BEN: a generalizable Brain Extraction Net for multimodal MRI data from rodents, nonhuman primates, and humans

Ziqi Yu[1,2,3], Xiaoyang Han[1,2], Wenjing Xu[1,2], Jie Zhang[1,2], Carsten Marr[4], Dinggang Shen[5,6], Tingying Peng[7,*], Xiao-Yong Zhang[1,2,3,*] , Jianfeng Feng[1,2,3]

[1] Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, 200433, China
[2] MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China
[3] MOE Frontiers Center for Brain Science, Fudan University, Shanghai, 200433, China
[4] Institute of AI for Health (AIH), Helmholtz Zentrum München, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany
[5] School of Biomedical Engineering, ShanghaiTech University, Shanghai, 201210, China
[6] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, 200030, China
[7] Helmholtz AI, Helmholtz Zentrum München, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany

*Corresponding authors:

Xiao-Yong Zhang, PhD, Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, 200433, China (Phone: +86 21 6564 8379; Fax: +86 21 5566 5563; E-mail:xiaoyong_zhang@fudan.edu.cn)

Tingying Peng， PhD, Helmholtz AI, Helmholtz Zentrum München, Germany (E-mail: tingying.peng@tum.de)

Abstract

Extracting brain tissue is usually a vital first step in processing brain magnetic resonance imaging (MRI) data in neuroimaging studies. However, a generalizable brain tissue extraction tool is still lacking. Here, we propose a domain-adaptive and semi-supervised deep neural network, named the Brain Extraction Net (BEN), to extract brain tissues across species, MRI modalities and platforms while requiring minimal or even no annotations. We have evaluated BEN on 18 independent datasets including 1,029 rodent and nonhuman primate scans as well as 4,601 human scans, covering five species, four modalities, and six platforms with various magnetic field strengths. Compared to conventional toolboxes, the superiority of BEN is illustrated by its accuracy, robustness, and generalizability. BEN provides uncertainty maps to quantify the confidence of the network's decisions and to model interrater variability. We design BEN as an open-source software that provides interfaces compatible with several popular neuroimaging toolboxes. Our results demonstrate that BEN has great promise to enhance neuroimaging studies at high throughput for both preclinical and clinical applications.

Keywords: brain tissue extraction; deep learning; MRI; neuroimaging; cross-species

Introduction

Magnetic resonance imaging (MRI) provides noninvasive, high-resolution, multicontrast, and multidimensional deep soft tissue imaging, making it an attractive modality for brain imaging in humans, nonhuman primates (NHPs), and rodents. Brain tissue extraction, also known as skull stripping, is a critical preprocessing step in almost all MRI-based neuroimaging studies, as inaccurate brain extraction can lead to non-brain signal contamination or misleading registration, compromising the accuracy of downstream analyses.

Conventionally, the automatic brain extraction tools that have been developed to date have been categorized into three groups based on their target species: i) humans, ii) NHPs, and iii) rodents. Most well-established tools developed for humans, such as FreeSurfer[1], Analysis of Functional NeuroImages (AFNI)[2], and the FMRIB Software Library (FSL)[3], have been routinely used and integrated into standard preprocessing pipelines, e.g., fMRIPrep[4] and Data Processing & Analysis for Brain Imaging (DPABI)[5]. In contrast, brain extraction in animals is far from being standardized or fully automated. Although a few atlas- or morphometry-based tools have been developed, e.g., for NHPs[6,7] and for rodents[8–11], their performance is still limited.

Recently, deep learning (DL)-based algorithms have been developed to enable more accurate brain tissue segmentation in animals. For example, U-Net-based algorithms have been proposed to perform skull stripping automatically for NHPs[12–15] and for rodents[16–18]. Although these DL-based approaches outperform traditional tools such as FreeSurfer or FSL in NHP and rodent brains, they have mostly been evaluated on a single species and a single MRI modality, and their performance could drop if they are applied to other species or modalities that differ from those represented in their training data. Despite the great diversity in neuroimaging studies in terms of species, MRI modalities, and platforms (e.g., different magnetic field strengths), accurately extracting brain tissue from images collected via multiple MRI modalities or platforms using a single tool remains very difficult, even without additionally attempting to cover various species comprising rodents, NHPs, and humans. In summary, a comprehensive and generalizable toolkit to address these complex challenges is still lacking because current methods 1) are inflexible, which means that the available toolboxes are designed for certain species, modalities or platforms and are not suitable for cross-species, cross-modality, or cross-platform application, and 2) require labor-intensive manual curation, which means that they frequently result in erroneous segmentation, necessitating time-consuming manual correction.

To address these issues, we propose a domain-adaptive and semi-supervised deep neural network, named the Brain Extraction Net (BEN, **Figure 1**), to perform the challenging task of accurately extracting brain tissues across different species (mouse, rat, marmoset, macaque, and human), across different MRI platforms with various magnetic field strengths (from 1.5T to 11.7T), and across different MRI modalities (structural MRI and functional MRI). Being flexible and generalizable, BEN is also considerably faster for brain extraction than using a traditional toolbox. Another feature of BEN is that it incorporates an uncertainty-based assessment module that allows it to autonomously control the quality of brain segmentation mask outputs. In contrast, when using a traditional toolbox, experts need to manually check hundreds of extracted brain scans which is very time-consuming. It is worth noting that this laborious manual inspection is not a one-time procedure; it must be repeated after making parameter changes or

switching to a new toolbox. We believe that an automated and generalizable brain extraction tool such as BEN can reduce intra- and interrater variability while also dramatically increasing throughput for the labor-intensive brain extraction task, hence facilitating many neuroimaging studies.

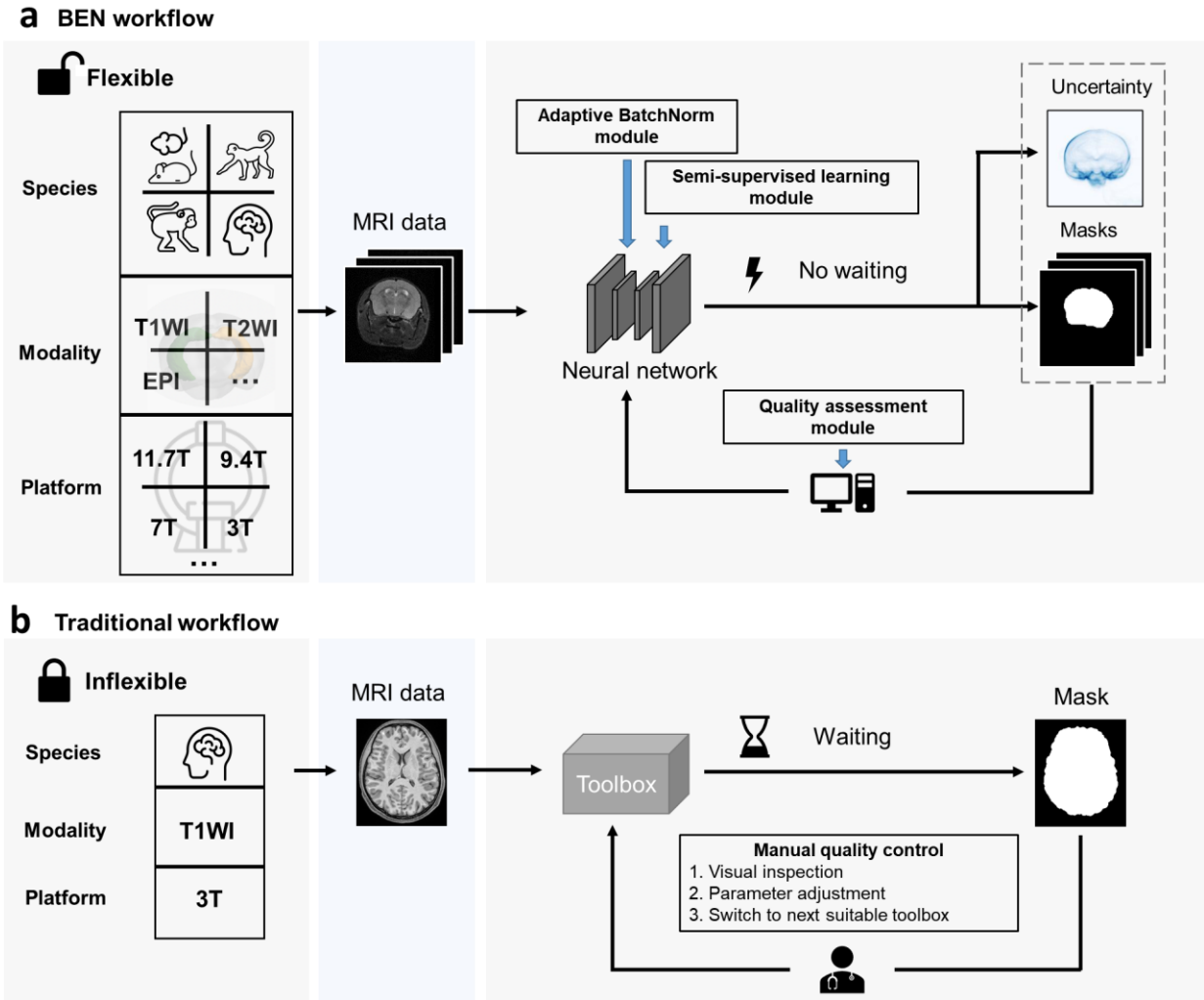

Figure 1. BEN renovates the brain extraction workflow to adapt to multiple species, modalities and platforms. The BEN workflow **(a)** has the following advantages over the traditional toolbox-based workflow **(b)**: (1) Transferability and flexibility: BEN can adapt to different species, modalities and platforms through its adaptive batch normalization module and semi-supervised learning module. (2) Automatic quality assessment: Unlike traditional toolboxes, which rely on manual inspection to assess the brain extraction quality, BEN incorporates a quality assessment module to automatically evaluate its brain extraction performance. (3) Speed: As a DL-based method, BEN can process an MRI volume faster (<1 s) than traditional toolboxes (several minutes or longer).

Results

Datasets

We evaluated BEN on eighteen datasets covering five distinct species, namely, mouse, rat, marmoset, macaque, and human; four MRI modalities, namely, T1- and T2-weighted imaging (T1WI/T2WI), echo-planar imaging (EPI), susceptibility-weighted imaging (SWI) and arterial spin labeling (ASL); and six MRI platforms spanning a wide range of magnetic field strengths, namely, 1.5 T, 3 T, 4.7 T, 7 T, 9.4 T, and 11.7 T. Each dataset contains 14–3,250 scans acquired from different research institutions/cohorts worldwide, including China, the United States, the United Kingdom, France, Canada, and the Netherlands. In total, these datasets comprise 1,029 rodent and NHP scans totaling 52,302 slices as well as 4,601 human scans totaling 873,453 slices. The details and characteristics of all datasets are provided in **Tables S1** and are further summarized in the *Methods* section. Because T2WI is the most commonly used modality for rodent brain imaging and we have accumulated extensive brain scans with high-quality annotations in our previous research studies [19,20], we chose the Mouse-T2WI-11.7T dataset, which contains 243 T2WI scans, as the source-domain dataset and the remaining 17 datasets as the target-domain datasets (one dataset corresponds to one domain). To the best of our knowledge, the datasets used in our study represent the largest-scale DL application of skull stripping reported for rodents and NHPs to date in terms of the diversity in species, field strengths, and modalities. Among all 18 datasets (**Table S1**), 2 macaque datasets are public datasets (https://fcon_1000.projects.nitrc.org/indi/indiPRIME.html), 2 human datasets (ABCD and UK Biobank) can be accessed through online applications, and the remaining 14 datasets are private. We will release Mouse-T2WI-11.7T alongside this paper to promote future studies in this field.

