Received 16 September 2023, accepted 14 October 2023, date of publication 20 October 2023, date of current version 26 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3326342

RESEARCH ARTICLE

PCSS: Skull Stripping With Posture Correction From 3D Brain MRI for Diverse Imaging Environment

KEI NISHIMAKI[1], KUMPEI IKUTA[1], SHINGO FUJIYAMA[1], KENICHI OISHI[2], AND HITOSHI IYATOMI[1], (Member, IEEE), for the Alzheimer's Disease Neuroimaging Initiative

[1]Department of Applied Informatics, Graduate School of Science and Engineering, Hosei University, Koganei, Tokyo 184-8584, Japan
[2]Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

Corresponding author: Kei Nishimaki (kei.nishimaki.1106@gmail.com)

This work was supported in part by the Ministry of Education, Science, Sports and Culture of Japan (JSPS KAKENHI), Grant-in-Aid for Scientific Research (C) (2021–2024), under Grant 21K12656.

ABSTRACT A subject's head position in magnetic resonance imaging (MRI) scanners can vary significantly with the imaging environment and disease status. This variation is known to influence the accuracy of skull stripping (SS), a method to extract the brain region from the whole head image, which is an essential initial step to attain high performance in various neuroimaging applications. However, existing SS methods have failed to accommodate this wide range of variation. To achieve accurate, consistent, and fast SS, we introduce a novel two-stage methodology that we call posture correction skull stripping (PCSS): the first involves adjusting the subject's head angle and position, and the second involves the actual SS to generate the brain mask. PCSS also incorporates various machine learning techniques, such as a weighted loss function, adversarial training from generative adversarial networks, and ensemble methods. Thorough evaluations conducted on five publicly accessible datasets show that the PCSS method outperforms current state-of-the-art techniques in SS performance, achieving an average increase of 1.38 points on the Dice score and demonstrating the contributions of each PCSS component technique.

INDEX TERMS Skull stripping, brain extraction, MRI, U-Net, GAN, ADNI, CC-12, LPBA40, NFBS, OASIS.

I. INTRODUCTION

Magnetic resonance imaging (MRI) is commonly used to diagnose various neurological diseases due to its superior ability to provide detailed images of the brain's anatomy. It offers excellent spatial and contrast resolution without exposing the patient to radiation. In daily clinical settings, radiologists interpret MRIs qualitatively, providing reports on disease-related findings visible in the images. On the other hand, numerous efforts have been made to derive quantitative measures from brain MRIs to characterize any disease-related changes in the brain and to understand its physiological status in relation to development, aging, and sex differences. Deep learning methodologies are frequently used in this regard, offering diagnostic labels for various conditions, including Alzheimer's disease [1], [2], [3], detecting tumors [4], [5], [6], and facilitating image-based searches for similar cases [7], [8], [9].

Such studies using MRI often require a preprocessing step known as skull stripping (SS). This process involves isolating the brain parenchyma from the whole head MRI by eliminating non-brain tissues such as the skull, skin, fat, and eyeballs. Manually extracting brain parenchyma from 3D MRIs is an extremely labor-intensive task. As a result, a variety of automated SS methods have been proposed to simplify this process.

The classical SS methods proposed in the first decade of this century [10], [11], [12] are expected to perform well

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

for parameter-optimized datasets. However, they have been reported to be less robust to changes in local anatomy and the type of disease under study and to perform significantly less accurately for different datasets [13]. Since the beginning of the 2010s, parameter-robust and versatile methods [14], [15], [16], [17], [18] and several open source software packages [19], [20] that can provide different kinds of image analysis, such as SS and anatomical segmentation, have been proposed. While these approaches achieve relatively high SS performance, they are often associated with long processing times.

On the other hand, several high-performance, rapid SS have recently been proposed as deep learning technology advances. U-Net [21], a symmetric arrangement of convolutional neural networks (CNNs) with a bypass structure corresponding to the same resolution, achieves accurate SS without requiring prior anatomical expertise of region extraction [22], [23], [24], [25], [26], [27], [28], [29], [30]. These methods can be categorized into two groups: vertically stack 2D SS results in each slice of the brain MR image [22], [23], [24], [25], and extend the U-Net architecture to three dimensions [26], [27], [28], [29], [30]. Although these SS methods based on deep learning show excellent performance, there are still significant concerns regarding their robustness across datasets. In many previous studies, the images used for training and evaluation were obtained from the same facility or datasets [22], [25], [26], [29]. Images taken within the same dataset or facility generally share common characteristics such as scanner, imaging protocol, subject posture, and other conditions. Therefore, training and evaluating a model using a single dataset is easier than evaluation using completely unknown cases. Robust performance for unknown environments is important in analyzing large-scale data that are collected across multiple sites or from multiple datasets. However, few studies have systematically investigated the influence of implicit differences in characteristics between datasets, which can be caused by variations in imaging environments, on SS performance.

One of the significant challenges in achieving robust SS is dealing with variability of the subject posture, which can differ significantly depending on different imaging environments and disease status between datasets. This variability can lead to geometric differences between the training and test data, making accurate extraction of brain structure more difficult, even with deep learning-based SS methods [23], [24]. Despite this known issue, no SS methodologies have been proposed that adequately account for the diversity of subject postures.

In this paper, we propose posture correction skull stripping (PCSS), a highly accurate and robust SS method that takes into account the diversity in subject postures. PCSS is a framework based on U-Net with the following four extensions: (i) preprocessing to estimate and correct the angle and position of the subject's head to suppress posture variation; (ii) weighted loss function, which considers the imbalance between the brain region and other tissues [22]; (iii) a discriminator network for adversarial training introduced in generative adversarial networks (GANs) [31] and used in an SS study [24]; and (iv) ensemble of three-way segmentation for the brain [29]. For a rigorous evaluation across multiple datasets, we use five T1-weighted brain MRI public datasets (ADNI, CC-12, LPBA40, NFBS, and OASIS) and discuss the impact of different datasets on SS performance, which has not been addressed previously. In addition to the effects of (i) posture correction, which is the main proposal in this paper, each of the technical elements introduced in (ii), (iii), and (iv) are also evaluated to discuss the key techniques involved in achieving a robust SS.

The main contributions in this paper are as follows:

- Proposal for posture estimation of subjects (head angle and position) and a connected correction method for constructing highly robust SS.
- Clarification of the elemental techniques required to realize accurate and robust SS based on appropriate evaluations.
- Proposal of a practical SS method with high speed (8.07 sec/case) and high accuracy (Dice score = 96.95) based on these effective elemental technologies.

The code and results are published at URL: https://github.com/IyatomiLab/Posture-Correction-Skull-Stripping.

II. RELATED WORK

Among the open source software employed for automated processing of MRIs, 3dSkullStrip is provided as a component of the Analysis of Functional NeuroImages (AFNI) [32][1] package. It uses a modified version of the Brain Extraction Tool (BET) [11]. FreeSurfer [19][2] is an open source software package employed for automated processing of MRIs, and includes SS among its functionalities. Within this package, the Hybrid Watershed Approach (HWA) [12] is used for SS; its efficacy has been evaluated with extensive datasets [33]. However, the process typically takes several hours to complete on a standard desktop computer. MRICloud [20],[3] which is currently recognized as one of the most efficient SS methods, generates the segmentation mask using a technique known as multi-atlas label fusion and arbitration algorithms. However, the segmentation and SS process for a single case typically requires approximately one hour to execute.

Salehi et al. [22] proposed a network architecture called Auto-Net, which introduced an auto-context CNN for SS of 3D brain MRIs. Their proposed method employs an auto-context CNN for this task. Auto-Net, based on the U-Net [21] framework, incorporates 26 convolutional layers. Their proposed method can implicitly train 3D images without using computationally expensive 3D convolution. It performs SS on each cross section and stacks them vertically to achieve SS for the entire brain MRI. The auto-context CNN in Auto-Net uses the auto-context algorithm [34].

[1] https://afni.nimh.nih.gov/

[2] https://surfer.nmr.mgh.harvard.edu/

[3] https://mricloud.org/

In this approach, the posterior probabilities obtained from the first training's segmentation are incorporated as contextual information through concatenation in the channel direction of the second training. The same thing occurs during evaluation. Salehi et al. achieved good results using Auto-Net with Dice scores of 97.73 and 97.62 on the LPBA40 and OASIS datasets, respectively. However, there is a possibility of overfitting due to the use of the same dataset for both training and evaluation. In addition, those authors did not separately evaluate the impact of the weighted loss function (cross entropy) on the ratio of non-brain region to brain region that they are introducing.

Jiang et al. [24] proposed SS using Wasserstein GAN (WGAN) [35] in conjunction with O-Net, which incorporates an attention module into the U-Net, establishing a new shortcut for the corresponding mapping between encoders and decoders. This approach effectively preserves detailed image features while leveraging deep semantic information to emphasize target regions for each channel. WGAN improves SS performance based on GAN by introducing the discrimination of SS results generated by O-Net. WGAN+O-Net, trained on the LPBA40 dataset, achieved a Dice score of 95.51 on the IBSR18 dataset. However, it was acknowledged that the IBSR18 dataset was of low quality and heavy artifact, and the evaluation was not performed on a variety of high-quality datasets. In addition, only 18 images were evaluated.

Fatima et al. [25] proposed MVU-Net, which performs separate SS on the coronal, sagittal, and transverse sections of input 3D MRIs and generates an ensemble probability map of the brain region. This approach reduces the ambiguity of SS that uses only single section. The architectural design of MVU-Net was inspired by U-Net and SCU-Net [36]. Its structure enables efficient training with a relatively small number of parameters (1.4 million). MVU-Net performed well on the NFBS and IBSR datasets, achieving Dice scores of 96.81 and 91.84, respectively. However, as in Salehi et al. there exists a potential risk of overfitting due to the use of the same datasets for both training and evaluation.

Fabian et al. [28] proposed nnU-Net, a deep learning-based segmentation method that automatically performs preprocessing, network architecture, training, and post-processing. nnU-Net extracts three types of parameters: —fixed, rule-based, and empirical— to construct model training. Fixed parameters include network architecture and training plan, rule-based parameters include normalization and resampling, and empirical parameters include ensemble and post-processing. Given a new segmentation task, nnU-Net determines these parameters and builds a pipeline connecting them, allowing users to generate segmentation models easily without domain knowledge. Theirs is one of the few papers to conduct a rigorous evaluation using 23 datasets of biomedical image segmentation and show robust and high performance. Therefore, we apply nnU-Net to SS and use it as a comparison in this paper.

