BOOSTING SKULL-STRIPPING PERFORMANCE FOR PEDIATRIC BRAIN IMAGES

William Kelley[1,2], **Nathan Ngo**[1,2], **Adrian V. Dalca**[1,2,3,5], **Bruce Fischl**[1,2,3], **Lilla Zöllei**[1,2,3,*], **Malte Hoffmann**[1,2,3,4,*]

[1]Athinoula A. Martinos Center for Biomedical Imaging, Charlestown, MA 02129, USA

[2]Department of Radiology, Massachusetts General Hospital, Boston, MA 02114, USA

[3]Department of Radiology, Harvard Medical School, Boston, MA 02115, USA

[4]Division of Health Sciences and Technology, MIT, Cambridge, MA 02139, USA

[5]Computer Science & Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA

Abstract

Skull-stripping is the removal of background and non-brain anatomical features from brain images. While many skull-stripping tools exist, few target pediatric populations. With the emergence of multi-institutional pediatric data acquisition efforts to broaden the understanding of perinatal brain development, it is essential to develop robust and well-tested tools ready for the relevant data processing. However, the broad range of neuroanatomical variation in the developing brain, combined with additional challenges such as high motion levels, as well as shoulder and chest signal in the images, leaves many adult-specific tools ill-suited for pediatric skull-stripping. Building on an existing framework for robust and accurate skull-stripping, we propose developmental SynthStrip (d-SynthStrip), a skull-stripping model tailored to pediatric images. This framework exposes networks to highly variable images synthesized from label maps. Our model substantially outperforms pediatric baselines across scan types and age cohorts. In addition, the <1-minute runtime of our tool compares favorably to the fastest baselines. We distribute our model at https://w3id.org/synthstrip.

Index Terms—

skull-stripping; brain extraction; newborn; infant; toddler; machine learning; pediatric MRI

1. INTRODUCTION

Skull-stripping is the isolation of the brain from surrounding anatomical features, noise, and background signal in neuroimaging data, for example acquired with magnetic resonance imaging (MRI). It is an essential pre-processing step for many neuroimaging analysis

---

*Equal contribution

COMPLIANCE WITH ETHICAL STANDARDS

The retrospective analysis of the MGH data required no ethical approval. We signed data use agreements for access to the publicly available BCP and dHCP data.

pipelines, in which downstream image processing tasks frequently rely on input images with non-brain tissue removed [1–3]. These pipelines automate labor-intensive processing steps and eliminate subjectivity, enabling researchers to focus on data interpretation and accelerating the pace of discovery in neuroscience.

As neuroanatomy differs substantially between infants and adults, methods developed for the latter are not generally well-suited for younger cohorts. For example, the brain undergoes rapid development during the first two years of life [4]. During this time the brain doubles in size and the gray-white matter tissue MRI contrast flips (6–9 months). Additionally, pediatric scans are prone to motion artifacts and commonly include parts of the shoulders and chest. These challenges motivate the development of dedicated algorithms for skull-stripping in pediatric populations.

Related Work.

There are many existing skull-stripping methods developed for adult brain scans, which leverage a variety of strategies. Some methods iteratively fit deformable brain surfaces to the image [5], while others determine the brain boundary using a combination of generative and discriminative models, such as Random-Forest classifiers [6]. More recently, deep-learning (DL) approaches train deep neural networks to segment the brain [7], often building on U-Net architectures [8].

Few skull-stripping algorithms are tailored specifically to pediatric populations. Typically, these more recent DL methods either target a single MRI contrast [9] or train a different network for each of the available contrasts [10]. One approach uses separate two-dimensional (2D) networks for axial, coronal, and sagittal views extracted from the same input volume before fusing predictions via a voting scheme [9]. Another method trains a 3D U-Net to operate on overlapping 3D patches of the input volume [10].

Synthesis Strategy.

A recent learning strategy trains neural networks without acquired images, producing models that robustly generalize across datasets and imaging modalities [11, 12]. Synthesizing diverse training images from label maps, prior work achieves state-of-the-art performance on registration [13–15] and segmentation [16,17]. SynthStrip [18] lever-ages this approach for robust skull-stripping. Despite demonstrated performance across a large variety of images including pediatric MRI, SynthStrip is an age-agnostic tool that does not specifically target this younger population.

