UniBrain: A Unified Model for End-to-End Brain Imaging Analysis

Anonymous Author(s)

ABSTRACT

Brain imaging analysis has emerged as a prevalent paradigm in neuroscience. The goal is to extract and analyze information from neuroimaging data, aiming to reveal the brain's structural and functional mysteries. However, efficiently discovering knowledge from neuroimaging data is impeded by complex and labor-intensive processing steps (*e.g.*, brain extraction, registration, segmentation, parcellation, network generation and classification). Conventionally, such steps are performed separately in a supervised manner, where their performance heavily relies on the quantity of training data and extensive visual inspection performed by experts for error correction. These processes are particularly burdensome for high-dimensional neuroimages (*e.g.*, 3D MRI), where acquiring detailed voxel-level annotations and conducting manual quality control are both expensive and time-consuming, presenting substantial obstacles in many medical studies. In this paper, we study the problem of end-to-end brain imaging analysis, leveraging only low-cost labels (*i.e.*, classification and extraction labels) and one labeled template image (*a.k.a.* atlas) for training guidance. We propose a unified end-to-end framework, called UniBrain, to jointly optimize all processing steps, allowing feedback among them. Specifically, UniBrain consists of interconnected modules to learn extraction mask, transformation, segmentation mask, parcellation mask, brain network and classification label, with all modules being mutually reinforced by collective learning. Experimental results on real-world datasets demonstrate that our proposed method excels in brain extraction, registration, segmentation, parcellation and classification tasks.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Computing methodologies** → **3D imaging**; Supervised learning by classification.

KEYWORDS

brain network, classification, brain extraction, registration, segmentation, parcellation, end-to-end learning, unified model

ACM Reference Format:

Anonymous Author(s). 2024. UniBrain: A Unified Model for End-to-End Brain Imaging Analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 15 pages. https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

KDD'24, August 25–29, 2024, Barcelona, Spain

ACM ISBN XXX-X-XXXX-XXXX-X/XX/XX…$15.00
https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

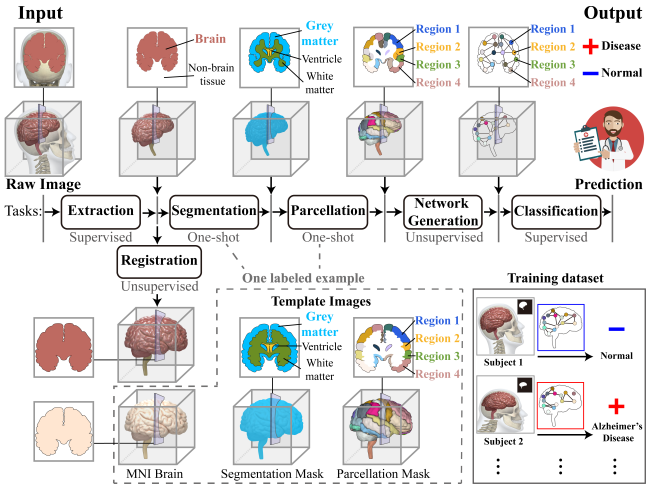


Figure 1: The problem of end-to-end brain imaging analysis. Given a set of raw images, each with its corresponding extraction mask and diagnosis label, along with a standard template brain image and its segmentation and parcellation labels, the goal is to train a model to perform brain extraction, registration, segmentation, parcellation, network generation and classification tasks simultaneously.

1 INTRODUCTION

Background. Brain imaging analysis has become a popular paradigm in the field of neuroscience with widespread applications, such as anatomical and functional studies [4, 10, 47, 69, 72], multi-modality fusion [7, 41], diagnostic assistance [21, 61], and clinical interpretation [36, 37, 71]. The analysis typically includes multiple processing tasks such as brain extraction (*a.k.a.* skull stripping), registration, segmentation, parcellation, network generation, and classification. Brain extraction involves removing non-brain tissues (*e.g.*, skull, dura, and scalp) from an imaging scan; registration aligns the extracted brain with a standard brain template; segmentation labels the brain tissue types (*e.g.*, gray matter, white matter, cerebrospinal fluid) in the raw imaging scan; parcellation further subdivides the brain into smaller, specific regions (*e.g.*, frontal lobe, temporal lobe) based on functional, structural, or connectivity patterns; network generation constructs a brain network based on the connectivity among parcellated regions; finally, the classification focuses making clinical predictions. These tasks are preliminary but essential in many neuroimaging studies. For example, in brain functional and anatomical analysis, extraction and registration assist in eliminating interference from non-cerebral tissues, imaging modalities, and differing viewpoints. Segmentation and parcellation enable the measurement of anatomical and functional variations based on specific tissues and regions. Network generation and classification facilitate graph theory for interpretation and prediction. In Alzheimer's disease diagnosis, the brain across subjects first needs to be extracted from raw brain imaging scans and then aligned with a standard template to eliminate inter-individual variations. Then, the intra-individual structural lesions (*e.g.*, brain atrophy) across

different pathological stages need to be monitored in anatomical analysis (*e.g.*, identify the corresponding brain tissue and measure its volume alteration). Meanwhile, group-level functional disorders (*e.g.*, frontal lobe decline) need to be discovered and interpreted by capturing the region connectivity through brain network analysis. Performing these comprehensive analyses is instrumental in aiding physicians to make thorough and precise diagnoses.

