ERNet: Unsupervised Collective Extraction and Registration in Neuroimaging Data

Yao Su
Worcester Polytechnic Institute
Worcester, MA, USA
ysu6@wpi.edu

Zhentian Qian
Worcester Polytechnic Institute
Worcester, MA, USA
zqian@wpi.edu

Lifang He
Lehigh University
Bethlehem, PA, USA
lih319@lehigh.edu

Xiangnan Kong
Worcester Polytechnic Institute
Worcester, MA, USA
xkong@wpi.edu

ABSTRACT

Brain extraction and registration are important preprocessing steps in neuroimaging data analysis, where the goal is to extract the brain regions from MRI scans (*i.e.*, extraction step) and align them with a target brain image (*i.e.*, registration step). Conventional research mainly focuses on developing methods for the extraction and registration tasks separately under supervised settings. The performance of these methods highly depends on the amount of training samples and visual inspections performed by experts for error correction. However, in many medical studies, collecting voxel-level labels and conducting manual quality control in high-dimensional neuroimages (*e.g.*, 3D MRI) are very expensive and time-consuming. Moreover, brain extraction and registration are highly related tasks in neuroimaging data and should be solved collectively. In this paper, we study the problem of unsupervised collective extraction and registration in neuroimaging data. We propose a unified end-to-end framework, called ERNet (Extraction-Registration Network), to jointly optimize the extraction and registration tasks, allowing feedback between them. Specifically, we use a pair of multi-stage extraction and registration modules to learn the extraction mask and transformation, where the extraction network improves the extraction accuracy incrementally and the registration network successively warps the extracted image until it is well-aligned with the target image. Experiment results on real-world datasets show that our proposed method can effectively improve the performance on extraction and registration tasks in neuroimaging data.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Computing methodologies** → **Neural networks**; **Image segmentation**; **Matching**.

KEYWORDS

brain extraction; skull stripping; registration; unsupervised; collective; multi-stage

ACM Reference Format:
Yao Su, Zhentian Qian, Lifang He, and Xiangnan Kong. 2022. ERNet: Unsupervised Collective Extraction and Registration in Neuroimaging Data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3534678.3539227

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539227

1 INTRODUCTION

Background. Brain extraction (*a.k.a.* skull stripping) and registration are preliminary but crucial steps in many neuroimaging studies. Examples of these studies include anatomical and functional analysis [3, 26], multi-modality fusion [5], diagnostic assistance [25]. The brain extraction step aims to remove the non-cerebral tissues such as the skull, dura, and scalp from the Magnetic Resonance Imaging (MRI) scan of a patient's head, while the registration step aims to align the extracted brain region with a template image of the standard brain. The extraction and registration steps are essential preprocessing procedures in many neuroimaging studies. For example, in anatomical and functional analysis, after removing and aligning the brain regions, the interference of non-neural tissues, imaging modalities, viewpoints can be eliminated, thus allowing precise quantification of changes in the shape, size, and position of anatomy and function. In brain atrophy diagnosis, a patient's brain region across different pathological stages needs to be first extracted from raw brain MRI scans and then aligned with a standard template to counteract the non-diagnostic changes. These essential processing steps help doctors accurately monitor the alteration of brain volume.

State-of-the-Art. In the literature, brain extraction and registration problems have been extensively studied [7, 14, 15, 23]. Conventional approaches focus on developing methods for extraction [14, 15] and registration [7, 23] separately under supervised settings, as shown in Figure 2(a). However, in many medical studies, obtaining annotations of brain regions and transformations between images is often expensive as expertise, effort, and time are needed to produce precise labels, especially for high-dimensional neuroimages, *e.g.*, 3D MRI. To address this limitation, recent works [4, 6, 19, 20, 22, 27] introduce a two-step approach

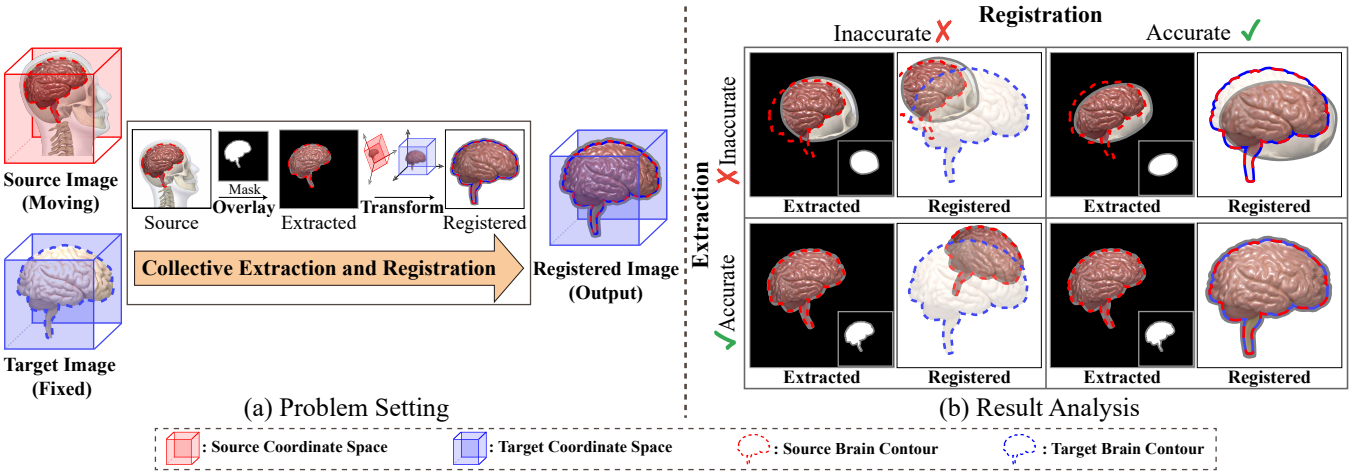


Figure 1: The problem of unsupervised collective extraction and registration in neuroimaging data. (a) Given a raw scan of a patient's head (the source image) and a template image of standard brain region (the target image), the goal is to extract the brain region from the source image, and transform it to align with the target image. Neither the extraction label (*i.e.*, the brain region in the source image) nor the registration label (*i.e.*, the transformation required to align the source with the target) is available. Examples of different possible results are shown in (b). The bottom-right box is the ideal result: the correct region of the brain in the source is extracted and is well-aligned with the target.

for unsupervised extraction and registration by using automated brain extraction tools[6, 19, 20, 22] and unsupervised registration models [4, 27], as shown in Figure 2(b). Nevertheless, these works typically rely on manual quality control to correct inaccurate extraction results before performing subsequent registration. Conducting such visual inspection is not only time-consuming and labor-intensive, but also suffers from intra- and inter-rater variability, thus limiting the efficiency and performance of both tasks. More importantly, most existing methods still conduct extraction and registration separetely and neglect the potential relationship between these two tasks.

Problem Definition. In this paper, we study the problem of unsupervised collective brain extraction and registration, as shown in Figure 1(a). Our goal is to capture the correlation of two tasks to boost their performance in an unsupervised setting. Specifically, the brain region needs to be extracted from the source image accurately and well-aligned to the target image without any labeled data.

Challenges. Despite its value and significance, the problem of unsupervised collective extraction and registration has not been studied before and is very challenging due to its unique characteristics listed below:

• *Lack of labels for extraction:* Conventional learning-based extraction approaches are trained with a large number of training samples with ground truth labels. However, collecting voxel-level labels is very expensive and time-consuming in high-dimensional neuroimaging data.

• *Lack of labels for registration:* The ground truth transformation between source and target images is difficult to obtain. Though there are unsupervised registration methods [4, 27] that optimize the transformation parameters by maximizing the similarity between images, these methods are only effective when the non-brain tissue of the source image is removed; otherwise, an erroneous transformation will be produced, rendering the registration invalid.

