, and their lower computational costs compared to regular convolutions.

MedicDeepLabv3+ utilizes dilated convolutions in the last layer of the Xception backbone and in the Atrous Spatial Pyramid Pooling (ASPP) module [42]. Dilated convolutions [43] expand the area considered by regular convolutions—the receptive field—by sampling a large padded image patch and multiplying the non-padded values with the kernel. These convolutions can control the receptive field by adjusting the padding/dilation rate. For a detailed description about dilated convolutions, we refer to [43]. After the Xception backbone, the ASPP module concatenates parallel dilated convolutional layers with different dilation rates and an average pooling followed by trilinear interpolation. Then, a pointwise convolution combines and reduces the number of channels. To this step, the described architecture reduces the data dimensionality by a factor of 16.

2.3.2. Decoder

We developed a new decoder for MedicDeepLabv3+ with more stages of skip connections and convolution blocks than DeepLabv3+, in which feature maps are upsampled via trilinear interpolation and concatenated to previous feature maps from the encoder. Subsequently, in each stage, a $3 \times 3 \times 3$ convolution halves the number of channels (Figure 2 blue blocks), and a ResNet block [40] further transforms the data (Figure 2 orange blocks). The consequent increase of skip-connections facilitates MedicDeepLabv3+ optimization [44]. Importantly, DeepLabv3+ produces segmentations at x4 less resolution than the original images that, to match their size, are upsampled via interpolation. In contrast, our MedicDeepLabv3+ incorporates more convolutional layers at the end of its architecture to directly produce segmentations at the original size.

Another key difference with respect to DeepLabv3+ is that MedicDeepLabv3+ utilizes spatial attention layers that learn and apply voxel-wise importance maps [15, 16]. These layers (Fig. 3) transform the inputs with a depthwise-separable convolution and, subsequently, average the resulting feature maps. Afterwards, a sigmoid activation function transforms the data non-linearly, producing spatial attention maps with values in the range $[0, 1]$. Then, these attention maps multiply the input feature maps voxel-wise. To encourage spatial attention maps that lead to the ground truth and to further facilitate the optimization, we added a branch in the first two attention layers for generating downsampled probability maps of the segmentation masks, enabling deep supervision (Fig 3 red and pink arrows).

2.3.3. Loss function

We trained MedicDeepLabv3+ with deep supervision [45], i.e., we minimized the sum of cross entropy and Dice
loss of all outputs of MedicDeepLabv3+ (Fig 2 pink arrow). Formally, we minimized \( L = \sum_{s \in S} L_{CE}^s + L_{Dice}^s \) with \( S = \{1, 2, 3\} \) indicating each MedicDeepLabv3+ output (see Fig. 2). Cross entropy treats the model predictions and the ground truth as distributions.

\[
L_{CE} = -\frac{1}{NC} \sum_{i=1}^{N} \sum_{c=1}^{C} p_{i,c} \log(q_{i,c}),
\]

(1)

where \( p_{i,c} \in [0, 1] \) represents whether voxel \( i \) belongs to class \( c \), and \( q_{i,c} \in [0, 1] \) its predicted Softmax probability. \( C = 3 \) for background, ipsilateral and contralateral hemisphere classes, and \( N \) is the total number of voxels. Dice loss estimates the Dice coefficient—that we ultimately aim to maximize—between the predictions and the ground truth:

\[
L_{Dice} = 1 - \frac{2}{C} \sum_{c=1}^{C} \frac{\sum_{i=1}^{N} p_{i,c} q_{i,c}}{\sum_{i=1}^{N} p_{i,c}^2 + q_{i,c}^2}.
\]

(2)

2.4. Experimental design

2.4.1. Metrics

We assessed the automatic segmentations with Dice coefficient [46], Hausdorff distance [47], precision, and recall. Dice coefficient measures the overlapping volume between the ground truth and the prediction:

\[
\text{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|}.
\]

(3)

where \( A \) and \( B \) are the segmentation masks. Hausdorff distance (HD) is a quality metric that calculates the distance to the misclassification located the farthest from the boundary masks. Formally:

\[
d(A, B) = \max_{\partial A} \min_{b \in \partial B} |a - b| \quad \text{and} \quad \min_{\partial B} \max_{a \in \partial A} |a - b|,
\]

(4)

where \( \partial A \) and \( \partial B \) are the boundary voxels of \( A \) and \( B \), respectively. In other words, HD provides the distance to the largest segmentation error. We provided HD values in mm and we accounted for voxel anisotropy. Finally,
precision is the percentage of voxels accurately classified as brain/hemisphere, and recall is the percentage of brain/hemisphere voxels that were correctly identified:

\[
\text{Prec} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}.
\]

### 2.4.2. Benchmarked methods

We compared our MedicDeepLabv3+ with DeepLabv3+ baseline [11], UNet [17], Demon [18], RATS [19], and RBET [20]. Since Demon, RATS, and RBET were exclusively designed for rodent brain segmentation, we computed contralateral hemisphere masks only with MedicDeepLabv3+, DeepLabv3+, and UNet. MedicDeepLabv3+, DeepLabv3+, UNet, and Demon were optimized with Adam [48] \((\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8})\), starting with a learning rate of \(10^{-5}\). MedicDeepLabv3+, DeepLabv3+, and UNet were trained for 300 epochs, and Demon was trained for an equivalent amount of time. We ensembled three models [49] since this strategy markedly improved segmentation performance in our previous work [36]. More specifically, for each architecture (MedicDeepLabv3+, DeepLabv3+, and UNet), we trained three models, starting from different random initializations, and formed the final segmentations based on the majority vote from the outputs of the three models.