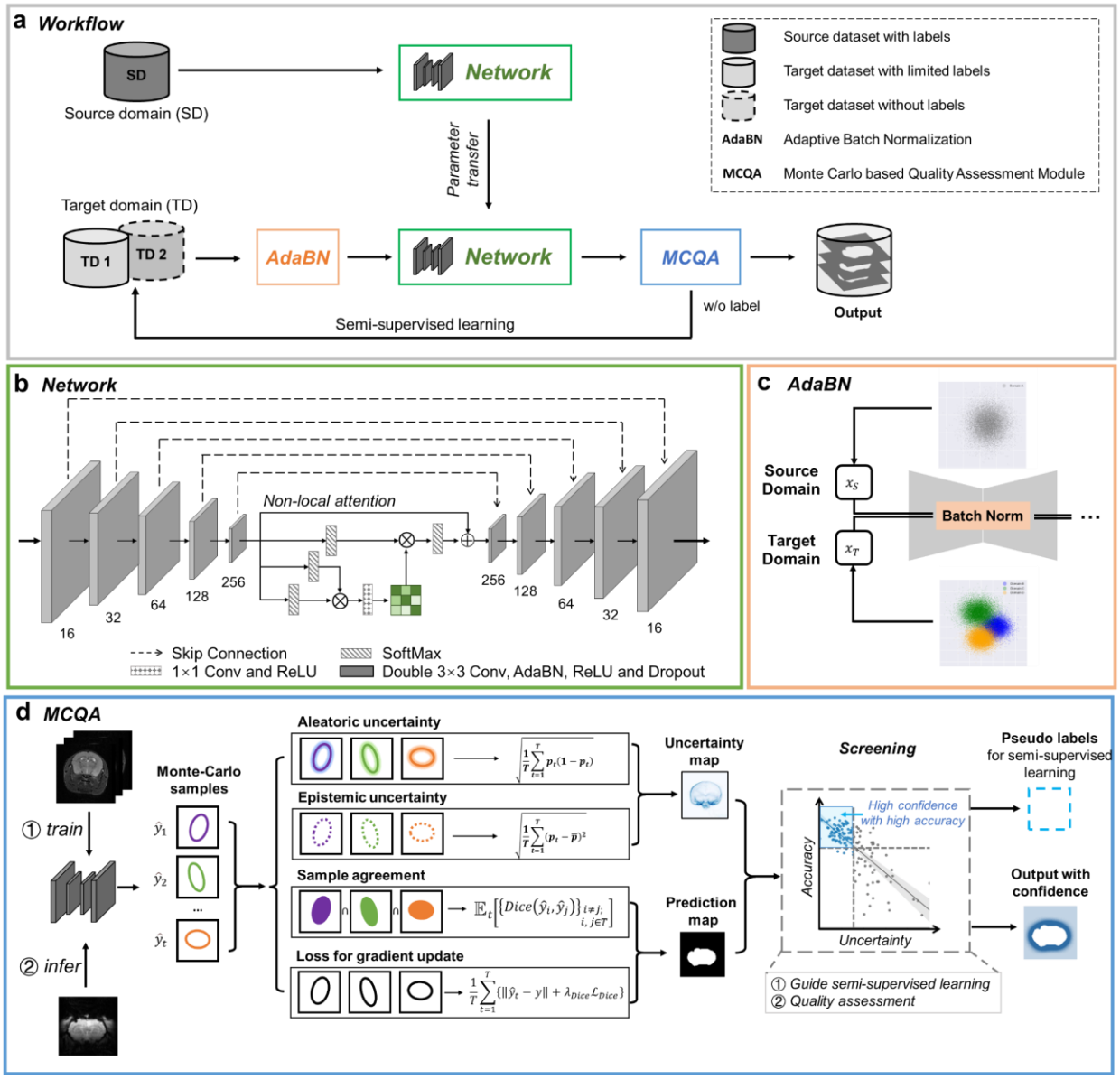

Figure 2. The architecture of our proposed BEN demonstrates its generalizability. (a) The domain transfer workflow. BEN is initially trained on the Mouse-T2-11.7T dataset (representing the source domain) and then transferred to many target domains that differ from the source domain in either the species, MRI modality, magnetic field strength, or some combination thereof. Efficient domain transfer is achieved via an adaptive batch normalization (AdaBN) strategy and a Monte Carlo quality assessment (MCQA) module. **(b)** The backbone of BEN is the nonlocal U-Net (NL-U-Net) used for brain extraction. Similar to the classic U-Net architecture, NL-U-Net also contains a symmetrical encoding and decoding path, with an additional nonlocal attention module to tell the network where to look, thus maximizing the capabilities of the model. **(c)** Illustration of the AdaBN strategy. The batch normalization (BN) layers in the network are first trained on the source domain. When transferring to a target domain, the statistical parameters in the BN layers are updated in accordance with the new data distribution in the target domain. **(d)** Illustration of the MCQA process. We use Monte Carlo dropout sampling during the inference phase to obtain multiple predictions in a stochastic

fashion. The Monte Carlo predictions are then used to generate aleatoric and epistemic uncertainties that represent the confidence of the segmentations predicted by the network, and we screen out the optimal segmentations with minimal uncertainty in each batch and use them as pseudolabels for semisupervised learning.

Model design

Based on the complexity of our datasets, inspired by transfer learning, we designed a domain-adaptive and semi-supervised network (BEN, **Figure 2a**) to segment brain MRI volumes into brain and nonbrain tissues. The backbone of BEN is a U-shaped network with nonlocal attention architecture (NL-U-Net, **Figure 2b**). Unlike previous approaches[12–14,16–18], the segmentation network of BEN is trained not from scratch but by leveraging domain transfer techniques achieved via two critical strategies: i) an adaptive batch normalization (AdaBN, **Figure 2c**) strategy for adapting the statistical parameters of the batch normalization (BN) layers in a network trained on the source domain to the new data distribution of the target domain and ii) an iterative pseudolabeling procedure for semisupervised learning in which a Monte Carlo quality assessment (MCQA, **Figure 2d**) method is proposed to assess the quality of the segmentations generated by the present network and to select the optimal pseudolabels to be added to the training set for the next iteration. These two strategies are fully automatic and do not require any human intervention. After being trained on a source domain with abundant available annotations, BEN incorporates above domain transfer modules to achieve high segmentation accuracy in each target domain while requiring zero or only a limited number of target-domain annotations. For a detailed explanation of each model component, refer to the *Methods* section.

We first trained the model on the Mouse-T2WI-11.7T dataset, which served as the source-domain dataset, following the conventional fully supervised training strategy. Here, we used fivefold cross-validation with a training/testing split of 80%/20% (194/49 out of a total of 243 scans) to evaluate the performance of our trained model. As shown in **Supplementary Table S2**, BEN obtained a Dice score of 0.99 and a 95% Hausdorff distance (HD95) of 0.14 mm, outperforming other state-of-the-art (SOTA) approaches. The success of BEN in the source domain ensures that it has the potential to excel at a variety of downstream tasks.

To evaluate the performance of BEN for domain generalization, we compared the performance of our method with two benchmark settings, namely, training from scratch (training the model only on the target-domain dataset) and fine-tuning (first pretraining on the source-domain dataset and then adjusting the parameters for the target domain), on three domain transfer tasks: across species, across modalities and across MRI scanners with various magnetic fields. Among these three tasks, cross-species domain transfer is the most challenging because of the large anatomical and structural variations between different species.