The SS method that Isensee et al. [28] proposed has shown a certain effectiveness under specific circumstances. However, for the practical assessment in SS performance, it is important to replicate and evaluate diverse imaging environments using multiple datasets, just as those authors did. The impacts of the weighted loss function in Salehi et al. [22], training introducing a discriminator in Jiang et al. [24], and ensembling techniques in Fatima et al. [25] must be evaluated equally under practical conditions. In this paper, we also evaluate the effects of these techniques on SS performance.

Isensee et al. [27] developed HD-BET, an accurate skull stripping technique, using a large EORTC-26101 dataset collected from 37 institutions. HD-BET performs skull stripping on a U-Net using detailed GTs obtained by manual correction based on the BET [11]. HD-BET outperforms the previous six SS methods in a rigorous evaluation where the evaluation data ($3,419$ images from 12 locations) are obtained from different locations than the training data ($6,586$ images from 25 locations) and on untrained CC359, LPBA40, and NFBS datasets. In addition, HD-BET outperformed the competition not only on T1-weighted images but also on processing different scan sequences such as contrast-enhanced T1-weighted, T2-weighted, and FLAIR. In our experiments, HD-BET was compared to our proposal and other comparative methods as a reference, although the number of data used for training is more than ten times different.

III. POSTURE CORRECTION SKULL STRIPPING

In this paper, we propose posture correction skull stripping (PCSS), a SS method that is designed to be robust to variations in the position and angle of the subject's head across different datasets. Figure 1 shows the overview of PCSS, which consists of two phases: (1) posture correction phase and (2) skull stripping (SS) phase. In the posture correction phase, a posture estimation network (PENet) consisting of CNNs is used to estimate and correct the head posture in MRIs. In the SS phase, the skull stripping network (SSNet), based on UNet, which has been widely used for segmentation tasks involving deep learning in recent years, is used to extract actual brain regions from original images.

The main contribution of this paper is to propose a robust and accurate SS method that includes posture correction. In addition, we compare and evaluate several technical elements used in the construction of SSNet using various datasets to provide guidance on the effective model structure for SS.

A. POSTURE CORRECTION PHASE

MRIs may also vary due to differences in the subject's head position at the time of imaging, which can significantly reduce SS performance. Therefore, in the posture correction phase, the position and angle of the head are estimated and corrected to prevent SS performance degradation. Figure 2 shows an overview of the process in this phase.

The variation in head position is pronounced in the pitch direction ($\tilde{\theta}$ in the sagittal cross section figure). Therefore,

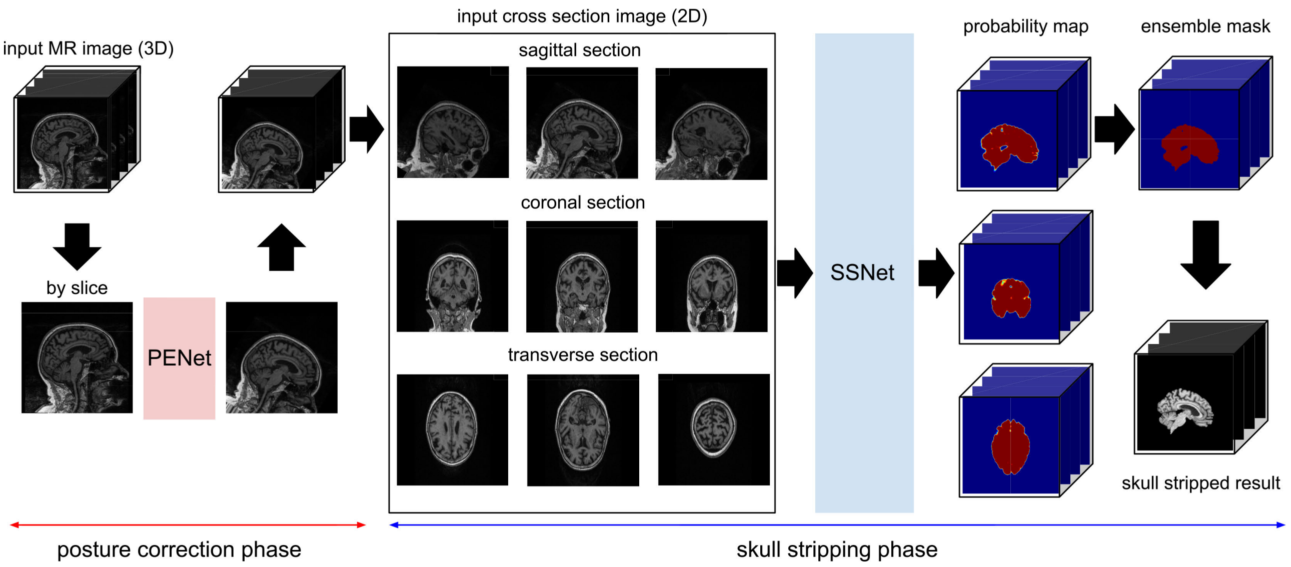


FIGURE 1. Overview of the posture correction skull stripping (PCSS).

we introduced a reference neck line (red line in Figure 2) in the sagittal cross section from the base of the nose, called the nasion, through the lowest point of the brain region, called the basion, and an alignment neck line (light blue line in Figure 2) as a horizontal line 24% of the way up from the bottom of the image. The nasion and basion were selected as references due to their clear visibility in the sagittal sections of the MRIs. Therefore, annotation of the reference neck line has high reproducibility. Since the reference neck line effectively works as a boundary between the soft tissues of the head and the neck, it provides a standard for correction for various head angles and positions. In the posture correction phase, the reference neck line of all cases is aligned with the alignment neck line to correct for variations in head position.

The reference neck line is represented by $y = ax + b$ in the Cartesian coordinate system (x, y) with the y coordinate at the top of the head. The tilt a corresponds to the angle $\tilde{\theta}$ of the subject's head, and the intercept b is the point at which this line intersects with the vertical axis of the images after passing through the neck area, which serves as the reference value for position adjustment.

First, the posture correction phase estimates the tilt a and the intercept b of the reference neck line in the central slice with the largest head cross sectional area in the sagittal cross section in two dimensions using a PENet consisting of a CNN and fully connected layers. Next, the original brain MRIs are rotated in three dimensions so that the reference neck line is horizontal (parallel to the transverse section) using the head angle $\tilde{\theta}$ calculated from the tilt a. Then, the images are shifted up or down so that the value of the intercept $\tilde{b}$, as modified by the rotation, is equal to the position of the alignment line b^* (i.e., 24% of the way up from the bottom of the image). These processes result in a similar posture of the subject's head across all cases.

B. SKULL STRIPPING PHASE

The 3D brain MRIs corrected by the posture correction phase are extracted by SSNet in the SS phase. Figure 3 shows an overview of the SS phase, including SSNet, which is based on U-Net and introduces weighted cross entropy as a loss function to account for the volume ratio between brain and non-brain regions [22] (weighted loss function), improves SS results by introducing the discriminator networks used in GANs [24] (adversarial training), and an ensemble of three-sections segmentation of the brain [25] (ensemble). The details of this technical component are described in III-C.

Medical images such as brain MRIs are difficult to collect large amounts of data due to privacy issues and acquisition costs. Moreover, the cost of annotating these images is very high, making it challenging to prepare a sufficient amount of 3D training data. In addition, SSNet, like many other SS methods, performs 2D SS on any cross section of a 3D brain MRI and vertically stacks them (and, if necessary, ensembles the SS results for each section) to achieve the final SS.

In the case of general T1-weighted brain MRIs of 256 × 256 × 256, the 2D SS model can use 256 images per case in one section and a total of 768 images in three sections for training. In addition, the 2D model has fewer model parameters than the 3D model, which can be expected to reduce the risk of overfitting.

SSNet takes a 2D cross sectional image x_i obtained from any section from the original 3D MRIs x as input for training and estimates the probability map $p(x_i)$ that each pixel in x_i is a brain region. The probability map $p(x_i)$ is stacked vertically to generate a probability map for entire brain regions, and regions with 50% or more are output as the final SS result. The training of SSNet updates the parameters to reduce the binary cross entropy, which is the reconstruction error with the mask image $m(y_i)$ of the gold standard corresponding to x_i.

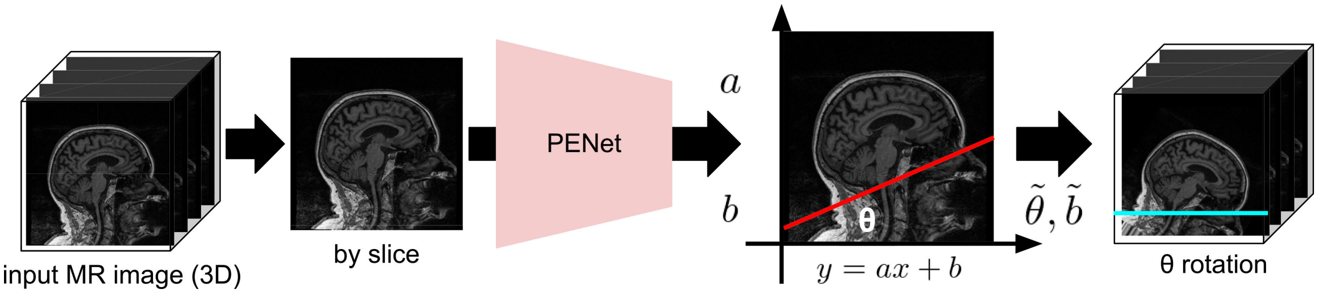


FIGURE 2. Overview of the posture correction phase.

