NeuroImage 229 (2021) 117734

Contents lists available at ScienceDirect

NeuroImage

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

Automated joint skull-stripping and segmentation with Multi-Task U-Net in large mouse brain MRI databases

Riccardo De Feo[a,b,c,*], Artem Shatillo[d], Alejandra Sierra[c], Juan Miguel Valverde[c], Olli Gröhn[c], Federico Giove[b,e], Jussi Tohka[c]

[a] Sapienza Università di Roma, Rome 00184, Italy
[b] Centro Fermi–Museo Storico della Fisica e Centro Studi e Ricerche Enrico Fermi, Rome 00184, Italy
[c] A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio 70210, Finland
[d] Charles River Discovery Services, Kuopio, Finland
[e] Fondazione Santa Lucia IRCCS, Rome 00179, Italy

ARTICLE INFO

Keywords:
MRI
Brain
Segmentation
Deep learning
U-Net
Mice

ABSTRACT

Skull-stripping and region segmentation are fundamental steps in preclinical magnetic resonance imaging (MRI) studies, and these common procedures are usually performed manually. We present Multi-task U-Net (MU-Net), a convolutional neural network designed to accomplish both tasks simultaneously. MU-Net achieved higher segmentation accuracy than state-of-the-art multi-atlas segmentation methods with an inference time of 0.35 s and no pre-processing requirements.

We trained and validated MU-Net on 128 T2-weighted mouse MRI volumes as well as on the publicly available MRM NeAT dataset of 10 MRI volumes. We tested MU-Net with an unusually large dataset combining several independent studies consisting of 1782 mouse brain MRI volumes of both healthy and Huntington animals, and measured average Dice scores of 0.906 (striati), 0.937 (cortex), and 0.978 (brain mask). Further, we explored the effectiveness of our network in the presence of different architectural features, including skip connections and recently proposed framing connections, and the effects of the age range of the training set animals.

These high evaluation scores demonstrate that MU-Net is a powerful tool for segmentation and skull-stripping, decreasing inter and intra-rater variability of manual segmentation. The MU-Net code and the trained model are publicly available at https://github.com/Hierakonpolis/MU-Net.

1. Introduction

Preclinical imaging studies serve a fundamental role in biological and medical research, relating research results at the molecular level to clinical application in diagnosis and therapy. Magnetic Resonance Imaging (MRI) represents approximately 23% of all small-animal imaging studies providing the opportunity to monitor the development of pathological conditions and responses to treatment in a non-invasive way (Cunha et al., 2014). Its unique qualities also include the availability of different imaging contrasts, rendering MRI extremely useful in the context of preclinical neuroscience with applications from drug development (Matthews et al., 2013) to basic research (Febo and Foster, 2016).

Skull-stripping and region segmentation represent an integral part of processing pipelines in murine MR imaging (Anderson et al., 2019; Calabrese et al., 2015). Skull-stripping refers to the identification of the brain within the MRI volume, and region segmentation refers to the labeling of specific anatomical regions of interest (ROIs) within the brain. In preclinical MRI, these tasks are often performed manually. While manual segmentation represents the gold standard and is employed as the ground truth when evaluating automated segmentation algorithms, it is time-consuming and depends on the expertise of the annotators performing the segmentation. Furthermore, manual segmentation suffers from both intra- and inter-rater variability, both in small animal (Ali et al., 2005) and human MRI (Entis et al., 2012; Yushkevich et al., 2006).

In preclinical MRI, state-of-the-art automated region segmentation pipelines are based on atlas registration: individual MRI volumes are aligned with a labeled template (atlas) and the labels propagated to the individual volumes (De Feo and Giove, 2019; Lerch et al., 2011; Pagani et al., 2016; Schwarz et al., 2006; Sharief et al., 2008). The accuracy of registration-based segmentation depends on both the suitability of the template and the registration algorithm. The segmentation accuracy can be improved by multi-atlas strategies, where multiple atlases are

* Corresponding author at: Sapienza Università di Roma, 00184 Rome, Italy.
E-mail address: riccardo.defeo@uniroma1.it (R. De Feo).

https://doi.org/10.1016/j.neuroimage.2021.117734
Received 17 August 2020; Received in revised form 9 December 2020; Accepted 7 January 2021
Available online 14 January 2021

registered to the same volume and the so-resulting segmentation maps are combined, for example, via majority voting. Regarding multi-atlas strategies in mouse MRI, Bai et al. (2012) compared different single and multi-atlas methods for atlas-based segmentation of the mouse brain and reported that the combination of a diffeomorphic registration algorithm and multi-atlas segmentation provided the most accurate results. Ma et al. (2014) demonstrated that the multi-atlas methods are superior to single-atlas methods and the STEPS procedure for combining segmentations (Cardoso et al., 2013) brings advantages over earlier combination methodologies. While multi-atlas segmentation accounts for individual variability more effectively than single-atlas segmentation, it also requires multiple labeled atlases and multiple registration steps, significantly increasing the segmentation time. Multi-atlas segmentation can be further combined with the construction of a Minimum Deformation Template (MDT) as an intermediate step in the processing pipeline (Avants et al., 2010; De Feo and Giove, 2019; Kovačević et al., 2004). An MDT minimizes the deformation required to adapt it to each individual volume, thus reducing errors when its labels are propagated to each target scan. Instead of directly employing one or more manually segmented atlases, deep neural networks (DNNs) (LeCun et al., 2015) can use these as training data to learn a mapping function from the images to the segmentation maps. In this way, the anatomical information is not explicitly represented in a set of maps but implicitly encoded in the trained network. DNNs, and in particular Convolutional Neural Networks (CNNs), have been successfully applied in a large number of computer vision tasks in medical imaging. For example, Wachinger et al. (2018) developed a region segmentation CNN significantly outperforming state-of-the-art, registration-based methods for the healthy human brain MRI, both in terms of inference time and accuracy. Roy et al. (2018a) further improved on both aspects with a network based on the U-Net architecture (Ronneberger et al., 2015), with a reported segmentation time of 20 s per brain scan. However, within small-animal MRI, the applications of CNNs have been limited to skull-stripping: Roy et al. (2018b) trained a CNN algorithm based on Google Inception (Szegedy et al., 2015) for the skull-stripping in humans and mice after traumatic brain injury, achieving better performance than other state-of-the-art methods (3D Pulse Coupled Neural Networks (3D-PCNN) (Chou et al., 2011) and Rapid Automatic Tissue Segmentation (RATS) (Oguz et al., 2014)).

A specific type of CNN architecture, U-Net, has proved to be valuable in biomedical image segmentation. U-Net is based on the encoder/decoder structure, adding skip connections between the encoder and the decoder branches, allowing it to easily integrate multi-scale information and better propagate the gradient during training. This architecture has been shown to generalize even from a limited amount of annotated data (Xie et al., 2015), and as such is well suited for medical imaging, where datasets as large as the ones commonly used for CNNs are rare. Valverde et al. (2019) recently demonstrated the effectiveness of U-Net-like architectures in preclinical research, designing the first DNN for the segmentation of ischemic lesions in rodents and achieving segmentation accuracy comparable or better to inter-rater agreement in manual segmentation.

In this work, we introduce multi-task U-Net (MU-Net) to simultaneously perform skull-stripping and region segmentation of the mouse brain, based on the U-Net architecture. We refer to our approach as multi-task as we consider skull-stripping and region segmentation as separate tasks, allowing for the complete delineation of the brain volume regardless of the choice of ROIs. While these tasks are often considered as separate in the context of murine brain segmentation, they are strongly related. Therefore, our approach is not multi-task learning in the stronger sense of providing two fundamentally different outputs, e.g., segmentation and classification (Yang et al., 2017).

Our main train and validation data consisted of 128 T_2 MRI volumes from 32 mice at 4 different ages as well as five manually annotated regions (cortex, hippocampi, ventricles, striati and brain mask) from these images. This dataset represents MR images typically employed in drug development. We demonstrate that with this data MU-Net achieves a significantly higher accuracy than state-of-the-art multi-atlas segmentation methods (Cardoso et al., 2013; Ma et al., 2014) in a fraction of the segmentation time (approximately 0.35 s). We trained MU-Net on 128 MRI volumes and tested on an independent dataset of 1782 volumes acquired over the course of four years from both wild type (WT) and Huntington (HT) C57BL/6J mice, allowing us to evaluate MU-Net in a variety of experimental conditions. Additionally, we trained MU-Net for the segmentation of mouse brain MRI with isotropic voxels into 37 ROIs and demonstrate that the segmentation accuracy of MU-Net was equal or better than a state-of-the-art multi-atlas segmentation method (Ma et al., 2014).

Table 1
Summary characteristics of the three datasets employed in this study. BM refers to brain mask. The test dataset included various genotypes of both sexes (see Supplementary Table S1 for details).

Dataset name	# Animals	# MRIs	# ROIs	Type
Train and validation	32	128	4 + BM	WT males
Test	817	1,782	2 + BM	various
MRM NeAt	10	10	37 + BM	WT males

2. Materials and methods

2.1. Materials

We utilized three different datasets in this work as summarized in Table 1 and detailed in the following subsections.

2.1.1. Animals: train, validation and test sets

A total of 849 mice (Charles River Laboratories, Germany) were used: 32 mice for the train and validation set and 817 mice for the test set. Train and validation set animals were scanned at four different ages (5 weeks, 12 weeks, 16 weeks, 32 weeks) resulting in 128 volumes. All train and validation set animals were WT males.