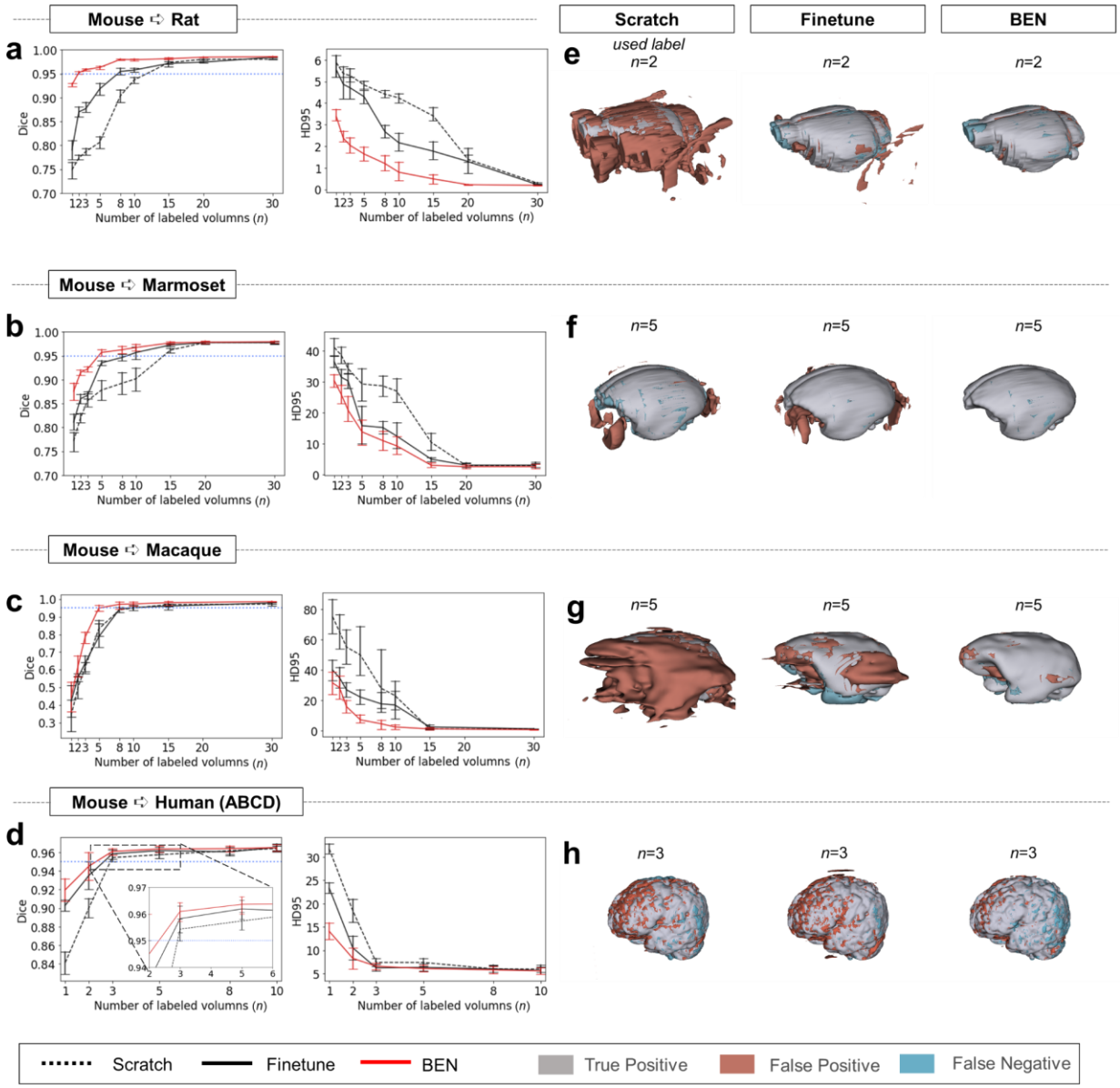

Figure 3. Performance comparison of BEN with two benchmark settings on the task of cross-species domain transfer. (a - d) Curve plots representing domain transfer tasks across species, showing the variation in the segmentation performance in terms of the Dice score and 95% Hausdorff distance (HD95) (y axis) as a function of the number of labeled volumes (x axis) for training from scratch (black dotted lines), fine-tuning (black solid lines) and BEN (red lines). From top to bottom: **(a)** mouse to rat, **(b)** mouse to marmoset, **(c)** mouse to macaque, and **(d)** mouse to human (ABCD) with an increasing amount of labeled training data (n=1, 2, 3, …, as indicated on the x axis in each panel). Both the Dice scores and the HD95 values of all three methods reach saturation when the number of labels is sufficient (n>20 labels); however, BEN outperforms the other methods, especially when limited labels are available (n≤5 labels). Error bars represent the mean with a 95% confidence interval (CI) for all curves. The blue dotted line corresponds to *y(Dice)*=0.95, which can be considered to represent qualified performance for brain extraction. **(e - h)** 3D renderings of representative segmentation results (the number *(n)* of labels used for each method is indicated in each panel). Images with fewer colored regions

represent better segmentation results. (gray: true positive; brown: false positive; blue: false negative.)

Transfer performance of BEN across species

We evaluated BEN's transferability from mouse MRI scans to scans from four other species: rat, marmoset, macaque and human. **Figure 3a-d** shows that the segmentation performance (quantified in terms of the Dice score and HD95) improves as the amount of labeled data increases in each of the four target species domains, as expected. In particular, BEN requires only 2 labeled rat MRI volumes, 3 labeled human volumes, and 5 labeled marmoset or macaque volumes to reach a Dice score of 95%, indicating satisfactory segmentation performance. In contrast, the two baseline methods, training from scratch and fine-tuning [15,18], require a minimum of 10–15 labeled volumes in the rat, marmoset or macaque domain to reach the same threshold, i.e., almost five times as many labels as BEN. The only exception is human brain extraction, a relatively simple task compared to animal brain extraction, for which all three methods need only three labeled volumes to reach a Dice score of 95%. The more rapid progression to this performance threshold indicates a lower annotation cost, which is achieved by BEN through more efficient transfer learning. Additionally, visual assessment of several representative segmented brain volumes (**Figure 3e-h**) illustrates that BEN's segmentations are more consistent with the reference ground truth than those of other methods. For example, either training from scratch or fine-tuning often leads to false positive errors in the cranial and caudal regions of the brain, which could be caused by regional heterogeneity of the scans.

Transfer performance across modalities

We further evaluated BEN's transferability across different MRI modalities, particularly between structural and functional modalities. Again, we used the structural Mouse-T2-11.7T dataset as the source domain, and we transferred the trained model to three other commonly used MRI modalities: EPI, SWI and ASL (**Figure S1**). Encouragingly, we found that even without labels, BEN could generalize well, with Dice scores of 0.93–0.95 on all three target domains, despite the different imaging sequences and parameters between these modalities (**Figure S1a-c**). This is an impressive result, as zero-shot learning is generally a challenging problem[21,22]. By comparison, a fine-tuning strategy with zero labeled volumes, i.e., a direct evaluation of the source-domain-trained model on the target domains, can achieve Dice scores of only 0.40–0.77, significantly lower than those of BEN ($p<0.001$ using the Mann–Whitney test). Note that training from scratch is impossible to implement without available labels in the target domain. Similar to the case of cross-species domain transfer, with fewer than 5 labeled volumes, BEN can achieve Dice > 0.95 in the target domains in these modality transfer tasks (2 volumes for T2WI to EPI, 1 volume for T2WI to SWI and 5 volumes for T2WI to ASL).

Transfer performance across MR scanners with various magnetic field strengths

We also examined BEN's transferability to MRI datasets acquired from multiple MRI scanners with various magnetic field strengths (**Figure S2**). The obtained image contrast varies depending on the magnetic field strength. Even with the same magnetic field strength, MRI scanners from different manufacturers can produce different image contrast, which further increases the complexity of the domain shift. Without domain transfer techniques, the direct application of a model trained on a certain source domain to a different target domain strongly suffers due to the differences in imaging contrast/quality produced by the different scanners; specifically, Dice scores of only 0.42, 0.81, and 0.61 are obtained for the 11.7 T to 9.4 T, 11.7 T to 7 T, and 9.4 T to 7 T transfer tasks, respectively. In comparison, without additional labeled data in the target domains, BEN outperforms these benchmarks by notable margins, with Dice scores of 0.97, 0.95, and 0.91, respectively. Apparently, BEN shows the capacity to generalize well to other domains without additional labeled data, while the baseline approaches overfit the source domain. With sufficient labeled data in the target domain, all three methods perform well, as anticipated.

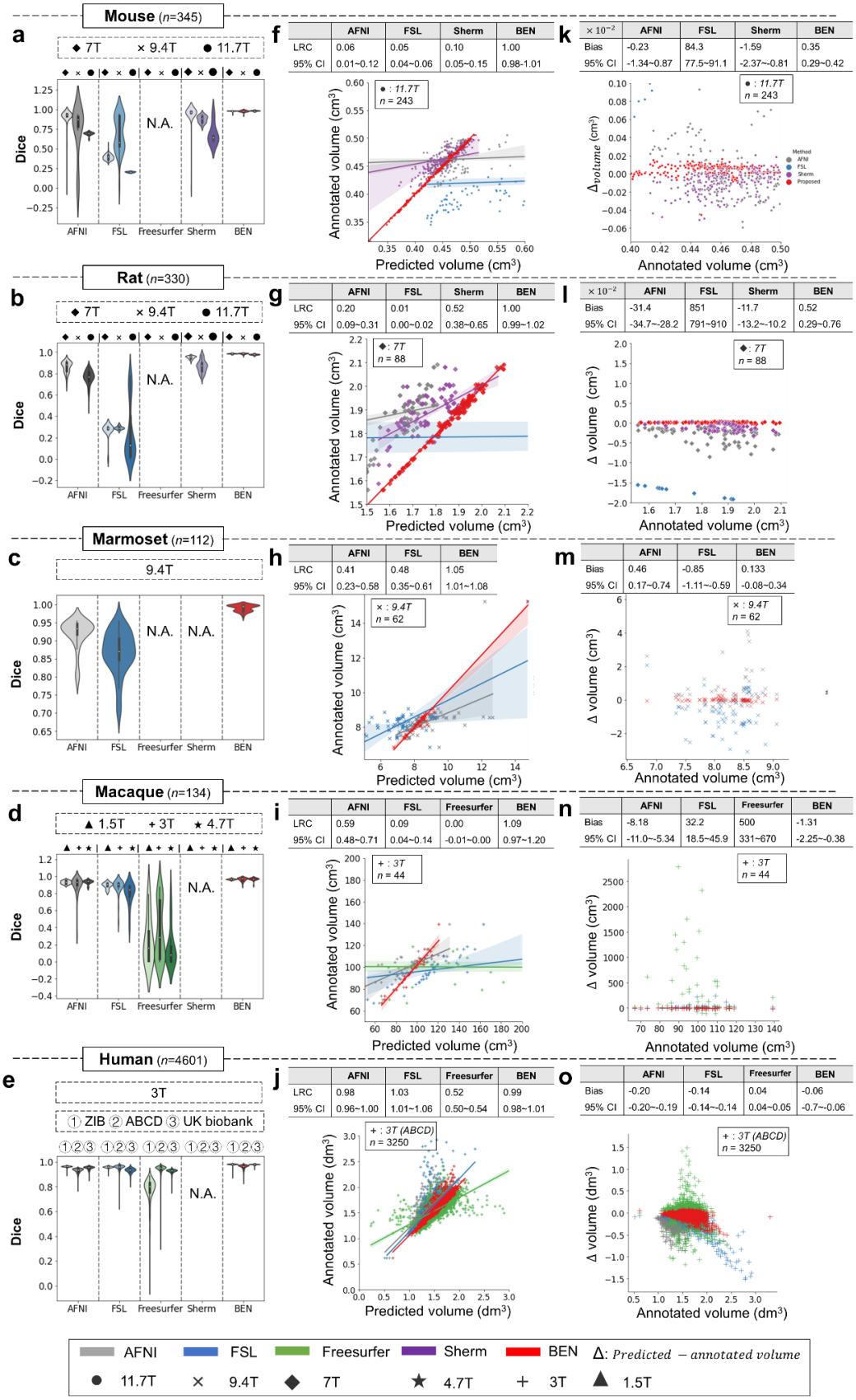
Mouse (n=345)
a
◆ 7T × 9.4T ● 11.7T
Dice
N.A.
AFNI FSL Freesurfer Sherm BEN
f
AFNI FSL Sherm BEN
LRC 0.06 0.05 0.10 1.00
95% CI 0.01~0.12 0.04~0.06 0.05~0.15 0.98-1.01
● : 11.7T
n = 243
Annotated volume (cm³)
Predicted volume (cm³)
k
× 10⁻² AFNI FSL Sherm BEN
Bias -0.23 84.3 -1.59 0.35
95% CI -1.34~0.87 77.5~91.1 -2.37~-0.81 0.29~0.42
● : 11.7T
n = 243
Δvolume (cm³)
Method
AFNI
FSL
Sherm
Proposed
Annotated volume (cm³)
Rat (n=330)
b
◆ 7T × 9.4T ● 11.7T
Dice
N.A.
AFNI FSL Freesurfer Sherm BEN
g
AFNI FSL Sherm BEN
LRC 0.20 0.01 0.52 1.00
95% CI 0.09~0.31 0.00~0.02 0.38~0.65 0.99~1.02
◆: 7T
n = 88
Annotated volume (cm³)
Predicted volume (cm³)
l
× 10⁻² AFNI FSL Sherm BEN
Bias -31.4 851 -11.7 0.52
95% CI -34.7~-28.2 791~910 -13.2~-10.2 0.29~0.76
◆: 7T
n = 88
Δ volume (cm³)
Annotated volume (cm³)
Marmoset (n=112)
c
9.4T
Dice
N.A. N.A.
AFNI FSL Freesurfer Sherm BEN
h
AFNI FSL BEN
LRC 0.41 0.48 1.05
95% CI 0.23~0.58 0.35~0.61 1.01~1.08
× : 9.4T
n = 62
Annotated volume (cm³)
Predicted volume (cm³)
m
AFNI FSL BEN
Bias 0.46 -0.85 0.133
95% CI 0.17~0.74 -1.11~-0.59 -0.08~0.34
× : 9.4T
n = 62
Δ volume (cm³)
Annotated volume (cm³)
Macaque (n=134)
d
▲ 1.5T + 3T ★ 4.7T
Dice
N.A.
AFNI FSL Freesurfer Sherm BEN
i
AFNI FSL Freesurfer BEN
LRC 0.59 0.09 0.00 1.09
95% CI 0.48~0.71 0.04~0.14 -0.01~0.00 0.97~1.20
+ : 3T
n = 44
Annotated volume (cm³)
Predicted volume (cm³)
n
AFNI FSL Freesurfer BEN
Bias -8.18 32.2 500 -1.31
95% CI -11.0~-5.34 18.5~45.9 331~670 -2.25~-0.38
+ : 3T
n = 44
Δ volume (cm³)
Annotated volume (cm³)
Human (n=4601)
e
3T
① ZIB ② ABCD ③ UK biobank
Dice
N.A.
AFNI FSL Freesurfer Sherm BEN
j
AFNI FSL Freesurfer BEN
LRC 0.98 1.03 0.52 0.99
95% CI 0.96~1.00 1.01~1.06 0.50~0.54 0.98~1.01
+ : 3T (ABCD)
n = 3250
Annotated volume (dm³)
Predicted volume (dm³)
o
AFNI FSL Freesurfer BEN
Bias -0.20 -0.14 0.04 -0.06
95% CI -0.20~-0.19 -0.14~-0.14 0.04~0.05 -0.7~-0.06
+ : 3T (ABCD)
n = 3250
Δ volume (dm³)
Annotated volume (dm³)
AFNI FSL Freesurfer Sherm BEN Δ: Predicted − annotated volume
● 11.7T × 9.4T ◆ 7T ★ 4.7T + 3T ▲ 1.5T