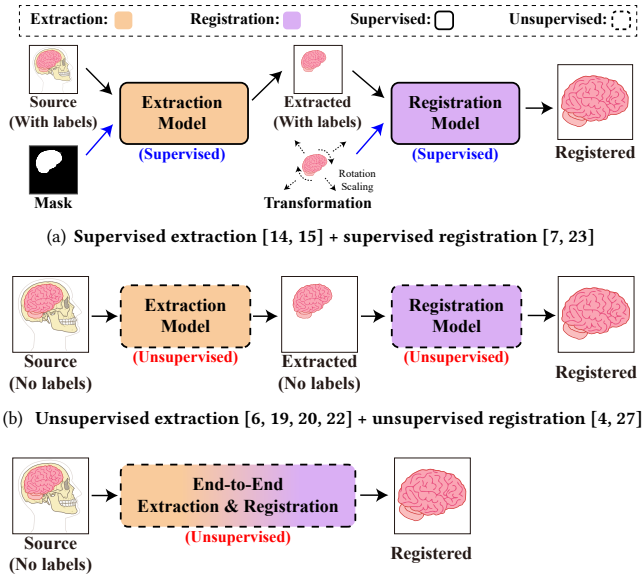


Figure 2: Related works in brain extraction and registration.

Thus, obtaining the accurate transformation between source and target images in an unsupervised setting remains largely unsolved.

• *Dependencies between extraction and registration:* Conventional research mainly focuses on conducting extraction and registration tasks separately. However, these two tasks are highly correlated. Brain extraction has a decisive impact on the accuracy of the registration task, as shown in Figure 1(b). Meanwhile, registration task can help extraction task to capture cerebral/non-cerebral information from the source and target images. Thereby, a holistic solution is desired to manage the interdependence between the two tasks.

Proposed Method. To tackle the above issues, we propose a unified end-to-end framework, called ERNet (Extraction-Registration

Network) for unsupervised collective brain extraction and registration. Figure 2 illustrates the comparison between our method and other state-of-the-art approaches. Specifically, ERNet contains a pair of multi-stage extraction and registration modules, where the multi-stage extraction module progressively removes the non-brain tissue from the source image to produce an extracted brain image, and the multi-stage registration module incrementally aligns the extracted image to the target image. These two modules help each other to boost extraction and registration performance simultaneously. The unalignable portion, *i.e.*, non-brain tissue, revealed in the registration module guides the refinement process in the extraction module. Meanwhile, the registration module benefits from accurate brain extraction generated in the extraction module. By bringing these two modules end-to-end, we achieve a joint optimization with no labels assistance.

We design a new regularization term to smooth the predicted brain mask during the training, which improves the extraction accuracy to a certain extent. Extensive experiments are performed on multiple public brain MRI datasets. The results indicate that our proposed method significantly outperforms state-of-the-art approaches in both extraction and registration accuracy.

2 PRELIMINARIES

In this section, we first introduce related concept and notations, then define the unsupervised collective brain extraction and registration problem formally.

2.1 Notations and Definitions

Definition 1 (Source and target images). Suppose we are given a training dataset $\mathcal{D} = \{(\mathbf{S}_i, \mathbf{T}_i)\}_{i=1}^{Z}$ that consists of Z pairs of training samples. Each pair contains a source image $\mathbf{S}_i \in \mathbb{R}^{W\times H\times D}$ (*e.g.*, the raw MRI scan of a patient's head) and a target $\mathbf{T}_i \in \mathbb{R}^{W\times H\times D}$ (*e.g.*, a standard template of the brain region). Here W, H and D denote the width, height and depth dimensions of the 3D images. For simplicity, we assume that the source and target images are resized to the same dimension, *i.e.*, $W \times H \times D$. Generally, in $\mathcal{D}$, the target images in different pairs can be different. For example, in the cross-modality studies [5], we need to align the functional MRI (*i.e.*, source image) with the structural MRI (*i.e.*, target image) for each patient in the study (*i.e.*, an image pair in $\mathcal{D}$), where different patients will have different structural MRI images. In many neuroimaging studies, however, all pairs in $\mathcal{D}$ can share a same target image, which is a special case of the dataset $\mathcal{D}$. For example, in brain network analysis, the functional MRI images (*i.e.*, source images) of all patients need to be aligned with a same template image (*i.e.*, target image), *e.g.*, MNI 152 [25]. For simplicity, in the following discussion, we omit the subscript i of $\mathbf{S}_i$ and $\mathbf{T}_i$.

Definition 2 (Brain extraction mask). Brain extraction mask $\mathbf{M} \in \{0,1\}^{W\times H\times D}$ is a binary tensor of the same dimensions as the source image $\mathbf{S}$. 1 in $\mathbf{M}$ corresponds to the cerebral tissues on $\mathbf{S}$ at the same location and 0 otherwise. The extracted image $\mathbf{E} = \mathbf{S} \circ \mathbf{M}$ is generated by applying the $\mathbf{M}$ on $\mathbf{S}$ via a element-wise product operator $\circ$.

Definition 3 (Affine transformation and warped image). Without loss of generality, here we assume that the transformation in the registration task is affine-based. However, this work can be easily extended to other types of registration, *e.g.*, nonlinear/deformable registration. The affine transformation parameters $\mathbf{a} \in \mathbb{R}^{12}$ is a vector used to parameterized an affine transformation matrix $\mathbf{A} \in \mathbb{R}^{4\times 4}$. The warped image $\mathbf{W} = \mathcal{T}(\mathbf{E}, \mathbf{a})$ is generated by applying the affine transformation on the extracted image $\mathbf{E}$, where $\mathcal{T}(\cdot,\cdot)$ is the affine transformation operator. The following relationship holds for $\mathbf{W}$ and $\mathbf{E}$ on the voxel level:

$$\mathbf{W}_{xyz} = \mathbf{E}_{x'y'z'}, \tag{1}$$

where the correspondences between coordinates x, y, z and x', y', z' are calculated based on the affine transformation matrix $\mathbf{A}$:

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \mathbf{A} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ a_5 & a_6 & a_7 & a_8 \\ a_9 & a_{10} & a_{11} & a_{12} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}. \tag{2}$$

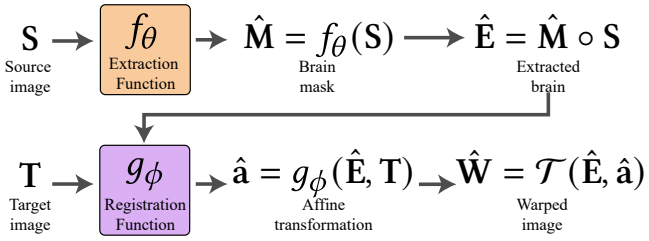


Figure 3: A demonstration of extraction and registration functions.

2.2 Problem Formulation

The goal of collective brain extraction and registration is to jointly learn the extraction function $f_\theta : \mathbb{R}^{W\times H\times D} \to \mathbb{R}^{W\times H\times D}$ and the registration function $g_\phi : \mathbb{R}^{W\times H\times D} \times \mathbb{R}^{W\times H\times D} \to \mathbb{R}^{12}$, as shown in Figure 3. Specifically, the extraction function $f_\theta(\cdot)$ takes the source image $\mathbf{S}$ as input to predicts a brain extraction mask $\hat{\mathbf{M}} = f_\theta(\mathbf{S})$. Then, the registration function $g_\phi(\cdot,\cdot)$ takes the extracted brain image $\hat{\mathbf{E}} = \hat{\mathbf{M}} \circ \mathbf{S}$ and the target image $\mathbf{T}$ to predict the affine transformation parameter $\hat{\mathbf{a}} = g_\phi(\hat{\mathbf{E}}, \mathbf{T})$. Finally, the warped image is $\hat{\mathbf{W}} = \mathcal{T}(\hat{\mathbf{E}}, \hat{\mathbf{a}})$. The optimal parameter θ^* and ϕ^* can be found by solving the following optimization problem:

$$\begin{aligned} \theta^*, \phi^* &= \arg\min_{\theta,\phi} \sum_{(\mathbf{S},\mathbf{T})\in\mathcal{D}} \left[\mathcal{L}\left(\hat{\mathbf{W}}, \mathbf{T}\right)\right] \\ &= \arg\min_{\theta,\phi} \sum_{(\mathbf{S},\mathbf{T})\in\mathcal{D}} \left[\mathcal{L}\left(\mathcal{T}\left(f_\theta(\mathbf{S}) \circ \mathbf{S},\ \ g_\phi\left(f_\theta(\mathbf{S}) \circ \mathbf{S}, \mathbf{T}\right)\right), \mathbf{T}\right)\right], \end{aligned} \tag{3}$$

where the image pair $(\mathbf{S}, \mathbf{T})$ is sampled from the training dataset $\mathcal{D}$, and $\mathcal{L}(\cdot,\cdot)$ is image dissimilarity criteria, *e.g.*, mean square error.