The test set animals were part of 10 studies scanned at a single or multiple ages from 4 up to 60 weeks, and included both WT and several HT genotypes: R6/2, Q175, Q175DN, Q111, Q50 and Q20 (Supplementary Table S1), for a total of 1782 MRI scans. The groups included both males and females. These volumes were acquired as part of ten studies of Huntington's disease, kindly provided by the CHDI 'Cure Huntington's Disease Initiative' foundation.

All mice were housed in groups of up to 4 per cage (single sex) in a temperature (22±1°C) and humidity (30–70%) controlled environment with a normal light-dark cycle (7:00–20:00).

2.1.2. MRI: train, validation and test sets

Mice were anesthetized using isoflurane (5% for induction, 1.5–2% maintenance) in 70%/30% mix of N_2/O_2 carrying gas, fixed to a head holder and positioned in the magnet bore in a standard orientation relative to gradient coils. Respiration rate and temperature were monitored using PC-SAMS software and Model 1030 Monitoring & Gating System, Small Animal Instruments, Inc., Stony Brook, NY. The temperature was maintained at ~ 37° C using Small Animal Instruments feedback water heating system.

All acquisitions were performed using a horizontal 11.7T magnet with a bore size of 160mm, equipped with a gradient set capable of maximum gradient strength of 750mT/m and interfaced to a Bruker Avance III console (Bruker Biospin GmbH, Ettlingen, Germany). A volume coil (Bruker Biospin GmbH, Ettlingen, Germany) was used for transmission and a surface phased array coil for receiving (Rapid Biomedical GmbH, Rimpar, Germany). T_2 weighted anatomical images were acquired using a TurboRARE sequence with effective TR/TE = 2500/36ms, 8 echoes, 12ms inter-echo distance, matrix size 256x256, FOV 20.0x20.0 mm^2, 31 0.6mm thick coronal slices, −0.15mm interslice gap, and 8 averages. Con-

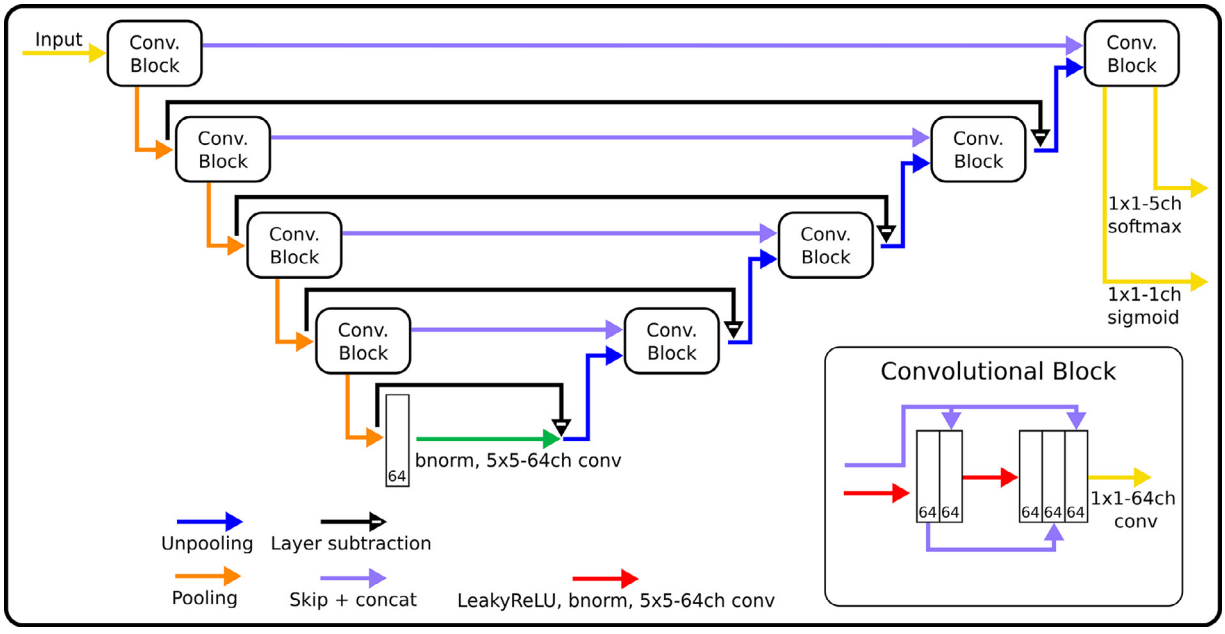

Fig. 1. General outline of the architectural features implemented and compared in the networks discussed, varying according to the presence or absence of the in-block dense connections (purple arrows in the convolutional block), presence or absence of the layer subtraction connections (black), and the use of 2D or 3D filters.

cerning the test data, MRI experimental parameters only differed in acquiring 19 0.7mm thick contiguous coronal slices.

Volumes within each study were manually segmented by an experienced rater, who had received a training and passed the qualification tests according to SOP (Standard Operating Procedure) for volumetric analysis in mice. Different studies were analyzed by different raters. Each training volume was manually segmented by a single rater drawing the brain mask and delineating 4 regions of interest: cortex, hippocampi, striati and ventricles. The brain mask did not include the olfactory bulb or the cerebellum. For the test set, only 3 regions were manually labeled: brain mask, cortex and striati. As each image was only segmented once by a single rater, intra- and inter-rater overlap statistics are not available for our dataset. Manual segmentation required from 10 to 15 min per ROI per image.

2.1.3. MRM NeAt dataset

The MRM NeAt dataset includes atlases of 10 individual T_2^*-weighted in vivo brain MR images of 12–14 weeks old C57BL/6J mice; each with 37 labelled anatomical structures (listed in Fig. 4) in addition to the brain mask (Ma et al., 2008). This dataset was downloaded from https://github.com/dancebean/mouse-brain-atlas, where an improved atlas is available (bias correction has been applied, left and right labels have been separated and 4th ventricle label added). This dataset was used to evaluate the STEPS algorithm by Ma et al. (2014) and is used here for the purpose of comparing MU-Net and STEPS on a larger number of ROIs on isotropic resolution MRI. As detailed in Ma et al. (2008), T_2-weighted MR data with a voxel-size of $0.1mm^3$ requiring about 2.8 h of scan time were acquired with a 3D large flip angle spin echo sequence using a super-conducting 9.4T/210 mm horizontal bore magnet (Magnex) controlled by an ADVANCE console (Bruker) and equipped with an actively shielded 11.6 cm gradient set (Bruker, Billerica, MA).

2.2. MU-Nets

2.2.1. Architectures

MU-Net (Fig. 1) presents an encoder-decoder U-Net-like architecture, with each branch articulated in four convolutional blocks. Unlike U-Net, the final block of the decoder branch further bifurcates into two different output maps representing our two tasks, sharing the same feature representation. Each convolutional block on the encoding path is followed by a 2x2 max-pooling layer. The last feature map feeds into the bottleneck layer, a 64 channel 5x5 convolutional layer with batch normalization (Ioffe and Szegedy, 2015) connecting the deepest layer of the encoding path with the decoding path.

The decoding path is composed of 4 more blocks alternating one un-pooling layer (Noh et al., 2015) and one convolutional block. Un-pooling operations effectively replace up-convolution layers in U-Net without any learnable parameters, while preserving spatial information. These layers operate by simply placing the elements of the un-pooled feature maps in the position of the respective maximum activation from the corresponding pooling operation, and setting the rest to zero. Skip connections concatenate the output of each dense layer in the encoding path with the respective un-pooled feature map of the same size before feeding it as input to the decoding convolutional block.

The output of the last decoding layer acts as the input of two different classification layers, which share the same feature representation up to this point: a 1x1 single channel convolution with a sigmoid activation function, and a 1x1 5 channels layer followed by a softmax activation function, for the skull-stripping task and the region classification task, respectively.

Convolutional block

Each convolutional block includes 3 convolutional layers preceded by leaky ReLU activation (Maas et al., 2013) layers and batch normalization. All 3 convolutions are padded and result in 64 output channels, in analogy with Roy et al. (2018a). The first and second convolutions employ 5x5 filters, while the third uses a 1x1 filter. This becomes especially relevant in the presence of dense connections, acting as a bottleneck for the 64x3 channels of the concatenated inputs and compressing the size of the feature maps.

2.2.2. Architectural variants

We study several variations to the basic network architecture.

Dense connections

In the models including dense connections (Huang et al., 2017) we modify each convolutional block by concatenating to the input of each convolution the outputs of the previous convolutions within the same block (Fig. 1).

Dual Framing connections

Dual framing connections refer to additional skip connections in the Dual Frame U-Net model. Han and Ye (2018) proposed this architecture for computed tomography reconstruction from sparse data based on signal processing arguments to reduce artifacts and improve recovery of high frequency edges. Dual framing connections consist in the subtraction of the input of each convolutional block on the encoding path from the output of the respective convolutional block of the same size on the decoding path, and as such the implementation of these connections does not increase the number of model parameters.

3D implementation

A 3D implementation could, in principle, provide better results by taking into account the features of the adjacent slices, whereas a 2D networks evaluates each coronal slice independently. However, the larger number of parameters also increases the risk of overfitting, and the lower resolution in the anterior-posterior axis compared to the in-plane resolution might constitute confounding factors in the presence of 3D pooling operations.

For these reasons, we compared 2D and 3D implementations of our network, using 5x5x5 filters and 2x2x2 max-pooling layers, replacing the filters and pooling layers described above. This results in 16,008,076 and 10,286,344 parameters for the 3D networks with and without in-block skip connections, respectively. Corresponding 2D networks contain 3,297,676 and 2,087,944 parameters, respectively. Thus, opting for a 3D architecture increases the number of parameters by factors of 4.85 and 4.93 as compared to the 2D architectures. The total number of parameters was measured by using the PyTorch instruction `sum(p.numel() for p in model.parameters())`. A complete breakdown of model parameters for each network is available in supplementary Table S2.