Figure 4. BEN outperforms traditional SOTA methods and advantageously adapts to datasets from various domains across multiple species, modalities, and field strengths.
(a - e) Violin plots and inner box plots showing the Dice scores of each method for **(a)** mouse (*n*=345), **(b)** rat (*n*=330), **(c)** marmoset (*n*=112), **(d)** macaque (*n*=134), and **(e)** human (*n*=4,601) MRI scans acquired with different magnetic field strengths. The field strength is illustrated with different markers above each panel, and the results for each method are shown in similar hues. The median values of the data are represented by the white hollow dots in the violin plots, the first and third quartiles are represented by the black boxes, and the interquartile range beyond 1.5 times the first and the third quartiles is represented by the black lines. 'N.A.' indicates a failure of the method on the corresponding dataset.
(f - j) Comparisons of the volumetric segmentations obtained with each method relative to the ground truth for five species. For better visualization, we select magnetic field strengths of 11.7 T for mouse scans **(f),** 7 T for rat scans **(g)**, 9.4 T for marmoset scans **(h)**, and 3 T for both macaque **(i)** and human **(j)** scans. Plots for other field strengths can be found in **Supplementary Figure S4**. The linear regression coefficients (LRCs) and 95% CIs are displayed above each graph. Each dot in a graph represents one sample in the dataset. The error bands in the plots represent the 95% CIs.
(k - o) Bland–Altman analysis showing high consistency between the BEN results and expert annotations. The biases between two observers and the 95% CIs for the differences are shown in the tables above each plot. Δvolume = predicted volume *minus* annotated volume.
Each method is shown using the same hue in all plots (gray: AFNI; blue: FSL; green: FreeSurfer; purple: Sherm; red: BEN). Different field strengths are shown with different markers (●: 11.7 T; ×: 9.4 T; ◆: 7 T; ★: 4.7 T; +: 3 T; ▲: 1.5 T).

Comparison with conventional neuroimage processing toolboxes

We compared BEN with several conventional SOTA toolboxes that are widely used in the neuroimaging research field, namely, AFNI, FSL, FreeSurfer, and Sherm[11], on all 18 datasets. Similar to the previous experiments, we used the Mouse-T2-11.7T dataset as the only source-domain dataset for BEN, with the remaining 17 datasets serving as the target-domain datasets. For each target domain, we used 5 labeled MRI volumes for the domain transfer task.

As shown in **Figure 4**, despite consistent performance on human MRI scans, the AFNI, FSL, and FreeSurfer toolboxes show large intra- and inter-dataset variations on the four animal species, suggesting that these methods are not generalizable to animals. Sherm, a toolbox that was specifically developed for rodent MRI, consequently achieves better results for mouse and rat scans but cannot be used for primates. In contrast, BEN is substantially superior to these conventional SOTA methods, and its performance remains consistent on both animal and human datasets, with median Dice scores of 0.97–0.99 for structural images (T2WI) (**Figure 4a-e**) and functional images (**Supplementary Figure S3**) acquired at MRI scanners with various magnetic field strengths (**Figure 4** & **Figure S4**). Better brain segmentation leads to a more accurate estimation of the brain volume: as shown in **Figure 4f-j**, BEN achieves almost perfect linear regression coefficients (LRCs) of approximately 1.00 (red lines) when compared to the corresponding expert-labeled brain volumes. In contrast, AFNI, FSL and Sherm exhibit poor consistency in animals (LRC= 0.06–0.10 for mouse scans, 0.01–0.52 for rat scans, 0.41–0.48 for

marmoset scans and 0.00–0.59 for macaque scans). Additionally, in Bland–Altman plots (**Figure 4k-n**), BEN shows higher agreement between the predicted and manually annotated volumes than the other toolboxes. The only exception is for human MRI data, with AFNI and FSL achieving excellent agreement with the ground truth on three human datasets, namely, the ABCD, UK Biobank and ZIB datasets (with LRC around 1.00). This is not surprising, as these conventional tools are well designed for human brain extraction. Encouragingly, BEN also exhibits comparable or even better performance for the human brain (**Figure 4e, j, & o**), with an accelerated speed of 0.6 s per scan on average. This speed is two orders of magnitude greater than that of conventional toolboxes (several minutes on average) and three orders of magnitude greater than that of manual annotation (an average of 20–30 minutes) (**Figure S5**), suggesting the potential of applying BEN in high-throughput studies.

In addition, we also compared BEN with published DL-based SOTA brain extraction methods [15,18,23] as well as non DL-based methods (RATS [10], Sherm [11], AFNI, FSL, and Freesurfer), on two public datasets, CARMI (https://openneuro.org/datasets/ds002870/versions/1.0.0) and PRIME-DE (https://fcon_1000.projects.nitrc.org/indi/indiPRIME.html). As shown in **Supplementary Table S3**, the segmentation performance of BEN, as evaluated using Dice score and the Hausdorff distance, surpasses other methods.

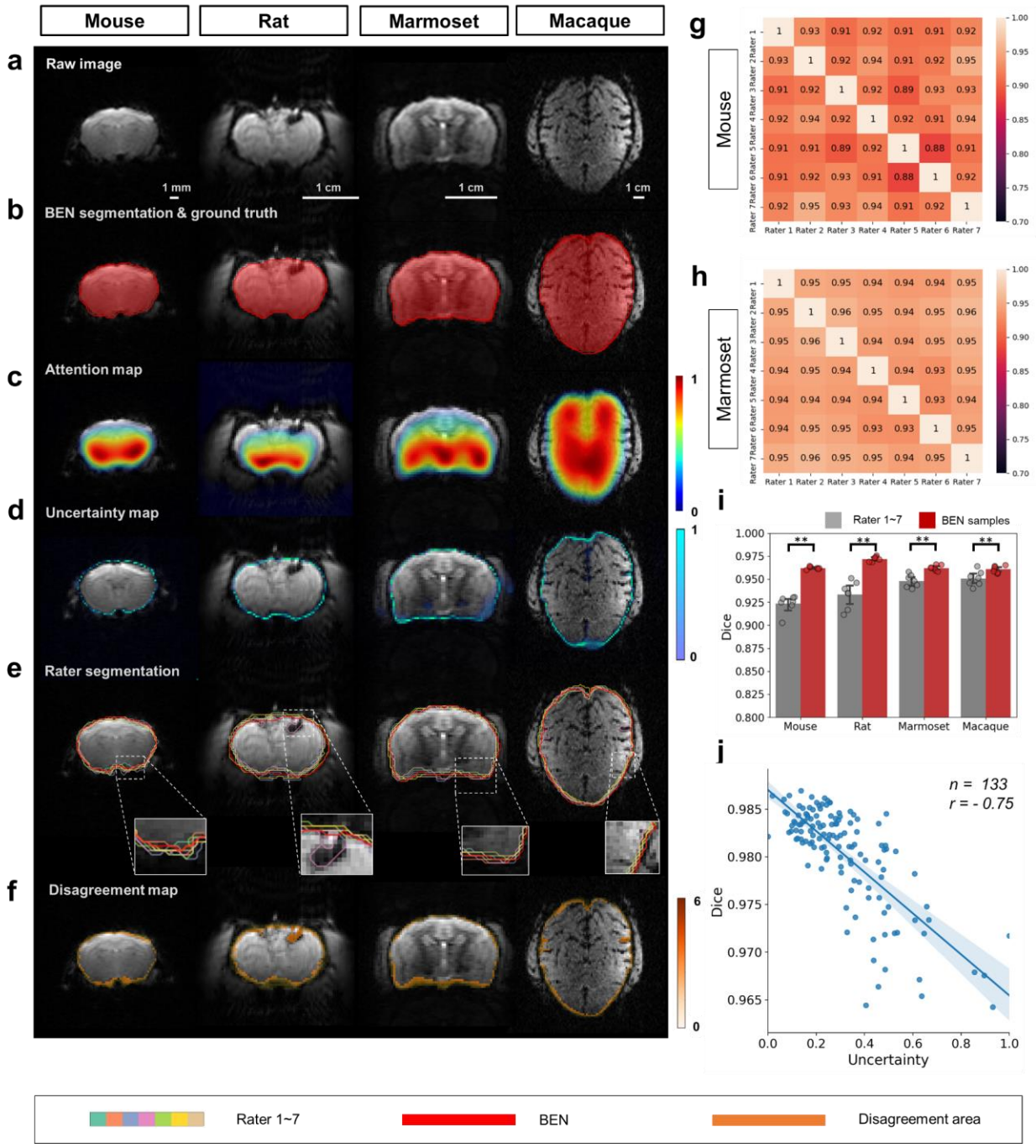

Figure 5. BEN provides a measure of uncertainty that potentially reflects rater disagreement. (a) Representative EPI images from four species (in each column from left to right, mouse, rat, marmoset, and macaque). **(b)** BEN segmentations (red semi-transparent areas) overlaid with expert consensus annotations (red opaque lines), where the latter are considered the ground truth. **(c)** Attention maps showing the key semantic features in images as captured by BEN. **(d)** Uncertainty maps showing the regions where BEN has less confidence. The uncertainty values are normalized (0–1) for better visualization. **(e)** Annotations by junior raters (*n*=7) shown as opaque lines of different colors. **(f)** Rater disagreement regions (0: consensus regions; 1–6: the number of raters who deem these regions as brain tissue). Note that these