We conducted a hyper-parameter grid search in RATS and RBET on the training and validation sets \(X_{\text{train/val}}\), and we utilized the best-performing hyper-parameters on the test set \(X_{\text{test}}\). With RATS, computing the brain mask \(\hat{y}\) of image \(x\) requires setting three hyper-parameters: intensity threshold \(t\), \(\alpha\), and rodent brain size \(s\), i.e., \(\hat{y} = \text{RATS}(x, t, \alpha, s)\). As rat brains have a similar size, we left this hyper-parameter with its default value, \(s = 1750\), thus, we only optimized for the threshold \(t\) and \(\alpha\) hyper-parameters. Since the intensity threshold must be \(t \in \mathbb{Z}^+\), we employed the unnormalized T2-weighted images with their intensity values reduced three orders of magnitude, enabling RATS to produce non-empty brain segmentations with its default hyper-parameters. We optimized RATS hyper-parameters by maximizing the Dice coefficient in the training and validation sets:

\[
\arg \max_{t, \alpha} \sum_{x \in X_{\text{train/val}}} \text{Dice}(y, \text{RATS}(x, t, \alpha, 1750)) : t = P_{\%i},
\]

where Dice is the Dice coefficient (Eq. (3)) between the ground-truth brain mask \(y\) and RATS’ output, \(\alpha \in [0, 1, \ldots, 10]\) balances the importance between gradients and intensity values, and \(P_{\%i}\) is the \(i\)th percentile of \(x\) with \(i \in [0.01, 0.02, \ldots, 0.99]\). Since finding \(t\) is potentially suboptimal due to the distribution variability across images, we optimized for the \(i\)th percentile, yielding imagespecific specific thresholds. In total, our hyper-parameter grid search in RATS comprised 1089 different parameter value combinations. For RBET, we optimized the Dice coefficient to find the optimal ellipse axes ratio \(w:h:d\) with \(w, h, d\) from 0.1 to 1 in steps of 0.05, accounting for 19\(^3\) different configurations.

Unlike ConvNets that can be optimized to segment specific brain areas, RATS and RBET perform skullstripping, segmenting also the cerebellum and olfactory bulb that were not annotated. As these brain areas were not part of our ground truth, RATS and RBET segmentations get unnecessarily penalized in those areas. Thus, before computing the metrics, we discarded the slices containing cerebellum and olfactory bulb. This evaluation strategy ignores potential misclassifications in the excluded slices, slightly favoring RATS and RBET results.

#### 2.4.3. Midline brain evaluation

We calculated the average Dice coefficients of contralateral and ipsilateral hemispheres around the brain midline—boundary between both hemispheres (see Fig. 4A). In contrast to brain vs. non-brain tissue boundaries, the brain midline area is more ambiguous to annotate due to the lower intensity contrast between hemispheres, hence the importance to assess the performance in this area. More specifically, we considered the area after expanding brain midline voxels in the coronal plane via morphological dilation \(n\) times, with \(n \in \{1, 2, \ldots, 10\}\). Note that, similarly to RATS and RBET evaluation, this experiment assessed the Dice coefficient only on the slices that were manually annotated, as finding the brain midline requires these manual annotations. Consequently, the evaluated masks...
excluded non-annotated slices that could have false positives.

2.4.4. Biomarkers based on hemisphere segmentation

Since the ratio between contra- and ipsilateral hemispheres size is an important biomarker for acute stroke \cite{2, 3}, we investigated which method produced the most similar hemispheric ratios to the ground truth. For this, we computed the effect size via Cohen’s d \cite{50} and the bias-corrected and accelerated (BCa) bootstrap confidence intervals \cite{51} with 100000 random iterations.

2.4.5. Performance with limited GPU memory and data

Motivated by potential GPU memory limitations, we studied the performance and computational requirements of multiple versions of MedicDeepLabv3+ with lower capacity and, consequently, lower GPU memory usage. To investigate this, we varied the number of kernel filters in all convolutions of MedicDeepLabv3+ that determines the number of parameters. For instance, decreasing the number of kernel filters by half in the encoder also decreases the number of kernel filters in the decoder to half. Separately, we evaluated the proposed MedicDeepLabv3+ on each study and time-point independently, simulating the typical scenario in rodent studies with extremely scarce annotated data. For each of the 17 groups containing no sham animals (Fig. 1-A), we trained an ensemble of three MedicDeepLabv3+ on only three images, employed another image for validation during the optimization, and we evaluated this ensemble on the remaining holdout images from the same group.