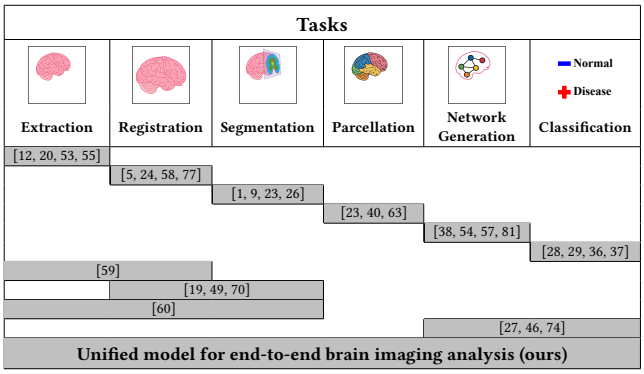
State-of-the-Art. In the literature, the related tasks in brain imaging analysis have been extensively studied, as shown in Table 1. Conventional methods primarily focus on designing methods for brain extraction [34, 43], registration [13, 56], segmentation [1, 9, 26], parcellation [40, 63], network generation [54, 81] and classification [28, 29, 36, 37] separately under supervised settings. However, in brain imaging studies, the collection of voxel-level annotations, transformations between images, and task-specific brain networks often prove to be expensive, as it demands extensive expertise, effort, and time to produce accurate labels, especially for high-dimensional neuroimaging data, *e.g.*, 3D MRI. To reduce this high demand for annotations, recent works have utilized automatic extraction tools [12, 52, 53, 55], unsupervised registration models [5, 24, 58, 77], inverse warping [23], and correlation-based metrics [38, 81] for performing extraction, registration, segmentation, parcellation and network generation. Nevertheless, these pipeline-based approaches frequently rely on manual quality control to correct intermediate results before performing subsequent tasks. Conducting such visual inspections is not only time-consuming and labor-intensive but also suffers from intra- and inter-rater variability, thereby impeding the overall efficiency and performance. More recently, joint extraction-registration [59], joint registration-segmentation [19, 49, 70], joint extraction-registration-segmentation [60], and joint network generation-classification [27, 46, 74] have been developed for collective learning. However, partial joint learning overlooks the potential interrelationships among all tasks, which can adversely affect overall performance.

Problem Definition. In this paper, we study the problem of end-to-end brain image analysis, as shown in Figure 1. The objective is to explore how tasks like brain extraction, registration, segmentation, parcellation, network generation, and classification interrelate, aiming to mutually boost their efficacy with minimal labeled data. Specifically, we utilize low-cost labels (*i.e.*, extraction mask, classification label) and only one labeled template (*a.k.a.* atlas) to perform all tasks simultaneously. Notably, it is our goal to avoid using any of the instance-level ground-truth labels of registration, segmentation, parcellation, or network connectivity in model training.

Challenges. Despite its value and significance, the problem of end-to-end brain image analysis has not been studied before and is very challenging due to its unique characteristics listed below:

• *Limited labeled information:* Conventional learning-based methods typically require a substantial collection of ground-truth labels for conducting these tasks. However, acquiring detailed voxel-level labels (*e.g.*, tissue type, region location) and transformation labels in high-dimensional neuroimaging data is notably expensive and time-consuming. We minimize reliance on extensive labeling, using only the necessary label (*i.e.*, classification label) and a relatively low-cost label (*i.e.*, brain extraction masks). Although we provide a template image with its segmentation and parcellation masks (in template image space), the segmentation and parcellation masks (in raw image space) for the raw image are not available. Furthermore, the ground-truth brain network connectivity is also not provided.

Table 1: Related works in brain imaging analysis. The gray boxes represent the scope of tasks supported by the methods in the boxes.



• *Dependencies among all tasks:* Conventional research often treats brain extraction, registration, segmentation, parcellation, network generation and classification tasks separately. However, these tasks exhibit significant interdependencies: 1) effective extraction (*i.e.*, successful removal of non-brain tissue) allows the brain to precisely align to template image, thereby enhancing registration accuracy; 2) precise registration can accurately guide segmentation and parcellation through inverse transformation (*i.e.*, inversely warping template masks to the raw image space); 3) accurate parcellation contributes to consistent region divisions across subjects, which is crucial for generating robust brain networks; 4) networks that precisely capture brain connectivity can substantially improve classification and diagnosis outcomes. Therefore, a holistic approach is essential to manage these interdependencies effectively in an end-to-end framework.

• *Heterogeneous inputs and outputs across tasks:* Developing a unified model for brain imaging analysis is markedly challenging since all tasks have incredibly diverse input and output representation. For example, extraction results in a binary mask that outlines the brain, while registration yields a transformation matrix for image coordinate mapping. Segmentation and parcellation generate multi-class masks to delineate brain tissues and regions. Network generation produces an adjacency matrix to capture brain region connectivity, and classification outputs categorical labels. This heterogeneity makes it very challenging to architect a single end-to-end model for all these tasks.

Proposed Method. To address the challenges outlined above, we introduce a unified framework, UniBrain, the first model for end-to-end brain imaging analysis. UniBrain comprises a group of modules for extraction, registration, segmentation, parcellation, network generation, and classification. The extraction module removes non-brain tissue from the raw image, yielding an extracted image. The registration module aligns the extracted image with a template. Based on the transformation provided by the registration module, the segmentation and parcellation module then inversely warps the template's segmentation and parcellation masks into the raw image space, facilitating learning segmentation and parcellation on the raw image. Concurrently, the network generation module learns representations of brain regions and constructs a network reflecting their connectivity. Finally, the classification module is responsible

for making the final predictions. By integrating these modules in an end-to-end fashion, UniBrain enables mutual boosting across tasks, achieving joint optimization with limited labeled data.

Extensive experiments on multiple public brain MRI datasets show that our method significantly surpasses existing state-of-the-art approaches in all six tasks.

2 PRELIMINARIES

In this section, we first introduce the relevant concepts and notations. We then formulate the problem of end-to-end brain imaging analysis. All notations are detailed in Appendix A.6 and Table 9.