disagreement maps share similar locations with the uncertainty maps. **(g, h)** Heatmaps representing the segmentation Dice scores of each junior rater as calculated using the labels from one of the others as the ground truth on **(g)** the mouse image and **(h)** the marmoset image. The corresponding heatmaps for the rat and macaque images are provided in **Supplementary Figure S8**. **(i)** Bar plots showing the Dice score comparisons between the ground truth and the BEN segmentation results (red) as well as the ground truth and the junior raters' annotations (gray) for all species (gray: raters, *n*=7, red: Monte Carlo samples from BEN, *n*=7; statistics: Mann–Whitney test, **p<0.01). Each dot represents one rater or one sample. Values are represented as the mean and 95% CI. **(j)** Correlation of the linear regression plot between the Dice score and the normalized uncertainty. The error band represents the 95% CI. Each dot represents one volume (species: mouse; modality: T2WI; field strength: 11.7 T; *n*=133; r=-0.75).

BEN represents interrater variations in the form of uncertainty maps

Due to different levels of annotation experience, disagreements or disputes between human annotators remain a problem. To quantitatively assess the bias of interrater disagreements, we carried out a case study of the annotation of brain functional MRI images from mice, rats, marmosets, and macaques. A total of 9 raters participated in the experiments: two senior raters (defined as having > 5 years of experience) and seven junior raters (defined as having 1-2 years of experience). The consensus region of the two senior raters' annotations was taken as the ground truth (reference) for comparison with the annotations of the seven junior raters. BEN generated an uncertainty map through Monte Carlo sampling during the inference process, without additional adjustment of the network structure.

As shown in **Figure 5a & b** and **Supplementary Figure S6**, the segmentations produced by BEN match the ground truth very well across different species. The attention maps and uncertainty maps display complementary information about the brain morphology; specifically, the attention maps focus on the image signal within the brain tissue in MRI images (**Figure 5c**), while the uncertainty maps provide information on the boundaries between the brain and nonbrain tissues (**Figure 5d**). Interestingly, for each rater's annotations, disputes always occur at the brain boundaries (**Figure 5e**), similar to BEN's uncertainty map. The concurrence between the projections of BEN's uncertainty maps and the rater disagreement maps (**Supplementary Figure S7**) further demonstrates that BEN has the potential to interrogate its own predictions and that its uncertainty measure can be deemed an alternative proxy for rater divergence.

Furthermore, to evaluate the annotation variability between different raters in a more quantitative fashion, we calculated the Dice scores and found that none of them was higher than 0.96 across the four species (**Figure 5g & h** and **Figure S8**). Specifically, the Dice scores between each junior rater and the ground truth provided by the senior raters are between 0.92 and 0.94 for the four species, and these are surpassed by BEN, with Dice scores of 0.96–0.97 (p<0.01) (**Figure 5i**). In addition to animal studies, we performed a parallel experiment to verify BEN's predictions for human brain MRI. As shown in **Supplementary Figure S9**, similar to the animal segmentation results, BEN produces attention and uncertainty maps to support interpretation of BEN's robustness compared with conventional toolboxes.

We further examined the correlation between BEN's uncertainty and segmentation quality. As shown in **Figure 5j**, the estimates of uncertainty at the volume level are negatively correlated with the Dice score ($r = -0.75$, $n = 133$ scans from the Mouse-T2WI-11.7T dataset), suggesting that BEN's uncertainty measure can be considered as an alternative metric for assessing the quality of segmentations at the time of inference, when the Dice score cannot be computed because the ground-truth segmentation is not available.

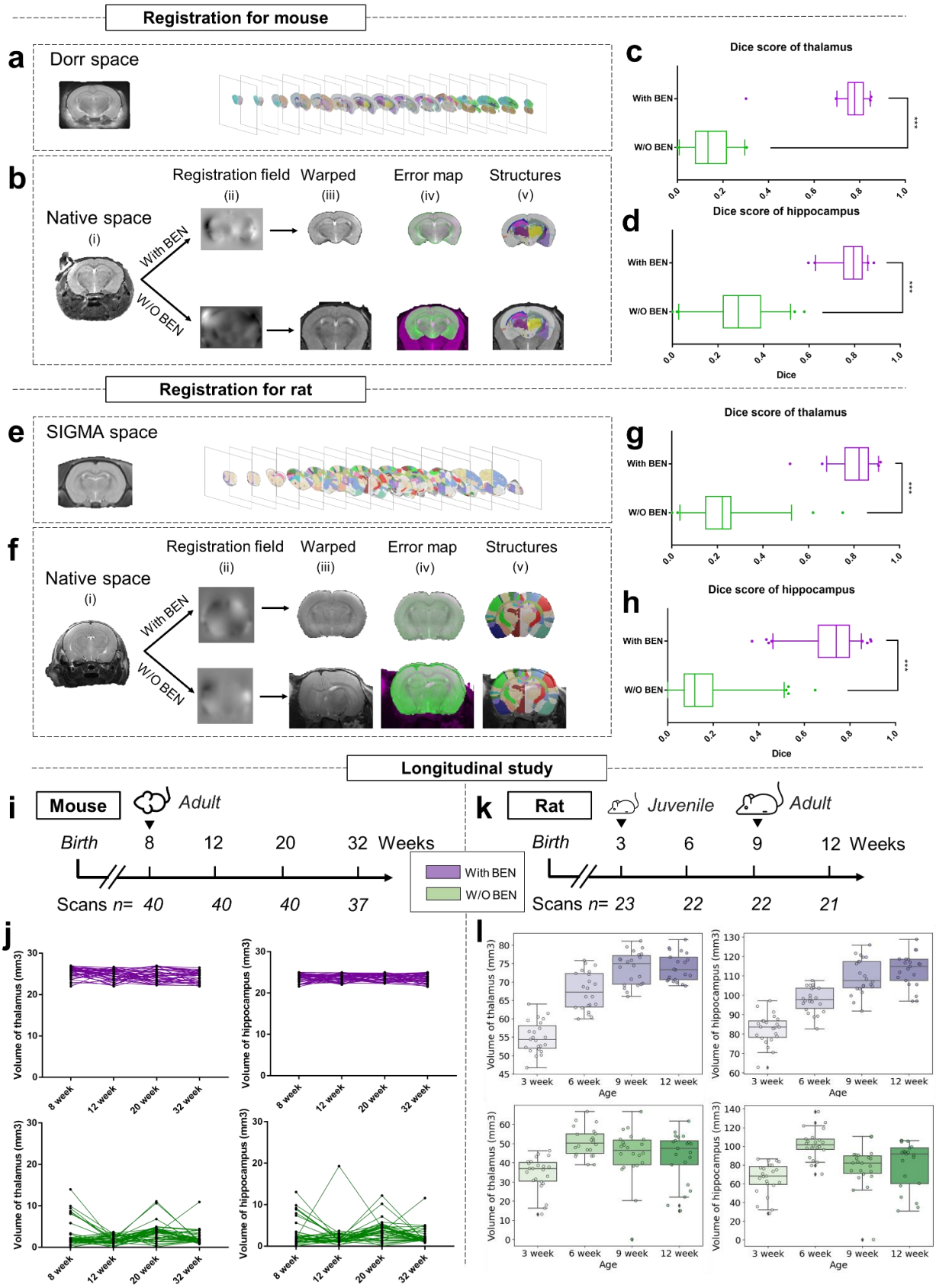
Registration for mouse
a
Dorr space
b
Native space
(i)
With BEN
W/O BEN
Registration field
(ii)
Warped
(iii)
Error map
(iv)
Structures
(v)
c
Dice score of thalamus
With BEN
W/O BEN
***
d
Dice score of hippocampus
Dice
Registration for rat
e
SIGMA space
f
g
Dice score of thalamus
h
Dice score of hippocampus
Longitudinal study
i
Mouse
Adult
Birth
8
12
20
32
Weeks
Scans n= 40 40 40 37
With BEN
W/O BEN
k
Rat
Juvenile
Adult
Birth
3
6
9
12
Weeks
Scans n= 23 22 22 21
j
Volume of thalamus (mm3)
Volume of hippocampus (mm3)
8 week
12 week
20 week
32 week
l
3 week
6 week
9 week
12 week
Age

Figure 6. BEN facilitates downstream tasks in neuroimaging by producing high-quality brain extraction results. (a-h) BEN improves the atlas registration accuracy for rodents. **(a)** The Dorr space and **(e)** the SIGMA space, corresponding to the mouse and rat atlases, respectively. **(b, f)** Integration of BEN into the registration workflow: (i) Two representative samples from a mouse (n=157) and a rat (n=88) in the native space. (ii) The BEN-segmented brain MRI volumes in the native space as registered into the Dorr/SIGMA space using the Advanced Normalization Tools (ANTs) toolbox, for comparison with the registration of original MRI volumes with respect to the corresponding atlas. (iii) The warped volumes in the Dorr/SIGMA spaces. (iv) Error maps showing the fixed atlas in green and the moving warped volumes in magenta. The common areas where the two volumes are similar in intensity are shown in gray. (v) The brain structures in the warped volumes shown in the atlas spaces. In our experiment, BEN significantly improves the alignment between the propagated annotations and the atlas, as confirmed by the improved Dice scores in the **(c, g)** thalamic and **(d, h)** hippocampal regions (box plots: purple for BEN, green for W/O BEN; volumes: n=157 for the mouse, n=88 for the rat; ***$p<0.001$). **(i - l)** An improvement in atlas registration is also beneficial for quantitative brain morphological analyses in longitudinal studies. In our two examples, **(i)** adult mice (n=40) were imaged at weeks 8, 12, 20, and 32. **(j)** With BEN (purple boxplots), the volume statistics of the thalamus and hippocampus remain almost constant, which is plausible, as these two brain regions do not further enlarge in adulthood. In comparison, **(l)** the volume changes of the thalamus and hippocampus in **(k)** juvenile rats (n = 23, imaged at weeks 3, 6, 9, and 12) suggest continuous growth with time, particularly between weeks 3 and 9. In contrast, without BEN, due to poor atlas registration and consequently inaccurate volume statistics (green box plots), these crucial brain growth trends during development may be missed.