2.4.6. Implementation

MedicDeepLabv3+, UNet, DeepLabv3+ and De-\textsc{mon} were implemented in Pytorch \cite{52} and were run on Ubuntu 16.04 with an Intel Xeon W-2125 CPU @ 4.00GHz processor, 64 GB of memory and an NVidia GeForce GTX 1080 Ti with 11 GB of memory. MedicDeepLabv3+ and the scripts for segmenting rodent MR images and to optimize new models are publicly available at https://github.com/jmlipman/MedicDeepLabv3Plus. These scripts are ready for use via command line interface with a single command, and users can easily adjust the number of initial filters that controls the model size, capacity, and GPU memory requirements. Additionally, we provided the optimized parameters (i.e., the weights) of MedicDeepLabv3+.

3. Results

3.1. Segmentation metrics comparison

Our MedicDeepLabv3+ produced brain and hemisphere masks with the highest Dice coefficients and precision, and the lowest HD (see Table 1). All ConvNets performed better than RATS and RBET and, particularly, 3D ConvNets (MedicDeepLabv3+, DeepLabv3+) consistently yielded lower HD than 2D ConvNets (UNet, Demo\textsc{mon}). Our MedicDeepLabv3+ produced finer segmentations that were more similar to the ground than the baseline DeepLabv3+ which generated masks with imprecise borders. UNet also produced segmentations with higher Dice and recall than DeepLabv3+, although UNet HD was considerably lower. Figure 22 illustrates these results on the MR image with the highest hemispheric volume im-\textsc{balance}. Figure 22 shows that RBET was incapable to find the brain boundaries; RATS produced segmentations with several holes and non-smooth borders; 2D ConvNets misclassified the olfactory bulb and cerebellum; and, in agreement with Table 1, MedicDeepLabv3+ produced the segmentation mask most similar to the ground truth. We included 17 images (one per study and time-point) in Appendix A that also corroborate the higher performance of MedicDeepLabv3+. The computation time to optimize these methods also varied notably: on average, ConvNets required 16 hours, and RATS and RBET needed six days. Furthermore, MedicDeepLabv3+ segmented the images in real time, requiring approximately 0.4 seconds per image.

3.2. Midline brain experiment

Regarding the brain midline area experiment (Section 2.4.3, Figure 4 B,C), MedicDeepLabv3+ outperformed the baseline DeepLabv3+ across different area sizes (average Dice coefficient difference of 0.07), and UNet provided slightly higher Dice coefficients than MedicDeepLabv3+ (average difference of 0.02). Additionally, Dice coefficients were similar across hemispheres regardless of the segmentation method.
Figure 4: A: Example of ground truth and its brain midline area after four (red) and ten (green) iterations of morphological dilation. B-C: Dice coefficients for the ipsi- and contralateral hemisphere classes in the brain midline area with different morphological dilation iterations (brain midline area sizes).

Table 1: Dice coefficients, Hausdorff distances (HD), precision, and recall of the brain and contralateral hemisphere (CH) masks derived from the evaluated methods. Bold: best scores.

| Approach        | Dice    | HD       | Prec | Recall |
|-----------------|---------|----------|------|--------|
| **Brain**       |         |          |      |        |
| MedicDeepLabv3+ | **0.952 ± 0.04** | **0.218 ± 0.11** | **0.94** | 0.97   |
| UNet            | 0.947 ± 0.05 | 0.407 ± 0.14 | 0.93 | 0.97   |
| DeepLabv3+      | 0.936 ± 0.04 | 0.252 ± 0.12 | 0.93 | 0.95   |
| Demon           | 0.934 ± 0.04 | 0.424 ± 0.14 | 0.92 | 0.96   |
| RATS            | 0.913 ± 0.01 | 2.221 ± 0.51 | 0.91 | 0.92   |
| RBET            | 0.781 ± 0.10 | 3.628 ± 0.46 | 0.89 | 0.70   |
| **CH**          |         |          |      |        |
| MedicDeepLabv3+ | **0.944 ± 0.04** | **0.242 ± 0.22** | **0.94** | 0.96   |
| UNet            | 0.941 ± 0.05 | 0.432 ± 0.19 | 0.92 | **0.97** |
| DeepLabv3+      | 0.921 ± 0.04 | 0.283 ± 0.21 | 0.91 | 0.94   |

3.3. Hemispheric ratio experiment

The computed Cohen’s d shows that, in terms of magnitude, all methods produced hemispheric ratio distributions not too different from the ground truth (Table 2). Among these methods, MedicDeepLabv3+ provided the smallest effect size—over four and six times smaller than UNet and DeepLabv3+, respectively, and its confidence interval was the most zero-centered. DeepLabv3+’s confidence interval was the largest and contained zero whereas UNet’s confidence interval was the smallest—slightly smaller than MedicDeepLabv3+’s—and did not contained zero.

Table 2: P-values from the permutation tests that assessed whether there was a significant difference between the hemisphere volume ratios in the derived segmentations and the ground truth. Cohen’s d that measured the effect size and its confidence intervals. Bold: smallest effect size.

| Approach        | Cohen’s d | Confidence Interval       |
|-----------------|-----------|---------------------------|
| MedicDeepLabv3+ | **0.008** | [-0.013, 0.035]           |
| DeepLabv3+      | 0.050     | [-0.008, 0.099]           |
| UNet            | -0.038    | [-0.054, -0.021]          |

3.4. Limited resources

Table 6 lists the characteristics, computational requirements, and performance of different versions of
MedicDeepLabv3+ on the contralateral hemisphere segmentation (performance on the brain can be found in Appendix B). Reducing the number of parameters by decreasing the number of initial filters reduced notably the required GPU memory and training time while it barely affected MedicDeepLabv3+’s performance. For instance, reducing the number of parameters by 93.5% (from 79.1M to 5.1M) decreased the required GPU memory and training time by 72% while it decreased the Dice coefficient in the contralateral hemisphere by only 1%.

