Fast Infant MRI Skullstripping with Multiview 2D Convolutional Neural Networks

Amod Jog$^a$, P. Ellen Grant$^c$, Sandra W. Jacobson$^{d,e}$, Joseph L. Jacobson$^{d,e}$, Andre van der Kouwe$^a$, Ernesta M. Meintjes$^d$, Bruce Fischl$^{a,b}$, Lilla Zöllei$^a$

$^a$Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital and Harvard Medical School  
$^b$MIT Computer Science and Artificial Intelligence Laboratory  
$^c$Fetal Neonatal Neuroimaging and Developmental Science Center, Division of Newborn Medicine, Boston Children’s Hospital and Harvard Medical School  
$^d$Department of Human Biology, University of Cape Town Faculty of Health Sciences, Cape Town, South Africa  
$^e$Department of Psychiatry and Mental Health, University of Cape Town Faculty of Health Sciences, Cape Town, South Africa  
$^f$Department of Psychiatry and Behavioral Neurosciences, Wayne State University School of Medicine, Detroit, USA

Abstract

Skullstripping is defined as the task of segmenting brain tissue from a full head magnetic resonance image (MRI). It is a critical component in neuroimage processing pipelines. Downstream deformable registration and whole brain segmentation performance is highly dependent on accurate skullstripping. Skullstripping is an especially challenging task for infant (age range 0–18 months) head MRI images due to the significant size and shape variability of the head and the brain in that age range. Infant brain tissue development also changes the $T_1$-weighted image contrast over time, making consistent skullstripping a difficult task. Existing tools for adult brain MRI skullstripping are ill equipped to handle these variations and a specialized infant MRI skullstripping algorithm is necessary. In this paper, we describe a supervised skullstripping algorithm that utilizes three trained fully convolutional neural networks (CNN), each of which segments 2D $T_1$-weighted slices in axial, coronal, and sagittal views respectively. The three probabilistic seg-

URL: lzollei@nmr.mgh.harvard.edu (Lilla Zöllei)
sentations in the three views are linearly fused and thresholded to produce a final brain mask. We compared our method to existing adult and infant skullstripping algorithms and showed significant improvement based on Dice overlap metric (average Dice of 0.97) with a manually labeled ground truth data set. Label fusion experiments on multiple, unlabeled data sets show that our method is consistent and has fewer failure modes. In addition, our method is computationally very fast with a run time of 30 seconds per image on NVidia P40/P100/Quadro 4000 GPUs.

Keywords: infant, brain, MRI, skullstripping, deep learning

1. Introduction

Skullstripping is a critical image preprocessing step in most neuroimage processing pipelines [1]. The goal of skullstripping is to take an input whole head magnetic resonance image (MRI) and output a binary mask with a value of one for the brain and zero for all the extra-cerebral tissues such as skin, muscle, fat, bone etc. Extra-cerebral tissues significantly vary in their size and shape across subjects. Pulse sequences for brain imaging are also typically optimized to provide the best image contrast for brain tissues such as gray (GM) and white matter (WM). This results in extra-cerebral tissues with large variation in their intensities and imaging features that can prove difficult to model in downstream whole brain segmentation and deformable registration algorithms [2], potentially reducing their accuracy. Thus, skullstripping becomes a necessary step for most neuroimage processing pipelines to mask out uninteresting background and help increase the accuracy of the complete segmentation and cortical reconstruction pipeline [1].

Skullstripping is all the more important when processing infant brain MRI. Infant heads almost double in size in the first two years of life, as opposed to adult brains, which are of a relatively fixed size. Moreover, tissue contrast in the infant brain changes significantly with different stages of development. In Fig. 1 we show MPRAGE (magnetization prepared gradient echo) $T_1$-weighted acquisitions [3] of three different infants in three stages of development: newborn (column 1), 6 months old (column 2), and 18 months old (column 3). The tissue contrast differs significantly in all these stages. The small dimensions of the infant head can also result in variations in the field of view and scanner distortions, as is apparent in Fig. 1.
There has been a significant amount of work in the development of skullstripping algorithms for adult head MRI. These include methods using spherical expansion [4], filtering and edge-detection-based approaches [5], deformable model-based segmentation [6], variations of watershed algorithms [7, 8, 9], a generative-discriminative framework [10], patch-based sparse reconstruction methods for multi-modal data [11], multi-atlas registration-based skullstripping [12] among many others. These methods tend to have a sub-optimal performance for infant MRI skullstripping due to inherent assumptions about skull shape, size, and appearance in adults [13]. Infant MRI skullstripping has received some attention in recent years. Methods have been developed to work simultaneously for pediatric and adult brain MRI [14, 15, 16]. These include model-based level set driven skullstripping [14], a statistical morphology-based tool [15], and non-local patch-based segmentation [16]. A specialized method for infant skullstripping using the prior shape of neonatal brains that is learned using a labeled set of atlas images, with final segmentation using graph cuts was described by [13]. A meta-learning algorithm that combines the outputs of brain extraction tool (BET) [6] and brain surface extractor (BSE) was described in [5]. A multi-atlas registration-based method focusing on apparent diffusion coefficient (ADC) images of infants has also been developed [17]. Recently there have been methods that post-process the BET outputs for infant data specifically [18]. Multi-atlas registration-based methods tend to be computationally expensive due to the many registrations needed to align the atlases to the subject image. The same holds true for meta learning-based methods that need to run multiple available skullstripping algorithms and combine their results. Intensity and morphology-based methods can also take up to 10–20 minutes to produce a brain mask. Segmentation accuracy can also reduce in the presence of intensity inhomogeneities that can vary the intensity gradient magnitudes between similar tissues across slices.