BEN improves the atlas-to-target registration and brain volumetric quantification

Accurate brain extraction is a key component of many neuroimaging pipelines, facilitating subsequent procedures in longitudinal or cohort analyses. We therefore designed experiments to demonstrate BEN's application for downstream processing and analysis in neuroimaging studies of rodents (**Figure 6**, **Figure S10**) and humans (**Figure S11**), including atlas registration (**Figure 6a-h**) and volumetric quantification (**Figure 6i-l**). The registration quality was evaluated in terms of the Dice scores between the brain structural segmentations of the warped MRI brain images and the atlas. **Figure 6b & f** and **Figure S11** plot the workflow of atlas registration, in which mice are registered to Dorr [24], rats to SIGMA [25], and humans to MNI [26]. Quantitative evaluations of two critical brain structures, the thalamus and hippocampus (**Figure 6c, d, g, & h** and **Figure S11**), demonstrate that BEN can tremendously enhance the registration quality; registration with AFNI achieves a Dice score of only 0.2–0.6, which can be improved to approximately 0.8 when using BEN for brain extraction ($p<0.001$).

As the next step, we examined the contribution of BEN to brain volumetric quantification in rodent longitudinal studies. Two longitudinal MRI datasets, representing adult mice (8, 12, 20, and 32 weeks old, **Figure 6i**) and adolescent rats (3, 6, 9, and 12 weeks old, **Figure 6k**), were collected. As shown in **Figure 6j**, after registration with BEN, atlas-based morphometry analysis

showed that the volumes of two representative brain regions, the thalamus and hippocampus, remained stable for the adult mice from 8 weeks old to 32 weeks old (purple lines). In contrast, calculating the volumes of these brain regions without BEN led to unreliable quantifications (green lines) due to poor atlas registration resulting in the propagation of error to the volumetric quantifications. Similarly, BEN improved brain volumetric quantification in adolescent rats and yielded a plausible pattern of growth from 3 weeks old to 12 weeks old (**Figure 6l**, purple boxes), suggesting rapid development of these two critical brain regions during this period, consistent with findings in the literature [27,28]. In comparison, the atlas-based segmentations were poor when without BEN, and no obvious trend in the volumetric statistics was apparent (green boxes). These results indicate that BEN not only is able to tremendously improve the registration accuracy and contribute to routine brain MRI processing but also is critical for longitudinal MRI studies, as it improves volumetric quantification.

Discussion

In this study, we introduce BEN, a generalizable brain extraction tool for MRI data from rodents, NHPs, and humans. Our results demonstrate BEN's transferability across different species, modalities and platforms thanks to several key modules: an AdaBN module, a semi-unsupervised learning module, and an uncertainty estimation module (the contribution of each module refers to ablation study in **Table S4 & S5**). To facilitate BEN's usage, we also provide interfaces to make BEN compatible with several popular toolboxes, including AFNI, FSL, and FreeSurfer, and therefore easy to integrate into conventional neuroimage processing pipelines.

Transferability

By virtue of its modules for domain adaptation and transfer learning, BEN is capable of handling the heterogeneity in MRI contrast associated with different subjects, MRI scanners, and imaging sequences. It is worth noting that the source domain for BEN can be freely chosen. In this study, we switched the source domain for BEN's deployment from a mouse dataset to a human dataset (**Figure S12**). Similar to the successful transfer of BEN from mouse data to human data (Dice scores of 0.95 and 0.97 with only 1 or 2 labeled scans), BEN can also achieve an impressive Dice score higher than 0.95 when only 1 labeled volume is used for transfer from human data to mouse data. This dual transferability between mouse and human data demonstrates that BEN can take advantage of the semantic features extracted from one domain and transfer the corresponding learned information to the other domain despite the enormous gap in brain morphometry between the two species (**Figure S13**). This also indicates that the source domain could be represented by any brain MRI dataset from any other species. On the project website, we have released trained model weights for five species (mouse, rat, marmoset, macaque, and human) to allow users to start domain transfer from the closest species to those of interest to them with minimal or even zero additional annotations.

To interpret how transfer learning affects network performance, we used uniform manifold approximation and projection (UMAP)[29] to visualize the feature distributions before and after domain transfer (details are described in the *Methods* section). **Figure S14** shows a representative cross-modality transfer task from Mouse-T2WI-11.7T to Mouse-EPI-11.7T,

which shows substantial perceptual differences. Although the semantic features are separated by the network in the source domain (**Figure S14a**), they are intermingled in the target domain without transfer learning (**Figure S14b**). In contrast, only after BEN's transfer modules are those features again separated in the target domain (**Figure S14c**), demonstrating the importance of using the domain transfer strategy.

Automatic quality assessment and "virtual" annotation

Just as divergences always exist among human annotators, the quality of neural-network-generated segmentations may also vary greatly among diverse sample data due to the inherent data distribution or noise [30–32]. However, the uncertainty of a network's decisions regarding its generated segmentations can be assessed via Monte Carlo sampling in many medical image analyses [33–35]. This is also the basis of BEN's uncertainty assessment module. Our results demonstrate that BEN's uncertainty map represents the expected interrater variations and that the uncertainty linearly increases as BEN's performance decreases. Based on the previously presented rater experiment (**Figure 5**), BEN generates segmentations that are comparable to the consensus segmentations of two senior experts with high consistency, indicating that BEN can apply consistent labeling rules alongside automatic quality assessments when constructing large datasets to prevent subjective bias in large-scale research.

In addition, BEN is able to produce uncertainty-guided pseudo-labels with high confidence by using a semi-supervised learning strategy. Note that on the ABCD human dataset, BEN achieves competitive or even superior performance with fewer than 5 labels, nearly perfectly delineating thousands of scans (n=3,250; **Figure 4e, j & o**); this further demonstrates that BEN can propagate annotations from limited labeled data to unlabeled data and produce a larger set of "virtual" labels for training, suggesting that it can be easily applied to large-scale datasets with limited annotations for rapid automatic deployment. Nevertheless, one limitation of our model is that BEN's quality assessment may result in deviations when the MRI images are of poor quality, such as motion artifacts and low signal-to-noise ratios (SNRs). In future work, this procedure may be updated by incorporating conditional random fields[36] or heuristic algorithms.

Compatibility with other toolboxes and deployment

BEN was developed in Python and is equipped with well-designed interfaces to support most types of software encountered in existing traditional pipelines (**Table S6**), including FSL, AFNI, FreeSurfer, Advanced Normalization Tools (ANTs), and statistical parametric mapping (SPM). We also provide tutorials and demo examples to flatten the learning curve and lower the entry barrier for researchers deploying our new algorithm. Moreover, BEN was built under the open-source paradigm from the beginning and has already been tested on various types of MRI data from many imaging centers. We have released packaged interfaces for BEN and source codes, including the implementations for domain transfer, semi-supervised training and quality control. Additionally, we have distributed the model weights trained on five species (mouse, rat, marmoset, macaque and human). Applications to other target domains can also benefit from our pretrained models, requiring the addition of zero or only a small number of labeled images to update the model.

For the introduction of a new dataset, we suggest that the user do as follows. 1) First, use BEN to load the pretrained weights for the closest corresponding species. 2) Consider labeling several scans to fine-tune the model if the pretrained weights do not yield satisfactory performance or the target species is beyond the scope of our datasets. The domain transfer procedures will be performed automatically, without human intervention. BEN follows the open-access paradigm, allowing users to save their updated models and share their weights for use by the neuroimaging community. The accumulation of additional imaging data will further improve the performance of BEN and support the exploration of complex neuroimaging research.

Conclusion

In summary, we have demonstrated the superiority of BEN in terms of accuracy, robustness and generalizability in a large-scale study involving eighteen different MRI datasets. BEN improves the robustness of atlas-to-target registration and brain volumetric quantification in neuroimaging. BEN is an open-source and modularly designed toolbox that provides uncertainty maps to quantify the confidence of the network's decisions and to model interrater variability. We believe that BEN has great potential to enhance neuroimaging studies at high throughput for both preclinical and clinical applications.

Methods

Eighteen datasets

Motivated by the need to develop one universal DL-based model for brain extraction in rodents and primates, we conducted experiments to evaluate BEN on 18 different domain datasets covering diverse species (mouse, rat, marmoset, macaque, and human), modalities (T2WI, T1WI, EPI, SWI, and ASL), and platforms (11.7 T, 9.4 T, 7 T, 4.7 T, 3 T, and 1.5 T) in order to assess the reliability, robustness and effectiveness of our network and its domain adaptation capabilities. Because T2WI is the most commonly used modality for rodent brain imaging and we have accumulated extensive brain scans in numerous previous research projects, we chose the Mouse-T2WI-11.7T dataset as the source-domain dataset, and the rest of the datasets collected across 17 sites/cohorts were used as separate target-domain datasets. Detailed information on each dataset is shown in **Table S1** and summarized below:

(1) Mouse: includes 7 datasets consisting of 453 scans (14,358 slices), namely, Mouse-T2WI-11.7T (243 scans), Mouse-T2WI-9.4T (14 scans), Mouse-T2WI-7T (14 scans), Mouse-EPI-11.7T (54 scans), Mouse-EPI-9.4T (20 scans), Mouse-SWI-11.7T (50 scans), and Mouse-ASL-11.7T (58 scans).
(2) Rat: includes 4 datasets consisting of 330 scans (11,264 slices), namely, Rat-T2WI-11.7T (132 scans), Rat-T2WI-9.4T (55 scans), Rat-T2WI-7T (88 scans) and Rat-EPI-9.4T (55 scans).
(3) Marmoset: includes 2 datasets consisting of 112 scans (4,060 slices), namely, Marmoset-T2WI-9.4T (65 scans) and Marmoset-EPI-9.4T (50 scans).
(4) Macaque: includes 2 datasets consisting of 134 scans (22,620 slices), namely, Macaque-T1WI (76 scans) and Macaque-EPI (58 scans). These data are publicly available from the nonhuman PRIMatE Data Exchange (PRIME-DE) (https://fcon_1000.projects.nitrc.org/indi/indiPRIME.html).
(5) Human: includes 3 datasets consisting of 4,601 scans (87,3453 slices): Human-ABCD (3,250 scans), Human-UKB (963 scans) and Human-ZIB (388 scans).