Contribution.

We demonstrate that optimizing SynthStrip for pediatric populations leads to performance gains, essential for downstream pediatric neuroimaging pipelines, and helps meet specific pipeline requirements, such as the exclusion of cerebrospinal fluid (CSF) from brain masks [2]. We build on SynthStrip's generative model and architecture to address the challenges of pediatric neuroimaging data. We create a novel set of pediatric label maps for training-image synthesis and use it to train a new skull-stripping model, developmental SynthStrip (d-SynthStrip). We thoroughly analyze d-SynthStrip's performance on real MRI scans across

MRI contrasts and pediatric age groups. We also investigate network-architecture variations to identify an optimal training configuration that surpasses state-of-the-art pediatric solutions in accuracy. Our baseline comparison focuses on publicly available and readily usable tools that can be run without retraining. We will freely distribute our model at w3id.org/synthstrip as a stand-alone tool and as part of the upcoming FreeSurfer and Infant FreeSurfer releases.

2. METHODS

We implement the supervised SynthStrip framework [18] for skull-stripping and tailor it to pediatric neuroimaging data. Let x be a 3D gray-scale image. A deep convolutional network (CNN) g_θ with trainable parameters θ predicts the binary brain mask $\hat{y} = g_\theta(x)$, such that a voxel-wise multiplication yields the skull-stripped image $x_{\hat{y}} = x \odot \hat{y}$.

Instead of training with real images, the framework draws a pre-computed whole-head label map s at each optimization step and synthesizes head scan x with randomized intensity features from it. Each step updates parameters θ to minimize a loss $\mathcal{L}(y, \hat{y})$ that encourages similarity between $\hat{y}$ and the target brain mask y, derived from the brain labels in s. Figure 1 provides an overview of the learning framework, while Figure 2 shows training-image examples.

Training and Validation Data.

We assemble a local dataset (MGH) from (i) 29 Infant FreeSurfer [2] training images (ii) 18 newborn scans [19, 20] and (iii) the M-CRIB atlas cohort (N=10) [21]. We select these 3 sources to cover a wide age range of 0–56 months (Table 1) and maximize variability across the included structural T1-weighted (T1w) and T2-weighted (T2w) structural scans as well as whole-brain manual label segmentations. We explicitly pool no training subjects from the test datasets (below) to assess generalizability to popular large-scale datasets unseen at training.

We emphasize that we train d-SynthStrip with images synthesized from label maps rather than the label maps themselves. We create training label maps by combining manually drawn brain labels with an additional six labels across the non-brain image content, produced by fitting a Gaussian mixture model (GMM) [18] to the intensities of each image. The added labels have no neuroanatomical significance – we include them in training to synthesize more variable image content. For a balanced distribution of the GMM labels across the image, we apply non-uniformity correction to the image intensities prior to the GMM fit [1]. For each image, we replace GMM-fitted labels that fall inside the brain boundary with the manual labels to produce a single label map.

Generative Model.

At each training step, we sample s from the set of training label maps [18]. First, we augment the spatial variability of s by applying the composition of a random affine (including translation, rotation, scaling, and shear) and nonlinear transform. Second, we sample a mean intensity value for each label and an overall variance. Then we sample intensity values for each voxel of the label from the corresponding normal distribution to generate gray-scale image x. Third, we apply an array of randomized corruptions

including a spatially-varying intensity bias field, global intensity exponentiation, cropping, downsampling, and Gaussian blurring. These steps produce highly variable training data with complex intensity patterns across the image voxels of each label, including and also far exceeding the variability seen in medical images (Figure 2).

From the spatially augmented label map, we also derive ground-truth brain mask y. First, we merge all brain labels excluding non-ventricular CSF to form a binary map. Second, we fill and include the space between brain folds into the brain mask, via 10 iterations of dilation followed by 10 iterations of erosion using nearest-neighbor connectivity. Third, we fill any remaining 3D holes. The resulting brain mask y serves as the target for the network prediction in the loss function.

Architecture and Loss.