Recently, deep learning architectures based on convolutional neural networks (CNN) have been successfully applied to a variety of medical image segmentation problems [19, 20, 21]. A 3D deep learning framework for skullstripping was described by [22] that used 3D patches of size $65 \times 65 \times 65$ as input to the network and the output was $7 \times 7 \times 7$-sized patches with probability values of each voxel belonging to the brain class. A great advantage of deep learning algorithms is the fast inference time. Inference using CNNs is very fast and fast skullstripping can significantly speed up neuroimage processing pipelines that typically take hours to complete for a single subject.
In this work, we describe SkullStripping CNN or SSCNN, a 2D multi-view CNN-based skullstripping approach. We train three independent 2D CNNs, one each for 2D slices in coronal, axial, and sagittal views. Predictions of these three networks are linearly combined to produce a final 3D mask. Past works have shown that using multi-view approaches can reduce disambigui-
eties as compared to a single view 2D segmentation [23]. 2D multi-view CNNs have also been used for whole brain segmentation [24]. 2D CNNs are more preferable for the skullstripping task because even for a human labeler, skull-
stripping is a visually intuitive task to perform on 2D slices instead of 3D
patches. Errors made in a single view prediction can potentially be corrected
by predictions from the other views. Additionally, skullstripping is usually
one of the first preprocessing tasks for brain MRI and the images can present
with uncorrected intensity inhomogeneities. The intensity bias is usually
assumed to be a slow-changing multiplicative field. A multiview 2D CNN
segments each slice independently, for each of the three independent views.
Intensity inhomogeneity may adversely affect segmentation of slices in one of
the views but the other views can compensate for that loss. In contrast, a
3D CNN would segment a 3D patch, which can have varying intensity inside
it making the segmentation less robust. Segmenting a full 3D image instead
of 3D patches is ideal but is not possible to achieve using CNNs due to GPU
memory constraints.

Our paper is organized as follows. Section 2 describes the method, train-
ing, and prediction using multi-view 2D CNNs. In Section 3 we describe
parameter selection experiments and comparison with other skullstripping
algorithms. Finally, in Section 4 we summarize our observations and con-
clude with possible avenues of future development.

2. Method

Let $\mathcal{A} = \{A_1, A_2, \ldots, A_M\}$ be a collection of $M$ images with a paired
expert manually labeled image set $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_M\}$. Each $A_i \in \mathcal{A}$
is a 3D image of size $256 \times 256 \times 256$ and is resampled to a voxel size of
$1 \times 1 \times 1$ mm$^3$. The paired collection $\{\mathcal{A}, \mathcal{Y}\}$ is referred to as the training
image set. Figures 1(a)–(f) show three training data subjects in three age
ranges and their manually labeled brain masks. From $\{\mathcal{A}, \mathcal{Y}\}$, we extract
corresponding 2D slices in the coronal orientation of size $256 \times 256$ to create
a 2D training data set $\{\mathcal{A}_{cor}, \mathcal{Y}_{cor}\}$. Similarly, we extract slices from the
Figure 1: Full head MPRAGE acquisitions and manually labeled skull masks for axial (rows 1 and 2), coronal (row 3 and 4), and sagittal (rows 5 and 6) of infants of (a) newborn, (b) 6 months and (c) 18 months of age.
Figure 2: The left half of the figure shows training of 3 2D U-Net segmentation architectures, each for slices oriented in one of the three cardinal orientations coronal, sagittal, and axial. The U-net architecture for all three with $D_l = 3$ is the one as shown on the right.

Sagittal and axial orientations from $\{A, Y\}$ to generate 2D training data sets, $\{A_{sag}, Y_{sag}\}$, and $\{A_{ax}, Y_{ax}\}$. Figure 2 (left) shows training examples for all three orientations.

2.1. SSCNN: Network Architecture and Training

For each orientation $r \in \{ax, cor, sag\}$, we train a 2D fully convolutional CNN $U_r$ that takes as input the $T_1$-weighted 2D slice in that orientation to predict the corresponding manually labeled slice. Paired intensity and label slice data $\{A_r, Y_r\}$ is used as training data for the CNN $U_r$. We are interested in the semantic voxel-wise segmentation of the input 2D slice and therefore use the well-known U-Net architecture [25] for $U_r$.

The U-Net consists of an encoder block with $L$ levels, followed by a symmetric decoder block. Figure 2 (right) shows an example U-net architecture with $L = 3$ levels. In each level $l$ of the encoder block, we have $D_l$ convolutional layers (orange layer in Fig. 2). The number of filters in each convolutional layer at level $l$ is $F_l$ and the number of filters in subsequent levels is $F_l = 2F_{l-1}$. All convolutional layers, except the last layer, have a rectified linear (ReLU) activation, with a filter size $f \times f$. We choose the optimal values for $L, f, F_1$, and $D_1$ by cross-validation experiments. Each convolutional layer (conv) is followed by a batch normalization (BN) layer [26] (green in Fig. 2). At the end of each level (except the last one) is a pooling layer (red in Fig. 2) that downsamples the output of that level by 2 in the image height-width directions. At the deepest level, the encoder block
ends and the decoder block begins. In the decoder, the deepest conv+BN layers are followed by an upsampling layer (blue in Fig. 2) that upsamples the output of the previous level by a factor of 2 in the image height-width directions. The upsampled output is followed by a symmetric number of levels as the encoder. The number of filters and number of conv+BN layers at each level of the decoder block is the same as the corresponding symmetric level on the encoder block. The final layer consists of \( n_{labels} = 2 \) 1 \times 1 \) conv layers with a softmax activation. To summarize, when a network \( U_r \) is applied to an input slice of the shape \((256, 256, 1)\) in orientation \( r \), the output is of the shape \((256, 256, 2)\), and stores the probabilities of each voxel in the input slice belonging to the background (class 0), or brain (class 1).