To the best of our knowledge, this work is the first endeavour to find an optimal solution to the problem of unsupervised collective brain image extraction and registration in an end-to-end neural network. Our approach excludes the necessity of labeling the brain extraction masks and transformation between images of the training dataset, as opposed to other supervised methods [7, 14, 15, 23].

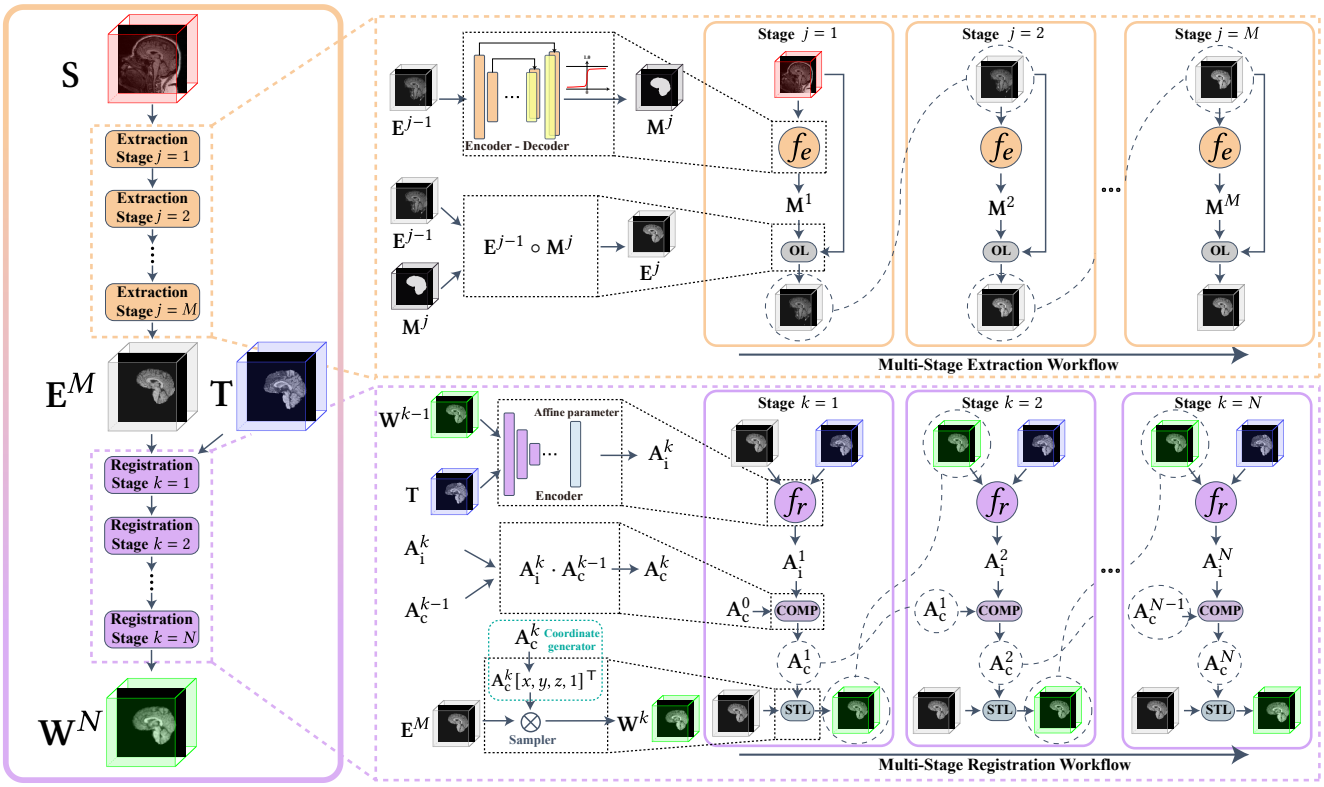


Figure 4: An overview of ERNet for collective brain extration and registration. Multi-stage extraction module: In each extraction stage j, the *Extraction Network* f_e predicts the current mask $\mathbf{M}^j$ from previous extracted image $\mathbf{E}^{j-1}$; the *Overlay Layer* (OL) performs an element-wise product between the previous extracted image $\mathbf{E}^{j-1}$ and the current mask $\mathbf{M}^j$ to generate the current extracted image $\mathbf{E}^j$. The final extracted image is $\mathbf{E}^M$. Multi-stage registration module: In each registration stage k, the *Registration Network* f_r predicts the current affine transformation $\mathbf{A}_i^k$ between the previous warped image $\mathbf{W}^{k-1}$ and target image T; the *Composition Layer* (COMP) fuses current affine transformation $\mathbf{A}_i^k$ and previous combined affine transformation $\mathbf{A}_c^{k-1}$ to generate the updated combined affine transformation $\mathbf{A}_c^k$; the *Spatial Transformation Layer* (STL) performs the transformation $\mathbf{A}_c^k$ on the final extracted image $\mathbf{E}^M$ to produce the warped image $\mathbf{W}^k$. The final registered image is $\mathbf{W}^N$.

3 OUR APPROACH

Overview. Figure 4 presents an overview of the proposed ERNet framework for the unsupervised collective brain extraction problem. Our method is a multi-stage deep neural network consisting of two main modules: 1) *Multi-Stage Extraction Module* takes the raw source image S as input, and gradually produces the extracted brain image $\mathbf{E}^M$ after M stages of extraction; 2) *Multi-Stage Registration Module* takes the extracted brain image $\mathbf{E}^M$ and the target image T as inputs, and incrementally aligns $\mathbf{E}^M$ with T through N stages of registration. The final output is the warped image $\mathbf{W}^N$. The whole framework is trained using backpropagation in an end-to-end unsupervised fashion, allowing feedback and collaboration between modules. Next we introduce the details of each module and the training process.

3.1 Multi-Stage Extraction Module

This module proposes to solve the extraction task in a multi-stage fashion to obtain high extraction accuracy. It contains M extraction stages with each stage j consisting of two main components: 1) *Extraction Network* takes the previous extracted brain image $\mathbf{E}^{j-1}$ as input, and generates a current brain mask $\mathbf{M}^j$; 2) *Overlay Layer* takes the previous extracted brain image $\mathbf{E}^{j-1}$ and the current extraction mask $\mathbf{M}^j$ as inputs, and generates an updated extracted image $\mathbf{E}^j$. The output of this module is the extracted brain image $\mathbf{E}^M$ at the final stage $j = M$.

3.1.1 Extraction Network: f_e. The extraction network $f_e(\cdot)$ serves to gradually remove the non-cerebral tissues in the source image S so that only cerebral tissues would remain on the extracted image at the final stage. At each stage j, based on the extracted brain image $\mathbf{E}^{j-1}$ from the previous stage, it would produce a current extraction mask $\mathbf{M}^j$ to remove the non-cerebral tissues believed to be still on $\mathbf{E}^{j-1}$. Specifically, we adopt 3D U-Net [17] as the base network to learn $f_e(\cdot)$, which is the state-of-the-art architecture widely used in image registration and semantic segmentation. The output of the U-Net would go through a Heaviside step function to obtain the binary mask $\mathbf{M}^j$ when performing inference:

$$H(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Note that the derivative of the Heaviside step function does not exist for $x = 0$ and is a constant 0 for $x \neq 0$. For the gradient to successfully backpropagate, we use a Sigmoid function with a large slope parameter γ to approximate the Heaviside step function when

performing training:

$$S(x) = \frac{1}{1 + e^{-\gamma x}} \tag{5}$$

$f_e(\cdot)$ follows a shared-weight design, which means that $f_e(\cdot)$ is repetitively applied across stages with the same parameters. It can be formalized as:

$$\mathbf{M}^j = f_e\left(\mathbf{E}^{j-1}\right), \tag{6}$$

where $\mathbf{M}^j$ is the outputted brain mask of the j-th stage for $j = [1, \cdots, M]$ and $\mathbf{E}^0 = \mathbf{S}$.