2.1 Notations and Definitions

Definition 1 (Training data). Given a training dataset $\mathcal{D} = \{(\mathbf{S}_i, \mathbf{M}_i, y_i)\}_{i=1}^{Z}$, along with a template $(\mathbf{T}, \mathbf{B}, \mathbf{P})$. The dataset contains Z source images $\mathbf{S}_i \in \mathbb{R}^{W\times H\times D}$ (*e.g.*, raw MRI scan), each paired with a brain extraction mask $\mathbf{M}_i \in \{0,1\}^{W\times H\times D}$ and a classification label $y_i \in \mathcal{Y}$. The template contains a target image $\mathbf{T} \in \mathbb{R}^{W\times H\times D}$, with its segmentation mask $\mathbf{B} \in \{0,1\}^{C\times W\times H\times D}$ and parcellation mask $\mathbf{P} \in \{0,1\}^{K\times W\times H\times D}$. Here, W, H, and D denote the width, height and depth dimensions of the 3D images, C denotes the number of segmentation labels (*i.e.*, the number of labeled brain tissue types), K denotes the number of labeled brain regions (*i.e.*, the number of ROIs). $\mathcal{Y}$ is the classification label space (*e.g.*, $\{0, 1\}$ for binary classification). Next, we omit the subscript i of $\mathbf{S}_i$, $\mathbf{M}_i$ and y_i for simplicity.

Definition 2 (Outputs of each task). 1) *The extraction task* outputs a binary extraction mask $\hat{\mathbf{M}} \in \{0,1\}^{W\times H\times D}$, representing cerebral tissues in source image $\mathbf{S}$ with a value of 1 and non-cerebral tissues with 0. The extracted image $\mathbf{E} = \mathbf{S} \circ \hat{\mathbf{M}}$ is obtained by applying $\hat{\mathbf{M}}$ on $\mathbf{S}$ via an element-wise product $\circ$. 2) *The registration task* outputs a 3D affine transformation matrix $\mathbf{A} \in \mathbb{R}^{4\times 4}$ with 12 degrees of freedom (*i.e.*, 3 for rotation, 3 for translation, 3 for scale, and 3 for shear), indicating the coordinate correspondence between the extracted image $\mathbf{E}$ and the target image $\mathbf{T}$. The warped image (*i.e.*, registered image) $\mathbf{W} = \mathcal{T}(\mathbf{E}, \mathbf{A})$ results from applying the affine transformation on the extracted image $\mathbf{E}$, where $\mathcal{T}(\cdot,\cdot)$ is the affine transformation operator. 3) *The segmentation task* outputs a multi-class segmentation mask $\mathbf{R} \in \{0,1\}^{C\times W\times H\times D}$, categorizing tissue types within the source image $\mathbf{S}$. 4) *The parcellation task* outputs a multi-class parcellation mask $\mathbf{U} \in \{0,1\}^{K\times W\times H\times D}$, subdividing various regions (*i.e.*, ROIs) within the source image $\mathbf{S}$. The parcellated image $\mathbf{F} = \mathbf{S} \circ \mathbf{U}$ is obtained by performing element-wise product $\circ$ between $\mathbf{S}$ and $\mathbf{U}$. 5) *The network generation task* outputs an adjacency matrix $\mathbf{C} \in \mathbb{R}^{K\times K}$ and a node feature (*i.e.*, ROI feature) matrix $\mathbf{H} \in \mathbb{R}^{K\times N}$, capturing the brain network connectivity in the source image $\mathbf{S}$ across K ROIs. N is the node feature vector length. 6) *The classification task* outputs a categorical label $\hat{y} \in \mathcal{Y}$, indicating the final predictive category.

2.2 Problem Formulation

Given the limited labeled data and heterogeneous inputs and outputs across tasks, we aim to minimize the need for extensive task-specific labels and enable efficient knowledge transfer across different tasks. Formulating the end-to-end brain imaging analysis as a joint learning problem, we allow interconnected and collective task optimization. Specifically, we designed multiple learnable functions to perform and bridge tasks, while our loss terms are jointly optimized by leveraging the limited labeled information. Without loss of generality, 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.

Formally, our learnable functions are expressed as: the extraction function $f_\theta : \mathbb{R}^{W\times H\times D} \to \mathbb{R}^{W\times H\times D}; \mathbf{S} \mapsto \hat{\mathbf{M}}$, the registration function $g_\phi : \mathbb{R}^{W\times H\times D} \times \mathbb{R}^{W\times H\times D} \to \mathbb{R}^{4\times 4}; (\mathbf{E}, \mathbf{T}) \mapsto \mathbf{A}$, the segmentation function $h_\psi : \mathbb{R}^{W\times H\times D} \to \mathbb{R}^{C\times W\times H\times D}; \mathbf{S} \mapsto \mathbf{R}$, the brain network generation function $n_\xi : \mathbb{R}^{K\times W\times H\times D} \to \mathbb{R}^{K\times N}; \mathbf{F} \mapsto \mathbf{H}$ and the classification function $c_\eta : (\mathbb{R}^{K\times K}, \mathbb{R}^{K\times N}) \to \mathcal{Y}; (\mathbf{C}, \mathbf{H}) \mapsto \hat{y}$. 1) *The extraction function* $f_\theta(\cdot)$ takes the source image $\mathbf{S}$ as input to predict the extraction mask $\hat{\mathbf{M}} = f_\theta(\mathbf{S})$. 2) *The registration function* $g_\phi(\cdot,\cdot)$ takes the extracted brain image $\mathbf{E} = \mathbf{S} \circ \hat{\mathbf{M}}$ and the target image $\mathbf{T}$ to predict the affine transformation $\mathbf{A} = g_\phi(\mathbf{E}, \mathbf{T})$, and thereby generating the warped image $\mathbf{W} = \mathcal{T}(\mathbf{E}, \mathbf{A})$. Then, the warped segmentation mask $\mathbf{V} = \mathcal{T}(\mathbf{B}, \mathbf{A}^{-1})$ and parcellation mask $\mathbf{U} = \mathcal{T}(\mathbf{P}, \mathbf{A}^{-1})$ are obtained by applying the inverse affine transform $\mathbf{A}^{-1}$ to target segmentation mask $\mathbf{B}$ and parcellation mask $\mathbf{P}$. 3) *The segmentation function* $h_\psi(\cdot)$ predicts source segmentation mask $\mathbf{R} = h_\psi(\mathbf{S})$ from the source image $\mathbf{S}$. 4) *The brain network generation function* $n_\xi(\cdot)$ processes parcellated image $\mathbf{F} = \mathbf{S} \circ \mathbf{U}$ to learn the node feature $\mathbf{H} = n_\xi(\mathbf{F})$. 5) *The classification function* $c_\eta(\cdot,\cdot)$ takes a learnable adjacency matrix $\mathbf{C} = \mathbf{H}\mathbf{H}^\top$, and node feature $\mathbf{H}$ to make the final prediction $\hat{y} = c_\eta(\mathbf{C}, \mathbf{H})$. The optimal parameter set $\mathcal{P}^* = \{\theta^*, \phi^*, \psi^*, \xi^*, \eta^*\}$ can be found by solving the following optimization problem:

$$\mathcal{P}^* = \arg\min_{\mathcal{P}} \sum_{(\mathbf{S},\mathbf{M},y)\in\mathcal{D}} \Big[\mathcal{L}_{cls}(\hat{y}, y) + \alpha\mathcal{L}_{ext}(\hat{\mathbf{M}}, \mathbf{M}) + \beta\mathcal{L}_{sim}(\mathbf{W}, \mathbf{T}) + \gamma\mathcal{L}_{seg}(\mathbf{R}, \mathbf{V})\Big], \tag{1}$$

where the image pair $(\mathbf{S}, \mathbf{M}, y)$ is sampled from the training dataset $\mathcal{D}$. $\mathcal{L}_{cls}(\cdot,\cdot)$ is classification loss term, $\mathcal{L}_{ext}(\cdot,\cdot)$ is extraction loss term, $\mathcal{L}_{sim}(\cdot,\cdot)$ is image dissimilarity loss term (*e.g.*, mean square error), and $\mathcal{L}_{seg}(\cdot,\cdot)$ is segmentation loss term. These four criteria leverage the limited labeled information, and guide a joint optimization of extraction, registration, segmentation, parcellation, network generation, and classification, enabling effective interaction and feedback among these tasks.

To our knowledge, this work is the first endeavor to find an optimal solution for the end-to-end brain imaging analysis problem. Our method unifies all tasks and significantly reduces the need for extensive annotations in brain imaging by utilizing minimal labeled data (*i.e.*, extraction mask, classification label, and one labeled template), as opposed to other fully supervised [1, 9, 13, 26, 40, 56, 63, 76, 78] and pipeline-based [5, 12, 38, 52, 53, 55, 58, 77] methods.

3 OUR APPROACH

Overview. Figure 2 presents the UniBrain framework, designed for the end-to-end brain imaging analysis problem. Our method is an end-to-end deep neural network consisting of five main modules: 1) *Extraction Module* processes the raw source image $\mathbf{S}$ to yield the extracted brain image $\mathbf{E}$; 2) *Registration Module* aligns extracted brain image $\mathbf{E}$ with the target image $\mathbf{T}$, resulting in the warped