2.2.3. Loss function

Recent literature suggests that Dice-based loss functions (Milletari et al., 2016; Roy et al., 2018a; Sudre et al., 2017) would constitute an improvement over cross-entropy losses for the segmentation of medical images (Karimi and Salcudean, 2019). We optimized a joint loss function L, that is the sum of two Dice loss functions corresponding to the the skull-stripping (L_{SS}) and the region classification task (L_{RS}). Let $p(i)$ be the predicted probability of voxel i of belonging to the brain mask, and $g(i)$ the ground truth for voxel i ($g(i) = 1$ if the voxel is in the brain mask). Further, let $p_l(i)$ and $g_l(i)$ be the same quantities for label l ($l = 1, \dots, K$) encoding the ground truth as a one-hot vector. Then, the loss function can be written as:

$$L = L_{SS} + L_{RS}, \tag{1}$$

$$L_{SS} = -\frac{2\sum_i p(i)g(i)}{\sum_i p^2(i) + \sum_i g^2(i)}, \tag{2}$$

$$L_{RS} = -\sum_{l=1}^{K} \frac{2\sum_i p_l(i)g_l(i)}{\sum_i p_l^2(i) + \sum_i g_l^2(i)}, \tag{3}$$

where K is the number of labels (ROIs) plus the background class.

2.2.4. Training

The networks were implemented using the PyTorch framework and trained with stochastic gradient descent using Adam optimizer (Kingma and Ba, 2014) with the default parameters (the initial learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and no weight decay) on an NVIDIA GeForce GTX 1080 GPU for up to 12 h (train and validation) or on an NVIDIA Volta V100 GPU for up to 24 h (MRM NeAt). Each network was trained with a batch size of one. Qualitatively, the training pace of 2D and 3D networks was substantially the same, as evidenced in supplementary Fig. S1.

We augmented the data online each time an image was loaded by scaling the volumes by a factor α randomly drawn from the interval $[0.95, 1.01]$ and rotating them around each axis by a random angle between $-5\circ$ and $5\circ$. Scaling factors smaller than one were preferred to decrease memory requirements. Each transformation was applied with 50% probability. To further decrease memory requirements, a bounding box was created for each volume using the annotated brain mask as a reference. Each volume was individually normalized to 0 mean and unit variance. Hyperparameters, optimizer and data augmentation scheme were fixed before training ensuring that each architecture would fit into memory, and applied to each network with no additional fine tuning.

2.2.5. Auxiliary bounding-box network

As MU-Net was trained after cropping the volumes to a bounding box, we trained a lighter 2D network to run a first estimate for the brain mask at inference time from the complete volume. This was then used to draw a bounding box around the brain with one voxel margin. This auxiliary network follows exactly the same architecture of MU-Net, omitting any framing or dense connections, and limiting the number of channels to 4, 8, 16 and 32, from the shallowest to the deepest layer. This results in a network with a total number of 122,455 trained parameters.

2.3. STEPS multi-atlas segmentation

STEPS is a state of the art label fusion algorithm to combine multiple registered templates to label a target volume (Cardoso et al., 2013). It takes into account the local and global image matching, combining an expectation-maximization approach with Markov Random Fields to improve on the segmentation based on the quality of the registration itself.

The registrations were performed as follows: before registration, each volume underwent non-parametric N3 bias field correction (Sled et al., 1998) implemented within the ANTS toolset (Avants et al., 2009). Taking each volume as reference, all other volumes were then registered with an affine transformation using FSL FLIRT (Jenkinson and Smith, 2001) and then nonlinearly registered via FSL FNIRT (Andersson et al., 2007; Jenkinson et al., 2012) with the aid of the manually drawn brain mask. Label fusion was achieved with the STEPS algorithm distributed in the NiftySeg package (Cardoso et al., 2013; 2012).

We used correlation ratio (corratio) as the cost function in FLIRT and FNIRT. We used the default FLIRT and FNIRT parameters with the following exceptions. The search range of angles in FLIRT was $[-70°, 70°]$ instead of the default $[-90°, 90°]$, because the orientations of the volumes were similar. In FNIRT, we used spline interpolation instead of the default linear interpolation.

STEPS depends on the number of templates employed and the standard deviation of its Gaussian kernel. We performed a grid search to select the optimal parameters, randomly selecting 10 volumes and labeling them using STEPS. We sampled the standard deviation of the Gaussian kernels between 0.5 and 6 with a stride of 0.5, and the number of templates ranged between 1 and 20 randomly selected volumes. This same process has been performed both using diffeomorphic registration and using affine registration only (supplementary Fig. S2), selecting 16 templates and kernel standard deviation of 1.5 for the diffeomorphic case, and 18 templates with kernel standard deviation of 2.5 for the affinely registered volumes. Exploring both grids required in total 287 h.

Each volume was then segmented using these parameters, randomly selecting an appropriate number of mice as templates for the STEPS algorithm as emerged from the parameter grid search outlined above. We repeated this procedure randomly selecting the same number of templates from mice of the same age only. The mice randomly selected as reference atlases were selected from the training set associated to each volume according to the same 5-fold cross validation scheme used to train the CNNs as outlined in Section 2.5.

When evaluating STEPS on MRM NeAt dataset, we used scripts provided by Ma et al. (2014) at https://github.com/dancebean/multi-atlas-segmentation as this implementation is optimized using this dataset.

The here described computations for the training and validation dataset were executed on a workstation equipped with a 6-core, 12-thread Intel Core i7-8700K CPU running at 3.70 GHz. To accelerate the computations generating several intermediate file outputs, we used RAMdisk to reduce the number of the disk operations. For the NeAt dataset, computations were performed on a 12-core, 24-thread AMD Ryzen 9 3900X Processor.

2.4. Post-processing

The only post-processing steps applied on the segmentation maps were the filling of holes in the resulting 3D volume, the selection of the largest connected component as the brain mask for the skull-stripping task, and assigning all voxels predicted as non-brain to the background class.

2.5. Validation and metrics

To assess the overlap between the ground truth and the predicted segmentation masks, we used the Dice coefficient as the primary performance measure (Dice, 1945). The Dice coefficient is defined as two times the size of the intersection over the sum of the sizes of the two regions:

$$D = \frac{2|Y_t \cap Y|}{|Y_t| + |Y|},$$

where by Y we indicate our prediction and by Y_t the ground truth. This coefficient ranges from 0, meaning no overlap, to 1, indicating a complete overlap between the two regions.

We further evaluated our results using the 95th percentile of the symmetric Hausdorff distance (HD95) (Huttenlocher et al., 1993). HD95 indicates the magnitude of the largest segmentation error compared to the ground truth, expressed in millimeters. We additionally computed precision (defined as $\frac{|Y_t \cap Y|}{|Y|}$) and recall (defined as $\frac{|Y_t \cap Y|}{|Y_t|}$). These measures provide complimentary information to the Dice overlap.

Each experiment on the train and validation dataset as well as the NeAt dataset (see Table 1) was validated according to a 5-fold cross validation (CV) scheme. Volumes were distributed in each fold according to the individual identity of each animal, preventing the use of the volumes from the validation animals for training. The animals were randomly assigned to each fold once, and the same animals remained assigned to their respective folds through all experiments. For train and validation dataset, this resulted in a training set of 25 or 26 mice and a validation set of 6 or 7 mice in each fold. For the MRM NeAt dataset, 5-fold CV resulted in 8 volumes used for training (or as registration atlases) and 2 for testing in each fold. The test dataset was used as an external test set to evaluate MU-Net trained on the train and validation dataset.

Unless otherwise specified, we used a paired permutation test to evaluate the significance of differences between the Dice scores obtained by different methods, pairing the Dice scores obtained on the same MRI volumes. The unpaired permutation test was used instead when comparing results obtained on different volumes, for example, when comparing the accuracy of a model on volumes from younger mice with that of the same model on older mice, and for all comparisons on the test set. We performed permutation tests using 100,000 iterations, and considered average differences to be significant when p was smaller than 0.05. The unpaired permutation tests of Dice coefficients between different animal groups were performed by permuting animals (not images) between the two groups. This ensures exchangeability when several images of the same animal existed due to longitudinal designs in the test set.

3. Results

Using the train and validation dataset, we compared the performance of different network architectures. Furthermore, we compared MU-Net with multi-atlas segmentation on both our data and the MRM NeAt dataset, and evaluated the impact of mouse age on the accuracy of our segmentation maps. The experiments reported in Sections 3.1–3.3 are based on 5-fold CV on the train and validation set, and experiments in Section 3.4 on 5-fold CV on the MRM NeAt dataset. Finally, in Section 3.5, we tested MU-Net trained on train and validation set on an independent test set that included 1782 MRI volumes from 817 mice.

3.1. Architecture comparison

We compared the performance of different networks trained with and without dense connections and dual framing connections, in both 2D and 3D implementations.

As shown in Table 2, all MU-Nets achieved Dice scores with the ground truth comparable to or higher than the typical inter-rater variability of manual segmentation in the mouse brain (Dice scores from 0.80 to 0.90 (Ali et al., 2005)). The skull-stripping task achieved an excellent Dice score of 0.984. The ventricles were characterized by the lowest segmentation performance (average Dice score 0.907), while the cortex displayed the highest overlap with the ground truth (average Dice score 0.966). Dice scores for each animal in all ROIs are provided as supplementary Table S3.

The network displaying the highest average Dice scores was, in fact, the simplest one, including no in-block skip connections nor framing connections, and using 2D convolutions. The accuracy of this network was significantly higher than the accuracy of other all other 2D networks ($p < 0.00003$). Because of its excellent performance and simplicity this network is our choice for the MU-Net architecture, which is the architecture we used for all experiments detailed in Sections 3.2 and 3.3.