We use the 3D SynthStrip U-Net [18] architecture. The U-Net g_θ has seven resolution levels with two leaky-ReLU activated $3 \times 3 \times 3$ convolutions per level. It outputs two softmax-activated channels j and k for brain and background, respectively. We optimize g_θ using a Dice-based loss $\mathcal{L}_{\text{Dice}}$, which measures the overlap between the target brain mask y and the predicted mask $\hat{y}$:

$$\mathcal{L}_{\text{Dice}}(y, \hat{y}) = -\frac{\sum_v y_i(v)\hat{y}_j(v) + \sum_v y_k(v)\hat{y}_k(v)}{\sum_v y_j(v)^2 + \sum_v \hat{y}_k(v)^2},$$

(1)

where we sum over all voxels $v \in \Omega$ of the spatial domain Ω of image x. In our experiments, we also analyze another model variant [18], which predicts a signed distance transform (SDT) $\hat{d}$ representing the distance to the brain boundary at each voxel. We optimize the mean squared error (MSE) from the target SDT d computed from y. To focus the optimization gradients on the brain boundary, we down-weight the MSE contribution of voxels farther than distance h from this boundary by a factor of b [18].

Training Details.

We use 50 label maps from the MGH dataset for synthesis-based training and the remaining 7 real MR images for validation. We train our d-SynthStrip models with stochastic gradient descent using Adam with a batch size of 1, until the loss on the validation set plateaus. We conform all images and label maps to 256^3 volume size with 1 mm^3 isotropic voxels and left-inferior-anterior orientation using linear interpolation.

3. EXPERIMENTS

To assess the skull-stripping performance of our models, we compare them against state-of-the-art baseline methods across MRI contrasts and age groups.

Test Data.

We select 20 subjects from the UNC/UMN Baby Connectome Project (BCP) [22] and another 20 subjects from the Developing Human Connectome Project (dHCP) [23] to form

a test cohort of N=40 subjects. For each subject, we source a T1w and T2w MR scan along with a label map which corresponds to both images (except 1 BCP subject, for which we have no T2w image). For the BCP cohort, we manually review and correct label maps generated with the Infant FreeSurfer pipeline [2]. We obtain label maps for the dHCP cohort using the dHCP minimal processing pipeline [24]. Table 1 displays the age distribution for each cohort.

Baselines.

We compare our tool to well established skull-stripping methods. First, we test SkullStripping CNN (SS-CNN) [9], which targets T1w pediatric MRI. Second, we test the skull-stripping module of the Infant Brain Extraction and Analysis Toolbox (iBEAT) [10] developed for T1w and T2w MRI (version 2.0, release 120). Third, we test SynthStrip [18] version 1.5, with the –no–csf flag in order to match the masks predicted by all other methods, which exclude non-ventricular CSF. Finally, we test deepbet [25] version 0.0.2. Although deepbet focuses on T1w adult MRI, we include it as another DL solution due to its demonstrated performance [25]. As deepbet and SSCNN are tailored specifically to T1w MRI, we do not evaluate them on T2w images.

Metrics.

We evaluate skull-stripping accuracy relative to binary ground truth masks using volumetric Dice overlap scores and Hausdorff distances between brain-mask boundaries.

Setup.

First, we assess the brain-masking accuracy of each tool across MRI contrasts and age groups. Second, we analyze the two different architectures: we compare a traditional segmentation model with a Dice loss to SDT prediction with an unweighted (uSDT, $b = 0$ mm) and a weighted SDT loss (wSDT, $b = 10^{-3}$, $h = 4$ mm) from Section 2.

Results.

Figure 3 shows that d-SynthStrip trained with a Dice loss outperforms other skull-stripping methods regardless of contrast or subject cohort. Figure 4 compares skull-stripping examples for all methods, and Figure 5 quantifies skull-stripping errors across each testset in a nonlinear mid-space. Our SDT models match or slightly under-perform SynthStrip for the BCP images. SSCNN and iBEAT underperform compared to SynthStrip and our model across cohorts except the T1w dHCP scans, where they match the performance of SynthStrip and our d-SynthStrip SDT models.