All three U-Nets are trained to minimize a soft Dice-based loss averaged over the whole batch. The loss is shown in Eqn. 1, where \( v \) denotes the voxels present in ground truth slice \( y_{true}(b) \) and predicted slice \( y_{pred}(b) \) of the \( b^{th} \) training sample in a batch of \( N_B \) samples in \( y_{true} \) and \( y_{pred} \). During training, \( y_{pred}(b) \) is a \( 256 \times 256 \times n_{labels} \)-sized tensor with each voxel storing the softmax probability of it belonging to a particular label. \( y_{true}(b) \) is a similarly-shaped one-hot encoding of the label present at each of the voxels.

\[
\text{Loss}(y_{true}, y_{pred}) = \frac{1}{N_B} \sum_b \left( 1 - \frac{2 \sum_v y_{true}(b) \cdot y_{pred}(b)}{\sum_v (||y_{true}(b)||^2 + ||y_{pred}(b)||^2)} \right)
\]

(1)

We use the Adam optimization algorithm [27] to minimize the loss with an initial learning rate of \( 10^{-5} \) that reduces by half if the training loss does not reduce for five continuous epochs. Each epoch uses 3000 training slice pairs extracted from 23 training subjects. We use \( \approx 600 \) slices extracted from a validation dataset that consists of three subjects. We train each of the \( U_{cor}, U_{sag}, \) and \( U_{ax} \) for 30 epochs. The training loss reduces monotonously as the number of epochs increases, however, we select the model with the minimum validation loss as our final trained model. This avoids using networks that overfit to the training data and do not generalize well for test data from other sources. All three networks \( U_{cor}, U_{sag}, U_{ax} \) are trained independent of each other.

2.2. SSCNN: Prediction

Given a 3D test subject volume \( S \), we reorient it in coronal, sagittal, and axial directions to create 3D volumes \( S_{cor}, S_{sag}, \) and \( S_{ax} \) respectively. Let \( S_r(z) \) be the \( z^{th} \) slice in \( S_r, r \in \{ax, cor, sag\} \). We apply the trained network \( U_r \) to \( S_r(z) \) to generate output \( \hat{Y}_r(z) = U_r(S_r(z)) \) for all \( z \) to form \( \hat{Y}_r \).
Each of $\hat{Y}_{ax}$, $\hat{Y}_{cor}$, $\hat{Y}_{sag}$ is reoriented to the coronal orientation to create $\hat{Y}_{cor-ax}$, $\hat{Y}_{cor-cor}$, $\hat{Y}_{cor-sag}$, respectively. Note that each $\hat{Y}_{cor-r}$ image stores the voxel-wise class probabilities. We linearly combine these probability images to generate a fused result in the coronal orientation, as shown in Eqn. 2:

$$\hat{Y} = w_{ax}\hat{Y}_{cor-ax} + w_{cor}\hat{Y}_{cor-cor} + w_{sag}\hat{Y}_{cor-sag}.$$  

The weights are constrained such that $w_{ax} + w_{cor} + w_{sag} = 1$. We found the optimal weights using cross-validation experiments to be $w_{ax} = 0.44$, $w_{cor} = 0.33$, and $w_{sag} = 0.23$. Figure 3 illustrates the SSCNN prediction workflow. The brainmask is obtained by identifying the class with the higher probability value by $\hat{Y}_{hard} = \text{arg max}_{\{0,1\}} \hat{Y}$. To remove any small non-brain blobs, we identify the connected components in the binary labeled image and select the largest connected component as the final labeled brain.

3. Experiments

In this section we describe the training data set (Section 3.1), the SSCNN parameter selection step (Section 3.2), and a set of quantitative experiments demonstrating the high quality performance of our tool: a set of leave-one-out cross validation experiments and comparison with other skullstripping algorithms on the manually labeled training data set (Section 3.3), validation...
on an independent manually labeled test data set (Section 3.3.2), and a comparison on multiple unlabeled data sets (Section 3.3.3).

3.1. Training Data Set:

Our training data set was generated by using the volumes described in [28]. A total of 26 T1-weighted MRI volumes, from subjects whose age is almost uniformly distributed in the age range of 0-18 months, together with their binary brain masks were used.

3.2. SSCNN Parameter Selection

The 2D U-Nets used in SSCNN have a number of free parameters that need to be carefully selected for optimal performance. For parameter selection, we chose 13 subjects distributed roughly equally in the age range of 0-18 months as our parameter selection training set. We used four subjects to generate a validation data set to prevent overfitting. We tested the trained network on the remaining nine subjects. While selecting the optimal value of a parameter, we iterated over a set of possible values by keeping the other parameters fixed and chose the parameter with the highest test segmentation accuracy over the nine subjects. The optimal parameters are therefore dependent on the order in which they were fixed.