C. ASSESSMENT OF EFFECTIVE TECHNIQUES FOR SKULL STRIPPING

In addition to posture correction, the proposed PCSS introduces three techniques (weighted loss function, adversarial training, and ensemble) used with U-Net in recent years to achieve accurate, robust, and practical SS. PCSS is available in two architectures, PCSS-1 and PCSS-3, depending on the objective. PCSS-1 is an SS model that includes two machine learning technical elements (weighted loss function and adversarial training) in addition to posture correction, and high-speed inference is expected because SS can obtained by estimating only one section. PCSS-3 is a SS model that includes all elements (posture correction, weighted loss function, adversarial training, and ensemble), and since it provides an ensemble of SS results in three directions during SS execution, higher accuracy can be expected, but that occurs at the expense of longer execution time. Each of the machine learning technical elements used in PCSS is described in detail below. In this paper, we evaluate and discuss the effectiveness of these techniques.

1) WEIGHTED LOSS FUNCTION

Since brain regions account for only about 10% of the volume of many original brain MRIs, the segmentation task can be viewed as an imbalanced classification problem. However, no paper has focused on this imbalance and discussed and evaluated it. Therefore, the SS performance in training using weighted cross entropy, which is a loss function commonly used for imbalanced data, and ordinary binary cross entropy was compared.

$\mathcal{L}_{\text{ss}}$ is the loss function of the reconstruction error based on weighted cross entropy (binary cross entropy), and α is the hyperparameter that addresses imbalance:

$$\begin{aligned}\mathcal{L}_{\text{ss}} = -\mathbb{E}_{x_i \sim x}[\alpha \cdot m(y_i) \cdot \log p(x_i) \\ + (1 - m(y_i)) \cdot \log(1 - p(x_i))].\end{aligned} \tag{1}$$

The first and second terms in the expectation clause indicate the reconstruction loss (i.e., the difference between the predicted SS result and its GT) for the brain region and non-brain region, respectively. The more correct the prediction probability $p(x_i)$ for a brain region, the smaller its value.

The value of α in normal binary cross entropy is 1. When the weighted cross entropy was used, the value of α was defined as the ratio of the volume of non-brain regions to the volume of brain regions in the entire dataset of all brain MRIs used for training, according to the general weighting method; α is determined by a unique value at the start of the training. We understand that there may be more optimal values for α, but our evaluation focuses solely on the impact of addressing imbalance on SS performance. Tuning the hyperparameter is outside the scope of this paper.

2) ADVERSARIAL TRAINING

Recently, models such as pix2pix [37], which applies the adversarial training introduced by GAN [31], have been proposed and reported to improve the performance of U-network-based segmentation techniques. This approach improves the performance of the original model (generator; G) by adding another model (discriminator; D) that determines the validity of the results generated by G and by training G and D in an adversarial manner. Specifically, in the SS task using U-Net, the U-Net becomes G, and the CNN model added to verify its validity becomes D. In this paper, the improvement in SS performance by introducing this adversarial learning into the SS task is evaluated; that is, by adding a discriminator network.

The loss function for the model due to the addition of the discriminator is obtained by a reconstruction error $\mathcal{L}_{\text{ss}}$ from the original U-Net plus the adversarial loss $\mathcal{L}_{\text{Adv}}$ of the generator and discriminator

$$\mathcal{L} = \mathcal{L}_{\text{ss}} + \lambda \mathcal{L}_{\text{Adv}}, \tag{2}$$

where λ is a hyperparameter, and $\mathcal{L}_{\text{Adv}}$ is the adversarial loss used in a general GAN and is expressed by the following equation:

$$\begin{aligned}\mathcal{L}_{\text{Adv}} = \mathbb{E}_{x_i \sim x, y_i \sim y}[\log(D(x_i, y_i) \\ + \log(1 - D(x_i, G(x_i)))].\end{aligned} \tag{3}$$

The first term is defined as the cost for the identification of the GT image, and the second term is the cost for the identification of the generated image. The G is a U-Net that performs SS from the input image x_i, and the D discriminates between the segmented $G(x_i)$ and the true segmentation result y_i given as the GT, and updates D and G parameters. By repeating this process, it is possible to generate a probability map that is accurate enough to deceive

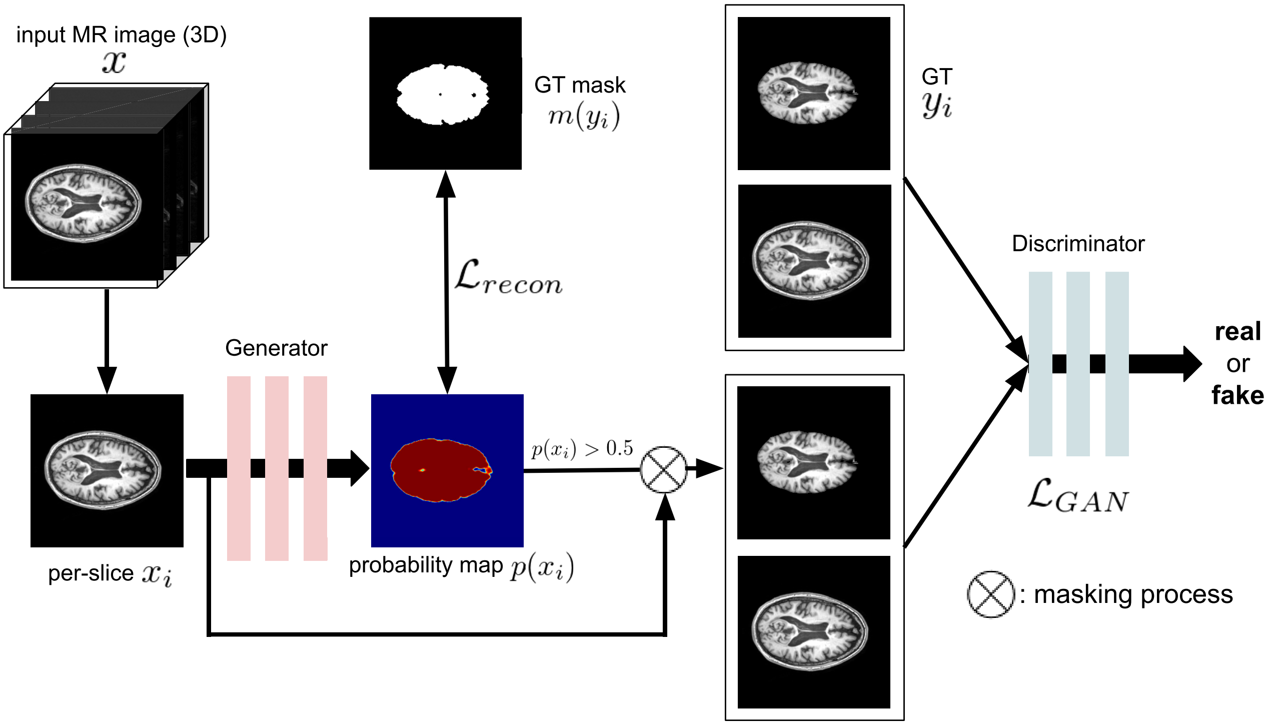


FIGURE 3. Overview of the skull stripping phase.

the trained *D* inside the trained *G* and to generate an image from which only brain regions are extracted.

3) ENSEMBLE OF EACH SECTION

Ensemble is a fundamental element of machine learning technology that improves performance by using multiple independent weak learners. The SSNet in this experiment obtained final SS results by stacking 2D SS results. We investigate the effect of the ensemble of SS results from different 2D sections on the final SS capability. Specifically, SS results were compared between stacks of only transverse sections and ensemble of three sections (i.e., the average of the brain probability maps). In the brain probability maps obtained for both approaches, regions with a probability of 50% or greater were extracted as the brain.

The ensemble typically combines the prediction results of separately trained models. However, preliminary experimental results showed no significant difference between the ensemble model trained with cross sections separately and the model trained with all three cross sections together. The ensemble adopted the latter implementation because it is undesirable in terms of computational resources and execution speed to invoke the three different models individually at runtime.

IV. DATASET

Table 1 shows an overview of the datasets used in this paper. The five public 3D brain MRI datasets used in this study are the Alzheimer's disease neuroimaging initiative 2 (ADNI2) dataset [38], the Calgary-campinas-12 (CC-12) dataset [39], the LONI probabilistic brain atlas (LPBA40) dataset [40], the Neurofeedback skull-stripped (NFBS) dataset [41], and the Disc-1 and Disk-2 of Open access series of imaging studies (OASIS) dataset [42]. Each image was acquired in NIfTI format.

For the training of the proposed PCSS (PENet and SSNet), images from ADNI2, which has the largest amount of recorded data and is a practical dataset with a wide range of head tilt, size, intensity, and other factors, were used. For the evaluation of SS performance, images from ADNI2 (data folded out in the 5-fold cross validation) were used. The remaining four datasets were used for evaluation.

In addition, only one case per patient was used for all datasets. Data from the same patient often have similar characteristics, and their inclusion may introduce bias in training and evaluation, affecting the validity and generalizability of the results. Moreover, if they are included in both the training and evaluation data, the model will overfit, resulting in higher values than the actual performance and incorrect evaluations. Therefore, in this experiment, they were eliminated to ensure rigorous evaluation.

Because ADNI2 does not have a ground truth (GT) of the SSed brain regions, the SS results were used from MRICloud [20], currently considered one of the most accurate SS methods, recognizing that it may not always be accurate in some cases. The evaluation of the proposed method is described below, but we have ensured the validity of the proposal by evaluating it with the CC-12, LPBA40, and NFBS datasets, which were assigned a GT of manual.

The GTs provided in the LPBA40 and NFBS datasets consist of manually edited masks based on automatic SS results from the BET [11] and brain extraction using nonlocal

TABLE 1. Dataset used in our study.

dataset	number of cases	size of image	GT method
ADNI2	612	(176, 240, 256)	MRICloud [20]
CC-12	12	(150, 288, 288) (170, 288, 288) (180, 240, 240) (180, 224, 224) (192, 256, 256) (224, 256, 256) (256, 196, 256) (200, 256, 256)	manual
LPBA40	40	(256, 124, 256) (256, 120, 256)	BET [11] + manual
NFBS	125	(192, 256, 256)	BEaST [16] + manual
OASIS	76	(128, 256, 256)	FreeSurfer [19]

segmentation technique (BEaST) [16], respectively. The GT provided in the OASIS dataset is obtained by FreeSurfer (HWA) [19] rather than manual results. Although this result may contain a certain percentage of inaccuracies, it was used as a reference result since it has been used in many other studies for performance evaluation.