The choice between 2D and 3D architectures was the most important factor in increasing performance, resulting in a marked increase in mean Dice scores for both tasks ($p < 0.00001$) between all 2D networks compared to the 3D ones. We further compared MU-Net with one featuring less channels per filter (49, 49, 50, 50, from the shallowest to the deepest convolutional block) to match the number of parameters to the number of parameters of the simplest 2D network. We registered a slightly (but not significantly, $p = 0.077$) lower accuracy compared to MU-Net, indicated as 2D SLP in Table 2.

To test whether the increased performance of 2D architectures compared to the 3D implementation depended on the reduced number of parameters or on an excessive loss of information when pooling in the anterior-posterior direction, we trained a network using 3D filters while limiting pooling operations to the coronal plane. This network achieved a segmentation accuracy in between the 3D and 2D implementations (Table 2), suggesting that both above mentioned aspects were relevant in increasing the algorithm's performance.

We studied the effect of bias field correction to the performance of MU-Net training it on images without bias-correction, and separately, on N3 bias-corrected MR images (Sled et al., 1998). The validation accuracy achieved with bias correction was indistinguishable from the accuracy of MU-Net trained without bias correction (see Table 2).

3.2. Age stratified training sets

We evaluated the performance of MU-Net when restricting the training set to mice of a specific age. Networks trained on data from mice of 12, 16 and 32 weeks achieved higher accuracy, both on their respective validation set and the overall ground truth, compared to the networks trained on 5 weeks mice ($p < 0.00001$). As shown in Fig. 5, while all networks trained on one specific age displayed a statistically significant ($p < 0.05$, unpaired) decrease in mean accuracy when validated on animals of a different age, this difference was highest between the 5 weeks data and the other datasets.

Limiting the training data to one specific age implies that these networks were trained only on a quarter of the data used to train the networks in Section 3.1. Irrespective of that, these networks still achieved average Dice score on the mixed-age validation dataset comparable with

Table 2
CNN and STEPS accuracies measured using Dice coefficient across different methodological choices. Cross-validation results on the train and validation dataset.

Dim	SC	FC	Brain mask	Cortex	Hippocampi	Ventricles	Striati	ROI mean
2D			**0.984±0.005**	**0.966±0.009**	**0.925±0.017**	**0.907±0.020**	**0.939±0.010**	**0.935±0.026**
2D	x	x	**0.984±0.006**	0.963±0.010	**0.924±0.016**	0.905±0.022	0.937±0.009	0.932±0.026
2D		x	**0.984±0.006**	0.963±0.011	**0.924±0.017**	0.905±0.022	0.938±0.009	0.932±0.026
2D	x		**0.984±0.005**	0.964±0.011	0.923±0.018	0.905±0.024	0.937±0.010	0.932±0.027
3D	x	x	0.982±0.007	0.956±0.016	0.914±0.033	0.900±0.025	0.926±0.045	0.924±0.038
3D		x	0.982±0.007	0.958±0.016	0.916±0.032	0.900±0.025	0.928±0.029	0.925±0.034
3D	x		0.982±0.006	0.957±0.016	0.913±0.041	0.899±0.028	0.926±0.042	0.924±0.040
3D			0.982±0.007	0.957±0.013	0.916±0.033	0.899±0.026	0.926±0.039	0.924±0.036
3DConv 2DPool			0.983±0.006	0.961±0.010	0.919±0.026	0.902±0.026	0.934±0.014	0.929±0.030
2D SLP			**0.984±0.005**	**0.965±0.009**	**0.924±0.016**	**0.907±0.021**	**0.939±0.010**	**0.934±0.026**
2D + N3			**0.984±0.005**	**0.965±0.009**	**0.924±0.020**	**0.907±0.020**	**0.939±0.009**	**0.934±0.026**
STEPS (affine)			\	0.920±0.058	0.827±0.079	0.761±0.090	0.873±0.062	0.845±0.093
STEPS (diffeo)			\	0.948±0.036	0.844±0.048	0.812±0.090	0.871±0.045	0.869±0.070
STEPS* (affine)			\	0.936 ± 0.013	0.831 ± 0.029	0.781 ± 0.049	0.887 ± 0.019	0.859 ± 0.066
STEPS* (diffeo)			\	0.954 ± 0.009	0.848 ± 0.025	0.826 ± 0.039	0.885 ± 0.016	0.879 ± 0.055
Majority Voting			\	0.889±0.179	0.780±0.232	0.677±0.208	0.816±0.245	0.791±0.230

Listed values are the average validation Dice scores between automatic and manual segmentation ± standard deviations of these Dice scores in 5-fold CV. ROI mean column refers to the mean Dice coefficient of the cortex, the hippocampi, the ventricles and the striati. SC and FC indicate the presence of skip connection and framing connections. MU-Net results are displayed in the first row. STEPS refers to STEPS using randomly selected templates; STEPS* refers to STEPS runs using randomly selecting mice of the same age only; affine indicates that only affine registration was used, whereas diffeo indicates this was followed by a diffeomorphic registration step; Majority voting refers to the selection of the most occuring label after diffeomorphic registration; 3DConv 2DPool: network featuring no in-block skip connections or framing connections, with 3D filtering and 2D pooling in the coronal plane; 2D SLP: 2D network with in-block skip connections and a limited number of parameters; 2D +N3: 2D network trained on data bias-corrected using the N3 algorithm. Boldface characters indicate the best performing network, achieving significantly higher Dice scores than all other networks for that ROI.

the accuracy of manual segmentation. The worst performing CNN was the network trained on 5 weeks old mice. Training on the 12, 16 and 32 weeks data and validating on mice of the same age, we observed Dice scores comparable with the overall performance of MU-Net trained on the entire dataset ($p > 0.15$, unpaired). However, we measured a lower overall performance when including mice of all ages in the validation data ($p < 0.00001$), slightly overfitting for each specific age.

3.3. Comparison with multi-atlas segmentation

We compared MU-Net with multi-atlas segmentation, applying the state-of-the-art STEPS (Cardoso et al., 2013; 2012) label fusion method to combine the labels obtained from the registration of multiple labeled volumes. This was implemented using the Niftyseg package as described in Section 2.3. We repeated this procedure using both diffeomorphic and affine registration methods, with randomly-selected templates restricted to same-age mice. The brain mask segmentation was not evaluated as the manually drawn mask was used during the diffeomorphic registration procedure.

MU-Net achieved higher Dice coefficients than all STEPS implementations ($p < 0.00001$, Cohen's d: 4.39, see Table 2). Also, there was a marked qualitative difference between STEPS segmentation and MU-Net (Fig. 2), the latter achieving results visually indistinguishable from manual segmentation. We computed HD95 distances further confirmed this difference, with an average of 0.084 ± 0.019mm for MU-Net against 0.251 ± 0.064mm for STEPS ($p < 0.00001$). We measured a mean precision of 0.962 ± 0.008 (MU-Net) vs 0.820 ± 0.025 (STEPS) ($p < 0.00001$) and a mean recall of 0.951 ± 0.011 (MU-Net) vs 0.952 ± 0.013 (STEPS) ($p = 0.65$).

MU-Net had an inference time of about 0.35s and a training time of 12 h. STEPS segmentation procedure required total inference time of 117 min for each labeled volume (on average 440s for each pairwise diffeomorphic registration and 7.85s for label fusion). Implementing STEPS segmentation using only templates of the same age led to a small but significant improvement in Dice coefficients over randomly choosing templates of any age ($p < 0.0007$, Cohen's d: 0.296). The employment of diffeomorphic registration was the most important factor affecting the performance of STEPS, as displayed in Table 2. A simple majority voting strategy led to significantly lower performance in all ROIs compared to all other label fusion strategies ($p < 0.003$).

Furthermore, we trained MU-Net on the outputs of the implemented STEPS procedures featuring diffeomorphic registration, and measured the Dice scores of each network's output with the ground truth (Table 3). As evidenced in Tables 2 and 3, and Fig. 3, MU-Net trained on STEPS segmentations achieved higher Dice score with the ground truth than the same STEPS segmentations constituting the training sets of MU-Net ($p < 0.00001$). With the exception of the network trained on 5 weeks old mice, these hybrid networks were still under-performing compared to training on manually segmented data ($p < 0.00001$).

3.4. Evaluation on a large number of ROIs with MRM NeAt dataset

We trained and evaluated MU-Net on the MRM NeAt datasets that includes atlases of 10 individual T_2 *-weighted in vivo brain MR images of 12–14 weeks old C57BL/6J mice; each with 37 manually labelled anatomical structures (Ma et al., 2008). This same database was selected by Ma et al. (2014) to evaluate the STEPS multi-atlas segmentation algorithm on mouse brain MRI. To compare MU-Net with STEPS, we followed the STEPS implementation by Ma et al. (2014) as released by the authors.