The parameters of interest for our proposed pipeline are the following: (1) kernel size \((f)\), (2) depth of network (number of pooling layers = \(L\)), (3) number of filters in the first level \((F_1)\), and (4) number of convolutional layers in each level \((D_l)\).

3.2.1. Kernel Size \((f)\)

This experiment was to select the most optimal kernel size of the 2D kernels used in each of the \(U_r\) CNNs. Larger kernel sizes have larger receptive fields that can act on larger image contexts. Therefore we expect larger kernel sizes to improve the segmentation accuracy, but they also have a much larger number of parameters to learn. We experimented with \(f = \{3, 5, 7\}\), by keeping the number of pooling layers \(L = 6\), the number of filters in the first layer \(F_1 = 16\), and the number of convolutional layers in each level \(D_l = 2\). Increasing the kernel size beyond \(7 \times 7\) resulted in out of memory errors. Figure [4] shows a boxplot of Dice overlap coefficients obtained on the nine test subjects when compared with the ground truth segmentation. The median Dice score increases as the kernel size increases, with \(f = 7\) resulting in the highest value. Therefore, we fixed the filter shape to be \(7 \times 7\).
3.2.2. Depth of Network

In this experiment we varied the depth of the network given by $2L + 1$, where $L$ is the number of pooling layers in the network. It has been observed that deeper networks lead to better performance. We varied $L \in \{4, 5, 6, 7\}$ and kept $D_l = 2$, $f = 7$, and $F_1 = 16$. In Fig. 5 we show boxplots for the four experiments, where we observe that the depth of the network does not change segmentation accuracy as it increases from four to seven. We set $L = 6$ as it produces marginally higher numbers than the rest, but it is not a significant difference.

3.2.3. Number of Filters

In this experiment, we chose the number of filters in the convolutional layers of the first level $F_1$. As described in Section 2, the number of filters in the subsequent levels $l$ is given by $F_l = 2F_{l-1}$, therefore the total number of filters in the entire network is a function of the number of filters in the first level. The more filters we have the more image variation can be captured at different network depths. We expect that more filters in the network will lead to higher segmentation accuracy. We experimented with $F_1 \in \{8, 16, 32\}$, by keeping $L = 6$, $f = 7$, and $D_l = 2$. $F_1 > 32$ was not possible due to GPU memory constraints. Figure 6 shows a boxplot of Dice coefficients obtained on the nine test subjects when compared with the ground truth segmentation.
The median Dice increases as $F_1$ increases from 8 to 32. $F_1 = 32$ produces the highest median Dice score and thus we set $F_1 = 32$. 

Figure 5: Dice coefficient with varying number of pooling layers

Figure 6: Dice coefficient with varying number of filters in the first level
3.2.4. Depth per Level of Network

In these experiments, we varied the number of convolutional layers $D_l$ at each level of the U-Net. We can expect that the higher the number of convolutional layers at each level, the better the image texture is captured. However, this also increases the number of parameters in their network leading to sub-optimal training and possible overfitting. We varied $D_l \in \{1, 2, 3, 4\}$ by keeping $f = 7$, $F_1 = 32$, $L = 6$. Figure 7 shows the Dice score boxplots calculated over the nine test subjects. We observe that the median Dice score increases from $D_l = 1$ to $D_l = 3$ and then falls for $D_l = 4$. Based on these experiments, we chose $D_l = 3$ for our optimal network.

![Figure 7: Dice coefficient with varying number of convolutional layers per level](image)

3.3. Evaluation of Segmentation Accuracy

3.3.1. Cross-validation on Training Data Set

We chose the best network with parameters selected from the experiments described in Section 3.2. We ran SSCNN in a leave-one-out-cross-validation framework (testing data was excluded from the training data) and evaluated the outcomes with the given ground truth segmentation using Dice and Jaccard overlap metrics. The mean and standard deviation of these overlap measures were: for Dice $(97.79; 1.43)$ and for Jaccard $(95.7; 2.66)$. Figure 8 displays the individual Dice score results.
In order to put such a performance into perspective, we selected a set of five publicly available and widely used skullstripping algorithms and evaluated them on the same data set. These tools were: ROBEX, BET, BSE, 3dSkullStrip and Watershed. We also optimized (where appropriate) the key parameters of these tools on the training data set, in order to run comparison as fair as possible. In the below description, we indicate the parameters, their optimization range and the optimized parameters.

Skullstripping tools: The RObust Brain Extraction (ROBEX) tool had been primarily designed for adult input images. It deforms a brain surface to brain boundary, which is found by the brain versus non-brain classification. The deformed brain surface is then locally refined by a graphcut algorithm to obtain the final brain mask. ROBEX had performed well among 10+ skullstripping algorithms in multi-site brain images of the adults [10] and in children as young as newborns [29]. The advantage of ROBEX is that it does not require users to modify any parameters. The Brain Extraction Tool (BET) is part of the FSL image analysis pipeline. It was primarily designed for images of adults. It evolves a deformable model to fit the brain surface. The performance of BET is often sensitive to parameter variations [30, 31]. BET was used for skullstripping neonatal brain images [32] [33]; but only upon careful and manual parameter tuning [31]. We varied the fractional intensity threshold (f) and vertical gradient in fractional intensity
| Method      | Dice (Mean; std. dev.) | Jaccard (Mean; std. dev.) |
|------------|-------------------------|---------------------------|
| SSCNN      | (97.8; 1.4)             | (95.7; 2.6)               |
| BET        | (76.0; 18.9)            | (64.9; 24.6)              |
| BSE        | (67.2; 24.1)            | (55.6; 29.0)              |
| ROBEX      | (93.4; 9.8)             | (88.7; 12.6)              |
| AFNI       | (86.7; 14.3)            | (78.7; 18.2)              |
| Watershed  | (84.9; 15.7)            | (76.4; 20.7)              |