In terms of Hausdorff distances, both our Dice and SDT models outperform all baselines tested, while the Dice model generally surpasses the SDT variants. SynthStrip closely follows SSCNN. While iBEAT struggles with the BCP data, it achieves the lowest Hausdorff distances among baseline methods for dHCP.

On an NVIDIA RTX 8000 GPU, d-SynthStrip, SynthStrip, and deepbet take less than 1 minute per image, including model setup. However, d-SynthStrip inference alone takes less

than 1 second. SSCNN takes approximately 15 minutes, while iBEAT requires up to 22 hours – skull-stripping results are not available before the full pipeline completes.

4. DISCUSSION

We present a pediatric brain extraction tool, d-SynthStrip, that outperforms specialized baseline skull-stripping methods on images acquired from newborns to toddlers.

While the synthesis strategy previously proved to produce networks that robustly generalize across patient populations, we demonstrate the benefit of synthesizing training data from label maps of a targeted population. d-SynthStrip outperforms SynthStrip by up to 10 Dice points and up to 20 mm Hausdorff distances on infant data. This difference in performance suggests that the synthetic scaling and deformations applied during synthesis may insufficiently cover the distribution of developing brain shapes.

While prior work shows similar skull-stripping accuracy between models trained with Dice and SDT losses [18], we find the Dice loss to lead to increased Dice scores at test time. This result is not surprising, and we plan to investigate receiver operating characteristic (ROC) curves in the future for a more comprehensive comparison of the two losses.

In addition, we will explore whether increasing the variability of the generative model, specifically the synthetic warps applied to input label maps, may bridge the performance gap to yield accurate masks across both pediatric and adult populations with a single model. We will also investigate whether a model trained with a dataset carefully balanced to cover the whole lifespan can robustly accommodate both pediatric and adult brain scans.

ACKNOWLEDGMENTS

REFERENCES

[1]. Fischl Bruce, "FreeSurfer," NeuroImage, vol. 62, no. 2, pp. 774–781, 2012. [PubMed: 22248573]

[2]. Zöllei Lilla et al. , "Infant FreeSurfer: An automated segmentation and surface extraction pipeline for T1-weighted neuroimaging data of infants 0–2 years," NeuroImage, vol. 218, pp. 116946, 2020. [PubMed: 32442637]

[3]. Gaser Christian et al. , "CAT–A computational anatomy toolbox for the analysis of structural MRI data," biorxiv, pp. 2022–06, 2022.

[4]. Knickmeyer Rebecca C et al. , "A structural MRI study of human brain development from birth to 2 years," Journal of Neuroscience, vol. 28, no. 47, pp. 12176–12182, 2008. [PubMed: 19020011]

[5]. Smith Stephen M, "Fast robust automated brain extraction," Hum Brain Mapp, vol. 17, no. 3, pp. 143–155, 2002. [PubMed: 12391568]

[6]. Iglesias Juan Eugenio, Liu Cheng-Yi, Thompson Paul M, and Tu Zhuowen, "Robust brain extraction across datasets and comparison with publicly available methods," IEEE TMI, vol. 30, no. 9, pp. 1617–1634, 2011.

[7]. Isensee Fabian et al. , "Automated brain extraction of multisequence MRI using artificial neural networks," Hum Brain Mapp, vol. 40, no. 17, pp. 4952–4964, 2019. [PubMed: 31403237]

[8]. Ronneberger Olaf et al., "U-net: Convolutional networks for biomedical image segmentation," in MICCAI. Springer, 2015, pp. 234–241.

[9]. Jog Amod et al. , "Fast Infant MRI Skullstripping with Multiview 2D Convolutional Neural Networks," arXiv preprint arXiv:1904.12101, 2019.

[10]. Wang Li et al. , "iBEAT V2.0: a multisite-applicable, deep learning-based pipeline for infant cerebral cortical surface reconstruction," Nature Protocols, vol. 18, 2023.

[11]. Hoffmann Malte, Billot Benjamin, Iglesias Juan E, et al., "Learning MRI contrast-agnostic registration," in 2021 IEEE ISBI. IEEE, 2021, pp. 899–903.