Transfer learning with BEN

We first trained an initial BEN model on a source domain with 243 annotated T2WI mouse brain volumes collected at 11.7 T to train our network. Although accurate for brain extraction in this source domain, this initial BEN model could not produce satisfactory segmentation results for another domain corresponding to images collected with a different MRI sequence (e.g., EPI functional imaging vs. T2WI structural imaging), using a different MRI scanner with a different magnetic field (e.g., 7 T vs. 11.7 T), or even in a different species (e.g., marmoset vs. mouse). To extend the usage of BEN to various species, MRI platforms and modalities, we incorporated two key domain transfer strategies: i) an AdaBN strategy to adapt the statistical parameters of the batch normalization (BN) layers in a network trained on a certain source domain to the new data distribution of a different target domain and ii) an iterative pseudolabeling procedure for semisupervised learning, for which a Monte Carlo quality assessment (MCQA) method is proposed to assess the quality of the segmentations generated by the present network and to

select the optimal pseudolabels to be added to the training set for the next iteration. These two strategies efficiently boost the network performance in a new target domain while requiring a minimal number of annotations in the target domain.

Backbone network architecture

The backbone of BEN is based on a U-shaped architecture [36], as is commonly used for biomedical volumetric segmentation [37]. We use 2D convolution kernels rather than 3D kernels here to accelerate the execution speed and reduce the demand for GPU memory. In addition, most rodent and NHP scans have an anisotropic spatial resolution, so 3D convolution is not necessary. As shown in **Figure 2b**, the backbone network consists of five symmetrical pairs of encoding and decoding blocks, which perform contraction and expansion operations, respectively, during the data feedforward process. Each encoding/decoding block is composed of two convolutional layers with 3*3 convolution kernels, followed by a BN layer, an exponential linear unit (ELU) function and a 2*2 max-pooling operation with stride 2. The first encoding block results in 16 feature channels, and every subsequent encoding block doubles the number of channels. Each decoding block is symmetric with respect to the encoding block at the corresponding level, except that the max pooling layer is replaced with transposed convolutional layers to upsample the feature map. In the last decoding block, we add a 1*1 convolutional layer with a 3*3 kernel size and a sigmoid activation function to map the 16 feature channels to a binary probability map that softly assigns each pixel to the brain and nonbrain tissue classes.

Additionally, we add a self-attention mechanism in the bottom layer of the U-Net to yield higher accuracy and smoother boundaries. A self-attention block [38] is used to improve the model performance for various domain tasks. Concretely, we use a nonlocal attention layer in the bottleneck of the network to model long-range and global contextual dependencies over feature representations. This attention mechanism calculates a weighted saliency map at each position in the feature map that accounts for the global contextual response, thus better assisting the network in distinguishing brain and nonbrain tissue in specific domains with different contrast characteristics and distributions.

The model takes preprocessed MRI images as input (for details, refer to the subsection titled “Model input preprocessing”) and merges successive mask slices to reconstruct the original 3D volumes. Here, we mainly focus on the structure and components of the model as well as the training and evaluation procedures in the source domain. The pretrained model weights will be retained for the transfer learning and domain adaptation stages, which are discussed in the next section.

Adaptive batch normalization (AdaBN)

To mitigate the domain shift problem encountered when DL models are applied to different domains, we adopt the AdaBN strategy [39]. Standard BN layers use fixed statistical descriptors for multiple domains; however, this approach is not valid, hence damaging the generalization ability of the trained model. In contrast, in AdaBN, the statistical descriptors are updated in accordance with the specific target domain in the inference phase, resulting in better model generalization.

For our brain segmentation task, AdaBN can help the model to automatically adapt to the different MRI intensity distributions associated with different species, modalities and platforms without requiring modifications to the model architecture or other model hyperparameters.

Semisupervised learning with pseudolabels using Monte Carlo quality assessment (MCQA)

Although AdaBN facilitates domain transfer, it alone cannot completely solve the challenging problem of transfer learning between species, as the gap between different species lies not only in the MRI intensities but also in the brain structure and geometry. In this case, additional supervision is still necessary to guide the model to achieve highly accurate brain segmentation in the new species. Instead of common approaches such as training from scratch for fine-tuning, we propose a novel semisupervised learning strategy for BEN. In addition to sparsely labeled data in the target domain, we make use of abundant unlabeled data by means of an iterative pseudolabeling process (**Figure 2d**): we first evaluate the model on all unlabeled data and then select predictions that are associated with the lowest uncertainty (highest confidence) based on MCQA (detailed in subsequent sections). The predictions selected through quality screening are taken as pseudolabels and are added to the training set in the next iteration. More specifically, we use a hybrid loss function:

$$\frac{1}{N}\sum_{i=1}^{n}[\|f(x_i;\theta_t)-y_i\|+\lambda_{Dice}\mathcal{L}_{dice}(f(x_i;\theta_t),y_i)]+\alpha(t)\frac{1}{M}\sum_{i=1}^{m}\mathcal{L}_{dice}(f(x_j;\theta_t),f(x_j;\theta_{t-1}\mid\varsigma)) \quad (1)$$

$$\alpha(t)=\begin{cases}0, & \mathrm{t}<\mathrm{T_{sl}}\\ \lambda_{sl}, & \mathrm{t}\geq\mathrm{T_{sl}}\end{cases} \quad (2)$$

where x_i represents a labeled data point with the ground-truth annotation y_i in the current minibatch (total number N) and x_j denotes a pseudolabeled data point whose model prediction $f(x_j;\ \theta_{t-1})$ from the previous iteration is below an uncertainty threshold ς based on MCQA (total number M). $\alpha(t)$ is an important parameter that balances the relative contributions of true labels and pseudolabels in the algorithm: we set this parameter to zero in the first several iterations, as our initial model is not good enough to make confident predictions, and then to a hyperparameter λ from iteration T_{sl} onward. The whole semisupervised training procedure can be summarized as follows: The total number of training epochs is denoted by T_{epoch}, and the current epoch is denoted by t. During the first phase ($0 < t < T_{sl}$), the network is trained directly on the target domain using labeled data. Then, in the next phase ($T_{sl} \leq t \leq T_{epoch}$), unlabeled data with pseudolabels additionally participate in the semisupervised learning process by contributing to the loss function as determined by the balance coefficient $\alpha(t)$. Empirically, we set $T_{sl} = 20$, $\lambda_{Dice} = 1$ and $\lambda_{sl} = 0.5$.

Model training implementation

The model was trained with the Dice loss function using the adaptive moment estimation (Adam) optimizer for 50 epochs with a batch size of 16 and learning rates of 10^{-4} and 10^{-5} in the source domain and all target domains, respectively. For training and testing on each domain, a separate cross-validation process was performed to assess the final performance of our method, with progressively larger numbers of observations (n = 1, 2, 3,…) assigned to the training set while the rest of the samples were assigned to the test set. The mechanisms for domain transfer and semisupervised learning have been described in previous sections.

Model input preprocessing

To lessen the variation across individual scans, we applied the following processes on all datasets. First, we applied the N4 algorithm [40] to remove bias fields in the MRI scans. Second, we normalized each image as follows (with V_{norm} denoting the normalized MRI volume and V^1 and V^{99} denoting the 1st and 99th percentiles, respectively, of the intensity values in the input volumes):

$$V_{norm}=\frac{V-V_{min}{}^{1}}{V_{max}{}^{99}-V_{min}{}^{1}} \quad (3)$$

For all functional MRI data, our model can process each time-point image and produce a corresponding binary mask. For consistency with the structural modalities, we used the results computed for the 5th time point to calculate the performance metrics.

The processed volumes were then fed into the network as input, slice by slice, after being resized or cropped to a matrix with dimensions of 256*256 for compatibility with the network. In addition, data augmentation was employed to empower the network to learn invariant characteristics from data exhibiting variations among different subjects and protocols. For each input slice, the following transformations were implemented: random rotation (±10), random scaling (90–110%) and random translation (up to ±10 pixels along each axis). For example, the rotation and translation operators can imitate conditions in which the subjects are represented in diverse coordinate systems and have different head orientations. For the case of inconsistent voxel sizes across different sites and even for the lifespan of one individual during brain maturation, scaling can simulate corresponding image data.

Model output postprocessing for uncertainty estimation and quality assessment

To understand the intrarater divergences as represented in BEN and provide valuable insights into model interpretability for clinicians, we utilize dropout layers in the network to obtain uncertainty measures for the model-generated segmentations.

There are two principal types of prediction uncertainties for deep neural networks: aleatoric uncertainty and epistemic uncertainty [30,41]. Aleatoric uncertainty captures the potential inherent noise of the input test data, and therefore, this type of uncertainty is unlikely to be lessened by increasing the amount of training data. In contrast, epistemic uncertainty refers to the uncertainty

in the model parameters, which represents the lack of knowledge of the best model. In DL networks, epistemic uncertainty can be caused by a lack of training data in certain areas of the input domain and will decrease with increased diversity of the training data distribution.

In our study, approximate Bayesian inference was used to estimate uncertainty. Specifically, keeping the dropout layers active in the inference phase, we collected N independent samples for each subject. The pixelwise uncertainties of each sample are given by the following formula:

$$\text{Uncertainty} = \underbrace{\sqrt{\frac{1}{T}\sum_{t=1}^{T} p_t(1-p_t)}}_{\text{aleatoric}} + \underbrace{\sqrt{\frac{1}{T}\sum_{t=1}^{T} (p_t-\bar{p})^2}}_{\text{epistemic}} \quad (4)$$

where p_t denotes the sigmoid probability output in the final layer of the network in the inference stage and $\underline{p}$ is the average prediction for all inferences. In this work, we chose $T = 10$, which offers a good trade-off between the reliability of uncertainty estimation and the time consumed for Monte Carlo sampling, except in the human expert disagreement study, for which we set $T = 7$ for consistency with the number of clinical raters.

Based on these estimated uncertainty observations, we can filter out relatively unreliable predictions (usually those with high uncertainty values) based on the guidance obtained through statistical analysis, thereby realizing quality assessment. Moreover, relatively convincing predictions can be further utilized as pseudolabels to guide the semisupervised learning process.

Baseline methods and traditional toolbox settings

To demonstrate the utility of domain transfer, we additionally tested two baseline methods of training the BEN model: training from scratch and fine-tuning. These methods share the same network architecture, learning rate, batch size and number of training epochs. For training from scratch, the weights of the layers in the network are randomly initialized, and then all weights are adjusted based on the training set. In the second baseline method, the training process is initialized on the source domain, and the weights are then fine-tuned to the new task in accordance with the target-domain dataset.

We also compared BEN with four widely used tools representing SOTA processing pipelines in their respective research fields. We adjusted the parameters provided by these traditional toolboxes to adapt to the different domain datasets to achieve better performance; however, for situations in which the demands were beyond the scope of the capability of the tools, the default parameters were used instead. These four methods are listed below, and their detailed parameter settings can be found in **Table S7**:

1. FSL: The Brain Extraction Tool (BET) provided by FSL is based on a deformable model of the brain surface, which is initialized as an ellipsoid or sphere centered on the brain and then grows.
2. AFNI: The 3dSkullStrip algorithm first preprocesses the image to correct for spatial changes in image intensity and then repositions the brain to the approximate center of the

image. Then, it uses an improved BET-based algorithm to expand a mesh sphere until it covers the entire brain surface. In our study, we used the '-monkey' parameter for the macaque data, the '-marmoset' parameter for the marmoset data and the '-rat' parameter for the rodent data.