Table 1: Mean and standard deviation of Dice and Jaccard overlap metrics computed on the training data set with SSCNN and five other commonly used skullstriping techniques.

threshold (g) parameters: f:[.2:.05:.8], g:[-.3:.05,.3]; (f,g)=(-.8,.05) Brain Surface Extractor (BSE) had been primarily designed for images of adults. It smooths the input image and uses edge detectors to find brain boundary, and refines the results by morphological operations. BSE was used in [32, 29] for neonatal images upon parameter tuning. We varied the diffusion (d) and edge detection constants (s): d:[10:5:60], s:[.42:.04:.82]; (d,s)=(-10,.58). 3dSkullStrip is part of the AFNI image analysis pipeline [4]. It replaces BET’s deformable model with a spherical surface expansion paradigm, and modifies BET in other parts to avoid the eyes and to reduce leakage into the skull. We varied the brain vs non-brain intensity threshold (shrink_fac) and speed of expansion (exp_frac) parameters: shrink_fac: [.4:.05:.8], exp_frac: [.05:.025:.15]; (shrink_fac, exp_frac)= (0.8, .05). Hybrid Watershed (WATERSHED) is part of the FreeSurfer neuroimaging analysis pipeline. It creates an initial brain surface using a watershed algorithm, and then evolves the brain surface to refine the result. We varied the preflooding height (h) parameter: h:[10:2:40]; h= 10.

The mean and standard deviation of these overlap measures are included in Table 3.3.1 and Figure 9 displays the individual Dice score results.

3.3.2. Evaluation on an Independent Newborn Data Set

Anonymized brain MRI images of eighteen newborns from a cohort of 43 non-sedated infants born to 32 heavy drinkers and 11 controls recruited prospectively during pregnancy for a brain imaging study [35] was selected and manually traced to be used as a second data set to quantitatively evaluate the performance of our new skullstripping algorithm. For imaging details, see Table 2. The segmenters were blind with respect to the newborn fetal
alcohol spectrum disorder diagnosis and prenatal alcohol and drug exposure. The mean and standard deviation of the Dice and Jaccard overlap scores was: (97.51, 0.35) and (95.15, 0.67).
| Data Set | N  | Age-at-scan          | Parameters                                                                                       | Scanner | MRI sequence                        |
|----------|----|----------------------|-------------------------------------------------------------------------------------------------|---------|-------------------------------------|
| dHCP     | 40 | Termination (37-44 weeks) | Hatsu TR = 150ms, TI = 134ms, TE = 8 ms, SENSE factor 2.27 (axial) and 2.66 (sagittal), 0.8x0.8mm^2 and 1.6mm slices overlapped by 0.8mm | 3T Philips Achieva | inversion recovery T1-weighted multislice fast spin-echo |
| BCH1     | 111+86 | 10-52 and 77-156 days (avg=2666, std=9.18 and avg=104.8, std=16.18) | TR = 250 ms, TE=1.74 ms, TI = 1450ms, flip angle=7, 1 mm^3 voxels, FOV=160mm | 3T Siemens Trio | T1-weighted mocoMEMPRAGE [17, 18] |
| BCH2     | 29+6  | 1-26, 106-202 days (avg=6.1, std=2.2 and avg=144.5, std=11.1) | TR = 250 ms, TE=1.74 ms, TI = 1450ms, flip angle=7, 1 mm^3 voxels, FOV=160mm | 3T Siemens Trio | T1-weighted mocoMEMPRAGE [17, 18] |
| BCH3     | 105  | 2-19 days (avg=9.32, std=3.59) | TR = 2500 ms, TE = 1.49 ms, flip angle = 7 degrees, 144 slices, 1 mm^3 voxels, FOV = 220x220 mm^2, GRAPPA=2 | 3T Siemens | T1-weighted mocoMEMPRAGE [17, 18] |
| BAN      | 54   | 68-70 days (avg=70, std=9.8) | TR = 250 ms, TE = 2.22 ms, flip angle = 7 degrees, 144 sagittal slices, 1 mm^3 voxels, FOV = 192 mm | 3T Siemens Verio | T1-weighted MPRAGE |
| UCP      | 18   | 7-47 days (avg=17.4, std=11.5) | MEF 5/20: TR = 20 ms, 8 echoes TE = 1.36, 68 ms echo, a = 0-7, 144 sagittal slices, 1 mm^3 voxels | 3T Siemens Allegia | multiflip angle, multi-echo FLASH (MEF) |
3.3.3. Evaluation on Unlabeled Data Sets

In addition to the previously described experiments, we also assembled a collection of anonymized and unlabeled test data set from five different initiatives in order to demonstrate the exceptional performance of our new tool on a variety of input images. Three of these originate from the Boston Childrens Hospital (BCH), one from Bangladesh and one from the recently released data set from the The Developing Human Connectome Project (dHCP) [36]. We compared the performance of SSCNN both qualitatively and quantitatively to the five already mentioned skull stripping solutions from Section 3.3.