Tables 4 and 5 show the performance of MedicDeepLabv3+ optimized on only three images for each study and acquisition time-point. MedicDeepLabv3+, on average, performed slightly worse than in our first experiment that utilized 17 times more annotated data. Performance across these groups varied notably: in the contralateral hemisphere segmentations (Table 5) Dice coefficients ranged from 0.876 to 0.951, HD from 0.162 to 0.688, precision from 0.871 to 0.962, and recall from 0.859 to 0.967. Additionally, in agreement with our previous experiment, performance on the contralateral hemisphere was slightly lower than on the brain.

4. Discussion

We presented MedicDeepLabv3+, the first method for hemisphere segmentation in rodent MR images with le-
Table 3: Comparison between multiple versions of MedicDeepLabv3+ with different capacity. Columns: proportion of kernel filters with respect to the default configuration, trainable ConvNet parameters (in millions), optimization time for 300 epochs in our workstation (see Section 2.4.6 for details) in hours, maximum GPU memory required during training and evaluation, Dice and HD in the contralateral hemisphere. Bold: default configuration, highest performance.

| Rate | Parameters | Time (h) | Mem. (train) | Mem. (eval) | Dice  | HD   |
|------|------------|----------|--------------|-------------|-------|------|
| 1    | 79.1M      | 16.2     | 8857 MiB     | 2935 MiB    | 0.944 | 0.242|
| 0.875| 60.7M      | 14.4     | 7571 MiB     | 2617 MiB    | 0.941 | 0.246|
| 0.750| 44.7M      | 12.1     | 6545 MiB     | 2319 MiB    | 0.941 | 0.248|
| 0.625| 31.1M      | 10.3     | 5619 MiB     | 2007 MiB    | 0.941 | 0.246|
| 0.500| 20.0M      | 7.8      | 4577 MiB     | 1717 MiB    | 0.939 | 0.246|
| 0.375| 11.3M      | 6.2      | 3531 MiB     | 1421 MiB    | 0.937 | 0.251|
| 0.250| 5.1M       | 4.5      | 2503 MiB     | 1121 MiB    | 0.933 | 0.243|

Table 4: Dice, Hausdorff distance (HD), precision, and recall on the brain masks derived with MedicDeepLabv3+ in each study and time-point (TP) separately.

| Study | TP | Images | Dice       | HD         | Prec | Recall |
|-------|----|--------|------------|------------|------|--------|
| 1     | 2h | 8      | 0.916 ± 0.03 | 0.236 ± 0.06 | 0.929 | 0.913  |
| 1     | 24h| 8      | 0.912 ± 0.10 | 0.303 ± 0.214 | 0.894 | 0.952  |
| 2     | 24h| 13     | 0.915 ± 0.02 | 0.439 ± 0.13 | 0.929 | 0.909  |
| 3     | D35| 16     | 0.949 ± 0.02 | 0.251 ± 0.05 | 0.956 | 0.942  |
| 4     | 24h| 40     | 0.957 ± 0.01 | 0.232 ± 0.10 | 0.985 | 0.930  |
| 5     | 24h| 23     | 0.947 ± 0.01 | 0.182 ± 0.05 | 0.935 | 0.960  |
| 6     | D3 | 60     | 0.973 ± 0.01 | 0.141 ± 0.04 | 0.968 | 0.978  |
| 6     | D28| 58     | 0.946 ± 0.02 | 0.198 ± 0.07 | 0.937 | 0.957  |
| 7     | D3 | 35     | 0.956 ± 0.01 | 0.202 ± 0.11 | 0.968 | 0.945  |
| 7     | D21| 35     | 0.950 ± 0.01 | 0.181 ± 0.09 | 0.937 | 0.964  |
| 8     | 24h| 29     | 0.956 ± 0.01 | 0.204 ± 0.11 | 0.956 | 0.956  |
| 8     | D3 | 26     | 0.953 ± 0.01 | 0.187 ± 0.04 | 0.949 | 0.957  |
| 8     | D14| 26     | 0.948 ± 0.01 | 0.197 ± 0.05 | 0.929 | 0.967  |
| 8     | D28| 23     | 0.943 ± 0.02 | 0.201 ± 0.04 | 0.940 | 0.946  |
| 9     | 24h| 77     | 0.951 ± 0.03 | 0.215 ± 0.12 | 0.970 | 0.935  |
| 10    | D7 | 36     | 0.919 ± 0.05 | 0.261 ± 0.10 | 0.880 | 0.970  |
| 11    | 24h| 28     | 0.937 ± 0.02 | 0.316 ± 0.09 | 0.954 | 0.923  |
| Average|     | 541    | 0.948 ± 0.03 | 0.215 ± 0.10 | 0.949 | 0.951  |
Table 5: Dice, Hausdorff distance (HD), precision, and recall on the contralateral hemisphere masks derived with MedicDeepLabv3+ in each study and time-point (TP) separately.