3.1.2 Overlay Layer: OL. The overlay layer would remove the non-cerebral tissues remaining in the image by applying the current brain mask $\mathbf{M}^j$ to the previous extracted image $\mathbf{E}^{j-1}$. The updated extracted image is:

$$\mathbf{E}^j = \mathbf{E}^{j-1} \circ \mathbf{M}^j$$

where $\circ$ is the element-wise product operator.

3.2 Multi-Stage Registration Module

Similar to the extraction module discussed in Section 3.1, we implement a multi-stage solution to address the registration task. This module consists of N cascaded stages with each stage k containing three main components: 1) *Registration Network* takes the previous warped image $\mathbf{W}^{k-1}$ and the target image $\mathbf{T}$ as inputs, and generates the current affine transformation $\mathbf{A}_{\text{i}}^k$; 2) *Composition Layer* takes the previous combined affine transformation $\mathbf{A}_{\text{c}}^{k-1}$ and the current affine transformation $\mathbf{A}_{\text{i}}^k$ as input, and generates an updated combined affine transformation $\mathbf{A}_{\text{c}}^k$; and 3) *Spatial Transformation Layer* transforms the extracted brain image $\mathbf{E}^M$ using $\mathbf{A}_{\text{c}}^k$ to produce the warped image $\mathbf{W}^k$. The output of this module is the warpped brain image $\mathbf{W}^N$ at the final stage.

3.2.1 Registration Network: f_r. The registration network $f_r(\cdot, \cdot)$ is designed to gradually transform the extracted brain image to maximize its similarity with the target image. At each stage k, it predicts a current affine transformation $\mathbf{A}_{\text{i}}^k$ relying only on the previous warped image $\mathbf{W}^{k-1}$ and the target image $\mathbf{T}$. Following a similar approach to the extraction network f_e in Section 3.1.1, we adopt a 3D CNN based encoder to learn $f_r(\cdot, \cdot)$ and a shared weight design to utilize $f_r(\cdot, \cdot)$ repetitively across stages with the same parameters. It can be formalized as:

$$\mathbf{A}_{\text{i}}^k = f_r\left(\mathbf{W}^{k-1}, \mathbf{T}\right), \tag{7}$$

where $\mathbf{A}_{\text{i}}^k$ is the output affine transformation of the k-th stage for $k = [1, \cdots, N]$ and $\mathbf{W}^0 = \mathbf{E}^M$.

3.2.2 Composition Layer: COMP. In each stage k, after obtaining the current affine transformation $\mathbf{A}_{\text{i}}^k$ from $f_r(\cdot, \cdot)$, we would combine all previous transformation:

$$\mathbf{A}_{\text{c}}^k = \mathbf{A}_{\text{i}}^k \cdot \mathbf{A}_{\text{c}}^{k-1}, \tag{8}$$

where $\cdot$ is matrix product. When $k = 1$, the initial affine transformation $\mathbf{A}_{\text{c}}^0$ is set to be an identity matrix representing no displacement. This layer serves as a bridge between the registration network and the spatial transformation layer. As such, it enables the transformation to be directly applied to the final extracted brain image $\mathbf{E}^M$ to avoid image sharpness loss caused by multiple interpolations.

3.2.3 Spatial Transformation Layer. An important step towards image registration is to reconstruct the warped image $\mathbf{W}^k$ from the extracted brain image $\mathbf{E}^M$ by affine transformation operator. Based on the combined transformation $\mathbf{A}_{\text{c}}^k$, we introduce a spatial transformation layer to resample the voxels into a uniform grid on the extracted image to acquire the warped image through $\mathbf{W}^k = \mathcal{T}(\mathbf{E}^M, \mathbf{A}_{\text{c}}^k)$. According to the definition of affine transformation operator in Eq. (1), we have

$$\mathbf{W}_{xyz}^k = \mathbf{E}_{x'y'z'}^M \ , \tag{9}$$

where $[x', y', z', 1]^\top = \mathbf{A}_{\text{c}}^k [x, y, z, 1]^\top$.

To ensure the success of gradient propagation in this process, we use a differentiable transformation based on trilinear interpolation introduced by [11]. That is,

$$\begin{aligned} \mathbf{W}_{xyz}^k = \sum_{o=1}^{W} \sum_{p=1}^{H} \sum_{q=1}^{D} \mathbf{E}_{opq}^M \cdot \max(0, 1 - |x' - o|) \\ \cdot \max(0, 1 - |y' - p|) \cdot \max(0, 1 - |z' - q|). \end{aligned} \tag{10}$$

Notice that Eq. (9) always performs transformation on the extracted image $\mathbf{E}^M$ instead of the previous warped images. Therefore, only one interpolation is required to produce the final warped brain image $\mathbf{W}^N$, which better preserves the sharpness of $\mathbf{W}^N$.

3.3 Unsupervised End-to-End Training

We train our ERNet model in an unsupervised setting by minimizing the following objective function

$$\min_{\mathbf{M}^1, \cdots, \mathbf{M}^M, \mathbf{A}_{\text{c}}^N} \mathcal{L}_{\text{sim}}\left(\mathbf{W}^N, \mathbf{T}\right) + \sum_{j=1}^{M} \lambda \mathcal{R}(\mathbf{M}^j), \tag{11}$$

where $\mathbf{W}^N = \mathcal{T}(\mathbf{E}^M, \mathbf{A}_{\text{c}}^N)$ and $\mathcal{L}_{\text{sim}}(\cdot, \cdot)$ is a loss function measuring the similarity between the final warped image $\mathbf{W}^N$ and the target image $\mathbf{T}$. Here we use the popular negative local cross-correlation loss, which is robust to voxel intensity variations often found across scans and datasets [4]. $\mathcal{R}(\cdot)$ is the regularization term for brain masks, and λ is a regularization parameter. Since the brain region is one connected entity, we would like our predicted brain masks to have the same properties across all stages. To put it more formally, if we view the brain mask as a 3D tensor with 6-connectivity, we would like it to have exactly one connected component. Though it is possible to count the number of connected components in a mask, such practice would be time-consuming ($O(W \times H \times D)$ by BFS algorithm) and not differentiable. For the purpose of both effective and efficient estimation, we use the ℓ_2-norm of the first-order derivative of $\mathbf{M}^j$ as the regularization term:

$$\mathcal{R}(\mathbf{M}^j) = \sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{z=1}^{D} \|\nabla \mathbf{M}_{xyz}^j\|^2. \tag{12}$$

This regularization term measures edge strength, *i.e.*, the likelihood of a voxel to be an edge voxel. By minimizing the regularization term, we can suppress the occurrence of edges, which in term suppress additional connected components. Specifically, we approximate the first-order derivative by measuring differences between neighboring voxels. For $\nabla \mathbf{M}_{xyz}^j = (\frac{\partial \mathbf{M}_{xyz}^j}{\partial x}, \frac{\partial \mathbf{M}_{xyz}^j}{\partial y}, \frac{\partial \mathbf{M}_{xyz}^j}{\partial z})$,

we have $\frac{\partial \mathbf{M}^j_{xyz}}{\partial x} \approx \mathbf{M}^j_{(x+1)yz} - \mathbf{M}^j_{xyz}$. The approximation of $\frac{\partial \mathbf{M}^j_{xyz}}{\partial y}$ and $\frac{\partial \mathbf{M}^j_{xyz}}{\partial z}$ follows.

Benefiting from the differentiability of each component of this design, our model can be cooperatively and progressively optimized across each stage in an end-to-end manner. Such a training scheme allows us to find a joint optimal solution to the collective brain extraction and registration task.