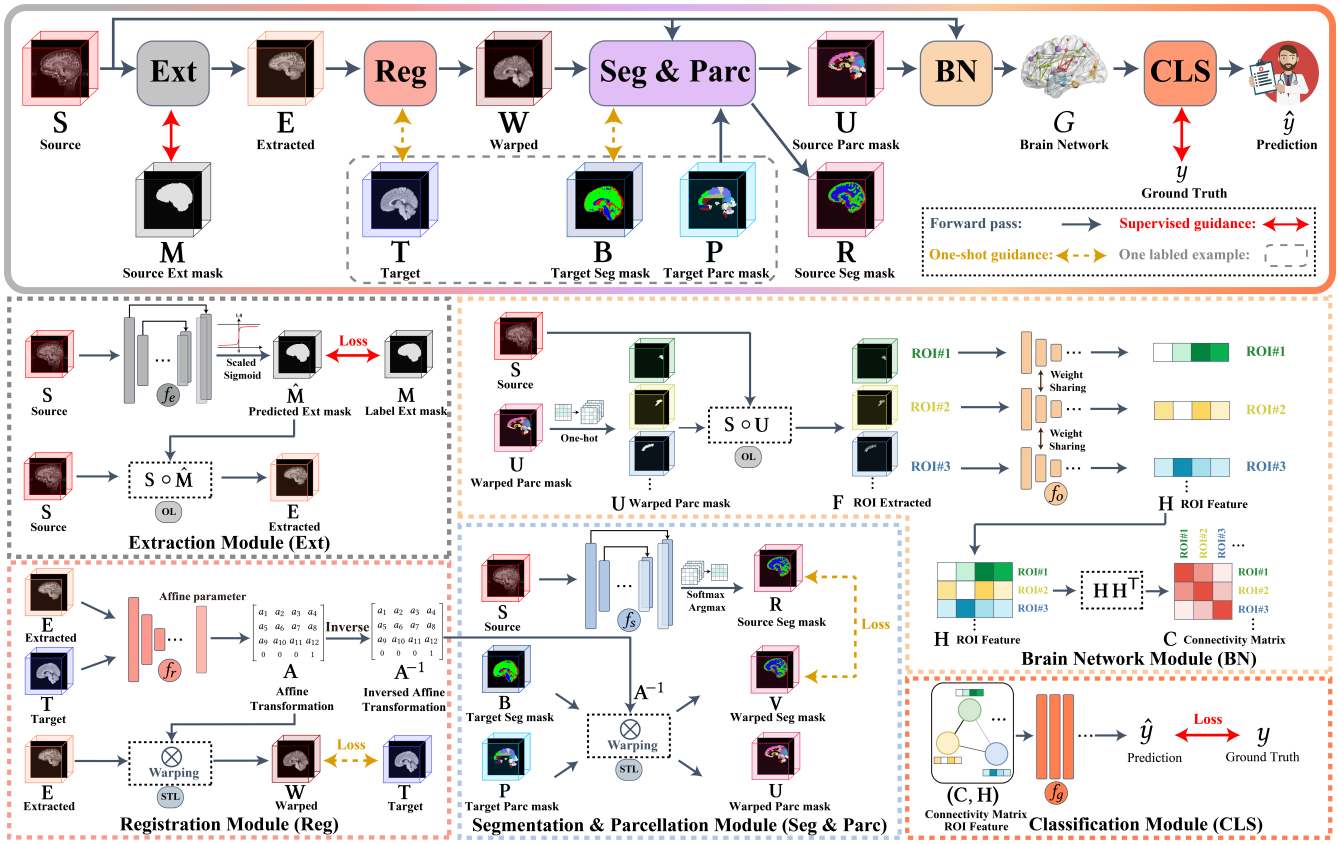


Figure 2: An overview of the proposed UniBrain. Extraction (Ext) module removes non-brain tissue of the raw source image S, producing the extracted brain image E. Registration (Reg) module aligns extracted brain E with the target image T through affine transformation A, resulting in the warped image W. Segmentation & parcellation (Seg & Parc) module takes the source image S, target segmentation mask B, target parcellation mask P, and inversed affine transformation A^{-1} to generate the segmentation mask R and the parcellation mask U. Brain network (BN) module uses the source image S and the parcellation mask U to construct brain network G by learning ROI feature H and connectivity matrix C. Classification (CLS) module then uses the ROI feature H and connectivity matrix C to produce the prediction $\hat{y}$. All modules are integrated to enable collective learning. UniBrain outputs include the extracted brain image E, the target-aligned warped image W, the source image's brain tissue segmentation mask R, the source image's brain ROI parcellation mask U, the brain network connectivity matrix C and the prediction $\hat{y}$.

image **W**; 3) *Segmentation & Parcellation Modules* leverage the target segmentation mask **B** and parcellation mask **P** to produce the source segmentation mask **R** and parcellation mask **U**; 4) *Brain Network Module* generate the brain network G based one the region information provided by source image **S** and parcellation mask **U**; 5) *Classification Module* employs the generated brain network G for prediction, outputting $\hat{y}$. The final output of UniBrain includes: extracted brain image **E** constituting only cerebral tissues; warped image **W** aligning with the target image; brain segmentation mask **R** and parcellation mask **U** indicating tissue types and regions of the source image; brain network G representing connectivity among regions, and prediction $\hat{y}$ signifying classification outcomes.

Key Insights. Given the intrinsic interdependence of tasks in brain imaging analysis, our method's design capitalizes on four synergistic mechanisms that facilitate collaboration among all modules:

• Performance in the initial extraction is crucial for preventing error propagation and ensuring the stability of all subsequent tasks.

• Accurate registration leads to generating precise segmentation and parcellation masks, aiding subsequent tasks.

• The tissue structure can provide auxiliary information for the extraction module to remove non-cerebral parts and for the registration module to find anatomical correspondences.

• Accurate classification not only guarantees predictive precision but also delivers task-aware feedback for network generation.

Consequently, we incorporate these four synergistic mechanisms into our loss design, training our network end-to-end to ensure efficiency and efficacy. Specifically, our loss function is structured to include four distinct loss terms, each aligning with one of these mechanisms: the extraction loss $\mathcal{L}_{ext}(\hat{\mathbf{M}}, \mathbf{M})$ aligns with the first mechanism, the registration loss $\mathcal{L}_{sim}(\mathbf{W}, \mathbf{T})$ corresponds to the second mechanism, the segmentation loss $\mathcal{L}_{seg}(\mathbf{R}, \mathbf{V})$ matches the third mechanism, and the classification loss $\mathcal{L}_{cls}(\hat{y}, y)$ pertains the fourth mechanism. Bidirectional supervision at both ends first envelops the entire network, ensuring positive forward propagation and controllable feedback across tasks. Then, the unsupervised and one-shot guidance inside the model effectively reduces the reliance on high-cost annotations. We jointly optimize the four loss terms to achieve collective learning on all tasks. Next, we introduce the details of each module and the training process.

3.1 Extraction Module

The extraction module extracts brain from the raw image, outputting an extracted image with assistance from two components:

3.1.1 Extraction Network: f_e. The extraction network $f_e(\cdot)$ acts as an annotator, intended to identify brain and non-brain tissues in the source image $\mathbf{S}$ and delineate their locations, thus providing the guidance for subsequent non-brain tissue elimination. Specifically, we employ the 3D U-Net [50] as the base network to learn $f_e(\cdot)$. The process can be formally expressed as:

$$\hat{\mathbf{M}} = f_e(\mathbf{S}), \tag{2}$$

where $\hat{\mathbf{M}}$ is predicted extraction mask. When conducting inference, the output $\hat{\mathbf{M}}$ is binarized by a Heaviside step function.

3.1.2 Overlay Layer: OL. The overlay layer serves to eliminate non-brain tissues by applying the predicted brain mask $\hat{\mathbf{M}}$ to the source image $\mathbf{S}$. The final extracted image is $\mathbf{E} = \mathbf{S} \circ \hat{\mathbf{M}}$, where $\circ$ denotes the operation of element-wise multiplication.

3.2 Registration Module

The registration module aims to align the extracted image with the target image and provide transformations for subsequent tasks (segmentation and parcellation), enabling the use of target mask information. This module comprises two main components:

3.2.1 Registration Network: f_r. The registration network $f_r(\cdot,\cdot)$ processes the extracted image $\mathbf{E}$ and target image $\mathbf{T}$ to learn the affine transformation $\mathbf{A}$, which establishes the coordinate correspondence between source and target image space. A 3D CNN-based encoder is used to learn $f_r(\cdot,\cdot)$, expressed as:

$$\mathbf{A} = f_r(\mathbf{E}, \mathbf{T}). \tag{3}$$

We leverage the multi-stage registration technique [58, 77] to boost registration performance, where $\mathbf{E}$ is recursively aligned with $\mathbf{T}$ though M stages. A study of M can be found in Figure 4(d).

3.2.2 Spatial Transformation Layer: STL. A key step in image registration is reconstructing the warped image $\mathbf{W}$ from the extracted image $\mathbf{E}$ using the affine transformation $\mathbf{A}$. This warping process is facilitated by a spatial transformation layer (STL), which resamples voxels from the extracted image $\mathbf{E}$ to produce the warped image $\mathbf{W} = \mathcal{T}(\mathbf{E}, \mathbf{A})$. Given the affine transformation operator, we hold

$$\mathbf{W}_{xyz} = \mathbf{E}_{x'y'z'}, \tag{4}$$

where coordinate correspondence $[x', y', z', 1]^\top = \mathbf{A}[x, y, z, 1]^\top$. To enable successful gradient propagation, we use a differentiable transformation based on trilinear interpolation proposed by [23].