| Study | TP  | Images | Dice       | HD         | Prec | Recall |
|-------|-----|--------|------------|------------|------|--------|
| 1     | 2h  | 8      | 0.883 ± 0.03 | 0.421 ± 0.10 | 0.871 | 0.904  |
| 1     | 24h | 8      | 0.886 ± 0.11 | 0.373 ± 0.21 | 0.874 | 0.915  |
| 2     | 24h | 13     | 0.876 ± 0.03 | 0.444 ± 0.19 | 0.902 | 0.859  |
| 3     | D3  | 16     | 0.927 ± 0.02 | 0.232 ± 0.11 | 0.948 | 0.907  |
| 4     | 24h | 40     | 0.928 ± 0.02 | 0.221 ± 0.09 | 0.962 | 0.898  |
| 5     | 24h | 23     | 0.899 ± 0.04 | 0.191 ± 0.06 | 0.912 | 0.888  |
| 6     | D3  | 60     | 0.951 ± 0.02 | 0.324 ± 0.27 | 0.936 | 0.967  |
| 6     | D28 | 58     | 0.935 ± 0.02 | 0.162 ± 0.06 | 0.932 | 0.939  |
| 7     | D3  | 35     | 0.930 ± 0.02 | 0.239 ± 0.16 | 0.959 | 0.904  |
| 7     | D21 | 35     | 0.939 ± 0.01 | 0.174 ± 0.10 | 0.927 | 0.952  |
| 8     | 24h | 29     | 0.935 ± 0.02 | 0.302 ± 0.27 | 0.936 | 0.936  |
| 8     | D3  | 26     | 0.903 ± 0.07 | 0.688 ± 0.10 | 0.902 | 0.912  |
| 8     | D14 | 26     | 0.933 ± 0.02 | 0.224 ± 0.16 | 0.901 | 0.967  |
| 8     | D28 | 23     | 0.932 ± 0.02 | 0.214 ± 0.14 | 0.929 | 0.935  |
| 9     | 24h | 77     | 0.917 ± 0.03 | 0.262 ± 0.20 | 0.926 | 0.911  |
| 10    | D7  | 36     | 0.908 ± 0.05 | 0.374 ± 0.26 | 0.871 | 0.958  |
| 11    | 24h | 28     | 0.894 ± 0.03 | 0.492 ± 0.27 | 0.900 | 0.892  |
| Average|      | 541    | 0.923 ± 0.04 | 0.291 ± 0.22 | 0.924 | 0.926  |
We compared MedicDeepLabv3+’s performance with the state-of-the-art DeepLabv3+, UNet, and three brain extraction algorithms (Demon, RATS, and RBET) used in preclinical neuroimaging studies on a large dataset of 723 rat MR images.

ConvNets performed markedly better and their training time was about 10 times shorter than RATS \cite{19} and RBET \cite{20}. RATS and RBET were incapable of learning from data (e.g., to distinguish between hemispheres, or to handle brain lesions). Therefore, the higher performance of ConvNets was unsurprising, as RATS and RBET were not designed to segment brains with lesions. This outperformance over non-ConvNet segmentation algorithms in rodent images aligns with recent research \cite{18, 25, 9}. Among the evaluated ConvNets, MedicDeepLabv3+ achieved the highest Dice coefficients, precision and recall, and the lowest HD (Table 1). Particularly, the outperformance of MedicDeepLabv3+ over the baseline DeepLabv3+ \cite{11} indicates that our proposed modifications (i.e., the incorporation of spatial attention layers and additional skip-connections) led to such improvements. Similar improvements after incorporating attention layers, such as the proposed spatial attention layers, have also been reported in the literature \cite{15, 16, 53, 54}. In the brain midline area experiment, UNet achieved slightly higher Dice coefficients than MedicDeepLabv3+. However, these Dice coefficients were computed only in the annotated slices, as finding the brain midline requires the manual annotations. As we showed in Table 1, Figure 2, and the 17 Figures in Appendix A, 2D ConvNets, including UNet, produced misclassifications in irrelevant areas (i.e., non-annotated slices), such as the cerebellum and the olfactory bulb, leading to notably higher HD. Therefore, the small difference between UNet and MedicDeepLabv3+ (Fig. 3) comes at the expense of those misclassifications that were disregarded during the evaluation. In contrast, the difference between MedicDeepLabv3+ and the baseline DeepLabv3+ (3D ConvNets) was three times larger. These results, altogether, show MedicDeepLabv3+ outperformance.

Our benchmark (Table 1) provides a valuable insight into whether 2D ConvNets produce better segmentations than 3D ConvNets on highly anisotropic data. In recent literature, 2D ConvNets appeared to be better \cite{53, 56, 57}, including in rodent images similar to our dataset \cite{2}. 2D ConvNets outperformance may arise because contiguous slices can differ significantly in anisotropic data, thus, three-dimensional information might be unnecessary, and slice appearance might suffice to segment the regions of interest. Our data and, particularly, our manual annotations, were specially challenging since our regions of interest had similar intensity values to the cerebellum and olfactory bulb that were not annotated. Therefore, three-dimensional information can be critical to learn the location in the Z-axis of those irrelevant areas to avoid them. Indeed, our results support this intuition. Although Dice coefficient, precision and recall varied across architectures (Table 1), HD was consistently lower with 3D ConvNets. In other words, 2D ConvNets produced more critical misclassifications. Thus, our data showcased a scenario in which, despite the anisotropy, 3D ConvNets were superior to 2D ConvNets, showing that the architectural choices need to consider more specific information and not just whether the data is anisotropic.