4 EXPERIMENTS

4.1 Datasets

We evaluate the effectiveness of our proposed method on three different real-world 3D brain MRI datasets: 1) *LPBA40* [21] consists of 40 raw T1-weighted 3D brain MRI scans along with their brain masks. It also provides the corresponding segmentation ground truth of 56 anatomical structures; 2) *CC-359* [24] consists of 359 raw T1-weighted 3D brain MRI scans and the corresponding brain masks. It also contains the labeled white matter as the ground truth; 3) *IBSR* [16] provides 18 raw T1-weighted 3D brain MRI scans along with the corresponding manually segmentation results. Due to the small sample size, We use this dataset only to test the model trained on CC359. The brain mask and anatomical segmentations are used to evaluate the accuracy of extraction and registration, respectively. Datasets are splited into training, validation, and test sets, respectively. For more details, please refer to Appendix A.

4.2 Compared Methods

We compare our ERNet with several representative brain extraction and registration methods, as shown in Table 1. To the best of our knowledge, there is no existing solution that can perform the brain extraction and registration simultaneously under an unsupervised setting. Therefore, we designed two-stage pipelines for comparison using the following brain extraction and registration methods.

• *Brain Extraction Tool (BET)* [22]: This is a skull stripping method included in FSL package. It uses a deformable approach to fit the brain surface by applying locally adaptive set models.

• *3dSkullStrip* [6]: This is a modified version of BET that is included in the AFNI package. It performs skull stripping based on the expansion paradigm of the spherical surface.

• *Brain Surface Extractor (BSE)* [20]: It extracts the brain region based on morphological operations and edge detection, which employs anisotropic diffusion filtering and a Marr Hildreth edge detector for brain boundary identification.

• *FMRIB's Linear Image Registration Tool (FLIRT)* [12]: This is a fully automated affine brain image registration tool in FSL package.

• *Advanced Normalization Tools (ANTs)* [2]: It is considered a state-of-the-art medical image registration toolkit. Here we utilize affine transformation model and cross-correlation metric for registration.

• *VoxelMorph (VM)* [4]: This unsupervised, deformable image registration method employs a neural network to predict the nonlinear transformation between images.

• *Cascaded Registration Networks (CRN)* [27]: It is an unsupervised multi-stage image registration method. In different stages, the source image is repeatedly deformed to align with a target image.

• *ERNet*: This is our proposed model which consists of both extraction and registration modules in an end-to-end manner.

Table 1: Summary of compared methods.

Methods	Extraction	Registration	Collaborative	Deep learning
BET [22]	✓	✗	✗	✗
3dSkullStrip [6]	✓	✗	✗	✗
BSE [20]	✓	✗	✗	✗
FLIRT [12]	✗	✓	✗	✗
ANTs [2]	✗	✓	✗	✗
VM [4]	✗	✓	✗	✓
CRN [27]	✗	✓	✗	✓
ERNet w/o Ext	✗	✓	✗	✓
ERNet (ours)	✓	✓	✓	✓

• *ERNet w/o Ext*: This is a variant of ERNet where we remove the extraction modules. Here it is a registration method only.

4.3 Evaluation Metrics

Our defined problem aims to identify the brain region within the source image and align the extracted cerebral tissues to the target image simultaneously. Thus, we evaluate the accuracy of extraction and registration to show the performance of our proposed method and compared methods as follows:

4.3.1 Extraction Performance. The brain MRI datasets contain the brain mask ground truth, which is the label of brain tissue in the source image. To evaluate the extraction accuracy, we measure the volume overlap of brain masks by Dice score, which can be formulated as:

$$\text{Dice}_{\text{ext}} = 2 \cdot \frac{|\hat{\mathbf{M}} \cap \mathbf{M}|}{|\hat{\mathbf{M}}| + |\mathbf{M}|}, \tag{13}$$

where $\hat{\mathbf{M}}$ is the predicted brain mask and $\mathbf{M}$ denotes the corresponding ground truth. If $\hat{\mathbf{M}}$ represents accurate extraction, we expect the non-zero regions in $\hat{\mathbf{M}}$ and $\mathbf{M}$ to overlap well.

4.3.2 Registration Performance. We evaluate the registration accuracy by measuring the volume overlap of anatomical segmentations, which are the location labels of different tissues in the brain MRI image. If two images are well aligned, then their corresponding anatomical structures should overlap with each other. Likewise, the Dice score can evaluate the performance of registration, as follows:

$$\text{Dice}_{\text{reg}} = 2 \cdot \frac{|\mathbf{G}_{\text{w}} \cap \mathbf{G}_{\text{t}}|}{|\mathbf{G}_{\text{w}}| + |\mathbf{G}_{\text{t}}|} \tag{14}$$

where $\mathbf{G}_{\text{w}} = \mathcal{T}(\mathbf{G}_{\text{s}}, \hat{\mathbf{a}})$. $\mathbf{G}_{\text{s}}$, $\mathbf{G}_{\text{t}}$ and $\mathbf{G}_{\text{w}}$ are anatomical structural segmentation of the source image, target image and warped image, respectively. $\hat{\mathbf{a}}$ is the predicted affine transformation parameters. A dice score of 1 means that the corresponding structures are well aligned after registration, a score of 0 indicates that there is no overlap. If the image contains multiple labeled anatomical structures, the final score is the average of the dice score of the each structure.

4.4 Experimental Results

We compare our ERNet with the baseline models regarding extraction and registration accuracy. For each task, we quantify the performance by its corresponding dice score. We also report the running time for each method to complete both tasks. Across all these metrics, we show that ERNet not only consistently achieves better extraction and registration performance than other alternatives, but is also more time-efficient.

Table 2: Results for brain extraction and registration in different datasets. The results are reported as performance(mean ± std) of extraction and registration of each compared method. The running time is reported as the average processing time for each image in its corresponding task. "↑" point out "the larger the better" and "↓" point out "the smaller the better".

Methods		Datasets						Running Time	
		LPBA40		CC359		IBSR			
Extraction	Registration	Extraction $Dice_{ext}$ ↑	Registration $Dice_{reg}$ ↑	Extraction $Dice_{ext}$ ↑	Registration $Dice_{reg}$ ↑	Extraction $Dice_{ext}$ ↑	Registration $Dice_{reg}$ ↑	Extraction Sec ↓	Registration Sec ↓
BET [22]	FLIRT [12]	0.935 ± 0.028	0.606 ± 0.026	0.811 ± 0.087	0.747 ± 0.060	0.911 ± 0.038	0.798 ± 0.010	2.45 (CPU)	4.57 (CPU)
3dSkullStrip [6]	FLIRT [12]	0.902 ± 0.032	0.594 ± 0.018	0.849 ± 0.037	0.790 ± 0.034	0.869 ± 0.039	0.787 ± 0.020	178.56 (CPU)	4.64 (CPU)
BSE [20]	FLIRT [12]	0.938 ± 0.022	0.614 ± 0.010	0.846 ± 0.112	0.801 ± 0.021	0.873 ± 0.064	0.798 ± 0.015	4.75 (CPU)	4.35 (CPU)
BET [22]	ANTs [2]	0.935 ± 0.028	0.609 ± 0.025	0.811 ± 0.087	0.764 ± 0.053	0.911 ± 0.038	0.796 ± 0.014	2.45 (CPU)	2.76 (CPU)
3dSkullStrip [6]	ANTs [2]	0.902 ± 0.032	0.616 ± 0.016	0.849 ± 0.037	0.807 ± 0.027	0.869 ± 0.039	0.794 ± 0.017	178.56 (CPU)	2.89 (CPU)
BSE [20]	ANTs [2]	0.938 ± 0.022	0.616 ± 0.013	0.846 ± 0.112	0.796 ± 0.027	0.873 ± 0.064	0.797 ± 0.017	4.75 (CPU)	2.52 (CPU)
BET [22]	VM [4]	0.935 ± 0.028	0.488 ± 0.092	0.811 ± 0.087	0.811 ± 0.015	0.911 ± 0.038	0.792 ± 0.010	2.45 (CPU)	0.02 (GPU)
3dSkullStrip [6]	VM [4]	0.902 ± 0.032	0.479 ± 0.094	0.849 ± 0.037	0.809 ± 0.018	0.869 ± 0.039	0.785 ± 0.014	178.56 (CPU)	0.02 (GPU)
BSE [20]	VM [4]	0.938 ± 0.022	0.512 ± 0.065	0.846 ± 0.112	0.810 ± 0.017	0.873 ± 0.064	0.794 ± 0.011	4.75 (CPU)	0.02 (GPU)
BET [22]	CRN [27]	0.935 ± 0.028	0.556 ± 0.046	0.811 ± 0.087	0.815 ± 0.008	0.911 ± 0.038	0.799 ± 0.017	2.45 (CPU)	0.10 (GPU)
3dSkullStrip [6]	CRN [27]	0.902 ± 0.032	0.528 ± 0.056	0.849 ± 0.037	0.813 ± 0.009	0.869 ± 0.039	0.796 ± 0.014	178.56 (CPU)	0.10 (GPU)
BSE [20]	CRN [27]	0.938 ± 0.022	0.547 ± 0.071	0.846 ± 0.112	0.812 ± 0.011	0.873 ± 0.064	0.799 ± 0.011	4.75 (CPU)	0.10 (GPU)
BET [22]	**ERNet** (w/o Ext)	0.935 ± 0.028	0.616 ± 0.021	0.811 ± 0.087	0.804 ± 0.017	0.911 ± 0.038	0.798 ± 0.011	2.45 (CPU)	0.04 (GPU)
3dSkullStrip [6]	**ERNet** (w/o Ext)	0.902 ± 0.032	0.606 ± 0.006	0.849 ± 0.037	0.815 ± 0.007	0.869 ± 0.039	0.791 ± 0.014	178.56 (CPU)	0.04 (GPU)
BSE [20]	**ERNet** (w/o Ext)	0.938 ± 0.022	0.613 ± 0.012	0.846 ± 0.112	0.807 ± 0.018	0.873 ± 0.064	0.795 ± 0.013	4.75 (CPU)	0.04 (GPU)
ERNet (ours)		**0.946 ± 0.009**	**0.626 ± 0.008**	**0.938 ± 0.008**	**0.818 ± 0.006**	**0.916 ± 0.013**	**0.800 ± 0.011**	**0.11 (GPU)**	