3.3 Segmentation & Parcellation Module

The Segmentation & Parcellation Module is designed to create segmentation and parcellation masks on the source image. Leveraging recent developments in one-shot learning [14, 60, 68, 75], our approach enables the generation of these masks using a single labeled template image. The module contains two main components:

3.3.1 Inverse Warping. Utilizing a single labeled example (*i.e.*, target image $\mathbf{T}$ with its corresponding segmentation mask $\mathbf{B}$ and parcellation mask $\mathbf{P}$) and the learned affine transformation $\mathbf{A}$, we apply the inverse transformation $\mathbf{A}^{-1}$ to effectively generate warped segmentation mask $\mathbf{V} = \mathcal{T}(\mathbf{B}, \mathbf{A}^{-1})$ and parcellation mask $\mathbf{U} = \mathcal{T}(\mathbf{P}, \mathbf{A}^{-1})$ in the source image space, expressed as:

$$\mathbf{V}_{cxyz} = \mathbf{B}_{cx'y'z'}, \forall c \in \{1, \ldots, C\}, \tag{5}$$

$$\mathbf{U}_{kxyz} = \mathbf{P}_{kx'y'z'}, \forall k \in \{1, \ldots, K\}, \tag{6}$$

where coordinate correspondence $[x', y', z', 1]^\top = \mathbf{A}^{-1}[x, y, z, 1]^\top$, c is the index for tissue class and k is the index for ROIs. Same as the STL in Section 3.2.2, we then apply a differentiable transformation based on trilinear interpolation.

3.3.2 Segmentation Network: f_s. The segmentation network $f_s(\cdot)$ aims to generate a segmentation mask for the source image $\mathbf{S}$ that matches the synthesized warped segmentation mask $\mathbf{V}$. Similar to the extraction network in Section 3.1.1, we employ the widely-used 3D U-Net as the base network to learn $f_s(\cdot)$. Formally, we have:

$$\mathbf{R} = f_s(\mathbf{S}). \tag{7}$$

In our study, we adopted different strategies for brain segmentation and parcellation: 1) For segmentation, neural networks are utilized to learn the source image's segmentation mask $\mathbf{R}$. This method is effective in identifying brain tissues with clear boundaries and adjusts well to variations between tissues, resulting in precise segmentation masks. The outcomes of this process provide positive feedback to the registration module, enhancing label accuracy [14, 60, 68, 75]; 2) For parcellation, given its complexity and the less defined boundaries in functional regions, we adopted inverse warping to obtain source's parcellation mask $\mathbf{U}$. Additionally, due to the variety of brain parcellation systems, the choice of parcellation atlas can vary (*e.g.*, AAL, Harvard-Oxford and Desikan) [2].

3.4 Brain Network Module

The brain network module generates the brain network using ROI information from parcellation mask $\mathbf{U}$ on the source image $\mathbf{S}$.

3.4.1 Overlay Layer: OL. Similar to Section 3.1.2, this component is responsible for isolating each ROI from the source image $\mathbf{S}$ using parcellation mask $\mathbf{U}$. The parcellated image $\mathbf{F} = \mathbf{S} \circ \mathbf{U}$ is generated by applying an element-wise product $\circ$ between $\mathbf{S}$ and $\mathbf{U}$.

3.4.2 Brain Network Function: f_o. The brain network function aims to learn the representation for each ROI within the parcellated image $\mathbf{F}$. A weight-sharing Multilayer Perceptron (MLP) is employed to learn $f_o(\cdot)$, ensuring consistent feature extraction and generalization. This procedure is formally expressed as:

$$\mathbf{H}_k = f_o(\mathbf{F}_k), \forall k \in \{1, \ldots, K\}, \tag{8}$$

where k is the index for the ROIs.

3.4.3 Brain Network Construction. The step aims to construct a brain network based on the similarity between ROI representation pairs. Without loss of generality, here we use inner-product to measure the edge weights of the brain network. However, this measurement can be easily extended to other types of differentiable similarity functions, *e.g.*, Mahalanobis distance and cosine similarity.

To compute the connectivity matrix $\mathbf{C}$, each ROI representation $\mathbf{H}_k$ is first normalized with the ℓ^2-norm, followed by the inner-product:

$$\mathbf{C} = \mathbf{H}\mathbf{H}^\top. \tag{9}$$

This normalization scales the values of $\mathbf{C}$ to the range of $[-1, 1]$, ensuring the stabilization of the learning process and maintaining consistent weight magnitudes across the network. The similarity score indicates the ROIs closer in embedding space are more likely to be connected, as suggested by [17, 27, 32, 82]. To further refine the network, connections in $\mathbf{C}$ with negative weights are screened out as previous works [31, 37, 51, 67], which reduces complexity and potential noise from less relevant connections.

3.5 Classification Module

The classification module makes the final diagnostic prediction.

3.5.1 Classification Network: f_g. The classification network $f_g(\cdot,\cdot)$ aims to make a prediction based on the generated brain network while also feeding task-specific insights to the preceding module, thus facilitating the brain network generation. We leverage the Graph Convolutional Network (GCN) [33] as the base network, which is adept at handling structured data and capturing relational dependencies. The prediction $\hat{y}$ can be obtained as:

$$\hat{y} = f_g(\mathbf{C}, \mathbf{H}), \tag{10}$$

where $\mathbf{H}$ is the initial node features and $\mathbf{C}$ is the learnable connectivity matrix provided by the preceding brain network module.