We measured the discrepancy magnitude between the hemispheric ratio distributions from the segmentations and from the ground truth (Table 2), and our MedicDeepLabv3+ yielded the smallest effect size, indicating that the hemispheric ratios of MedicDeepLabv3+ segmentations were more similar to the ground truth than UNet and DeepLabv3+. We want to emphasize the importance of accurate hemispheric ratios as they are biomarkers for predicting acute stroke \cite{2, 3}. DeepLabv3+ effect size and confidence intervals were much larger whereas UNet’s confidence intervals were the smallest and did not include zero, indicating that UNet’s hemispheric ratios were biased, being consistently larger than the ground truth.

ConvNets, and especially our MedicDeepLabv3+, produced segmentations more similar to the ground truth than the other methods (Table 1). Since these ConvNets were high capacity—requiring large GPU memory—and they were optimized with several images, their outperformance is in line with recent research \cite{58}. However, annotated data are often scarce, and large GPU memory to optimize ConvNets is not necessarily available. Motivated by these constraints, we showed in two separate experiments that MedicDeepLabv3+ performed remarkably well with few annotated data and very limited GPU memory (see Tables 6 \cite{4} and 5). In other words, our method can handle different scenarios without excessively sacrificing performance, which showcases MedicDeepLabv3+ generaliza-
tion capabilities and can encourage its usage to a wider audience.

MedicDeepLabv3+ is publicly available, and it can be easily incorporated into existing pipelines, reducing human workload and accelerating rodent neuroimaging analyses. Furthermore, MedicDeepLabv3+ is fast, requires no preprocessing and postprocessing, and it can be optimized on MR images with different contrast, voxel resolution, field of view, lesion appearance, and limited GPU memory and annotated data. As hemisphere segmentation masks can be utilized in diverse studies, our work is relevant for multiple applications involving brain lesions in rodent images.

Acknowledgements

The work of J.M. Valverde was funded from the European Union’s Horizon 2020 Framework Programme (Marie Skłodowska Curie grant agreement #740264 (GENOMMED)). This work has also been supported by the grant #316258 from Academy of Finland (J. Tohka) and grant S21770 from the European Social Fund (R. De Feo). Part of the computational analysis was run on the servers provided by Bioinformatics Center, University of Eastern Finland, Finland.

References

[1] L. Carbone, Estimating mouse and rat use in american laboratories by extrapolation from animal welfare act-regulated species, Scientific Reports 11 (1) (2021) 1–6.

[2] R. A. Swanson, M. T. Morton, G. Tsao-Wu, R. A. Savalos, C. Davidson, F. R. Sharp, A semi-automated method for measuring brain infarct volume, Journal of Cerebral Blood Flow & Metabolism 10 (2) (1990) 290–293.

[3] T. Gerriets, E. Stolz, M. Walberer, C. Muller, A. Kluge, A. Bachmann, M. Fisher, M. Kaps, G. Bachmann, Noninvasive quantification of brain edema and the space-occupying effect in rat stroke models using magnetic resonance imaging, Stroke 35 (2) (2004) 566–571.

[4] T. Freret, L. Chazalviel, S. Roussel, M. Bernaudin, P. Schumann-Bard, M. Boulouard, Long-term functional outcome following transient middle cerebral artery occlusion in the rat: correlation between brain damage and behavioral impairment., Behavioral neuroscience 120 (6) (2006) 1285.

[5] D. W. McBride, D. Klebe, J. Tang, J. H. Zhang, Correcting for brain swelling’s effects on infarct volume calculation after middle cerebral artery occlusion in rats, Translational stroke research 6 (4) (2015) 323–338.

[6] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge, arXiv preprint arXiv:1811.02629.

[7] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al., Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?, IEEE transactions on medical imaging 37 (11) (2018) 2514–2525.

[8] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han, et al., The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge, Medical Image Analysis 67 (2021) 101821.

[9] R. De Feo, A. Shatillo, A. Sierra, J. M. Valverde, O. Gröhn, F. Giove, J. Tohka, Automated joint skull-stripping and segmentation with multi-task u-net in large mouse brain mri databases, NeuroImage (2021) 117734.

[10] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, I. B. Ayed, Constrained-cnn losses for weakly supervised segmentation, Medical image analysis 54 (2019) 88–99.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable
convolution for semantic image segmentation, in: 
Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.

[12] Y. Xie, H. Lu, J. Zhang, C. Shen, Y. Xia, Deep segmentation-emendation model for gland instance segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 469–477.

[13] C. Ma, Z. Ji, M. Gao, Neural style transfer improves 3d cardiovascular mr image segmentation on inconsistent data, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 128–136.

[14] Z. Khan, N. Yahya, K. Alsaih, S. S. A. Ali, F. Meriaudeau, Evaluation of deep neural networks for semantic segmentation of prostate in t2w mri, Sensors 20 (11) (2020) 3183.