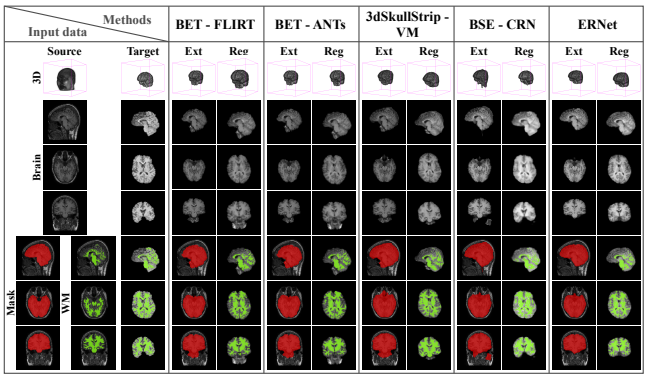


Figure 5: Visual comparisons for brain extraction and registration task. We render a 3D visualization of the image and display middle slice in three different planes: sagittal, axial and coronal. The right side contains the source and target images and their corresponding ground truth labels. We show extraction and registration results of each method and its corresponding predictive labels used for performance evaluations. To evaluate the brain extraction task, a predicted brain Mask (red) should coincide as much as possible with the ground truth brain Mask (red) of the source image. Likewise, in the brain registration task, a warped White Matter (green) should well-overlap with the White Matter (green) of the target image.

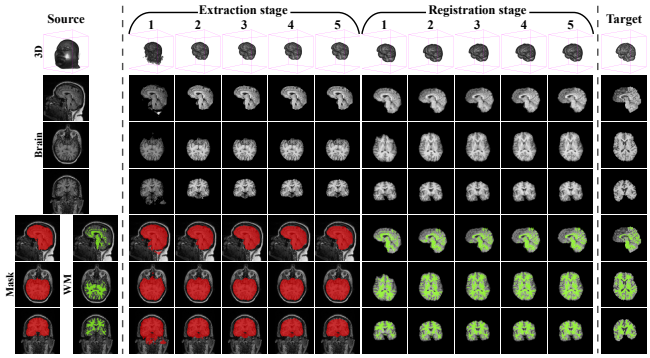


Figure 6: A demonstration of the multi-stage extraction and registration process. The example shows the extraction and registration result at each stage using an ERNet with 5-stage extraction and 5-stage registration.

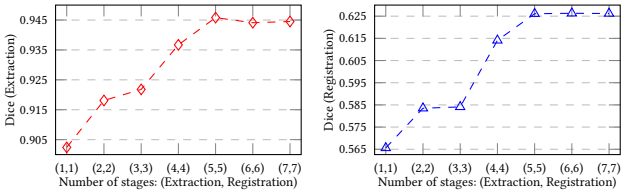
4.4.1 Experiment Setting. We split the datasets into training, validation, and test sets as described in Appendix A. Note that the IBSR dataset is used for test only. The training set is used to learn model parameters and the validation set is used to evaluate the performance of hyperparameter settings (e.g., the number of stages or smoothing regularization term). We use the test set only once to report the final evaluation results for each model.

4.4.2 Extraction and Registration Results. Table 2 summarized the results of designed two-stage pipelines and proposed ERNet in both extraction and registration tasks. Based on the global competition in three datasets, ERNet outperforms the existing methods in all metrics. For the extraction task, the performance of ERNet is superior to that of the other compared methods, especially on the CC359 dataset. Specifically, we observed a gain in extraction dice score up to 10.5% compared to the best extraction method 3dSkullStrip. Besides, ERNet is more robust in the extraction task than other alternatives, as it performs consistently well and obtains the smallest standard deviation across all datasets.

When observing registration performance, once again, ERNet outperforms all other methods across all datasets. Most notably, we find that the registration result of almost every method is bounded by the result of its corresponding extraction method. This proves that the accuracy of extraction has a significant impact on the quality of the subsequent registration task. ERNet captures this property to deliver an improved result via end-to-end collective learning. In addition, the overall performance of ERNet w/o Ext, the variant of ERNet, is slightly worse than that of ERNet, indicating

Table 3: Influence of number of stages on extraction and registration performance. "↑" point out "the larger the better".

Number of Stages		Extraction	Registration
Extraction	Registration	Dice_{ext} ↑	Dice_{reg} ↑
0	0	0.216 ± 0.018	0.252 ± 0.158
0	1	0.216 ± 0.018	0.269 ± 0.101
0	5	0.216 ± 0.018	0.264 ± 0.103
1	0	0.040 ± 0.026	0.007 ± 0.007
1	1	0.902 ± 0.010	0.566 ± 0.040
1	5	0.927 ± 0.007	0.604 ± 0.017
5	0	0.095 ± 0.040	0.024 ± 0.016
5	1	0.919 ± 0.010	0.550 ± 0.033
5	**5**	**0.946 ± 0.009**	**0.626 ± 0.008**



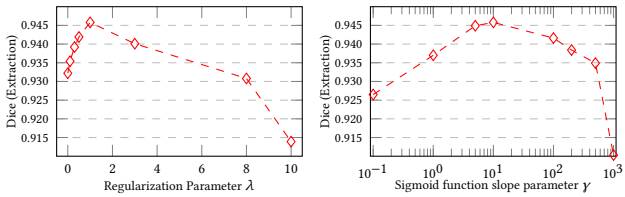
(a) Extraction Accuracy (b) Registration Accuracy

Figure 7: Performance of ERNet with different number of extraction and registration stages.

that the extraction network in ERNet is crucial and beneficial to improve both extraction and registration accuracy.

4.4.3 Running Efficiency. To evaluate the efficiency of ERNet, we compare its running time with other baselines. The run time is measured on the same machine with a Intel® Xeon® E5-2667 v4 CPU and an NVIDIA Tesla V100 GPU. As shown in Table 2, ERNet is roughly 20 to 200 times faster than existing methods. This is because ERNet can perform both extraction and registration tasks end-to-end on the same device efficiently. In contrast, other alternatives handle these two tasks separately using two-step approaches, which is time-consuming. No GPU implementations of BET, 3dSkullStrip, BSE, FLIRT and ANTs are made available [2, 6, 12, 20, 22].