Below is the description of all the data sets that were used for our experiments. The participating infants underwent structural MRI imaging and T1-weighted scans were acquired either on a 3T Siemens or a Phillips scanner during natural sleep. Human subject approval was obtained from all respective Institutional Review Boards and written informed parental consent was obtained for imaging. For details of the imaging protocol refer to Table 2.

**Unlabeled Data Sets:**

- **dHCP** (*N* = 40): Imaging data of forty newborn subjects was released by the developing human connectome project [36]. Even though the consortium processed the T2w images of these subjects in their initial release, the corresponding T1w images were also made available and we processed these in our study. The original data sets of $0.8 \times 0.8 \times 0.8$ mm$^3$ were downsampled to 1mm isotropic for our processing.

- **BCH1** (*N* = 199): Healthy, full-term neonates were recruited at the Brigham and Womens Hospital (BWH) and Beth Israel Deaconess Medical Center (BIDMC) as part of an ongoing prospective data collection study. The protocol was reviewed and approved by the institutional review boards at Boston Childrens Hospital (BCH), BWH and BIDMC. The participating infants were all singletons with normal Apgar scores and have no clinical concerns regarding perinatal brain injury or congenital or metabolic abnormalities. All subjects were full-term infants scanned within their first month of life and a subset called back for a second scan at about 4 months of age.

- **BCH2** (*N* = 35): Parents of neonates with congenital heart disease were approached for consent in the Cardiac ICU at BCH. This prospective study was approved by the institutional review board of BCH and was performed in compliance with the Health Insurance Portability and Accountability Act. Criteria for inclusion were: diagnosis of CHD confirmed by echocardiogram or cardiac MRI and ability to safely tolerate the brain MRI examination.
without sedation, prior and in some cases post to surgery. Neonates were excluded if there was evidence of a syndrome or genetic disease. BCH3 (N = 105): Native English-speaking children with and without a family history of DD were studied. All children were enrolled in a longitudinal dyslexia study which was approved by BCH institutional review. The data set used in this study is from the first timepoint acquisitions. BEAN (N = 54): Imaging data were collected in a set of infants from the Bangladesh Early Adversity Neuroimaging (BEAN) study investigating the effect of early biological and psychosocial adversity on children’s neurocognitive development among infants and children growing up in Dhaka, Bangladesh [37, 38]. For the set of unlabeled data sets, we used the optimized version of SSCNN and the five other skullstripping tools. Figures 10 and 11 display outcomes on the dHCP data, where SSCNN clearly outperforms the rest. Given that we did not have access to a ground truth solution, we STAPLe [39] all the outcomes together and then compared the SSCNN solution to it. The expectation was that the more SSCNN outperformed the other solutions the lower the Dice overlap score was between it and the STAPLEd labels, as well as the higher standard deviation. Both of these are well demonstrated in Figure 12.

4. Discussion and Conclusion

We have described SSCNN, a 2D multi-view CNN-based skullstripping method for infant MRI. SSCNN trains three independent networks to extract the brain mask from slices in three cardinal orientations—coronal, axial, and sagittal. The outputs of the three networks are linearly combined to produce a final brain mask. We have demonstrated in Section 3.3 and Section 3.3.2 that SSCNN is a highly accurate skullstripping algorithm and is significantly better than existing methods.

We ran SSCNN on hundreds of subjects from diverse, multi-site, multi-scanner MRI studies (Section 3.3.3) and showed an improved skullstripping performance compared to five other tools. SSCNN did not have a single case of gross skullstripping failure on any of these datasets, demonstrating its generalizability and robustness. Choosing a training dataset encompassing the infant age range ensured that SSCNN was robust to the contrast changes due to brain development.

SSCNN is computationally fast with a run time of less than 30 seconds with a GPU (NVidia P100, P40, Quadro P6000) and less than 2 minutes
Figure 10: Skullstripping solutions on all 40 data sets from the dHCP project using six different tools. The images are aligned in an unbiased affine coordinate space for visualization purposes and the central coronal slice is selected from each of the MRIs. Skullstripping contours are indicated in color: BET (pink), BSE (yellow), robex (white), AFNI (blue), watershed (black) and SSCNN (red).

Figure 11: Skullstripping solutions on three representative data sets from the dHCP project using six different tools. The images are aligned in an unbiased affine coordinate space for visualization purposes and the central coronal slice is selected from each of the MRIs. Skullstripping contours are indicated in color: BET (pink), BSE (yellow), robex (white), AFNI (blue), watershed (black) and SSCNN (red).
Figure 12: Dice overlap coefficients on unlabeled test data between solutions of our tool and the STAPLeEd version of all tested automated algorithms.

Presently, SSCNN is designed to work on $T_1$-weighted acquisitions. In the next version we plan to update the training with cross-sequence augmentation that will enable to it skullstrip acquisitions with $T_2$-weighted contrasts as well [40]. The linear combination of the three predictions in coronal, axial, and sagittal orientations was learned from cross-validation experiments. In the future, we will learn the weights of this combination by adding a custom layer that will collects the slice predictions in each orientation, reorients them, and combines them. This will ensure an end-to-end learning scheme for combining all three orientations and jointly optimizing their weights.