[15] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999.

[16] G. Wang, J. Shapey, W. Li, R. Dorent, A. Demitriadis, S. Bisdas, I. Paddick, R. Bradford, S. Zhang, S. Ourselin, et al., Automatic segmentation of vestibular schwannoma from t2-weighted mri by deep spatial attention with hardness-weighted loss, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 264–272.

[17] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[18] S. Roy, A. Knutsen, A. Korotcov, A. Bosomtwi, B. Dardzinski, J. A. Butman, D. L. Pham, A deep learning framework for brain extraction in humans and animals with traumatic brain injury, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 687–691.

[19] I. Oguz, H. Zhang, A. Rumple, M. Sonka, Rats: rapid automatic tissue segmentation in rodent brain mri, Journal of neuroscience methods 221 (2014) 175–182.

[20] T. C. Wood, D. J. Lythgoe, S. C. Williams, rbet: making bet work for rodent brains, in: Proc. Intl. Soc. Mag. Reson. Med. Vol. 21, 2013, p. 2706.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[22] I. A. Mulder, A. Khmelinskii, O. Dzyubachyk, S. de Jong, N. Rieff, M. J. Wermer, M. Hoehn, B. P. Lelieveldt, A. M. van den Maagdenberg, Automated ischemic lesion segmentation in mri mouse brain data after transient middle cerebral artery occlusion, Frontiers in neuroinformatics 11 (2017) 3.

[23] M. Pagani, M. Damiano, A. Galbusera, S. A. Tsafarlis, A. Gozzi, Semi-automated registration-based anatomical labelling, voxel based morphometry and cortical thickness mapping of the mouse brain, Journal of neuroscience methods 267 (2016) 62–73.

[24] D. Ma, M. J. Cardoso, M. Modat, N. Powell, J. Wells, H. Holmes, F. Wiseman, V. Tybulewicz, E. Fisher, M. F. Lythgoe, et al., Automatic structural parcellation of mouse brain mri using multi-atlas label fusion, PloS one 9 (1) (2014) e86576.

[25] Y. Liu, H. S. Unsal, Y. Tao, N. Zhang, Automatic brain extraction for rodent mri images, Neuroinformatics (2020) 1–12.

[26] M. H. Bae, R. Pan, T. Wu, A. Badea, Automated segmentation of mouse brain images using extended mrf, NeuroImage 46 (3) (2009) 717–725.

[27] T. Wu, M. H. Bae, M. Zhang, R. Pan, A. Badea, A prior feature svm-mrf based method for mouse brain segmentation, NeuroImage 59 (3) (2012) 2298–2306.

[28] N. Chou, J. Wu, J. B. Bingren, A. Qiu, K.-H. Chuang, Robust automatic rodent brain extraction
using 3-d pulse-coupled neural networks (pcnn), IEEE Transactions on Image Processing 20 (9) (2011) 2554–2564.

[29] M. Murugavel, J. M. Sullivan Jr, Automatic cropping of mri rat brain volumes using pulse coupled neural networks, Neuroimage 45 (3) (2009) 845–854.

[30] L.-M. Hsu, S. Wang, P. Ranadive, W. Ban, T.-H. H. Chao, S. Song, D. H. Cerri, L. R. Walton, M. A. Broadwater, S.-H. Lee, et al., Automatic skull stripping of rat and mouse brain mri data using u-net, Frontiers in Neuroscience.

[31] Y. Wang, P.-T. Cheung, G. X. Shen, I. Bhatia, E. X. Wu, D. Qiu, P.-L. Khong, Comparing diffusion-weighted and t2-weighted mr imaging for the quantification of infarct size in a neonatal rat hypoxic–ischemic model at 24 h post-injury, International journal of developmental neuroscience 25 (1) (2007) 1–5.

[32] C.-H. Choi, K. S. Yi, S.-R. Lee, Y. Lee, C.-Y. Jeon, J. Hwang, C. Lee, S. S. Choi, H. J. Lee, S.-H. Cha, A novel voxel-wise lesion segmentation technique on 3.0-t diffusion mri of hyperacute focal cerebral ischemia at 1 h after permanent mcao in rats, Journal of Cerebral Blood Flow & Metabolism 38 (8) (2018) 1371–1383.

[33] A. Dervieux, F. Thomasset, A finite element method for the simulation of a rayleigh-taylor instability, in: Approximation methods for Navier-Stokes problems, Springer, 1980, pp. 145–158.

[34] S. Osher, J. A. Sethian, Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations, Journal of computational physics 79 (1) (1988) 12–49.

[35] A. Arnaud, F. Forbes, N. Coquery, N. Collomb, B. Lemasson, E. L. Barbier, Fully automatic lesion localization and characterization: Application to brain tumors using multiparametric quantitative mri data, IEEE transactions on medical imaging 37 (7) (2018) 1678–1689.

[36] J. M. Valverde, A. Shatillo, R. De Feo, O. Gröhn, A. Sierra, J. Tohka, Ratslesnetv2: A fully convolutional network for rodent brain lesion segmentation, Frontiers in neuroscience 14 (2020) 1333.