4.4.4 Qualitative Analysis. Figure 5 shows visualized brain extraction and registration results of our ERNet compared with other two-step approaches on the LPBA40 test set. Upon inspection, we can see that extracted image of ERNet is more accurate than those of BET, 3dSkullStrip, and BSE. The brain mask predicted by ERNet overlaps best with the ground truth mask of the source image, while the masks predicted by other extraction methods contain obvious non-brain tissues. In terms of registration results, ERNet also clearly outperforms other compared methods. The final registered image of ERNet is more similar to the target image than that of alternatives. Most notably, inaccurate extraction results with non-brain tissue also appear in the following registration results and hamper the final performance. This supports our claim that a failed extraction can propagate to the following registration task, rendering an irreversible error. Furthermore, we demonstrate the intermediate extraction and registration results of ERNet in Figure 6 using a test sample of CC359 dataset. It is clear that brain tissue is progressively extracted from the source image with the help of a multi-stage design. Likewise, the extracted image is transformed multiple times to align with the target image incrementally.



(a) Varying of λ (b) Varying of γ

Figure 8: Effect of varying the mask smoothing regularization parameter λ and sigmoid function slope parameter γ.

4.4.5 Influence of Parameters. We study three important hyperparameters of our ERNet, *i.e.*, the number of stages for extraction and registration, the value of mask smoothing regularization parameter λ and the value of Sigmoid function slope parameter γ.

Number of stages of extraction and registration. In our multi-stage design, the number of stages corresponds to network depth and the number of repeated extraction and registration. In other words, more stages mean more refinements of the stripping and alignment, which is usually beneficial to the improvement of the extraction and registration accuracy. As shown in Table 3, the ablation study demonstrates that both the extraction network and the registration network of ERNet are essential, and removing either one of them causes invalid results. As illustrated in Figure 7, we vary the number of stages of extraction and registration to learn their effects. The results indicate that the performance of extraction and alignment improves with additional stages. This supports the idea that a multi-stage design yields improved overall performance.

Mask smoothing parameter λ. As mentioned in Section 3.3, we introduce an regularization term to smooth the predicted masks. To show the effectiveness of the smoothing regularization, we vary different values of the smoothing parameter λ as shown in Figure 8(a). The optimal Dice score occurs when λ =1, while the performance gets worse as the λ increases more or decreases. This indicates that our ERNet benefits from mask smoothing regularization.

Sigmoid function slope parameter γ. To show the effectiveness of the shifted sigmoid function we introduced in Eq. (5), we evaluate the performance of our model under different γ settings in a range from 10^{-1} to 10^3. As shown in Figure 8(b), the model achieves the best performance when $\gamma = 10^1$, which is better than using the standard sigmoid function ($\gamma = 10^0$). As the γ increases to 10^2 and 10^3, the dice score drops significantly because the large flat region of the sigmoid function prevents the error from backpropagation.

5 RELATED WORK

Brain extraction. Over the past decade, myriad methods have been proposed, emphasizing the importance of the brain extraction problem. Smith et al. [22] proposed a deformable model to fit the brain surface using a locally adaptive set model. 3dSkullStrip [6] is a modified version of [22], which uses points lying outside the brain surface to guide the evolution of the mesh. Shattuck et al. [20] employs anisotropic diffusion filtering and a 2D Marr Hildreth edge detector to identify the brain boundary. Apart from methods, several other traditional approaches [8, 10, 19] are also commonly used for brain extraction. However, the aforementioned methods rely heavily on parameter setting and manual quality control, which are time-consuming and labor-intensive. Recently, deep learning-based methods have been introduced for brain extraction due to

their superior capability and extreme speed. Kleesiek et al. [14] proposed a voxel-wise 3D CNN for skull stripping. Hwang et al. [9] suggested that 3D-UNet yields highly competitive results for skull stripping. The above learning-based methods often demand a large amount of adequately labeled data for effective training. However, neuroimage datasets are usually small and expensive to annotate.

Image registration. Traditional image registration methods [1, 2, 12, 18] often try to maximize the similarity between images by iteratively optimizing the transformation parameters, where normalized cross-correlation (NCC) and mutual information (MI), *etc.* , are commonly used as intensity-based similarity measures. However, iteratively optimizing each pair of images tends to face the drawbacks of high computational cost and being trapped in local optima, resulting in failing to yield an efficient and robust registration result. Recently, many deep learning-based methods have been proposed due to their superior computational efficiency and registration performance. Sokooti et al. [23] proposed a multi-scale 3D CNN termed as RegNet to learn the artificially generated displacement vector field (DVF) for 3D chest CT registration. While these methods present competitive results, they are all supervised. In other words, the training procedure is guided by ground truth transformations. In practice, obtaining high-quality ground truth is often expensive in medical imaging. To address this limitation, unsupervised registration methods [4, 27] received much attention and delivered promising results. However, the above methods rely on accurate skull stripping results to perform the registration, where manual visual inspection is required to remove the inaccurate extracted image. Such human involvement is not only time-consuming but also brings in biases. Unlike these works, we pursue unsupervised joint learning for extraction and registration.

6 CONCLUSION

In this paper, we propose a novel unified end-to-end framework, called ERNet, for unsupervised collective extraction and registration. Different from previous work, our proposed method seamlessly integrated two tasks into one system to achieve joint optimization. Specifically, ERNet contains a pair of multi-stage extraction and registration modules. These two modules help each other boost extraction and registration performance simultaneously without any annotation information. Moreover, the multi-stage design allows each task to proceed incrementally, thus refining their respective performance to a better extent. The experimental results demonstrate that ERNet not only outperforms state-of-the-art approaches in both extraction and registration accuracy but is also more robust and time-efficient.

7 ACKNOWLEDGMENTS

Lifang He was supported by Lehigh's accelerator grant S00010293.

REFERENCES

[1] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis* 12, 1 (2008), 26–41.

[2] Brian B Avants, Nick Tustison, and Gang Song. 2009. Advanced normalization tools (ANTS). *Insight j* 2, 365 (2009), 1–35.

[3] Zilong Bai, Peter Walker, Anna Tschiffely, Fei Wang, and Ian Davidson. 2017. Unsupervised network discovery for brain imaging data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 55–64.

[4] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. 2018. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9252–9260.

[5] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1158–1166.

[6] Robert W Cox. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research* 29, 3 (1996), 162–173.

[7] Xin Dai, Xiangnan Kong, Xinyue Liu, John Boaz Lee, and Constance Moore. 2020. Dual-Attention Recurrent Networks for Affine Registration of Neuroimaging Data. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 379–387.

[8] Simon F Eskildsen, Pierrick Coupé, Vladimir Fonov, José V Manjón, Kelvin K Leung, Nicolas Guizard, Shafik N Wassef, Lasse Riis Østergaard, D Louis Collins, Alzheimer's Disease Neuroimaging Initiative, et al. 2012. BEaST: brain extraction based on nonlocal segmentation technique. *NeuroImage* 59, 3 (2012), 2362–2373.

[9] Hyunho Hwang, Hafiz Zia Ur Rehman, and Sungon Lee. 2019. 3D U-Net for skull stripping in brain MRI. *Applied Sciences* 9, 3 (2019), 569.

[10] Juan Eugenio Iglesias, Cheng-Yi Liu, Paul M Thompson, and Zhuowen Tu. 2011. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE transactions on medical imaging* 30, 9 (2011), 1617–1634.

[11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. *Advances in neural information processing systems* 28 (2015).

[12] Mark Jenkinson and Stephen Smith. 2001. A global optimisation method for robust affine registration of brain images. *Medical image analysis* 5, 2 (2001), 143–156.

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[14] Jens Kleesiek, Gregor Urban, Alexander Hubert, Daniel Schwarz, Klaus Maier-Hein, Martin Bendszus, and Armin Biller. 2016. Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *NeuroImage* 129 (2016), 460–469.

[15] Oeslle Lucena, Roberto Souza, Leticia Rittner, Richard Frayne, and Roberto Lotufo. 2019. Convolutional neural networks for skull-stripping in brain MR imaging using silver standard masks. *Artificial intelligence in medicine* 98 (2019), 48–58.