[12]. Billot Benjamin et al. , "Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets," PNAS, vol. 120, no. 9, pp. e2216399120, 2023. [PubMed: 36802420]

[13]. Hoffmann Malte et al., "Anatomy-specific acquisitionagnostic affine registration learned from fictitious images," in Medical Imaging 2023: Image Processing. SPIE, 2023, vol. 12464, p. 1246402.

[14]. Hoopes Andrew et al. , "Learning the Effect of Registration Hyperparameters with HyperMorph," Machine Learning for Biomedical Imaging, vol. 1, no. IPMI 2021 special issue, pp. 1–30, 2022.

[15]. Hoffmann Malte, Hoopes Andrew, Douglas N Greve Bruce Fischl, and Dalca Adrian V, "Anatomy-aware and acquisition-agnostic joint registration with Synth-Morph," arXiv preprint arXiv:2301.11329, 2024.

[16]. Billot Benjamin, Robinson Eleanor, Dalca Adrian V, and Iglesias Juan Eugenio, "Partial volume segmentation of brain MRI scans of any resolution and contrast," in MICCAI. Springer, 2020, pp. 177–187.

[17]. Billot Benjamin et al. , "SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining," Medical image analysis, vol. 86, pp. 102789, 2023. [PubMed: 36857946]

[18]. Hoopes Andrew, Mora Jocelyn S, Dalca Adrian V, et al. , "SynthStrip: Skull-stripping for any brain image," NeuroImage, vol. 260, pp. 119474, 2022. [PubMed: 35842095]

[19]. Warton Fleur L et al. , "Prenatal methamphetamine exposure is associated with reduced subcortical volumes in neonates," Neurotoxicology and teratology, vol. 65, pp. 51–59, 2018. [PubMed: 29069607]

[20]. Warton FL et al. , "Maternal choline supplementation mitigates alcohol exposure effects on neonatal brain volumes," Alcoholism: Clinical and Experimental Research, vol. 45, no. 9, pp. 1762–1774, Sep 2021. [PubMed: 34342017]

[21]. Alexander Bonnie et al. , "A new neonatal cortical and subcortical brain atlas: the Melbourne Children's Regional Infant Brain (M-CRIB) atlas," NeuroImage, vol. 147, pp. 841–851, 2017. [PubMed: 27725314]

[22]. Howell Brittany R et al. , "The UNC/UMN Baby Connectome Project (BCP): An overview of the study design and protocol development," NeuroImage, vol. 185, pp. 891–905, 2019. [PubMed: 29578031]

[23]. Hughes Emer J et al. , "A dedicated neonatal brain imaging system," Magn Reson Med, vol. 78, no. 2, pp. 794–804, 2017. [PubMed: 27643791]

[24]. Makropoulos Antonios et al. , "The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction," NeuroImage, vol. 173, pp. 88–112, 2018. [PubMed: 29409960]

[25]. Fisch Lukas et al. , "Deepbet: Fast brain extraction of T1-weighted MRI using Convolutional Neural Networks," arXiv preprint arXiv:2308.07003, 2023.
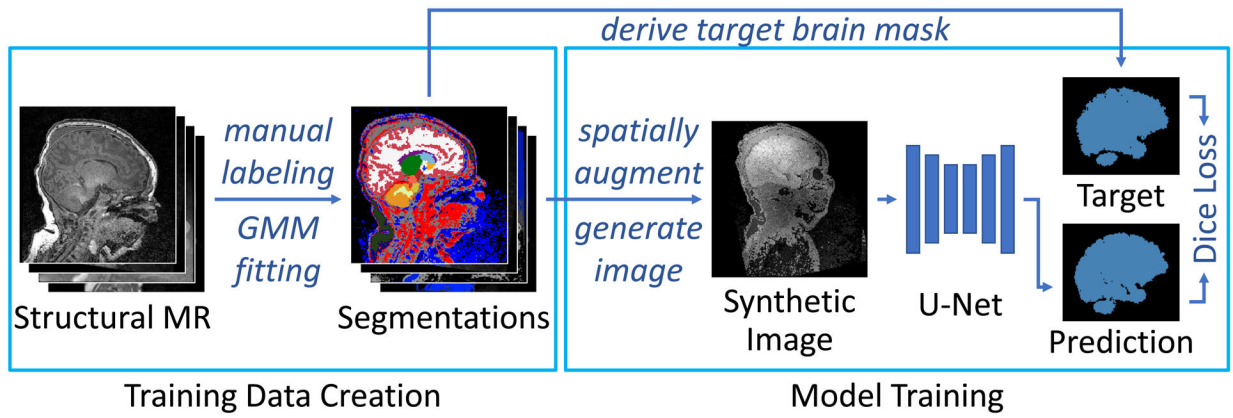
Fig. 1.