3.6 End-to-End Training

We train UniBrain by minimizing the following objective function:

$$\mathcal{P}^* = \min_{\mathcal{P}} \mathcal{L}_{cls}(\hat{y}, y) + \alpha \mathcal{L}_{ext}(\hat{\mathbf{M}}, \mathbf{M}) + \beta \mathcal{L}_{sim}(\mathbf{W}, \mathbf{T}) + \gamma \mathcal{L}_{seg}(\mathbf{R}, \mathbf{V}), \tag{11}$$

where $\mathcal{P} = \{\theta, \phi, \psi, \xi, \eta\}$ is the parameter set. θ, ϕ, ψ, ξ, η are the parameters for the extraction network $f_e(\cdot;\theta)$, registration network $f_r(\cdot,\cdot;\phi)$, segmentation network $f_s(\cdot;\psi)$, brain network function $f_o(\cdot;\xi)$, and classification network $f_g(\cdot,\cdot;\eta)$, respectively. The classification loss $\mathcal{L}_{cls}(\cdot,\cdot)$ is a cross-entropy error, measuring the label similarity between prediction $\hat{y}$ and ground-truth y. The extraction loss $\mathcal{L}_{ext}(\cdot,\cdot)$ is also a cross-entropy error, quantifying the similarity between predicted extraction mask $\hat{\mathbf{M}}$ and ground-truth extraction mask $\mathbf{M}$. The registration loss $\mathcal{L}_{sim}(\cdot,\cdot)$ is a negative cross-correlation, assessing the image similarity between warped image $\mathbf{W}$ and target image $\mathbf{T}$ on voxel level. The segmentation loss $\mathcal{L}_{seg}(\cdot,\cdot)$ is also a cross-entropy error, measuring the similarity between the predicted segmentation mask $\mathbf{R}$ and the warped segmentation mask $\mathbf{V}$. α, β, and γ scale the numerical value of each loss term to the same order of magnitude, balancing their impacts. The influence of these hyperparameters is attached in Appendix A.1.1.

By leveraging the differentiability in each component of this design, our model achieves joint optimization in an end-to-end manner. All tasks are unified within a single model for collective learning, mutually boosting their performance with limited labels.

Table 2: Summary of compared methods.

Methods	Extraction	Registration	Segmentation	Parcellation	Network Generation	Classification
BET [55]	✓	✗	✗	✗	✗	✗
SynthStrip [20]	✓	✗	✗	✗	✗	✗
FLIRT [24]	✗	✓	✗	✗	✗	✗
VM [5]	✗	✓	✗	✗	✗	✗
ABN [58]	✗	✓	✗	✗	✗	✗
DW [23]	✗	✗	✓	✓	✗	✗
DeepAtlas [70]	✗	✓	✓	✗	✗	✗
ERNet [59]	✓	✓	✗	✗	✗	✗
JERS [60]	✓	✓	✓	✗	✗	✗
KNN [80]	✗	✗	✗	✗	✓	✗
GCN [33]	✗	✗	✗	✗	✗	✓
BGN [37]	✗	✗	✗	✗	✗	✓
BNT [28]	✗	✗	✗	✗	✗	✓
UniBrain	✓	✓	✓	✓	✓	✓

4 EXPERIMENTS

4.1 Datasets and Compared Methods

We evaluate the effectiveness of our proposed method on two public real-world 3D brain sMRI datasets: 1) *ADHD* [11] is collected from ADHD-200 global competition dataset. The dataset contains records for 776 subjects, labeled as real patients (positive) and normal controls (negative). The original dataset is unbalanced, following [35], we randomly sampled 100 ADHD patients and 100 normal controls from the dataset for performance evaluation ; 2) *ABIDE* [64] is collected from Autism Brain Imaging Data Exchange dataset. The dataset contains 1112 subjects, labeled as real patients and normal controls. Same to ADHD dataset, we randomly sampled 500 ASD patients and 500 normal controls from the dataset for performance evaluation. We use MNI 152 with the AAL atlas [65] as the template image. Additional details for datasets can be found in Appendix A.3.

We compare our UniBrain with several representative brain extraction, registration, segmentation, parcellation, network generation, and classification methods, as shown in Table 2. Notably, there are no existing solutions that can simultaneously perform all tasks in an end-to-end framework. Thus, for comparison, we designed a pipeline-based solution by combining different state-of-the-art methods for each task. Baselines are detailed in Appendix A.5.

4.2 Experimental Results

We compare UniBrain with the baseline methods on extraction, registration, segmentation, parcellation, and classification accuracy. Based on the experiment results, we find that UniBrain not only consistently outperform other alternatives in terms of extraction, registration, segmentation, parcellation and classification, but is also time-efficient. Additional experiments and evaluation metrics are detailed in Appendix A.1 and Appendix A.2, respectively.

4.2.1 Experiment Setting. We split the datasets into training, validation and test sets as described in Appendix A.3. The training set is used to learn model parameters, while the validation set is employed to evaluate the performance of hyperparameter settings (*e.g.*, the weight of each loss term). The test set is used only once to report the final evaluation results for each model. We describe the details of the hyperparameter settings of UniBrain in Appendix A.4. The source code is available at https://github.com/Anonymous7852/UniBrain.

4.2.2 Overall Results. Table 3 and Table 4 present the results of the compared methods and the proposed UniBrain in extraction, registration, segmentation, parcellation, and classification tasks. Based on the comprehensive evaluation of two datasets, UniBrian

Table 3: Results on ADHD dataset. The results are reported as performance (mean ± std) of extraction, registration, segmentation, parcellation and classification of each compared method. "↑" point out "the larger the better". The best results are highlighted in bold.