V. EXPERIMENT

In this experiment, we quantitatively evaluated the performance of the posture correction proposed to achieve an accurate and robust SS. To verify the effectiveness of the proposed PCSS and various SS techniques, we compared SS performance with previously reported methods using publicly available datasets (ADNI [38], CC-12 [39], LPBA40 [40], NFBS [41], and OASIS [42]).

A. PREPROCESSING

The datasets with a resolution of approximately 1 mm × 1 mm × 1 mm (ADNI2, CC-12, NFBS, and OASIS) were preprocessed by zero-padding and standardizing the image size to 256 × 256 × 256. For the LPBA40 dataset, 38 of the 40 images are 0.86 mm × 1.5 mm × 0.86 mm, and 2 are 0.78 mm × 1.5 mm × 0.78 mm. The resolution in the coronal section is lower than in the other two directions. Therefore, spline interpolation was used to align the 38 images with a resolution of 0.86 *mm* and the remaining 2 images with a resolution of 0.78 *mm* in the higher resolution direction. Then, the size was changed to 256 × 256 × 256 by zero padding. Outliers in each image were considered, and pixels with pixel values less than 0 or greater than four times the standard deviation were excluded as outliers and linearly normalized to a range of −1 to 1. The removed pixels were replaced with the minimum and maximum values after normalization.

B. DETAILS OF PENet AND ITS EVALUATION

The PENet that performs the posture correction consists of three convolution layers and two fully connected layers. The details of the PENet configuration are described in Figure 4.

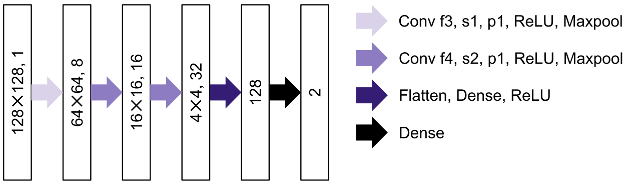


FIGURE 4. Architecture of PENet.

Note that the network models presented in this paper are not limited to the configuration shown. The PENet took as input the slice of the sagittal section with the largest number of pixels in the non-zero brightness (non-background). These slices were also extracted from the center of the matrix (i.e., the 128th slice) to 15% on each side (i.e., 38 slices each) to eliminate the possibility that the detected slices would be outside the brain region. The selected slices were reduced to 128 × 128 by bi-cubic interpolation, and the reference neck line was predicted. The PENet outputs two variables (a, b), the tilt and position values of the reference neck line, which is the reference for the posture correction.

To evaluate the performance of PENet, the mean absolute error (MAE) was calculated between the estimated angle $\tilde{\theta}$ obtained by tilt a and its GT θ^* and intercept $\tilde{b}$ after rotation and intercept $\tilde{b}^*$ of the alignment line. The performance of PENet in estimating head angle and position was evaluated by 5-fold cross validation from the ADNI2 dataset. Note that 10% of the training data was used as validation data.

In some MRIs, the facial region may be blacked out for patient privacy reasons, or the additional regions below the head may have already been removed. In order to achieve robust posture correction even for such images as data augmentation, the cutOut approach [43] was used between 0.6% and 25% of the whole image, and images that had already had the skull stripped by MRICloud with a probability of 20% (i.e., 60% were normal images). In addition, a random rotation of −30 degrees to +30 degrees and a random shift of −20 pixels to +20 pixels were added as data augmentation with a probability of 70% to account for the imaging environment. The early stopping was used to avoid overfitting of the PENet. It ends training when the loss of validation data can be considered as virtually no updates for 100 epochs. The PENet was trained on an RTX3090 with the Adam optimizer, using a learning rate of 5×10^{-4} and a batch size of 32.

To evaluate the robustness of PENet, evaluations were also performed when the amount of training data was reduced to 1/2, 1/4, 1/8, and 1/16. In this case, the remaining images not used for training were evaluated as test data for the ADNI2 evaluation.

C. DETAILS OF SSNet AND ITS EVALUATION

The U-Net based on SSNet consists of 16 layers of encoder and decoder CNNs; the detailed structure of the U-Net is shown in Figure 5. Only the ADNI2 dataset was evaluated for SS by 5-fold cross validation, while all images from the

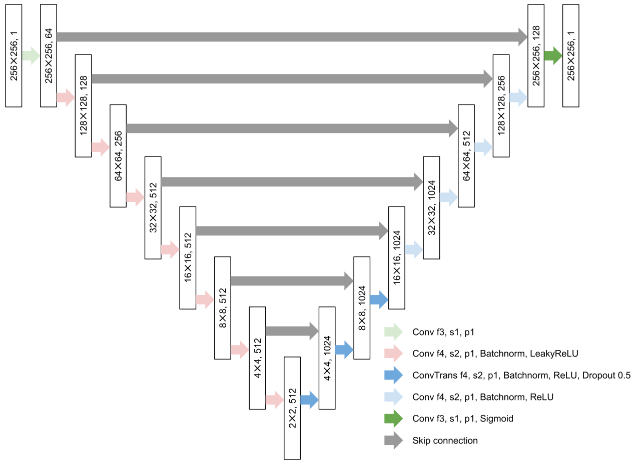


FIGURE 5. Architecture of the baseline of SSNet (U-Net).

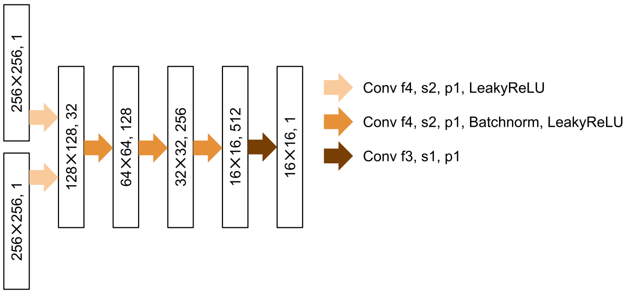


FIGURE 6. Architecture of the discriminator.

TABLE 2. Technical component of skull-stripping.

	Posture correction	Weighted cross entropy	Discriminator	Ensemble
U-Net				
+PC	✓			
+W		✓		
+Adv			✓	
+Ens				✓
PCSS-1	✓	✓	✓	
PCSS-3	✓	✓	✓	✓

ADNI2 dataset were used as training data for the evaluation of the other datasets (CC-12, LPBA40, NFBS, and OASIS).

When training the model (i.e., training on the transverse cross sections), online data augmentation was added by randomly rotating the image from −20 degrees to +20 degrees and shifting it by −20 pixels to +20 pixels to mimic the variation in the subject's posture during actual imaging. The same augmentation was applied in each cross section when training the ensemble model. For the sagittal cross sections, however, a wider range of random rotation of −30 to +30 degrees was applied, as it was assumed that the range of motion for the posture at the time of imaging was wider.

A five-layer fully convolutional network (FCN) was used as the discriminator to introduce GANs training in SS. The detailed structure of the discriminator is shown in Figure 6.

D. EVALUATION OF SKULL STRIPPING PERFORMANCE

In order to discuss the effectiveness of the proposed PCSS (PCSS-1 and PCSS-3), we compared and evaluated the SS performance for the five datasets described above with Auto-Net [22], MVU-Net [25], and HD-BET [27], the three state-of-the-art studies that have reported the best SS performance. In addition, to compare with 3D U-Net, nnU-Net [28] architecture was created by 3D U-Net. Auto-Net and nnU-Net were reproduced using the author's public implementation, and MVU-Net was reproduced to the best of our ability based on the original paper. In addition, HD-BET, which uses 6, 586 training images (about ten times more than PCSS) from 25 sites as a publicly trained model, was included in the comparison. The GT of this model was manually modified based on BET [11] and is not publicly available. Therefore, we cannot compare its performance under fair conditions using the same training images, but we included it for reference. Note that the ADNI2 result from HD-BET is not 5-fold cross validation.

Table 2 shows the technical elements summary of the PCSS. To evaluate the technical components of the proposed PCSS, U-Net + posture correction (+PC), U-Net + weighted loss function (+W), U-Net + discriminator (+Adv), and U-Net + ensemble (+Ens) were evaluated (i.e., using an ablation study.). In order to fix the experimental conditions, U-Net, +PC, +W, +Adv, and +Ens were given the same data augmentation as PCSS.

Recall, precision, and their harmonic score (the Dice score) were used as evaluation metrics. To quantitatively evaluate the number of SS failures, the number of cases with a Dice score less than 0.95 was tabulated.

VI. RESULTS

A. RESULT OF POSTURE CORRECTION

Figure 7 shows four examples of posture correction results for the ADNI2 test case. The top row compares the predicted reference neckline for the input image (red dashed line) with the GT displayed as a reference (yellow line), and the bottom row shows the results with the image rotated and shifted to the position of the alignment neck line (light blue line) from the predicted reference neck line; that is, the final posture correction result.

In the ADNI2 dataset, the GT measurements of the reference neck line had a standard deviation of 7.81 degrees in angle and 14.69 pixels (equivalent to 29 mm in physical size) in vertical position. These findings indicate the presence of such variations in head posture in the ADNI2 datasets.

The proposed PENet had an accurate estimation for the reference neck line, with an average error of 3.61 ± 2.72 degrees in the head angle and 6.42 ± 5.17 pixels (equivalent to 13 ± 10 mm in physical size) in the vertical direction, based on the average of 5-fold cross validation. As a result, the variation in the angle and position of the head was significantly reduced. The rightmost image in Figure 7 shows an example of large estimation error compared with three other examples.