SynthStrip-based training framework. Starting with manual brain label maps, we synthesize widely variable brain images and matching ground-truth brain masks, which we then use to train the model.
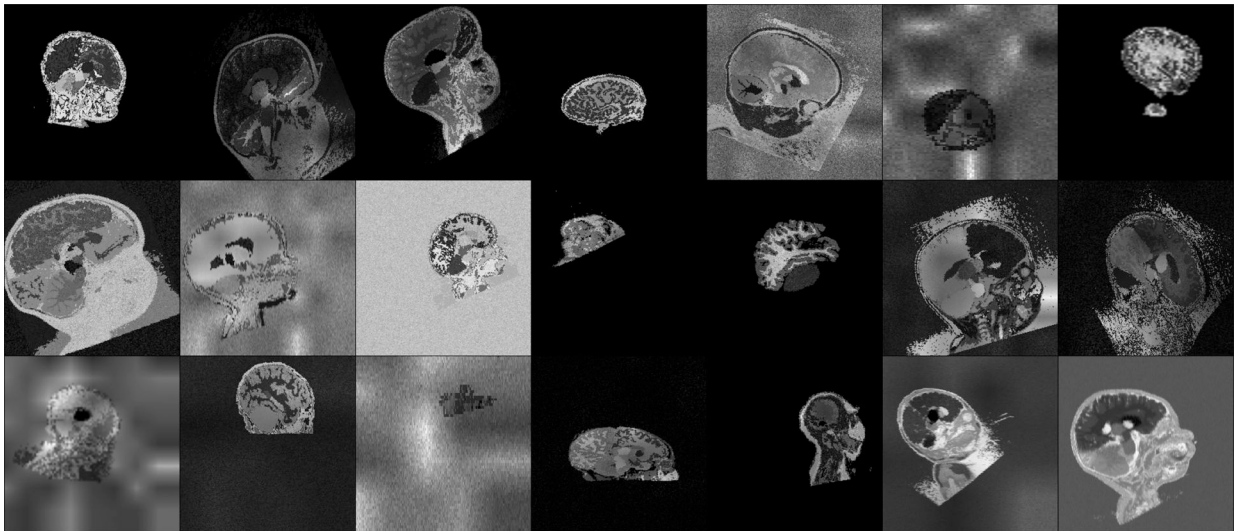
Fig. 2.
Synthetic training images generated from pediatric label maps. The spatial and intensity variability deliberately exceeds the range of medical images to encourage d-SynthStrip to generalize across MRI contrasts and age groups.
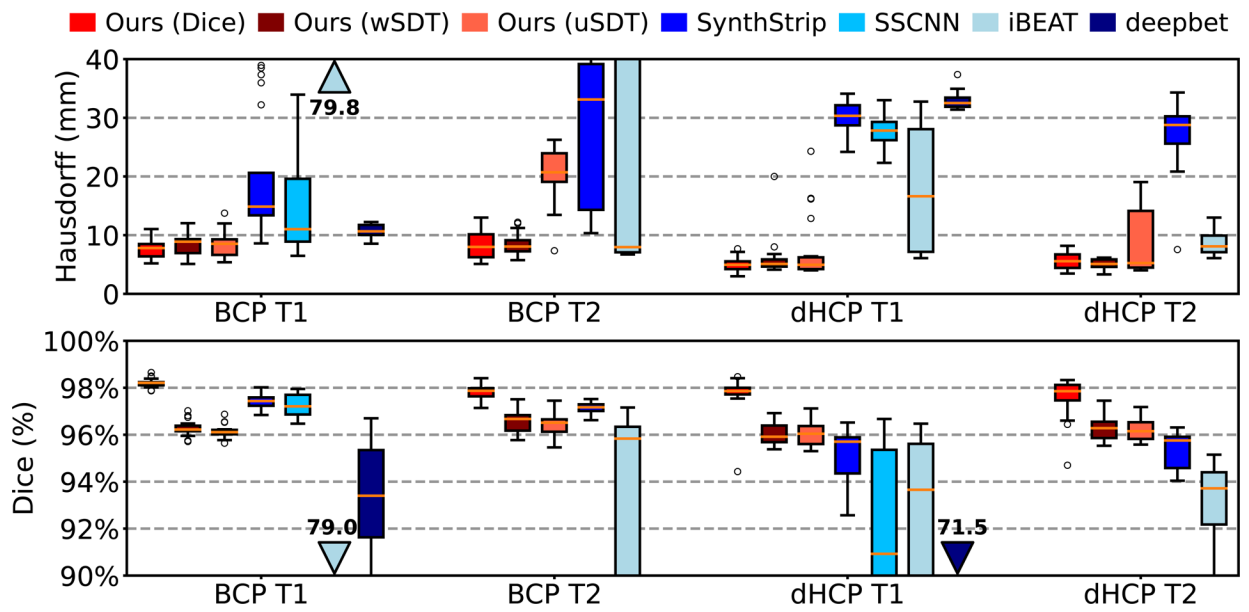
Fig. 3.