We also need to evaluate SSCNN for infant subjects with pathologies such as tumors, large ventricles that can significantly change the brain shape and boundary characteristics of brain and skull. Our training dataset does not have such subjects and we would need to augment the existing training dataset or add newer training datasets to enhance this feature for SSCNN.
In summary, we have described SSCNN, a fast, robust, infant MRI skullstripping framework. The code will be made available as a part of the FreeSurfer development version repository (https://github.com/freesurfer/freesurfer). Further validation and testing will be necessary before incorporating it into a release version.

5. Acknowledgements

Support for this research was provided in part by the BRAIN Initiative Cell Census Network grant U01MH117023, the National Institute for Biomedical Imaging and Bioengineering (P41EB015896, 1R01EB023281, R01EB006758, R21EB018907, R01EB019956), the National Institute on Aging (5R01AG008122, R01AG016495), the National Institute of Mental Health, the National Institute for Neurological Disorders and Stroke (R01NS0525851, R21NS072652, R01NS070963, R01NS083534, 5U01NS086625, 5U24NS10059103), the Eunice Kennedy Shriver National Institute of Child Health & Human Development (5R01HD065762), the National Institute on Alcohol Abuse and Alcoholism (R21 AA020037) and was made possible by the resources provided by Shared Instrumentation Grants 1S10RR023401, 1S10RR019307, and 1S10RR023043. Additional support was provided by the NIH Blueprint for Neuroscience Research (5U01-MH093765), part of the multi-institutional Human Connectome Project. In addition, BF has a financial interest in CorticoMetrics, a company whose medical pursuits focus on brain imaging and measurement technologies. BF’s interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies.

References

[1] B. Fischl, D. H. Salat, et al., Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain, Neuron 33 (2002) 341 – 355.

[2] Y. Ou, H. Akbari, M. Bilello, X. Da, C. Davatzikos, Comparative evaluation of registration algorithms in different brain databases with varying difficulty: Results and insights, IEEE Transactions on Medical Imaging 33 (2014) 2039–2065.
[3] J. P. Mugler III, J. R. Brookeman, Rapid three-dimensional T1-weighted MR imaging with the MP-RAGE sequence, Journal of Magnetic Resonance Imaging 1 (1991) 561–567.

[4] R. W. Cox, AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages, Computers and Biomedical Research 29 (1996) 162 – 173.

[5] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, R. M. Leahy, Magnetic resonance image tissue classification using a partial volume model, NeuroImage 13 (2001) 856 – 876.

[6] S. M. Smith, Fast robust automated brain extraction, Human Brain Mapping 17 (2002) 143–155.

[7] H. K. Hahn, H.-O. Peitgen, The skull stripping problem in MRI solved by a single 3D watershed transform, in: Proceedings of the Third International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI ’00, Springer-Verlag, London, UK, UK, 2000, pp. 134–143.

[8] F. Segonne, A. Dale, E. Busa, M. Glessner, D. Salat, H. Hahn, B. Fischl, A hybrid approach to the skull stripping problem in MRI, NeuroImage 22 (2004) 1060 – 1075.

[9] A. Carass, J. Cuzzocreo, M. B. Wheeler, P.-L. Bazin, S. M. Resnick, J. L. Prince, Simple paradigm for extra-cerebral tissue removal: Algorithm and analysis, NeuroImage 56 (2011) 1982 – 1992.

[10] J. E. Iglesias, C. Liu, P. M. Thompson, Z. Tu, Robust brain extraction across datasets and comparison with publicly available methods, IEEE Transactions on Medical Imaging 30 (2011) 1617–1634.

[11] S. Roy, J. A. Butman, D. L. Pham, Robust skull stripping using multiple MR image contrasts insensitive to pathology, NeuroImage 146 (2017) 132 – 147.

[12] J. Doshi, G. Erus, Y. Ou, B. Gaonkar, C. Davatzikos, Multi-atlas skull-stripping, Academic Radiology 20 (2013) 1566 – 1576.
[13] D. Mahapatra, Skull stripping of neonatal brain MRI: Using prior shape information with graph cuts, Journal of Digital Imaging 25 (2012) 802–814.

[14] A. H. Zhuang, D. J. Valentino, A. W. Toga, Skull-stripping magnetic resonance brain images using a model-based level set, NeuroImage 32 (2006) 79 – 92.

[15] J. Chiverton, K. Wells, E. Lewis, C. Chen, B. Podda, D. Johnson, Statistical morphological skull stripping of adult and infant MRI data, Computers in Biology and Medicine 37 (2007) 342 – 357.

[16] S. F. Eskildsen, P. Coup, V. Fonov, J. V. Manjin, K. K. Leung, N. Guizard, S. N. Wassef, L. R. Stergaard, D. L. Collins, Beast: Brain extraction based on nonlocal segmentation technique, NeuroImage 59 (2012) 2362 – 2373.

[17] Y. Ou, R. L. Gollub, K. Retzepi, N. Reynolds, R. Pienaar, S. Pieper, S. N. Murphy, P. E. Grant, L. Zllei, Brain extraction in pediatric ADC maps, toward characterizing neuro-development in multi-platform and multi-institution clinical images, Neuroimage 122 (2015) 246–61.

[18] A. Alansary, M. Ismail, A. Soliman, F. Khalifa, M. Nitzken, A. Elnakib, M. Mostapha, A. Black, K. Stinebruner, M. F. Casanova, J. M. Zurada, A. El-Baz, Infant brain extraction in t1-weighted mr images using bet and refinement using lcedg and mgrf models, IEEE Journal of Biomedical and Health Informatics 20 (2016) 925–935.

[19] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, R. M. Summers, Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 556–564.