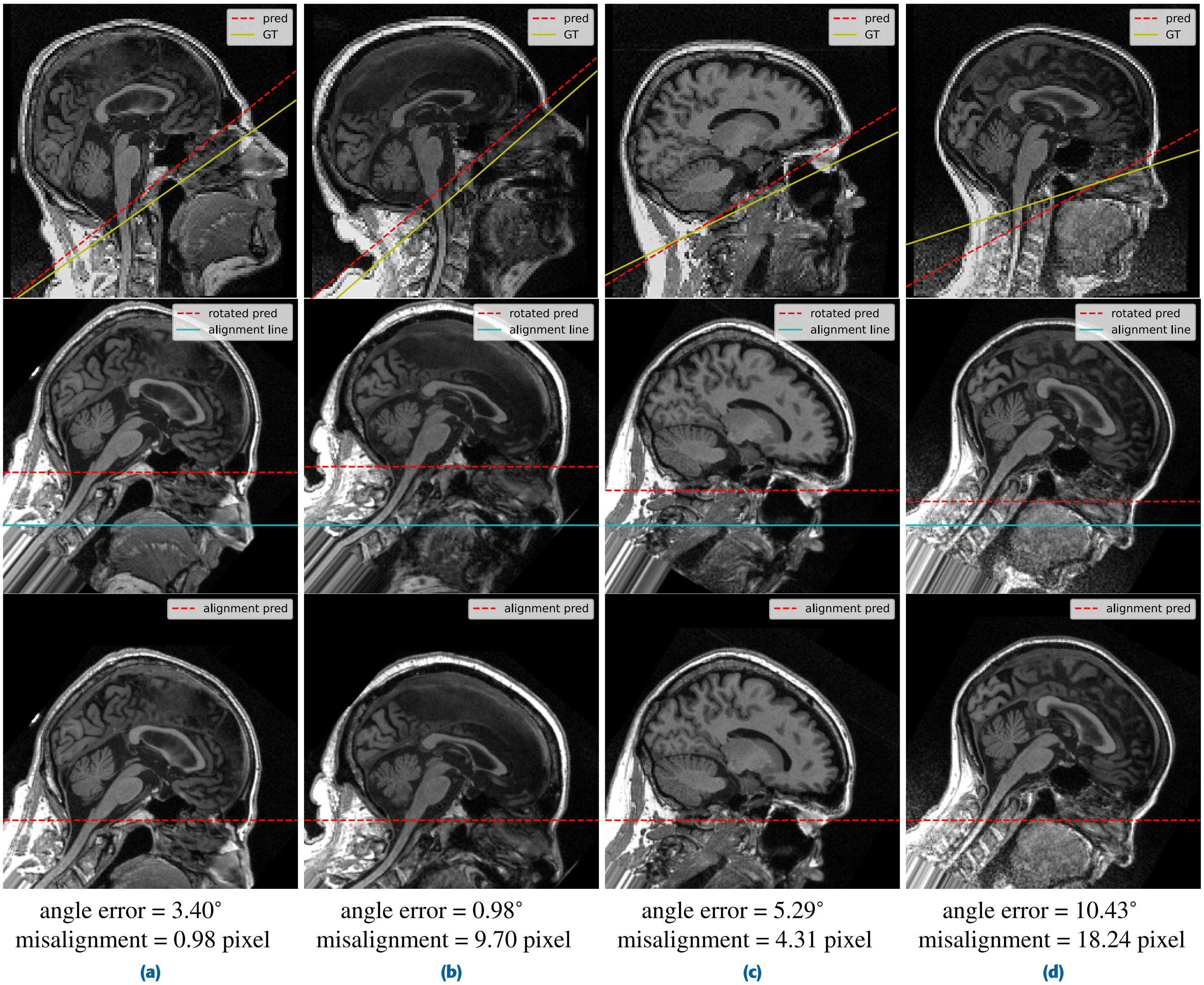


FIGURE 7. **Example of posture correction for ADNI2.**

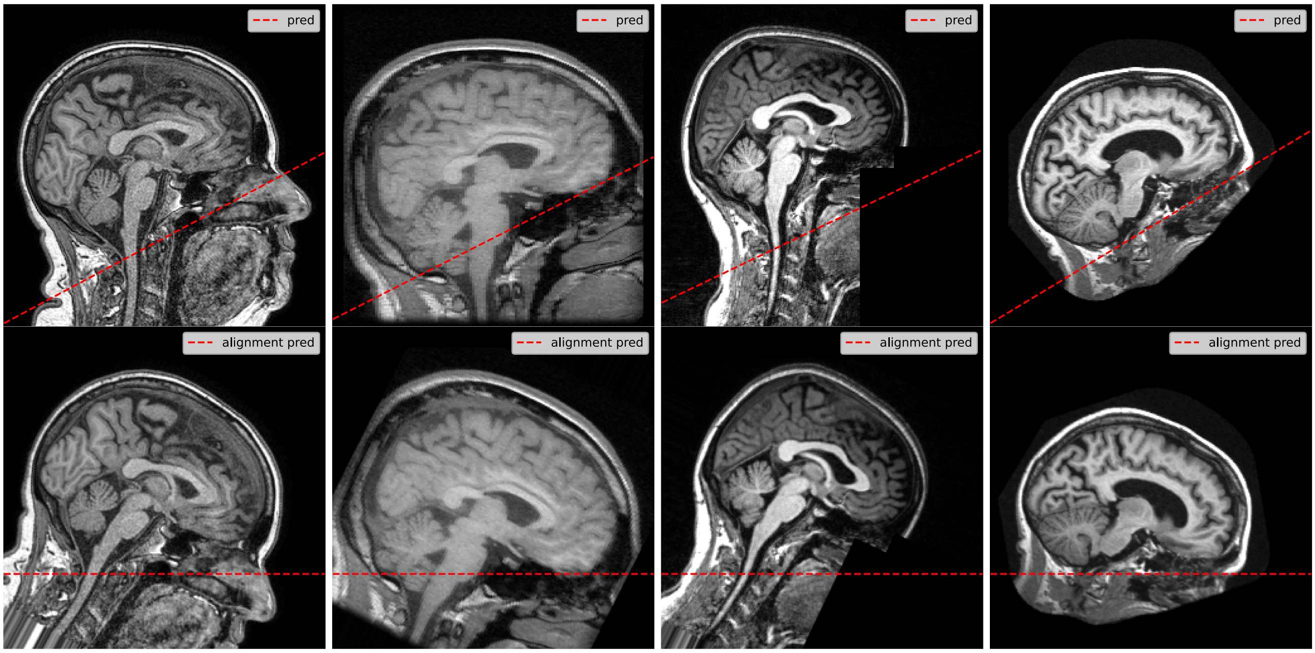
Figure 8 shows the example result of a posture correction for the CC-12, LPBA40, NFBS, and OASIS datasets in panels (a) to (d), respectively. Although a quantitative evaluation is not available for these datasets, it was visually confirmed that the posture correction was performed appropriately, even for the datasets that were not used for training. In addition, this posture correction improved the SS performance in the later stages for all datasets, indirectly suggesting that this posture correction was also successful. Details are discussed below.

Figure 9 shows the relationship between the amount of training data for the PENet (612 cases in total) and the posture correction error (red, angle error; blue, misalignment). This result shows that PENet can accurately estimate posture even when trained with only about 150 cases (i.e., 1/4 in Figure 9) and that its performance gradually deteriorates with further reductions.

B. RESULT OF SKULL STRIPPING

Table 3 shows the score for all methods compared. In the evaluation of the ADNI2 dataset, where the training and evaluation data are from the same source, there was not much difference in the scores for each method. Only the proposed PCSS resulted in a lower precision than the other methods, resulting in slightly lower Dice scores. This is not due to inherently low SS performance, as discussed below, but rather because it was determined to be overdetected in the evaluation based on the GT defined by MRICloud. This overdetection is the fact that a GT for the determination of the cerebral sickle, which is very difficult to identify, was not determined to be a brain region. See discussion below for details.

On the other hand, the results of LPBA40, CC-12, NFBS, and OASIS show that the proposed PCSS achieves significantly higher SS performance than existing methods, including the state-of-the-art Auto-Net, MVU-Net, and nnU-Net. In particular, PCSS has a very small number of cases that could be considered SS failures (#Dice < 0.95). In addition, PCSS performs almost as well as HD-BET, which is trained on about ten times more data and outperforms HD-BET for CC-12 and NFBS with manual labels.



(a) CC-12 **(b)** LPBA40 **(c)** NFBS **(d)** OASIS

FIGURE 8. Posture correction for other datasets.

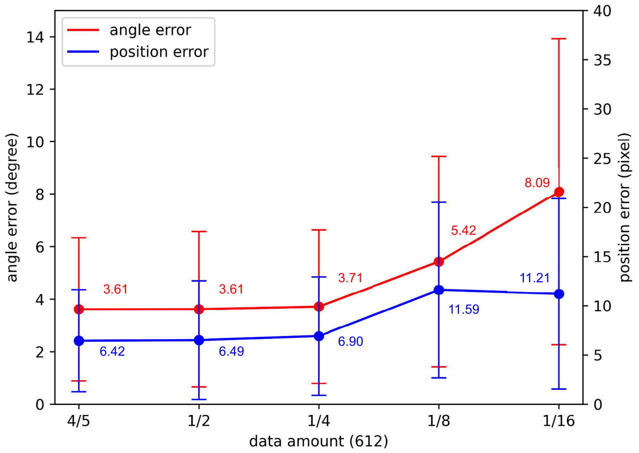


FIGURE 9. Relationship of posture correction error to the amount of training data.

The performance difference between PCSS-1 and PCSS-3 was only 0.36 points on average in terms of Dice score, while the time required for SS was 8.07 and 20.24 seconds for PCSS-1 and PCSS-3, respectively. Of these, the time required for posture correction of one section was 1.96 seconds (approximately 24.2% and 9.6% of the total SS processing, respectively).

Figure 10 compares examples of SS results using the baseline Auto-Net [22] and the proposed PCSS-3 observed from the sagittal plane. Note that Auto-Net offered the best performance of the three state-of-the-art methods that were used for comparison and were reproduced in the authors' implementation. The red line shows the predicted mask by PCSS-3, and the yellow line shows the mask by GT. The proposed PCSS effectively performed SS for the images in all datasets. In particular, recall is significantly improved compared to the existing methods, and the reproducibility of brain regions in PCSS-3 is confirmed to be very high. The relatively low numerical results in OASIS for all methods are due to incomplete GT by FreeSurfer, as mentioned above; PCSS actually detects the appropriate regions.

Table 4 summarizes the effect of each technical element introduced in this paper on final SS performance. These are the averages of the differences in SS performance, as measured by Dice score, between the baseline (U-Net) and the case where each element X is implemented alone (i.e., +X). The average is the macro average of the scores of the datasets, excluding ADNI2 (used for training) and OASIS (which has GT reliability concerns). We could confirm that our posture correction proposal contributed to the improvement of SS performance for all datasets. In addition, we confirmed that the other three techniques also contributed to the improvement of SS performance.