[37] J. Koizumi, Y. Yoshida, T. Nakazawa, G. Ooneda, Experimental studies of ischemic brain edema. 1. a new experimental model of cerebral embolism in rats in which recirculation can be introduced in the ischemic area, Jpn J stroke 8 (1986) 1–8.

[38] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[39] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167.

[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[41] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861.

[42] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE transactions on pattern analysis and machine intelligence 40 (4) (2017) 834–848.

[43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, arXiv preprint arXiv:1412.7062.

[44] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, C. Pal, The importance of skip connections in biomedical image segmentation, in: Deep Learning and Data Labeling for Medical Applications, Springer, 2016, pp. 179–187.
[45] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in: Artificial intelligence and statistics, PMLR, 2015, pp. 562–570.

[46] L. R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) (1945) 297–302.

[47] G. Rote, Computing the minimum hausdorff distance between two point sets on a line under translation, Information Processing Letters 38 (3) (1991) 123–127.

[48] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980.

[49] T. G. Dietterich, Ensemble methods in machine learning, in: International workshop on multiple classifier systems, Springer, 2000, pp. 1–15.

[50] D. Lakens, Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas, Frontiers in psychology 4 (2013) 863.

[51] B. Efron, Better bootstrap confidence intervals, Journal of the American statistical Association 82 (397) (1987) 171–185.

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019, pp. 8024–8035.

[53] Q. Tao, Z. Ge, J. Cai, J. Yin, S. See, Improving deep lesion detection using 3d contextual and spatial attention, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 185–193.

[54] X. Xu, C. Lian, S. Wang, A. Wang, T. Royce, R. Chen, J. Lian, D. Shen, Asymmetrical multi-task attention u-net for the segmentation of prostate bed in ct image, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 470–479.

[55] Y. Jang, Y. Hong, S. Ha, S. Kim, H.-J. Chang, Automatic segmentation of lv and rv in cardiac mri, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2017, pp. 161–169.

[56] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, K. H. Maier-Hein, Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features, in: International workshop on statistical atlases and computational models of the heart, Springer, 2017, pp. 120–129.

[57] C. F. Baumgartner, L. M. Koch, M. Pollefeys, E. Konukoglu, An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2017, pp. 111–119.

[58] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.
Appendix

A. Supplementary Figures

Figure 6: T2-weighted image, ground truth and automatic segmentations of a rat from Study 1, 2h.
Figure 7: T2-weighted image, ground truth and automatic segmentations of a rat from Study 1, 24h.
Figure 8: T2-weighted image, ground truth and automatic segmentations of a rat from Study 2, 24h.
Figure 9: T2-weighted image, ground truth and automatic segmentations of a rat from Study 3, D35.
Figure 10: T2-weighted image, ground truth and automatic segmentations of a rat from Study 4, 24h.
Figure 11: T2-weighted image, ground truth and automatic segmentations of a rat from Study 5, 24h.
Figure 12: T2-weighted image, ground truth and automatic segmentations of a rat from Study 6, D3.
Figure 13: T2-weighted image, ground truth and automatic segmentations of a rat from Study 6, D28.
Figure 14: T2-weighted image, ground truth and automatic segmentations of a rat from Study 7, D3.
Figure 15: T2-weighted image, ground truth and automatic segmentations of a rat from Study 7, D21.
Figure 16: T2-weighted image, ground truth and automatic segmentations of a rat from Study 8, 24h.
Figure 17: T2-weighted image, ground truth and automatic segmentations of a rat from Study 8, D3.
Figure 18: T2-weighted image, ground truth and automatic segmentations of a rat from Study 8, D14.
Figure 19: T2-weighted image, ground truth and automatic segmentations of a rat from Study 8, D28.
Figure 20: T2-weighted image, ground truth and automatic segmentations of a rat from Study 9, 24h.
Figure 21: T2-weighted image, ground truth and automatic segmentations of a rat from Study 10, D7.
Figure 22: T2-weighted image, ground truth and automatic segmentations of a rat from Study 11, 24h.
## B. Supplementary Table

Table 6: Comparison between multiple versions of MedicDeepLabv3+ with different capacity. Columns: number of initial filters, trainable ConvNet parameters (in millions), optimization time for 300 epochs in our workstation in hours, maximum GPU memory required during training and evaluation, Dice and HD in the brain mask. Bold: default configuration.

| Filters | Parameters | Time (h) | Mem. (train) | Mem. (eval) | Dice  | HD   |
|---------|------------|----------|--------------|-------------|-------|------|
| 32      | 79.1M      | 16.2     | 8857 MiB     | 2935 MiB    | 0.952 | 0.218 |
| 28      | 60.7M      | 14.4     | 7571 MiB     | 2617 MiB    | 0.950 | 0.210 |
| 24      | 44.7M      | 12.1     | 6545 MiB     | 2319 MiB    | 0.950 | 0.206 |
| 20      | 31.1M      | 10.3     | 5619 MiB     | 2007 MiB    | 0.950 | 0.207 |
| 16      | 20.0M      | 7.8      | 4577 MiB     | 1717 MiB    | 0.949 | 0.200 |
| 12      | 11.3M      | 6.2      | 3531 MiB     | 1421 MiB    | 0.948 | 0.205 |
| 8       | 5.1M       | 4.5      | 2503 MiB     | 1121 MiB    | 0.947 | 0.199 |