[20] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Medical Image Analysis 36 (2017) 61 – 78.
[21] E. Gibson, W. Li, C. Sudre, L. Fidon, D. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat, D. C. Barratt, S. Ourselin, M. J. Cardoso, T. Vercauteren, Niftynet: a deep-learning platform for medical imaging, volume 158, pp. 113–122.

[22] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendzus, A. Biller, Deep MRI brain extraction: A 3D convolutional neural network for skull stripping, NeuroImage 129 (2016) 460 – 469.

[23] A. J. Bekker, M. Shalhon, H. Greenspan, J. Goldberger, Multi-view probabilistic classification of breast microcalcifications, IEEE Transactions on Medical Imaging 35 (2016) 645–653.

[24] A. G. Roy, S. Conjeti, N. Navab, C. Wachinger, Quicknat: A fully convolutional network for quick and accurate segmentation of neuroanatomy, NeuroImage 186 (2019) 713 – 727.

[25] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 9351 of LNCS, Springer, 2015, pp. 234–241. (available on arXiv:1505.04597 [cs.CV]).

[26] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, PMLR, Lille, France, 2015, pp. 448–456.

[27] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

[28] K. de Macedo Rodrigues, E. Ben-Avi, D. D. Sliva, M.-s. Choe, M. Drottar, R. Wang, B. Fischl, P. E. Grant, L. Zollei, A FreeSurfer-compliant consistent manual segmentation of infant brains spanning the 0–2 year age range, Frontiers in Human Neuroscience 9 (2015) 21.

[29] A. Serag, M. Blesa, E. J. Moore, R. Pataky, S. A. Sparrow, A. G. Wilkinson, G. Macnaught, S. I. Semple, J. P. Boardman, Accurate learning
with few atlases (ALFA): an algorithm for MRI neonatal brain extraction and comparison with 11 publicly available methods, Scientific Reports 6 (2016).

[30] J.-M. Lee, U. Yoon, S. H. Nam, J.-H. Kim, I.-Y. Kim, S. I. Kim, Evaluation of automated and semi-automated skull-stripping algorithms using similarity index and segmentation error, Computers in Biology and Medicine 33 (2003) 495–507.

[31] V. Popescu, M. Battaglini, W. Hoogstrate, S. Verfaillie, I. Sluimer, R. van Schijndel, B. van Dijk, K. Cover, D. Knol, M. Jenkinson, F. Barkhof, N. de Stefano, H. Vrenken, Optimizing parameter choice for FSL-brain extraction tool (BET) on 3D T1 images in multiple sclerosis, NeuroImage 61 (2012) 1484–1494.

[32] F. Shi, L. Wang, Y. Dai, J. H. Gilmore, W. Lin, D. Shen, Label: Pediatric brain extraction using learning-based meta-algorithm, NeuroImage 62 (2012) 1975–1986.

[33] A. Makropoulos, I. S. Gousias, C. Ledig, P. Aljabar, A. Serag, J. V. Hajnal, A. D. Edwards, S. J. Counsell, D. Rueckert, Automatic whole brain MRI segmentation of the developing neonatal brain, IEEE Transactions on Medical Imaging 33 (2014) 1818–1831.

[34] A. Serag, P. Aljabar, G. Ball, S. J. Counsell, J. P. Boardman, M. A. Rutherford, A. D. Edwards, J. V. Hajnal, D. Rueckert, Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression, NeuroImage 59 (2012) 2255–2265.

[35] S. W. Jacobson, J. L. Jacobson, C. D. Molteno, C. M. Warton, P. Wintemark, H. E. Hoyme, G. De Jong, P. Taylor, F. Warton, N. M. Lindinger, R. C. Carter, N. C. Dodge, E. Grant, S. K. Warfield, L. Zollei, A. J. van der Kouwe, E. M. Meintjes, Heavy prenatal alcohol exposure is related to smaller corpus callosum in newborn MRI scans, Alcoholism: Clinical and Experimental Research 41 (2017) 965–975.

[36] A. Makropoulos, E. C. Robinson, A. Schuh, R. Wright, S. Fitzgibbon, J. Bozek, S. J. Counsell, J. Steinweg, K. Vecchiato, J. Passerat-Palmbach, G. Lenz, F. Mortari, T. Tenev, E. P. Duff, M. Bastiani,
L. Cordero-Grande, E. Hughes, N. Tusor, J.-D. Tournier, J. Hutter, A. N. Price, R. P. A. Teixeira, M. Murgasova, S. Victor, C. Kelly, M. A. Rutherford, S. M. Smith, A. D. Edwards, J. V. Hajnal, M. Jenkinson, D. Rueckert, The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction, NeuroImage 173 (2018) 88 – 112.

[37] S. K. Jensen, S. Kumar, W. Xie, F. Tofail, R. Haque, W. A. Petri, C. A. Nelson, Neural correlates of early adversity among bangladeshi infants, Scientific reports 9 (2019) 3507.

[38] C. Storrs, How poverty affects the brain, 2017. [Online; posted 12-July-2017].

[39] S. K. Warfield, K. H. Zou, W. M. Wells, Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation, IEEE Transactions on Medical Imaging 23 (2004) 903–921.

[40] A. Jog, A. Hoopes, D. N. Greve, K. Van Leemput, B. Fischl, PSACNN: Pulse sequence adaptive fast whole brain segmentation, arXiv preprint arXiv:1901.05992 (2019).