We used a 5-fold cross validation scheme for evaluation (8 templates for training and 2 templates for testing in each fold). The only adaptation required to train MU-Net on MRM NeAT dataset was to expand the number of output channels to 37 (plus one for the brain mask) to equal that of the number of ROIs. As displayed in Fig. 4, Dice coefficient of MU-Net was greater or comparable to STEPS: while in a majority of regions MU-Net's accuracy was higher than the accuracy of STEPS, this was statistically significant only for the brain mask, external capsule, hypothalamus and brain stem. In the left inferior colliculi, STEPS achieved significantly higher Dice coefficient than MU-Net. Averaging the Dice coefficients across all ROIs, we measured an average Dice score of 0.820 ± 0.031 for MU-Net and 0.814 ± 0.023 for STEPS. While this aver-

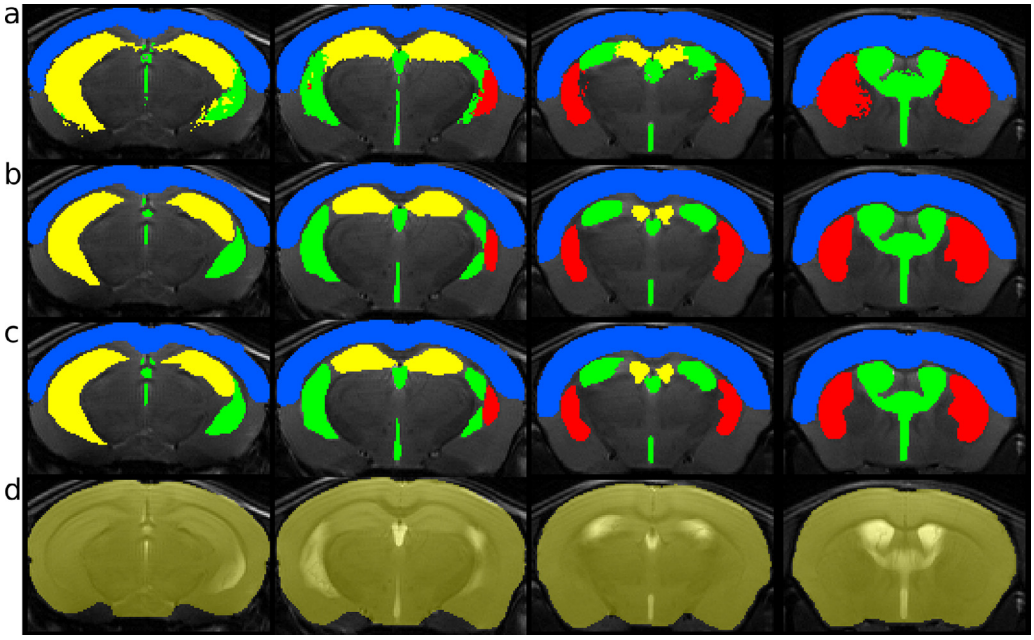

Fig. 2. Segmentation comparison in four slices from a single animal: (a) STEPS, (b) MU-Net, and (c) manual annotation. In (a)–(c), the regions highlighted are the cortex (blue), ventricles (green), striati (red), and hippocampi (yellow). Panel (d) shows the inferred brain mask by MU-Net.

Table 3
Mean and standard deviation of average Dice scores evaluating the accuracy of MU-Net trained on volumes segmented via STEPS.

Training Set	Cortex	Hippocampus	Ventricles	Striatum	ROI mean
STEPS*	0.954±0.011	0.867±0.027	0.866±0.035	0.898±0.017	0.896±0.043
STEPS	0.953±0.009	0.872±0.022	0.849±0.041	0.885±0.016	0.890±0.046

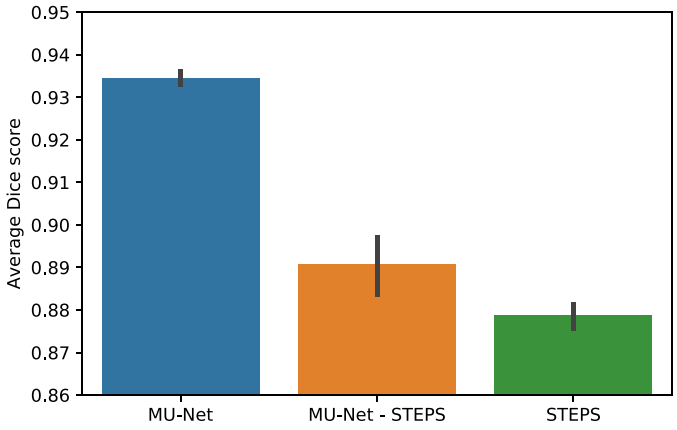

Fig. 3. Average Dice score comparison between different segmentation methods, across all ROIs. MU-Net: MU-Net trained on the manually segmented data; MU-Net - STEPS: MU-Net trained on volumes segmented employing same-age diffeomorphic STEPS; STEPS: same-age diffeomorphic STEPS segmentation. The error bar represents standard deviation.

age Dice coefficient for MU-Net was higher, the difference was not statistically significant ($p = 0.170$, Cohen's d: 0.134). Similarly, we measured an higher (but not statistically significant, $p = 0.07$) average HD95 distance for MU-Net (0.360 ± 0.252mm vs 0.240 ± 0.038mm). In contrast, we measured a significantly higher average precision with MU-Net (0.823 ± 0.033 vs 0.786 ± 0.024, $p = 0.0009$) and a significantly lower recall (0.815 ± 0.032 vs 0.853 ± 0.023, $p = 0.001$). A full breakdown of these metrics is available in supplementary Fig. S3. The computation time required by STEPS to segment a single volume was of approximately 20 min while MU-Net required less than one second per volume.

3.5. Evaluation with a large test dataset

We optimized the MU-Net model on the train and validation dataset and tested on a large test set of 1782 MRI volumes, acquired from 817 mice with ages ranging from 4 to 60 weeks, and including both WT and HT mice. As the 5-fold cross-validation experiment produced five different MU-Net models, the segmentation maps for the test set were obtained by averaging the five prediction maps produced by the five models. To outline the brain mask, we averaged sigmoid-activated predictions from five networks and thresholded them at 0.5. For region segmentation, we averaged the softmax-activated output maps, and for each voxel, we selected the class yielding the maximal averaged value as our predicted label.

Out of the entire test set, segmentation failed completely on two volumes, where no brain mask was detected. The remaining 1780 volumes were successfully segmented with an average Dice score of 0.978 $\pm$ 0.012 for the brain mask, 0.906 $\pm$ 0.041 for the striati, and 0.937 $\pm$ 0.035 for the cortex, distributed as illustrated in Fig. 7. There was no significant difference between the segmentation accuracy of male and female animals ($p > 0.1$, unpaired). However, there was a significant difference in accuracy between HT and WT mice ($p < 0.00001$, unpaired) for all ROIs. Dice scores of WT animals were 0.4% higher for the brain mask, 1.7% higher for the cortex, and 1.9% higher for the striati. Applying N3 bias correction on all volumes before segmentation did not result in a significant Dice score difference. A detailed list of Dice scores, HD95, precision and recall, for each animal and each ROI, is available in supplementary Table S4.

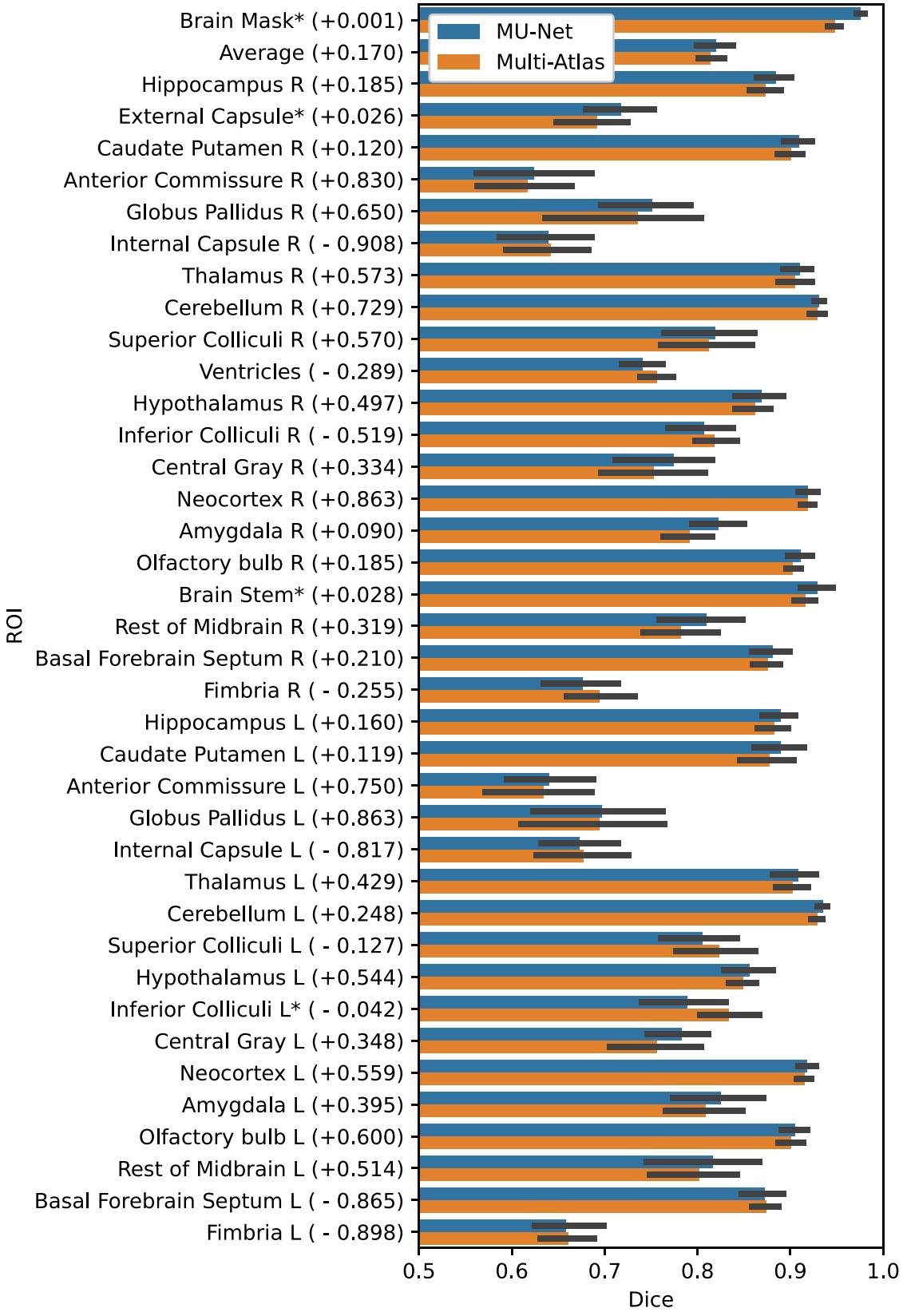

Fig. 4. Comparison between the average Dice coefficients of MU-Net and STEPS multi-atlas algorithm by Ma et al. Error bars correspond to standard deviation for the average accuracy. Permutation-test based p-values for each comparison are provided in parentheses after the ROI name, + indicates that the average Dice coefficient for MU-Net was higher and - indicates that the average Dice coefficient for STEPS was higher, * indicates a statistically significant difference.