[16] Torsten Rohlfing. 2011. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE transactions on medical imaging* 31, 2 (2011), 153–163.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[18] Ziad S Saad, Daniel R Glen, Gang Chen, Michael S Beauchamp, Rutvik Desai, and Robert W Cox. 2009. A new method for improving functional-to-structural MRI alignment using local Pearson correlation. *Neuroimage* 44, 3 (2009), 839–848.

[19] Florent Ségonne, Anders M Dale, Evelina Busa, Maureen Glessner, David Salat, Horst K Hahn, and Bruce Fischl. 2004. A hybrid approach to the skull stripping problem in MRI. *Neuroimage* 22, 3 (2004), 1060–1075.

[20] David W Shattuck and Richard M Leahy. 2002. BrainSuite: an automated cortical surface identification tool. *Medical image analysis* 6, 2 (2002), 129–142.

[21] David W Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L Narr, Russell A Poldrack, Robert M Bilder, and Arthur W Toga. 2008. Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage* 39, 3 (2008), 1064–1080.

[22] Stephen M Smith. 2002. Fast robust automated brain extraction. *Human brain mapping* 17, 3 (2002), 143–155.

[23] Hessam Sokooti, Bob De Vos, Floris Berendsen, Boudewijn PF Lelieveldt, Ivana Išgum, and Marius Staring. 2017. Nonrigid image registration using multi-scale 3D convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 232–239.

[24] Roberto Souza, Oeslle Lucena, Julia Garrafa, David Gobbi, Marina Saluzzi, Simone Appenzeller, Letícia Rittner, Richard Frayne, and Roberto Lotufo. 2018. An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage* 170 (2018), 482–494.

[25] Liang Sun, Rinkal Patel, Jun Liu, Kewei Chen, Teresa Wu, Jing Li, Eric Reiman, and Jieping Ye. 2009. Mining brain region connectivity for alzheimer's disease study via sparse inverse covariance estimation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1335–1344.

[26] Shen Wang, Lifang He, Bokai Cao, Chun-Ta Lu, Philip S Yu, and Ann B Ragin. 2017. Structural deep brain network mining. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 475–484.

[27] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. 2019. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10600–10610.

A APPENDIX FOR REPRODUCIBILITY

This section provides more details to support the reproducibility of the results in this paper. We have released our code and data publicly available at https://github.com/ERNetERNet/ERNet.

A.1 Details of Data Preprocessing

We evaluate the effectiveness of our proposed method on three different public brain MRI datasets, LPBA40, CC-359 and IBSR. Table 4 summarizes the properties of the datasets.

Table 4: Summary of datasets

	LPBA40	CC359	IBSR
Raw size	$256 \times 124 \times 256$	$171 \times 256 \times 256$	$256 \times 256 \times 128$
Cropped size	$96 \times 96 \times 96$	$96 \times 96 \times 96$	$96 \times 96 \times 96$
Training	30	298	-
Validation	5	30	-
Test	4	30	18

• *LONI Probabilistic Brain Atlas (LPBA40)* [21]: This dataset consists of 40 raw T1-weighted 3D brain MRI scans (40 different patients) along with their brain masks and the corresponding segmentation ground truth of 56 anatomical structures. The brain mask and anatomical segmentations are used to evaluate the accuracy of extraction and registration, respectively. Same to [4, 27], we focus on atlas-based registration, where the first scan is the target image and the remaining scans need to align with it. Among the 39 scans, we use 30, 5, and 4 scans for training, validation, and test, respectively. All scans are resized to $96 \times 96 \times 96$ after cropping.

• *Calgary-Campinas-359 (CC-359)* [24]: This dataset consists of 359 raw T1-weighted 3D brain MRI scans (359 different patients) and the corresponding brain masks. It also contains the labeled white matter as the ground truth. We use the brain masks and white-matter masks to evaluate the accuracy of extraction and registration, respectively. Same to LPBA40, we concentrate on atlas-based registration and split CC359 into 298, 30, and 30 scans for training, validation, and test sets. All scans are resized to $96 \times 96 \times 96$ after cropping.

• *Internet Brain Segmentation Repository (IBSR)* [16]: This dataset provides 18 raw T1-weighted 3D brain MRI scans (18 different patients) along with the corresponding segmentation results. We merge all segmentation results to construct the brain mask. Due to the small sample size, We use this dataset only to test the model trained on CC359. Thus, all 18 scans need to align with the first scan of CC359. All scans are resized to $96 \times 96 \times 96$ after cropping.

A.2 Details Settings of ERNet

Training settings of ERNet. Our experiments are running on Red Hat Enterprise Linux 7.3 with an Intel® Xeon® E5-2667 v4 CPU and an NVIDIA Tesla V100 GPU. The implementation of neural networks is based on PyTorch 1.7.1. During the training process, we apply batch gradient descent with each training batch consists of one pair of images to address GPU memory limitation. Models are optimized using Adam optimizer [13] with a learning rate of 1×10^{-6}. We also leverage the image augmentation technique to expand the datasets, where a transformation with random translation, rotation, and scaling is applied to source images. We detail it in Table 5.

Parameters settings of ERNet. Since we design a multi-stage extraction and registration network, the extraction and registration stages are both set to 5 in this work. The best mask smoothing parameter λ in Eq. (11) is 1, and the best sigmoid function slope parameter γ in Eq. (5) is 10^1. The extraction network contains 10 convolutional layers with 16, 32, 32, 64, 64, 64, 32, 32, 32 and 16 filters. The registration network adopt 3D CNNs and fully-connected layers to map the input to the dimension of 1×12. It contains 6 convolutional layers with 16, 32, 64, 128, 256 and 512 filters. The output dimensions of the 2 fully-connected layers are 128 and 12.

Table 5: Range of random transformation.

Datasets	Transformation		
	Translation (Voxels)	Rotation (Degree)	Scale (Times)
LPBA40	± 5	± 5	0.98 ~ 1.02
CC359	± 3	± 3	0.99 ~ 1.01

A.3 Settings of Baselines

Brain Extraction Tool (BET) [22]: This is a skull stripping method included in FSL package. It uses a deformable approach to fit the brain surface by applying locally adaptive set models. The command we use for BET is `bet <input> <output> -f 0.5 -g 0 -m`, where `f` and `g` are fractional intensity threshold and gradient in fractional intensity threshold, respectively. We set them to default values.

3dSkullStrip [6]: This is a modified version of BET that is included in the AFNI package. It performs skull stripping based on the expansion paradigm of the spherical surface. The command we use for 3dSkullStrip is `3dSkullStrip -input <input> -prefix <output> -mask_vol -fac 1000`. `fac` is set to the default value.

Brain Surface Extractor (BSE) [20]: It extracts the brain region based on morphological operations and edge detection, which employs anisotropic diffusion filtering and a Marr Hildreth edge detector for brain boundary identification. The command we use for BSE is `bse -i <input> -o <output> –mask <mask> -p –trim –auto –timer` . Hyperparameters are set to default values.

FMRIB's Linear Image Registration Tool (FLIRT) [12]: This is a fully automated affine brain image registration tool in FSL package. The command we use for FLIRT is `flirt -in <source> -ref <target> -out <output> -omat <output parameter> -bins 256 -cost corratio -searchrx -90 90 -searchry -90 90 -searchrz -90 90 -dof 12 -interp trilinear`.

Advanced Normalization Tools (ANTs) [2]: It is a state-of-the-art medical image registration toolkit. Here we utilize affine transformation model and cross-correlation metric for registration.

VoxelMorph (VM) [4]: This unsupervised, deformable image registration method employs a neural network to predict the nonlinear transformation between images. For network architectures, we use the latest version, VoxelMorph-2, and configure 10 convolutional layers with 16, 32, 32, 64, 64, 64, 32, 32, 32 and 16 filters. The ratio of deformation regularization is set to 10.

Cascaded Registration Networks (CRN) [27]: It is an unsupervised multi-stage registration method. In different stages, the source image is repeatedly deformed to align with a target image. Same to ERNet, the number of stages is set to 5. In each stage, we configure 10 convolutional layers with 16, 32, 32, 64, 64, 64, 32, 32, 32 and 16 filters. The ratio of deformation regularization is set to 10.

ERNet w/o Ext: This is a variant of ERNet where we remove the extraction modules. Here it is a registration method only.