VII. DISCUSSION

A. EFFECTS OF POSTURE CORRECTION

The proposed posture correction method is a simple and robust method that requires only a small amount of training data (Figure 9). Our postural correction reduced head angle variability (i.e., standard deviation) by 4.2 degrees, from 7.81 degrees to 3.61 degrees, and vertical deviation by 16 mm, from 29 mm to 13 mm, each less than 50% of their pre-correction magnitude. In other words, correcting for pitch

TABLE 3. Summaries of SS performance.

	ADNI2 (#data = 612)				LPBA40 (#data = 40)			
	Recall	Precision	Dice	(#Dice < 95)	Recall	Precision	Dice	(#Dice < 95)
U-Net	98.63±0.34	98.38±0.48	98.50±0.30	0,0,1,0,0	90.54±1.06	99.34±0.30	94.73±0.56	31
+PC	98.64±0.30	98.50±0.45	98.57±0.26	0,0,1,0,0	93.09±0.98	96.78±2.46	94.84±1.07	15
+W	99.77±0.13	95.44±0.75	97.55±0.40	0,1,1,0,0	90.91±3.05	98.51±0.51	94.53±1.75	18
+Adv	98.92±0.32	97.42±0.61	98.15±0.32	0,0,1,0,0	91.69±1.01	99.03±0.53	95.22±0.57	7
+Ens	98.76±0.31	98.68±0.45	98.72±0.26	0,1,0,0,0	88.73±2.52	99.87±0.06	93.95±1.45	36
PCSS-1	99.82±0.13	94.62±0.78	97.14±0.42	0,0,1,0,0	97.08±0.75	96.19±1.46	96.62±0.63	1
PCSS-3	99.92±0.11	94.33±0.93	97.04±0.50	0,1,1,0,0	95.74±0.95	98.76±0.38	97.27±0.39	0
Auto-Net	98.80±0.28	97.71±0.52	98.25±0.31	0,0,1,0,0	89.44±2.76	98.69±0.42	93.82±1.58	29
MVU-Net	98.61±0.39	98.24±0.48	98.42±0.30	0,1,1,0,0	88.76±1.47	99.84±0.08	93.97±0.81	39
nnU-Net (3D)	98.93±0.21	98.83±0.41	**98.88±0.26**	0,1,0,0,0	89.04±2.33	98.98±2.32	93.75±2.29	35
HD-BET*	99.94±0.11	87.47±1.32	93.28±0.76	610**	98.41±0.53	96.59±0.82	**97.49±0.27**	0

	CC-12 (#data = 12)				NFBS (#data = 125)			
	Recall	Precision	Dice	(#Dice < 95)	Recall	Precision	Dice	(#Dice < 95)
U-Net	89.34±2.46	97.12±4.41	93.30±2.31	8	92.18±1.25	99.72±0.09	95.80±0.65	15
+PC	93.05±0.01	97.70±2.54	95.29±1.16	3	95.55±0.80	99.11±0.29	97.29±0.36	0
+W	93.82±3.33	97.70±0.83	95.69±1.76	1	97.24±0.72	98.50±0.35	97.87±0.27	0
+Adv	92.80±1.94	96.15±6.28	94.28±2.91	5	93.57±1.04	99.55±0.27	96.46±0.53	2
+Ens	91.34±1.64	97.72±2.82	94.39±1.35	7	93.51±0.98	99.72±0.11	96.51±0.50	3
PCSS-1	97.91±0.82	96.16±1.00	**97.02±0.26**	0	99.44±0.23	94.78±0.80	97.05±0.35	0
PCSS-3	96.45±1.31	97.34±0.98	96.88±0.58	0	98.66±0.52	97.51±0.58	**98.08±0.16**	0
Auto-Net	92.56±1.22	97.18±2.30	94.75±1.11	6	94.71±0.94	99.23±0.28	96.91±0.44	0
MVU-Net	91.00±1.43	98.95±0.47	94.80±0.68	5	92.09±1.16	99.79±0.10	95.78±0.61	12
nnU-Net (3D)	92.59±1.06	98.56±0.59	95.47±0.51	2	96.11±0.89	96.63±7.14	96.22±3.93	14
HD-BET*	99.17±0.44	94.17±1.39	96.60±0.58	0	99.69±0.14	93.93±0.73	96.72±0.35	0

	OASIS (#data = 77)†			
	Recall	Precision	Dice	(#Dice < 95)
U-Net	77.82±1.83	99.97±0.08	87.50±1.16	76
+PC	82.08±1.29	99.84±0.23	90.09±0.79	76
+W	84.35±1.49	99.84±0.46	91.43±0.89	76
+Adv	82.96±1.41	99.88±0.27	90.63±0.84	76
+Ens	79.19±1.32	99.98±0.09	88.37±0.82	76
PCSS-1	89.34±1.51	99.44±0.88	94.11±0.90	65
PCSS-3	88.13±1.67	99.89±0.15	93.63±0.91	69
Auto-Net	78.69±1.39	99.91±0.12	88.03±0.86	76
MVU-Net	77.58±1.38	99.97±0.77	87.36±0.87	76
nnU-Net (3D)	88.78±1.42	99.96±0.93	89.35±0.86	76
HD-BET*	89.80±2.05	99.76±0.24	**94.50±1.05**	48

† These results are for reference only, as the OASIS GTs are from FreeSurfer, which is less reliable.

* HD-BET has been trained on 6, 586 images from 25 locations. This is about ten times more than other methods, but is included for reference.

** Reasons for this large number for ADNI2 are described in the main text.

TABLE 4. Summary of the impact of each elemental tequniques on SS performance in Dice score.

	ADNI2	CC-12	LPBA40	NFBS	OASIS	average†
+PC	+0.07	+1.99	+0.11	+1.49	+2.59	+1.19
+W	-0.95	+2.39	-0.20	+2.07	+3.93	+1.42
+Adv	-0.35	+0.98	+0.49	+0.66	+3.13	+0.71
+Ens	+0.22	+1.09	-0.78	+1.09	+0.87	+0.46

† Averages are macro averages excluding ADNI2 (used for training), and OASIS results (which has GT reliability concerns).

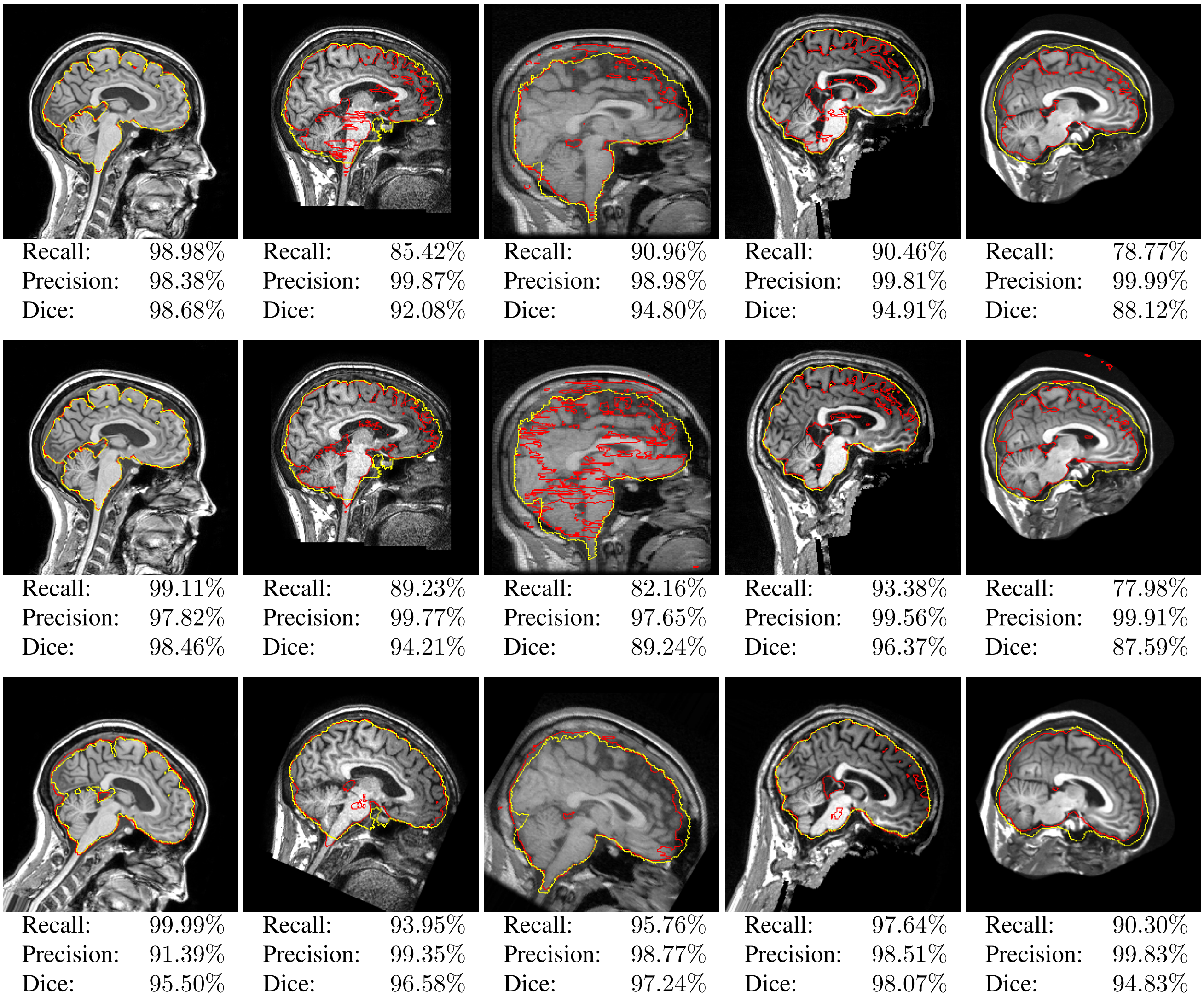
direction, which varied widely across the datasets, standardized the appearance across slices. As a result, this process improves SS performance for datasets that are not used for training and shows that our posture correction contributed to the improvement in SS performance. In addition, this process is computationally efficient (1.96 sec/case; approximately 24.2% of the entire SS process in PCSS-1).