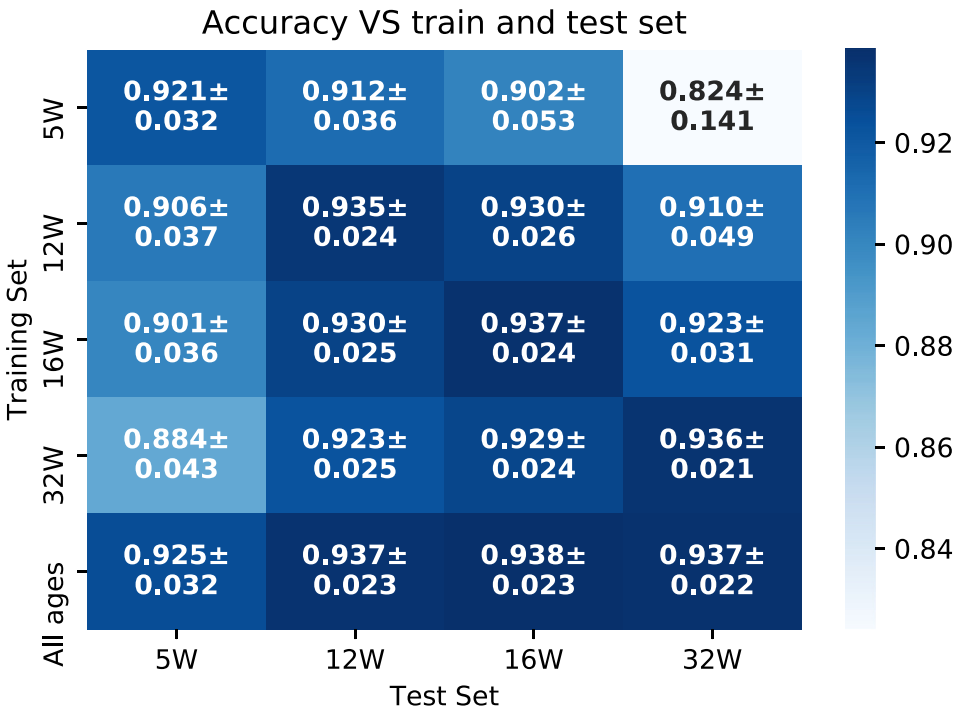

Fig. 5. Mean accuracy ± standard deviation for the average accuracy of MU-Net trained and evaluated on different datasets according to mouse age. Networks exclusively trained on older animals achieved lower accuracy when attempting to generalize to the youngest animals, and vice-versa.

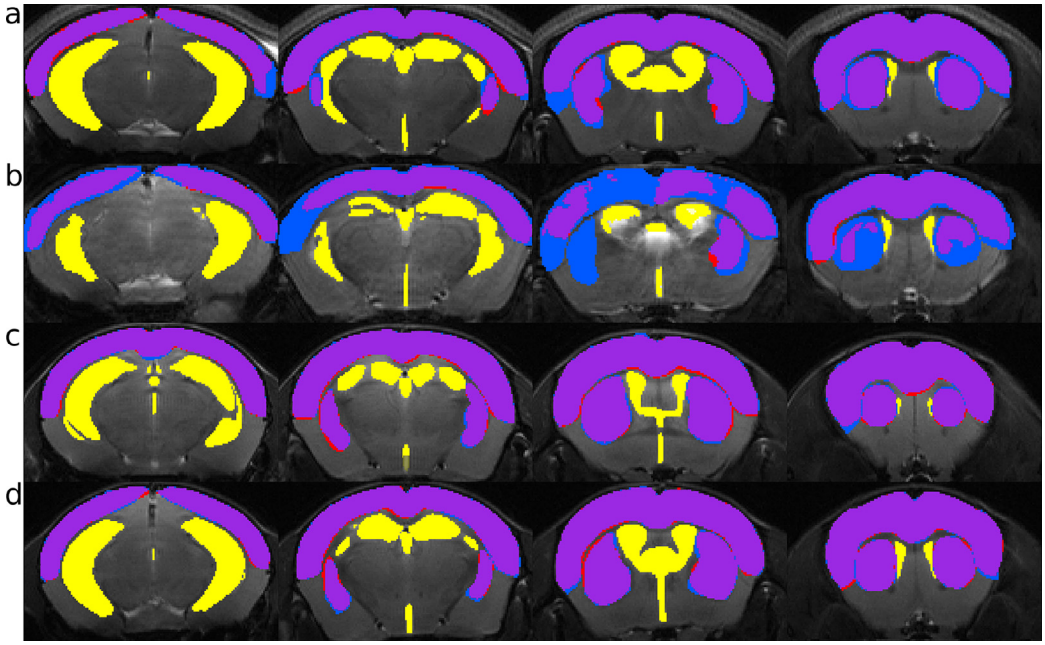

Fig. 6. MU-Net segmentation compared to the manual segmentation in four slices of four volumes of the test set. Blue and red indicate, respectively, ground truth and inferred segmentation, purple their overlap (striati and cortex); yellow ROIs (ventricles and hippocampi) are inferred ROIs for which manual annotations were not available. Rows indicate (a) the highest performing volume (mean Dice 0.964, 8 weeks old R6/2 mouse); (b) the lowest performing volume (mean Dice 0.685, 12 weeks old R6/2 mouse); (c) the volume displaying performance closest to the mean performance on the entire test set (Dice 0.923, 12 weeks old Q175DN mouse); (d) one randomly selected volume (Dice 0.919, 8 weeks old Q175DN mouse)

A visual inspection of the segmentation maps (Fig. 6) revealed that ROIs were qualitatively similar to those obtained on the validation set and displayed in Fig. 2. We observed, however, a visible decrease in performance in the presence of strong ringing artifacts (Fig. 6.b) This is further reflected in the higher average HD95 distances in the test dataset than in the validation dataset (Table 4).

4. Discussion

We have presented a multi-task deep neural network, MU-Net, for the simultaneous skull-stripping and segmentation of mouse brain MRI. We selected the best performing network among a number of architectures and found it to achieve better segmentation accuracy on the validation set compared to state-of-the-art multi-atlas segmentation procedures, with a markedly lower segmentation time (0.35s vs 117min). We then evaluated the performance of MU-Net on a large and hetero-