3. FreeSurfer: The mri_wartershed command call is a hybrid method that combines the watershed algorithm with a deformable surface algorithm. Under the assumption of the connectivity of white matter, an initial mask is first generated via the watershed algorithm. Then, the deformable surface model is used to merge geometric constraints into the mask.
4. Sherm: In this method, a brain template mask is used as the reference. Then, shape descriptors are computed to select target candidates that match the shape of the reference template. We used the 'rat' and 'mouse' parameters for the corresponding species and used anisotropic spatial resolution. In NHP applications, this method fails.

Ground-truth generation

The ground-truth brain and nonbrain regions for each domain dataset were manually annotated by two experts in the field of brain anatomy. Discrepancies were addressed through consensus. In addition, the labels for the macaque dataset were provided by a previous study [15], and we made no further modifications. The human brain masks of the ABCD, UKB and ZIB datasets used in this work were obtained from efforts and contributions reported in previous literature [42–46]. All of the labels have been visually inspected.

Evaluation metrics

To obtain quantitative performance results, we used the following metrics to compare the methods from different perspectives.
The Dice score measures the overlap between the segmentation results and the ground truth. The definition of this measure is formulated as follows:

$$Dice = \frac{2|S \cap G|}{|S| + |G|} \qquad (5)$$

where G and S denote the predicted and ground-truth label maps, respectively. Higher scores indicate better consistency.

The 95% Hausdorff distance (HD95) measures how far the segmentation results and ground truth for a metric space are from each other as given by the following equation:

$$\mathrm{HD}_{95}(S, G) = \max\left(E^{95th}_{s \in \partial(S)} \min_{g \in \partial(G)} \{\| s - g \|\}, E^{95th}_{g \in \partial(G)} \min_{s \in \partial(S)} \{\| g - s \|\} \right) \qquad (6)$$

where $\partial(\cdot)$ represents the boundary of the set and $E^{95th}{}_{s \in \partial(S)}$ and $E^{95th}{}_{g \in \partial(G)}$ are the 95th percentiles of the Euclidean distances between pixels in S and G, respectively.

Statistics

Statistical analyses were performed using SciPy (version 1.3.3) in Python (version 3.6.10). The exact sample values and results of statistical tests, including the Mann–Whitney test, Wilcoxon signed-rank test, and paired t-test, are provided in the figure captions or within the figure panels themselves. Differences are considered statistically significant when $p<0.05$.

Acknowledgements

This work was supported by the Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01), ZJLab, Shanghai Center for Brain Science and Brain-Inspired Technology, National Natural Science Foundation of China (81873893, 82171903, 92043301), the Office of Global Partnerships (Key Projects Development Fund) at Fudan University.

Declaration of Competing Interests

All authors declare no competing interests.

Author contributions

Dr. Xiao-Yong Zhang and Ziqi Yu conceived and designed the study. Dr. Xiao-Yong Zhang and Wenjing Xu performed some experiments. Ziqi Yu and Xiaoyang Han analyzed and interpreted the data. Dr. Jie Zhang and Dr. Jianfeng Feng provided resources. Ziqi Yu, Dr. Tingying Peng, Dr. Xiao-Yong Zhang wrote the manuscript. Dr. Jie Zhang, Dr. Dinggang Shen and Dr. Jianfeng Feng discussed and revised the draft. All authors reviewed and commented on the manuscript, and approved its final submission.

Code availability

We have released BEN as open-source code (https://github.com/yu02019/BEN).

Data availability

The MRI data will be released via https://github.com/yu02019/BEN.

References

1. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
2. Cox, R. W. AFNI: What a long strange trip it's been. *NeuroImage* vol. 62 743–747 (2012).
3. FSL. *Neuroimage* **62**, 782–790 (2012).
4. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
5. Yan, C.-G., Wang, X.-D., Zuo, X.-N. & Zang, Y.-F. DPABI: Data Processing & Analysis for (Resting-State) Brain Imaging. *Neuroinformatics* **14**, 339–351 (2016).
6. Beare, R. *et al.* Brain extraction using the watershed transform from markers. *Front. Neuroinform.* **7**, 32 (2013).
7. Lohmeier, J., Kaneko, T., Hamm, B., Makowski, M. R. & Okano, H. atlasBREX: Automated template-derived brain extraction in animal MRI. *Sci. Rep.* **9**, 12219 (2019).
8. Nie, J. & Shen, D. Automated segmentation of mouse brain images using multi-atlas multi-ROI deformation and label fusion. *Neuroinformatics* **11**, 35–45 (2013).
9. Chang, H.-H., Yeh, S.-J., Chiang, M.-C. & Hsieh, S.-T. Automatic brain extraction and hemisphere segmentation in rat brain MR images after stroke using deformable models. *Med. Phys.* (2021) doi:10.1002/mp.15157.
10. Oguz, I., Zhang, H., Rumple, A. & Sonka, M. RATS: Rapid Automatic Tissue Segmentation in rodent brain MRI. *J. Neurosci. Methods* **221**, 175–182 (2014).
11. Liu, Y., Unsal, H. S., Tao, Y. & Zhang, N. Automatic Brain Extraction for Rodent MRI Images. *Neuroinformatics* **18**, 395–406 (2020).
12. Zhao, G. *et al.* Bayesian convolutional neural network based MRI brain extraction on nonhuman primates. *Neuroimage* **175**, 32–44 (2018).
13. Zhong, T. *et al.* DIKA-Nets: Domain-invariant knowledge-guided attention networks for brain skull stripping of early developing macaques. *Neuroimage* **227**, 117649 (2021).
14. Garcia-Saldivar, P. *et al.* PREEMACS: Pipeline for preprocessing and extraction of the macaque brain surface. *Neuroimage* **227**, 117671 (2021).
15. Wang, X. *et al.* U-net model for brain extraction: Trained on humans for transfer to non-human primates. *Neuroimage* **235**, 118001 (2021).
16. De Feo, R. *et al.* Automated joint skull-stripping and segmentation with Multi-Task U-Net in large mouse brain MRI databases. *Neuroimage* **229**, 117734 (2021).
17. Valverde, J. M. *et al.* RatLesNetv2: A Fully Convolutional Network for Rodent Brain Lesion Segmentation. *Front. Neurosci.* **14**, 610239 (2020).
18. Hsu, L.-M. *et al.* Automatic Skull Stripping of Rat and Mouse Brain MRI Data Using U-Net. *Front. Neurosci.* **14**, 568614 (2020).
19. Yu, Z., Zhai, Y., Han, X., Peng, T. & Zhang, X.-Y. MouseGAN: GAN-based multiple MRI modalities synthesis and segmentation for mouse brain structures. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* 442–450 (Springer International Publishing, 2021).
20. Han, X., Zhai, Y., Yu, Z., Peng, T. & Zhang, X.-Y. Detecting extremely small lesions in mouse brain MRI with point annotations via multi-task learning. in *Machine Learning in Medical Imaging* 498–506 (Springer International Publishing, 2021).
21. Wang, W., Zheng, V. W., Yu, H. & Miao, C. A survey of zero-shot learning. *ACM Trans. Intell. Syst. Technol.* **10**, 1–37 (2019).
22. Ma, J. *et al.* Toward data-efficient learning: A benchmark for COVID-19 CT lung and

infection segmentation. *Med. Phys.* **48**, 1197–1210 (2021).
23. Chou, N., Wu, J., Bai Bingren, J., Qiu, A. & Chuang, K.-H. Robust automatic rodent brain extraction using 3-D pulse-coupled neural networks (PCNN). *IEEE Trans. Image Process.* **20**, 2554–2564 (2011).
24. Dorr, A. E., Lerch, J. P., Spring, S., Kabani, N. & Henkelman, R. M. High resolution three-dimensional brain atlas using an average magnetic resonance image of 40 adult C57Bl/6J mice. *Neuroimage* **42**, 60–69 (2008).
25. Barrière, D. A. *et al.* The SIGMA rat brain templates and atlases for multimodal MRI data analysis and visualization. *Nat. Commun.* **10**, 1–13 (2019).
26. Evans, A. C. *et al.* 3D statistical neuroanatomical models from 305 MRI volumes. in *1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference* (IEEE, 2005). doi:10.1109/nssmic.1993.373602.
27. Calabrese, E., Badea, A., Watson, C. & Johnson, G. A. A quantitative magnetic resonance histology atlas of postnatal rat brain development with regional estimates of growth and variability. *Neuroimage* **71**, 196–206 (2013).
28. Hamezah, H. S. *et al.* Volumetric changes in the aging rat brain and its impact on cognitive and locomotor functions. *Exp. Gerontol.* **99**, 69–79 (2017).
29. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).
30. Zeng, D. *et al.* Segmentation with Multiple Acceptable Annotations: A Case Study of Myocardial Segmentation in Contrast Echocardiography. *Lecture Notes in Computer Science* 478–491 (2021) doi:10.1007/978-3-030-78191-0_37.
31. Hüllermeier, E. & Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning* vol. 110 457–506 (2021).
32. Baumgartner, C. F. *et al.* PHiSeg: Capturing uncertainty in medical image segmentation. in *Lecture Notes in Computer Science* 119–127 (Springer International Publishing, 2019).
33. Jungo, A. *et al.* On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* 682–690 (Springer International Publishing, 2018).
34. Wang, G. *et al.* Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* **335**, 34–45 (2019).
35. Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P. & Kapur, T. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. *IEEE Trans. Med. Imaging* **39**, 3868–3878 (2020).
36. Xie, X. *et al.* A survey on incorporating domain knowledge into deep learning for medical image analysis. *Med. Image Anal.* **69**, 101985 (2021).
37. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Lecture Notes in Computer Science* 234–241 (Springer International Publishing, 2015).
38. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
39. Wang, X., Girshick, R., Gupta, A. & He, K. Non-local Neural Networks. (2017).
40. Li, Y., Wang, N., Shi, J., Hou, X. & Liu, J. Adaptive Batch Normalization for practical domain adaptation. *Pattern Recognition* vol. 80 109–117 (2018).

41. Tustison, N. J. *et al.* N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* **29**, 1310–1320 (2010).
42. Kwon, Y., Won, J.-H., Kim, B. J. & Paik, M. C. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. *Comput. Stat. Data Anal.* **142**, 106816 (2020).
43. Accurate brain age prediction with lightweight deep neural networks. *Med. Image Anal.* **68**, 101871 (2021).
44. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* **166**, 400–424 (2018).
45. The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).
46. Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523–1536 (2016).
47. Gong, W., Rolls, E. T., Du, J., Feng, J. & Cheng, W. Brain structure is linked to the association between family environment and behavioral problems in children in the ABCD study. *Nat. Commun.* **12**, 1–10 (2021).