B. PERFORMANCE OF SKULL STRIPPING

1) DISCUSSION ON GT AND PERFORMANCE EVALUATION

The proposed PCSS (PCSS-1 and PCSS-3) generally achieved the best SS performance (96.95 and 97.31 in Dice score, excepting OASIS) but was numerically lower than the other methods when evaluated on the test cases of the ADNI2 dataset used for training.

Firstly, we discuss why PCSS has lower SS (precision) scores only on ADNI2 test cases. Figure 11 shows images



(a) ADNI2 (b) CC-12 (c) LPBA40 (d) NFBS (e) OASIS

FIGURE 10. Example of SS results by PCSS; from top to bottom, SS results for U-Net, Auto-U-Net, and PCSS-3.

of the worst five Dice scores (as well as precision) in one fold on the PCSS-3, from left to right. The red line shows the region predicted by PCSS-3, and the yellow line shows that by MRICloud used as GT. In case (a), which has the lowest Dice score, MRICloud failed to detect a portion of the medulla oblongata and the cerebellum's tonsil (1). In contrast, PCSS-3 accurately extracted this region. Hence, the lower Dice score was attributed to an error in the GT, not a failure in PCSS-3. In cases (b)-(e), the disparities between the brain masks generated by PCSS-3 and MRICloud were primarily concerned with the inclusion or exclusion of the cerebral longitudinal fissure, the space is occupied by the cerebrospinal fluid and the falx cerebri, a membrane that separates the left and right cerebral hemispheres. While MRICloud generally excluded this space and membrane from the brain mask, PCSS-3 was inclined to include them. Delineation of this space, which has a complex boundary, can be a challenging task even for neuroimaging experts. Considering that the GT of other databases includes this area in their brain masks, we believe a Dice score slightly lower than nnU-Net (97.04 vs. 98.88) is practically acceptable. Even in the worst case (b), the Dice score is 96.44, which represents sufficient accuracy for the ADNI2 dataset.

2) COMPARISON AND DISCUSSION WITH OTHER METHODS

The proposed PCSS achieves the best SS performance, including state-of-the-art 2D and 3D methods, in two datasets not used for training. In addition, excluding HD-BET, which is about ten times more training data than PCSS from the comparison, PCSS shows the best performance on all four datasets. This is evidence of its inherently high SS performance. The PCSS shows robust and accurate SS results

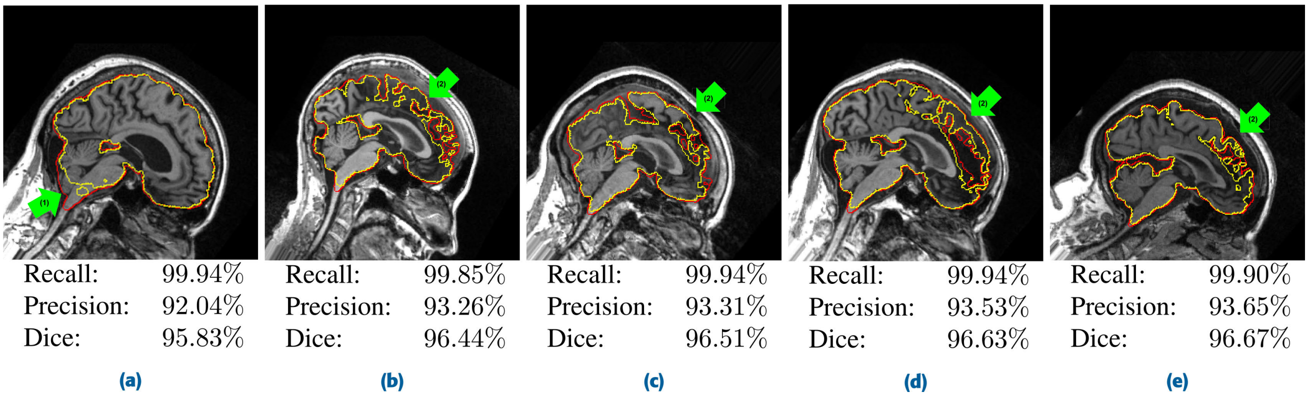


FIGURE 11. Worst five SS results (red) on ADNI2 test cases using PCSS-3; yellow is the GT provided by MRICloud [20].

for these unknown datasets, indicating that it does not overfit with the GT given by MRICloud, which is also confirmed by the results in Figure 10.

Especially for LPBA40 with manual GT, most of the cases (31 of 40) in the baseline U-Net have a Dice score of less than 0.95 due to under detection (low recall). By contrast, PCSS achieved extremely accurate SS for almost all cases despite using the same training data. In addition, LPBA40 has a resolution of 0.86 mm and 0.88 mm per pixel, which is different from the resolution used in the training. This result shows that PCSS is robust to data with different resolutions than the training data. The same trend is observed for other datasets, which confirms the excellent SS performance and robustness of PCSS.

The advanced SS methods Auto-Net [22], MVU-net [25], and nnU-Net [28] show good results for the ADNI2 dataset. However, the results for the other datasets are lower than PCSS, especially due to the lower recall (averaging 4.68, 5.94, and 4.39 points, respectively, with the exception of OASIS). The reason for this is that Auto-Net is a method that aims to improve accuracy by feeding back the trained probability maps of brain regions to the input, which may lead to overfitting. In addition, MVU-Net employs an architecture with a small number of parameters to improve efficiency, but this architecture may not have been expressive enough to achieve accurate SS even for multiple datasets. Meanwhile, nnU-Net uses 3D spatial information from 3D U-Net. Therefore, it has the best Dice score on the training data, ADNI2, but it is thought that overfitting prevents it from corresponding images, such as NFBS, which are cut off under the neck. HD-BET [27] shows the most robust SS result among previous studies. The low Dice score of the ADNI2 dataset is due to the difference in GT between HD-BET and MRICloud. The GT of MRICloud does not include the cerebral longitudinal fissure, whereas HD-BET does. As a result, the accuracy of HD-BET in ADNI2 is reduced. As we described earlier, it is difficult even for experts to create an accurate GT for this space. Therefore, the low score of HD-BET in ADNI2 is not inherently a problem with its SS performance, and the result shown in other datasets suggests that HD-BET is capable of achieving highly accurate and robust SS for diverse data. However, they used 6, 586 images as training data and semi-manually created masks for 1, 568 T1-weighted images. It takes 15 minutes to create a mask, which is very expensive to build a larger network in the future. PCSS achieved almost the same performance as HD-BET even though it only trained 612 images. This result shows that PCSS can achieve highly robust SS with few training data.

Although the original paper [22] and [25] showed better performance than other comparison methods when SS ability was evaluated by splitting within the same dataset, we found concerns about the robustness to data from different environments in this result. These differences in datasets are thought to be due to the stronger effects of overfitting caused by differences in subjects' postures. By contrast, the proposed PCSS, which introduces posture correction and other techniques, is able to achieve robust and accurate SS results for a large number of datasets.

C. DISCUSSION OF TECHNICAL ELEMENTS FOR SKULL STRIPPING

The main proposal in this paper, posture correction, improves SS performance on all five datasets and is an important factor in achieving robust and accurate SS. The introduction of weighted loss functions to address the imbalance between brain and non-brain regions and the introduction of the discriminator to further improve SS performance are of significant importance in this achievement. In particular, the three datasets labeled manual (i.e., excluding ADNI2 and OASIS) showed average improvements of 1.42 and 0.71 points, respectively. These results show that PCSS suppresses overfitting for ADNI's GT and contributes to making SS robust to unknown data in different environments. This suppression of overfitting is also true for posture correction. In addition, The PCSS results in Table 3 show that these elements perform better when combined.

The trained PCSS-3 using ensemble of three sections improves the Dice score by only 0.36 points on average, compared to PCSS-1. On the other hand, PCSS-3 takes about 2.5 times longer than PCSS-1 per SS case. For maximum performance, PCSS-3, which uses a three-section ensemble, is preferred. However, PCSS-1 also outperforms previous state-of-the-art SS methods through its proposed posture correction. Therefore, PCSS-1 is generally preferable for actual operations due to its combination of speed and performance therefore suitable for high-throughput brain MRI analysis.

VIII. LIMITATION

The evaluation of PCSS in this study was limited to T1-weighted Images. However, MRIs have multiple types, including contrast-enhanced T1-weighted, T2-weighted, and FLAIR. There is also a process in MRI called fat saturation, which reduces the fat signal to show other tissues and structures more clearly. We plan to evaluate the performance and usefulness of PCSS for these different image types and processing.

IX. CONCLUSION

In this paper, we proposed and published posture correction skull stripping, a highly accurate and robust skull stripping method for T1-weighted brain magnetic resonance imaging that accounts for the diversity of subjects' postures. Using five publicly available datasets, we confirmed that PCSS outperforms existing state-of-the-art methods and the effectiveness of each of its technical components. This paper discusses and evaluates the use of larger and more diverse data than previous SS papers, thus setting the standard for future SS papers. We hope that our published PCSS will contribute to future research on brain MRI.

ACKNOWLEDGMENT

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

The MRI data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research and Development LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Company Inc.; Meso Scale Diagnostics LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education and the study is coordinated by the Alzheimer's Therapeutic Research Institute, University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging, University of Southern California.

Data were provided 1 by OASIS Cross-Sectional: Principal Investigators: D. Marcus, R, Buckner, J, Csernansky J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, and U24 RR021382.

REFERENCES

[1] E. Hosseini-Asl, G. Gimel'farb, and A. El-Baz, "Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network," 2016, *arXiv:1607.00556*.

[2] S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova, "Residual and plain convolutional neural networks for 3D brain MRI classification," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 835–838.

[3] S. Esmaeilzadeh, D. I. Belivanis, K. M. Pohl, and E. Adeli, "End-to-end Alzheimer's disease diagnosis and biomarker identification," in *Machine Learning in Medical Imaging*. Granada, Spain: Springer, 2018, pp. 337–345.

[4] M. B. Naceur, R. Saouli, M. Akil, and R. Kachouri, "Fully automatic brain tumor segmentation using end-to-end incremental deep neural networks in MRI images," *Comput. Methods Programs Biomed.*, vol. 166, pp. 39–49, Nov. 2018.

[5] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, Jan. 2017.