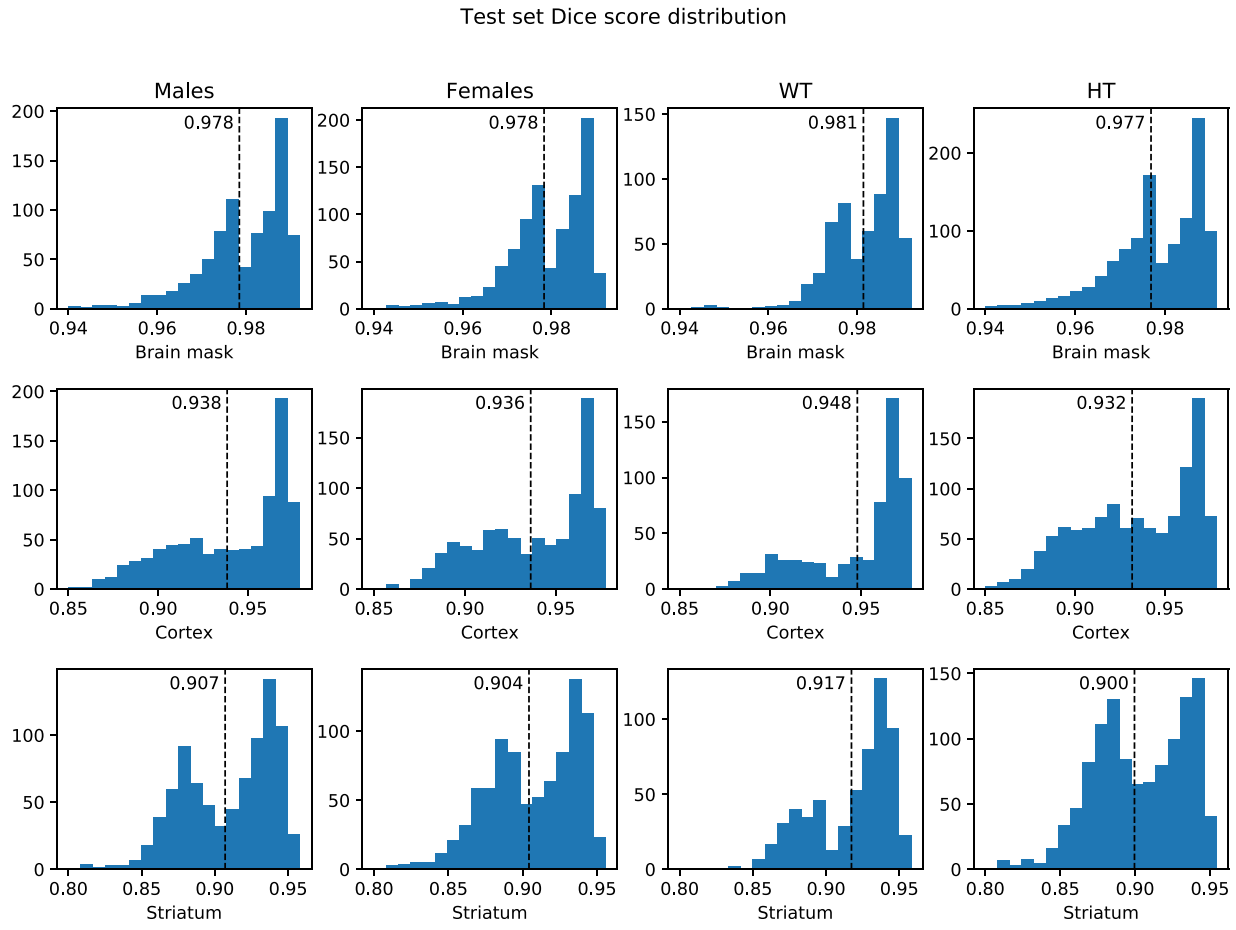

Fig. 7. Test set Dice score distribution for the brain mask, cortex and striati ROIs. Males and Females include all mice of each gender, both WT and TG. Likewise, WT and TG include both males and females.

Table 4
Average test set metrics (see Supplementary Table S4 for details).

Metric	Brain Mask	Cortex	Striati
Dice	0.978 ± 0.012	0.937 ± 0.035	0.906 ± 0.041
HD95 (mm)	0.345 ± 0.303	0.223 ± 0.231	0.180 ± 0.167
Precision	0.989 ± 0.006	0.939 ± 0.050	0.929 ± 0.045
Recall	0.969 ± 0.022	0.939 ± 0.054	0.888 ± 0.062

geneous test set of 1782 mice from 10 different studies of Huntington disease, with varying ages and genetic backgrounds (WT as well as HT Q175 and R6/2 variants). In this test set, we measured average Dice scores of 0.978, 0.906 and 0.937 for the brain mask, striati and cortex, rivaling human-level performance. We additionally trained MU-Net for the segmentation of high resolution mouse MRIs of the MRM Neat atlas into 37 ROIs measuring an average Dice score of 0.820. Hence, we argue that the employment of deep neural networks for the segmentation of animal MRI is a promising strategy for the reduction of both rater bias and segmentation time.

To put the Dice scores we have reported in context, Dice scores between two human experts have ranged from 0.80 to 0.90, depending on ROI, for mouse brain MRI segmentation (Ali et al., 2005). For different segmentation tasks in brain MRI in general, including human data, inter- and intra-rater Dice score have ranged between 0.75 and 0.96 (Ali et al., 2005; Entis et al., 2012; Yushkevich et al., 2006). The Dice scores of MU-Net exceeded the above mentioned scores between two human experts, suggesting human-level segmentation performance. In addition, the Dice score of MU-Net for skull-stripping was higher than Dice score from the skull-stripping CNN implemented by Roy et al. (2018b) (0.949). Obviously, comparing previously reported Dice scores to our segmentation accuracy measures must be done with care as these vary across different studies, segmentation tasks, and datasets, and the confounding factors include image resolution, presence of artifacts and noise, rater expertise, and the choice of ROIs.

While Roy et al. (2018b) proposed a CNN for skull-stripping for mouse MRI, to our knowledge this work represents the first CNN performing both region segmentation and skull-stripping in mouse brain MRI. The advantages of CNNs with respect to atlas-based region segmentation (Bai et al., 2012; De Feo and Giove, 2019; Ma et al., 2014) are clear. First, compared to atlas-based segmentation MU-Net is much faster and produces accurate results without pre-processing. Second, we found MU-Net to be significantly more accurate than the state-of-the-art STEPS multi-atlas segmentation (Ma et al., 2014) on anisotropic, relatively quick to acquire MR images favored in pre-clinical drug and biomarker discovery applications. Third, we found MU-Net to perform better than or equally well compared to STEPS on isotropic, high-resolution MR images with relatively long acquisition times, favored in basic research.

We observed that the segmentation accuracy of atlas-based methods can vary markedly, based on the specific use case depending on the number of manually drawn ROIs, voxel-size, and image quality. The best performance was achieved using advanced registration-based methods (Ma et al., 2014) on the high resolution data (Ma et al., 2008) with a densely labeled atlas of 37 ROIs, and the lowest using a majority voting rule on a sparsely outlined atlas with a low resolution along the fronto-caudal direction.

With a dense segmentation of high resolution images (NEaT dataset), we measured slightly higher average Dice coefficients with MU-Net than with STEPS, but the difference was not statistically significant. Therefore, it appears that for this case the main advantage of MU-Net over STEPS would be in terms of segmentation time. The performance of MU-Net on the NeAt dataset was likely hampered by the small number of training images available (8 images for training in each fold). This also provides an explanation for the higher standard deviation for HD95 distances for MU-Net compared to STEPS. Interestingly, MU-Net achieved Dice coefficients similar to STEPS with a larger average precision but a lower average recall. This would indicate that STEPS prediction contained more false positives, labeling background voxels as belonging to ROIs, and conversely MU-Net's prediction favored false negatives. For sparsely segmented images, typical in drug development, where only specific structures are of interest, STEPS appears to be markedly less effective than MU-Net, and the time required for manual annotation is notably decreased. This also means that it might be feasible to annotate a small number of volumes as required by the specific study, and then use MU-Net to automate the segmentation of the remaining data.

Interestingly, MU-Nets trained on automatic STEPS multi-atlas segmentations achieved higher Dice score with the ground truth than STEPS, highlighting the generalization ability of MU-Net. This supports the use of atlas based segmentation methods to augment MRI segmentation datasets suggested in Roy et al. (2018a), leveraging unlabeled data. The results obtained by training on STEPS segmentations alone remain, however, of insufficient quality to eliminate the need for manual annotations in the training data, as the CNN attempts to replicate any form of systematic error present in the atlas-based labeling procedure.

In literature both 3D and 2D implementations of CNNs are available for different segmentation tasks (Çiçek et al., 2016; Milletari et al., 2016; Roy et al., 2018a), and other architectural variants have been proposed: Roy et al. (2018a) added dense connections (Huang et al., 2017) in the convolution blocks of U-Net while keeping the number of output channels constant; Han and Ye (2018) proposed two variants based on signal processing arguments for the reduction of artifacts in a sparse image reconstruction task. We, however, found that a more complex model did not improve and in fact lowered the accuracy of our results, perhaps given the simplicity of the task. Thus, in agreement with Isensee et al. (2018), we found that a 2D approach was preferable to 3D approach in the presence of anisotropic voxels. We also found the Dice loss to be sufficient to effectively train our model without the addition of a cross-entropy loss. As we did not perform any fine tuning of hyperparameters for any of our models, it is possible that after sufficient fine tuning the performance of one of these alternative approaches might be improved.

Much like the human eye, MU-Net was not significantly affected by the presence of the bias field, and did not benefit from N3 bias correction. Correcting for the bias field might still be beneficial as it depends on the specific experimental setup, and thus N3 bias correction might avoid specializing the network to one particular acquisition procedure. For this reason, we release the trained parameters of the model for MU-Net trained on both the non-corrected and the N3-corrected data.

To ensure the network generalizes to a wide age range, our results indicate that the distinctive features present before adulthood need to be adequately represented in the training data. This is evidenced by the degraded performance observed when testing networks trained on 5-week old mice on the volumes acquired from older ones, and vice-versa. As mice are typically weaned at 3–4 weeks and attain sexual maturity at 8–12 weeks (Dutta and Sengupta, 2016), 5-week old mice are not adults. In contrast, training solely on male mice did not significantly influence MU-Net performance on female animals. We studied why the Dice coefficient distributions were bi-modal with the large test set (see Fig. 7). The bi-modal nature of the distributions appears not to be explained by differences between different studies, genders, or genotypes (see supplementary Figs. S4 and S5). We cannot offer a definitive explanation for the cause of these bi-modal distributions, however, we speculate that it is a sum of several factors, including intra-rater segmentation variability.

An obvious limitation of our approach is its specialization for the specific MRI contrast the algorithm is trained on. Making MU-Net to be more robust to marked changes in the image acquisition could be achieved by expanding the training data to be more variable or/and utilizing techniques such as domain adaptation, transfer learning or image translation to minimize the amount of new training data for the model to generalize to new type of MRI acquisition (Armanious et al., 2020; Zhuang et al., 2020). This research line is one of the most important areas for future research in MRI segmentation with deep learning. However, MU-Net successfully generalized to a variety of transgenic mice in an age range wider than that of the training set, thus offering a valuable way to automate segmentation tasks. Another limitation of this study is the number of ROIs as mouse brain atlases with extremely detailed segmentation featuring over 700 ROIs currently exist (Nie et al., 2019). However, atlases such as (Nie et al., 2019) are constructed by specialized procedures and do not contain manual segmentations of all images used in the atlas construction. Therefore, these atlases are not directly applicable for training segmentation neural networks.

The employment of CNNs for the segmentation of mouse brain MRI provides a number of benefits for preclinical researchers. Beyond allowing for the employment of large datasets in a time-efficient manner, the ability to generalize and abstract from the training data results in more robust and reproducible predictions. We can thus expect these methods to reduce the confounding effect of intra- and inter-rater variability inherent in manual segmentation procedures while streamlining animal MRI experimental pipelines.