Brain extraction accuracy in terms of Hausdorff distance and volumetric Dice overlap.
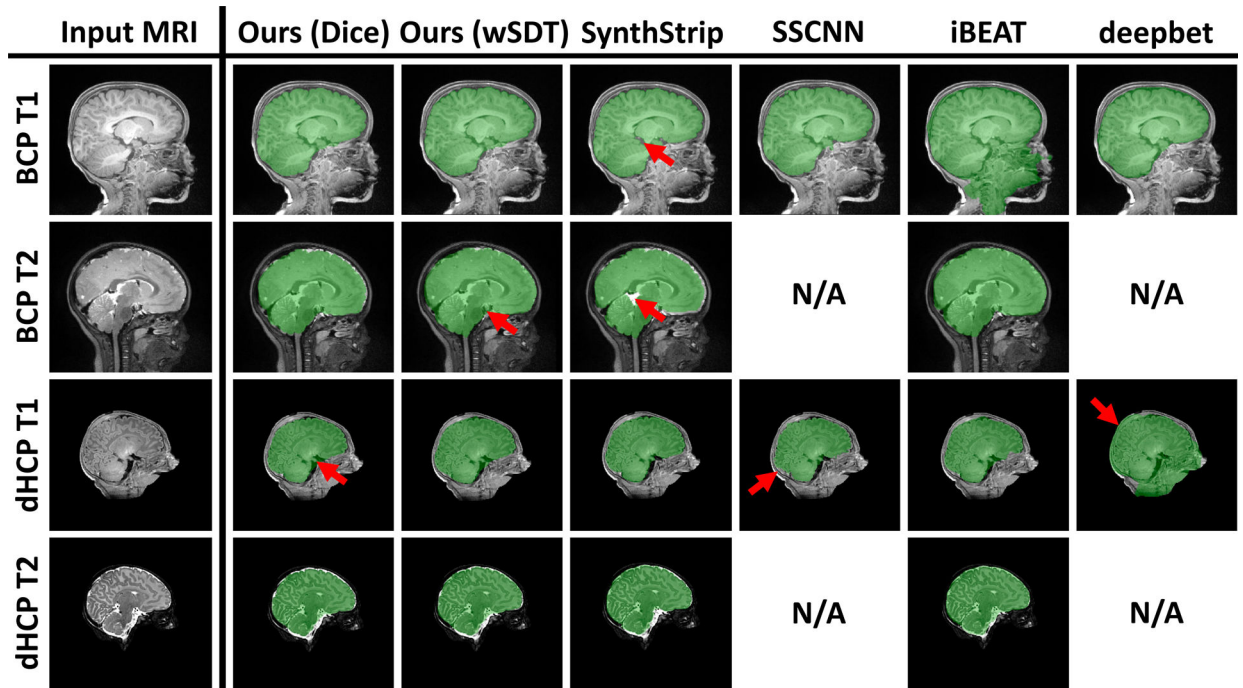
Testsets listed in Table 1.