[6] R. A. Zeineldin, M. E. Karar, J. Coburger, C. R. Wirtz, and O. Burgert, "DeepSeg: Deep neural network framework for automatic brain tumor segmentation using magnetic resonance FLAIR images," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 6, pp. 909–920, Jun. 2020.

[7] Z. N. K. Swati, Q. Zhao, M. Kabir, F. Ali, Z. Ali, S. Ahmed, and J. Lu, "Content-based brain tumor retrieval for MR images using transfer learning," *IEEE Access*, vol. 7, pp. 17809–17822, 2019.

[8] Y. Onga, S. Fujiyama, H. Arai, Y. Chayama, H. Iyatomi, and K. Oishi, "Efficient feature embedding of 3D brain MRI images for content-based image retrieval with deep metric learning," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2019, pp. 3764–3769.

[9] M. Owais, M. Arsalan, J. Choi, and K. R. Park, "Effective diagnosis and treatment through content-based medical image retrieval (CBMIR) by using artificial intelligence," *J. Clin. Med.*, vol. 8, no. 4, p. 462, Apr. 2019.

[10] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy, "Magnetic resonance image tissue classification using a partial volume model," *NeuroImage*, vol. 13, no. 5, pp. 856–876, May 2001.

[11] S. M. Smith, "Fast robust automated brain extraction," *Hum. Brain Mapping*, vol. 17, no. 3, pp. 143–155, Nov. 2002.

[12] F. Ségonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, and B. Fischl, "A hybrid approach to the skull stripping problem in MRI," *NeuroImage*, vol. 22, no. 3, pp. 1060–1075, Jul. 2004.

[13] C. Fennema-Notestine, I. B. Ozyurt, C. P. Clark, S. Morris, A. Bischoff-Grethe, M. W. Bondi, T. L. Jernigan, B. Fischl, F. Segonne, D. W. Shattuck, R. M. Leahy, D. E. Rex, A. W. Toga, K. H. Zou, and G. G. Brown, "Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: Effects of diagnosis, bias correction, and slice location," *Hum. Brain Mapping*, vol. 27, no. 2, pp. 99–113, Feb. 2006.

[14] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," *NeuroImage*, vol. 54, no. 3, pp. 2033–2044, Feb. 2011.

[15] J. E. Iglesias, C.-Y. Liu, P. M. Thompson, and Z. Tu, "Robust brain extraction across datasets and comparison with publicly available methods," *IEEE Trans. Med. Imag.*, vol. 30, no. 9, pp. 1617–1634, Sep. 2011.

[16] S. F. Eskildsen, P. Coupé, V. Fonov, J. V. Manjón, K. K. Leung, N. Guizard, S. N. Wassef, L. R. Østergaard, and D. L. Collins, "BEaST: Brain extraction based on nonlocal segmentation technique," *NeuroImage*, vol. 59, no. 3, pp. 2362–2373, 2012.

[17] R. Beare, J. Chen, C. L. Adamson, T. Silk, D. K. Thompson, J. Y. M. Yang, V. A. Anderson, M. L. Seal, and A. G. Wood, "Brain extraction using the watershed transform from markers," *Frontiers Neuroinform.*, vol. 7, p. 32, Dec. 2013.

[18] E. S. Lutkenhoff, M. Rosenberg, J. Chiang, K. Zhang, J. D. Pickard, A. M. Owen, and M. M. Monti, "Optimized brain extraction for pathological brains (optiBET)," *PLoS ONE*, vol. 9, no. 12, Dec. 2014, Art. no. e115551.

[19] B. Fischl, "FreeSurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, Aug. 2012.

[20] S. Mori, D. Wu, C. Ceritoglu, Y. Li, A. Kolasny, M. A. Vaillant, A. V. Faria, K. Oishi, and M. I. Miller, "MRICloud: Delivering high-throughput MRI neuroinformatics as cloud-based software as a service," *Comput. Sci. Eng.*, vol. 18, no. 5, pp. 21–35, Sep. 2016.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Munich, Germany: Springer, 2015, pp. 234–241.

[22] S. S. Mohseni Salehi, D. Erdogmus, and A. Gholipour, "Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging," *IEEE Trans. Med. Imag.*, vol. 36, no. 11, pp. 2319–2330, Nov. 2017.

[23] O. Lucena, R. Souza, L. Rittner, R. Frayne, and R. Lotufo, "Silver standard masks for data augmentation applied to deep-learning-based skull-stripping," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1114–1117.

[24] S. Jiang, L. Guo, G. Cheng, X. Chen, C. Zhang, and Z. Chen, "Brain extraction from brain MRI images based on Wasserstein GAN and O-Net," *IEEE Access*, vol. 9, pp. 136762–136774, 2021.

[25] A. Fatima, T. M. Madni, F. Anwar, U. I. Janjua, and N. Sultana, "Automated 2D slice-based skull stripping multi-view ensemble model on NFBS and IBSR datasets," *J. Digit. Imag.*, vol. 35, no. 2, pp. 374–384, Apr. 2022.

[26] H. Hwang, H. Z. U. Rehman, and S. Lee, "3D U-Net for skull stripping in brain MRI," *Appl. Sci.*, vol. 9, no. 3, p. 569, Feb. 2019.

[27] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H. Schlemmer, S. Heiland, W. Wick, M. Bendszus, K. H. Maier-Hein, and P. Kickingereder, "Automated brain extraction of multisequence MRI using artificial neural networks," *Hum. Brain Mapping*, vol. 40, no. 17, pp. 4952–4964, Dec. 2019.

[28] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "NnU-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[29] L. Pei, M. Ak, N. H. M. Tahon, S. Zenkin, S. Alkarawi, A. Kamal, M. Yilmaz, L. Chen, M. Er, N. Ak, and R. Colen, "A general skull stripping of multiparametric brain MRIs using 3D convolutional neural network," *Sci. Rep.*, vol. 12, no. 1, p. 10826, Jun. 2022.

[30] L. Wang, Z. Wu, L. Chen, Y. Sun, W. Lin, and G. Li, "IBEAT V2.0: A multisite-applicable, deep learning-based pipeline for infant cerebral cortical surface reconstruction," *Nature Protocols*, vol. 18, no. 5, pp. 1488–1509, May 2023.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[32] R. W. Cox, "AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages," *Comput. Biomed. Res.*, vol. 29, no. 3, pp. 162–173, Jun. 1996.

[33] A. B. Waters, R. A. Mace, K. S. Sawyer, and D. A. Gansler, "Identifying errors in freesurfer automated skull stripping and the incremental utility of manual intervention," *Brain Imag. Behav.*, vol. 13, no. 5, pp. 1281–1291, Oct. 2019.

[34] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1744–1757, Oct. 2010.

[35] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[36] Y. Zhang, J. Wu, W. Chen, Y. Liu, J. Lyu, H. Shi, Y. Chen, E. X. Wu, and X. Tang, "Fully automatic white matter hyperintensity segmentation using U-Net and skip connection," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 974–977.

[37] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[38] P. S. Aisen, R. C. Petersen, M. Donohue, and M. W. Weiner, "Alzheimer's disease neuroimaging initiative 2 clinical core: Progress and plans," *Alzheimer's Dementia*, vol. 11, no. 7, pp. 734–739, Jul. 2015.

[39] R. Souza, O. Lucena, J. Garrafa, D. Gobbi, M. Saluzzi, S. Appenzeller, L. Rittner, R. Frayne, and R. Lotufo, "An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement," *NeuroImage*, vol. 170, pp. 482–494, Apr. 2018.

[40] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, and A. W. Toga, "Construction of a 3D probabilistic atlas of human cortical structures," *NeuroImage*, vol. 39, no. 3, pp. 1064–1080, Feb. 2008.

[41] B. Puccio, J. P. Pooley, J. S. Pellman, E. C. Taverna, and R. C. Craddock, "The preprocessed connectomes project repository of manually corrected skull-stripped T1-weighted anatomical MRI data," *GigaScience*, vol. 5, no. 1, Dec. 2016, Art. no. s13742.

[42] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *J. Cognit. Neurosci.*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007.

[43] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

KEI NISHIMAKI received the B.E. degree in applied informatics from Hosei University, Tokyo, Japan, in 2022, where he is currently pursuing the M.E. degree. His research interests include machine learning and medical image analysis.

KUMPEI IKUTA received the B.E. and M.E. degrees in applied informatics from Hosei University, Tokyo, Japan, in 2020 and 2022, respectively. He is currently a Machine Learning Engineer with DWANGO Company Ltd. His research interests include machine learning and medical image analysis.

SHINGO FUJIYAMA received the B.E. and M.E. degrees in applied informatics from Hosei University, Tokyo, Japan, in 2019 and 2021, respectively. He is currently a Machine Learning Engineer with ChillStack Company Ltd. His research interests include machine learning and medical image analysis.

KENICHI OISHI received the M.D. and Ph.D. degrees in neuroscience from the Kobe University School of Medicine, Kobe, Japan, in 1997 and 2005, respectively. After working as a Research and Teaching Assistant with the Department of Clinical Molecular Medicine, Kobe University, he joined as an Assistant Professor of neurology, in 2005. He accepted a position as an Assistant Professor with the Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, in 2009, and became an Associate Professor, in 2016. He is the author of three books and more than 120 peer-reviewed articles and holds six patents. His research interests include applying deep learning technology to extract features of the human brain, developing multi-modal brain atlases based on structural and diffusion MRI and rsfMRI, and applications of the atlas-based image quantification. He has been a Board Certified Neurologist of the Japanese Society of Neurology, since 2003, and a fellow of the Japanese Society of Internal Medicine, since 2006.

HITOSHI IYATOMI (Member, IEEE) received the B.E., M.E., and Ph.D. (Eng.) degrees from Keio University, Japan, in 1998, 2000, and 2004, respectively, and the second Ph.D. degree in medical science from Tokyo Women's Medical University, in 2011. From 2000 to 2004, he was a Technical Consultant with Hewlett-Packard Japan. In 2004, he joined Hosei University as a Research Associate, where he is currently a Professor with the Department of Applied Informatics. From 2016 to 2017, he was a Visiting Scholar with Johns Hopkins University. He has authored or coauthored more than 130 peer-reviewed journals and conference papers in various research areas based on machine learning, such as computer vision, medical applications, and natural language processing.