Declarations

Data availability statement

MU-Net code and trained models are freely available at https://github.com/Hierakonpolis/MU-Net. A tutorial of usage of MU-Net is available at https://github.com/Hierakonpolis/NN4Kubiac The training and validation dataset is property of Charles River Discovery Services, and the test dataset is property of CHDI 'Cure Huntington's Disease Initiative' foundation. The MRM NeAt dataset is freely available at https://github.com/dancebean/mouse-brain-atlas. All the Dice scores between MU-Net and manual segmentations are available as supplementary files to this manuscript.

Ethics statement

All animal experiments were carried out according to the United States National Institute of Health (NIH) guidelines for the care and use of laboratory animals, and approved by the National Animal Experiment Board.

Credit authorship contribution statement

Riccardo De Feo: Methodology, Software, Formal analysis, Writing - original draft. **Artem Shatillo:** Data curation. **Alejandra Sierra:** Methodology, Formal analysis. **Juan Miguel Valverde:** Methodology. **Olli Gröhn:** Conceptualization. **Federico Giove:** Conceptualization. **Jussi Tohka:** Conceptualization, Software, Writing - original draft.

Acknowledgments

R.D.F.'s work has received funding from the European Union's Horizon 2020 Framework Programme under the Marie Skłodowska Curie grant agreement No #691110 (MICROBRADAM) and J.M.V.' work was founded from Marie Skłodowska Curie grant agreement No #740264 (GENOMMED). The content is solely the responsibility of the

authors and does not necessarily represent the official views of the European commission.

The authors wish to acknowledge CSC - IT Center for Science, Finland, for computational resources.

We also extend our thanks to the Academy of Finland, grants (#275453 to A.S. and #298007 to O.G. #316258 to J.T.) and to the CHDI 'Cure Huntington's Disease Initiative' foundation, for kindly providing us with the test data employed in this work. We acknowledge a grant S21770 from European Social Fund to J.T.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.neuroimage.2021.117734

References

Ali, A.A., Dale, A.M., Badea, A., Johnson, G.A., 2005. Automated segmentation of neuroanatomical structures in multispectral mr microscopy of the mouse brain. Neuroimage 27 (2), 425–435.

Anderson, R.J., Cook, J.J., Delpratt, N., Nouls, J.C., Gu, B., McNamara, J.O., Avants, B.B., Johnson, G.A., Badea, A., 2019. Small animal multivariate brain analysis (samba)–a high throughput pipeline with a validation framework. Neuroinformatics 17 (3), 451–472.

Andersson, J.L., Jenkinson, M., Smith, S., et al., 2007. Non-linear Registration AKA Spatial Normalisation FMRIB Technial Report tr07ja2. FMRIB Analysis Group of the University of Oxford.

Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., Yang, B., 2020. Medgan: medical image translation using GANS. Comput. Med. Imaging Graph. 101684.

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

Avants, B.B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., Gee, J.C., 2010. The optimal template effect in hippocampus studies of diseased populations. Neuroimage 49 (3), 2457–2466.

Bai, J., Trinh, T.L.H., Chuang, K.-H., Qiu, A., 2012. Atlas-based automatic mouse brain image segmentation revisited: model complexity vs. image registration. Magn. Reson. Imaging 30 (6), 789–798.

Calabrese, E., Badea, A., Cofer, G., Qi, Y., Johnson, G.A., 2015. A diffusion MRI tractography connectome of the mouse brain and comparison with neuronal tracer data. Cereb. Cortex 25 (11), 4628–4637.

Cardoso, M.J., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N.C., Ourselin, S., Initiative, A.D.N., et al., 2013. Steps: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. Med. Image Anal. 17 (6), 671–684.

Cardoso, M.J., Modat, M., Ourselin, S., Keihaninejad, S., Cash, D., 2012. Steps: multi--label similarity and truth estimation for propagated segmentations. In: Mathematical Methods in Biomedical Image Analysis (MMBIA), 2012 IEEE Workshop on. IEEE, pp. 153–158.

Chou, N., Wu, J., Bingren, J.B., Qiu, A., Chuang, K.-H., 2011. Robust automatic rodent brain extraction using 3-d pulse-coupled neural networks (PCNN). IEEE Trans. Image Process. 20 (9), 2554–2564.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, pp. 424–432.

Cunha, L., Horvath, I., Ferreira, S., Lemos, J., Costa, P., Vieira, D., Veres, D.S., Szigeti, K., Summavielle, T., Máthé, D., et al., 2014. Preclinical imaging: an essential ally in modern biosciences. Mol. Diagn. Ther. 18 (2), 153–173.

De Feo, R., Giove, F., 2019. Towards an efficient segmentation of small rodents brain: a short critical review. J. Neurosci. Methods 323, 82–89.

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

Dutta, S., Sengupta, P., 2016. Men and mice: relating their ages. Life Sci. 152, 244–248.

Entis, J.J., Doerga, P., Barrett, L.F., Dickerson, B.C., 2012. A reliable protocol for the manual segmentation of the human amygdala and its subregions using ultra-high resolution MRI. Neuroimage 60 (2), 1226–1235.

Febo, M., Foster, T.C., 2016. Preclinical magnetic resonance imaging and spectroscopy studies of memory, aging, and cognitive decline. Front. Aging Neurosci. 8, 158.

Han, Y., Ye, J.C., 2018. Framing u-net via deep convolutional framelets: application to sparse-view ct. IEEE Trans. Med. Imaging 37 (6), 1418–1429.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the hausdorff distance. IEEE Trans. Pattern Anal. Mach.Intell. 15 (9), 850–863.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al., 2018. nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv:1809.10486.

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

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156.

Karimi, D., Salcudean, S. E., 2019. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. arXiv:1904.10030.

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

Kovačević, N., Henderson, J., Chan, E., Lifshitz, N., Bishop, J., Evans, A., Henkelman, R., Chen, X., 2004. A three-dimensional MRI atlas of the mouse brain with estimates of the average and variability. Cereb. Cortex 15 (5), 639–645.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436.

Lerch, J.P., Sled, J.G., Henkelman, R.M., 2011. Mri phenotyping of genetically altered mice. In: Magnetic Resonance Neuroimaging. Springer, pp. 349–361.

Ma, D., Cardoso, M.J., Modat, M., Powell, N., Wells, J., Holmes, H., Wiseman, F., Tybulewicz, V., Fisher, E., Lythgoe, M.F., et al., 2014. Automatic structural parcellation of mouse brain MRIusing multi-atlas label fusion. PLoS One 9 (1), e86576.

Ma, Y., Smith, D., Hof, P.R., Foerster, B., Hamilton, S., Blackband, S.J., Yu, M., Benveniste, H., 2008. In vivo 3d digital atlas database of the adult c57bl/6j mouse brain by magnetic resonance microscopy. Front. Neuroanat. 2, 1.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models. In: Proc. ICML, 30, p. 3.

Matthews, P.M., Coatney, R., Alsaid, H., Jucker, B., Ashworth, S., Parker, C., Changani, K., 2013. Technologies: preclinical imaging for drug development. Drug Discov. Today 10 (3), e343–e350.

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

Nie, B., Wu, D., Liang, S., Liu, H., Sun, X., Li, P., Huang, Q., Zhang, T., Feng, T., Ye, S., et al., 2019. A stereotaxic MRI template set of mouse brain with fine sub-anatomical delineations: application to memri studies of 5xfad mice. Magn. Reson. Imaging 57, 83–94.

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

Oguz, I., Zhang, H., Rumple, A., Sonka, M., 2014. Rats: rapid automatic tissue segmentation in rodent brain MRI. J. Neurosci. Methods 221, 175–182.

Pagani, M., Damiano, M., Galbusera, A., Tsaftaris, S.A., Gozzi, A., 2016. Semi-automated registration-based anatomical labelling, voxel based morphometry and cortical thickness mapping of the mouse brain. J. Neurosci. Methods 267, 62–73.

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

Roy, A. G., Conjeti, S., Navab, N., Wachinger, C., 2018a. Quicknat: segmenting MRI neuroanatomy in 20 seconds. arXiv:1801.04161.

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

Schwarz, A.J., Danckaert, A., Reese, T., Gozzi, A., Paxinos, G., Watson, C., Merlo-Pich, E.V., Bifone, A., 2006. A stereotaxic MRI template set for the rat brain with tissue class distribution maps and co-registered anatomical atlas: application to pharmacological MRI. Neuroimage 32 (2), 538–550.

Sharief, A.A., Badea, A., Dale, A.M., Johnson, G.A., 2008. Automated segmentation of the actively stained mouse brain using multi-spectral mr microscopy. Neuroimage 39 (1), 136–145.

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

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 240–248.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.

Valverde, J.M., Shatillo, A., De Feo, R., Gröhn, O., Sierra, A., Tohka, J., 2019. Automatic rodent brain MRI lesion segmentation with fully convolutional networks. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 195–202.

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

Xie, L., Qi, Y., Subashi, E., Liao, G., Miller-DeGraff, L., Jetten, A.M., Johnson, G.A., 2015. 4d MRI of polycystic kidneys from rapamycin-treated glis3-deficient mice. NMR Biomed. 28 (5), 546–554.

Yang, X., Zeng, Z., Yeo, S. Y., Tan, C., Tey, H. L., Su, Y., 2017. A novel multi-task deep learning model for skin lesion segmentation and classification. arXiv:1703.01025.

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

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., 2020. A comprehensive survey on transfer learning. Proc. IEEE inpress.