Fig. 4.
Representative brain masks predicted by each skull-stripping method. SSCNN and deepbet focus on T1w MRI.
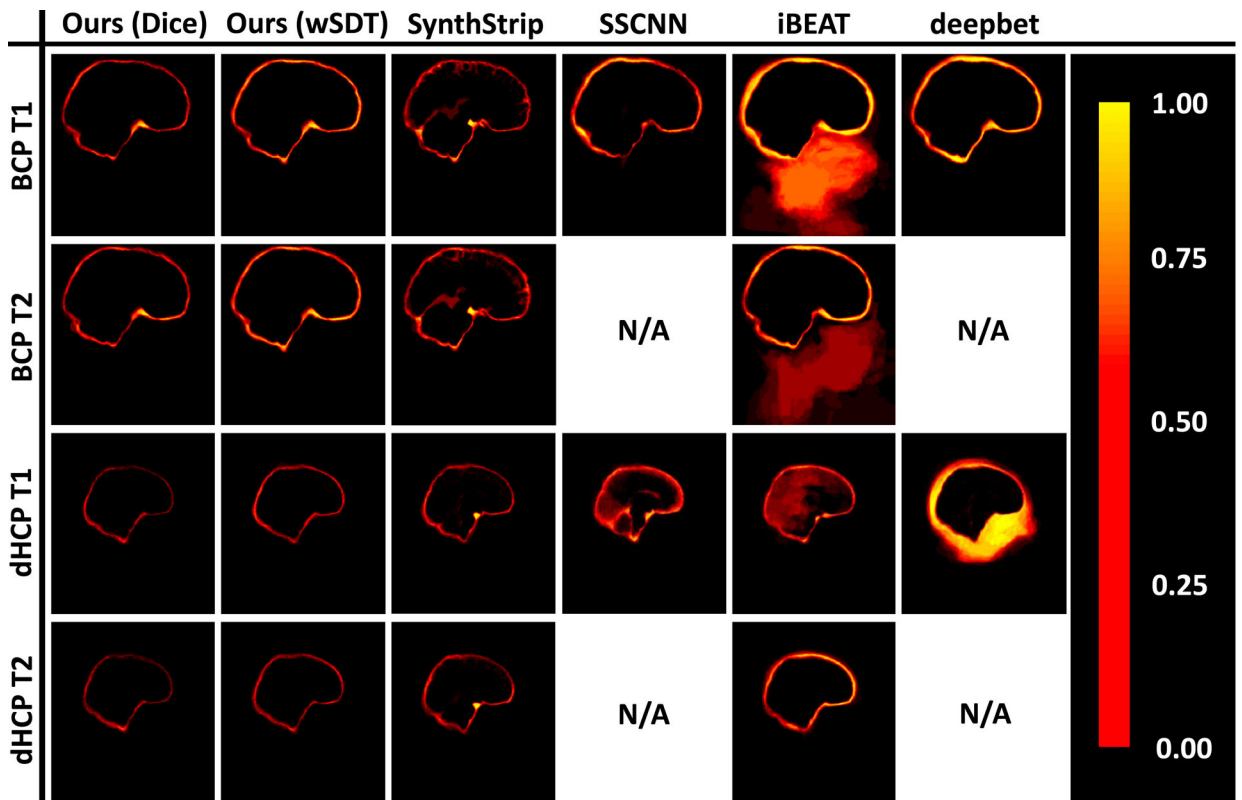
Fig. 5.
Proportion of absolute skull-stripping errors per voxel in a nonlinear mid-space, across all images of each testset.

Table 1.

Age distribution. The dHCP cohort includes preterm and term newborns, listed with gestational age (GA) at scan.

Cohort	Contrast	No.	Min	Max	Mean	St.Dev.
			Age (months)			
BCP	T1w	20	5	34	17	8
BCP	T2w	19	|	|	|	|
MGH	*mixed*	57	0	56	6	12
			GA at scan (weeks)			
dHCP	T1w	20	30	43	38	4
dHCP	T2w	20	|	|	|	|