Methods						Extraction		Registration		Segmentation		Parcellation		Classification	
Ext	Reg	Seg	Parc	NG	Cls	Dice ↑	Jaccard ↑	MI ↑	CC ↑	Dice ↑	Jaccard ↑	Dice ↑	Jaccard ↑	ACC ↑	AUC-ROC ↑
BET [55]	FLIRT [24]	DW [23]		KNN [80]	GCN [33]	0.830 ± 0.058	0.713 ± 0.079	0.585 ± 0.031	0.882 ± 0.041	0.431 ± 0.058	0.293 ± 0.049	0.510 ± 0.172	0.375 ± 0.142	0.582 ± 0.034	0.546 ± 0.028
Synth [20]	FLIRT [24]	DW [23]		KNN [80]	GCN [33]	0.920 ± 0.012	0.853 ± 0.021	0.621 ± 0.018	0.942 ± 0.006	0.494 ± 0.015	0.347 ± 0.013	0.678 ± 0.040	0.525 ± 0.040	0.595 ± 0.043	0.612 ± 0.024
BET [55]	VM [5]	DW [23]		KNN [80]	GCN [33]	0.830 ± 0.058	0.713 ± 0.079	0.584 ± 0.037	0.874 ± 0.043	0.432 ± 0.029	0.296 ± 0.026	0.599 ± 0.070	0.442 ± 0.066	0.578 ± 0.027	0.568 ± 0.016
Synth [20]	VM [5]	DW [23]		KNN [80]	GCN [33]	0.920 ± 0.012	0.853 ± 0.021	0.632 ± 0.020	0.940 ± 0.007	0.447 ± 0.014	0.309 ± 0.013	0.619 ± 0.041	0.463 ± 0.039	0.582 ± 0.055	0.598 ± 0.015
BET [55]	ABN [58]	DW [23]		KNN [80]	GCN [33]	0.830 ± 0.058	0.713 ± 0.079	0.585 ± 0.036	0.877 ± 0.043	0.446 ± 0.031	0.308 ± 0.027	0.653 ± 0.051	0.497 ± 0.051	0.526 ± 0.036	0.571 ± 0.017
Synth [20]	ABN [58]	DW [23]		KNN [80]	GCN [33]	0.920 ± 0.012	0.853 ± 0.021	0.635 ± 0.021	0.943 ± 0.009	0.455 ± 0.015	0.317 ± 0.013	0.675 ± 0.026	0.521 ± 0.027	0.595 ± 0.039	0.612 ± 0.012
ERNet [59]		DW [23]		KNN [80]	GCN [33]	0.935 ± 0.016	0.879 ± 0.028	0.636 ± 0.014	0.952 ± 0.009	0.498 ± 0.014	0.350 ± 0.014	0.677 ± 0.045	0.523 ± 0.047	0.582 ± 0.070	0.612 ± 0.015
BET [55]	DeepAtlas [70]	DW [23]		KNN [80]	GCN [33]	0.830 ± 0.058	0.713 ± 0.079	0.587 ± 0.037	0.874 ± 0.041	0.478 ± 0.029	0.344 ± 0.028	0.591 ± 0.069	0.434 ± 0.065	0.599 ± 0.017	0.579 ± 0.013
Synth [20]	DeepAtlas [70]	DW [23]		KNN [80]	GCN [33]	0.920 ± 0.012	0.853 ± 0.021	0.632 ± 0.021	0.940 ± 0.007	0.480 ± 0.016	0.348 ± 0.015	0.654 ± 0.030	0.497 ± 0.031	0.621 ± 0.047	0.647 ± 0.012
JERS [60]		DW [23]		KNN [80]	GCN [33]	0.938 ± 0.014	0.883 ± 0.025	0.637 ± 0.014	0.952 ± 0.009	0.504 ± 0.013	0.369 ± 0.013	0.681 ± 0.043	0.527 ± 0.045	0.626 ± 0.039	0.584 ± 0.009
JERS [60]		DW [23]		KNN [80]	BGN [37]	0.938 ± 0.014	0.883 ± 0.025	0.637 ± 0.014	0.952 ± 0.009	0.504 ± 0.013	0.369 ± 0.013	0.681 ± 0.043	0.527 ± 0.045	0.548 ± 0.085	0.582 ± 0.094
JERS [60]		DW [23]		KNN [80]	BNT [28]	0.938 ± 0.014	0.883 ± 0.025	0.637 ± 0.014	0.952 ± 0.009	0.504 ± 0.013	0.369 ± 0.013	0.681 ± 0.043	0.527 ± 0.045	0.535 ± 0.039	0.585 ± 0.034
UniBrain (ours)						**0.970 ± 0.003**	**0.942 ± 0.006**	**0.652 ± 0.008**	**0.957 ± 0.008**	**0.520 ± 0.013**	**0.381 ± 0.013**	**0.708 ± 0.019**	**0.557 ± 0.022**	**0.652 ± 0.027**	**0.712 ± 0.030**

Table 4: Results on ABIDE dataset. The results are reported as performance (mean ± std) of extraction, registration, segmentation, parcellation and classification of each compared method. "↑" point out "the larger the better". The best results are highlighted in bold.

Methods						Extraction		Registration		Segmentation		Parcellation		Classification	
Ext	Reg	Seg	Parc	NG	Cls	Dice ↑	Jaccard ↑	MI ↑	CC ↑	Dice ↑	Jaccard ↑	Dice ↑	Jaccard ↑	ACC ↑	AUC-ROC ↑
BET [55]	FLIRT [24]	DW [23]		KNN [80]	GCN [33]	0.786 ± 0.173	0.673 ± 0.180	0.575 ± 0.029	0.890 ± 0.036	0.411 ± 0.074	0.279 ± 0.061	0.525 ± 0.183	0.388 ± 0.152	0.521 ± 0.013	0.567 ± 0.004
Synth [20]	FLIRT [24]	DW [23]		KNN [80]	GCN [33]	0.909 ± 0.026	0.835 ± 0.042	0.613 ± 0.026	0.938 ± 0.010	0.474 ± 0.026	0.331 ± 0.023	0.670 ± 0.043	0.516 ± 0.043	0.551 ± 0.046	0.610 ± 0.007
BET [55]	VM [5]	DW [23]		KNN [80]	GCN [33]	0.786 ± 0.173	0.673 ± 0.180	0.577 ± 0.026	0.897 ± 0.039	0.385 ± 0.072	0.257 ± 0.059	0.466 ± 0.182	0.334 ± 0.148	0.524 ± 0.018	0.574 ± 0.005
Synth [20]	VM [5]	DW [23]		KNN [80]	GCN [33]	0.909 ± 0.026	0.835 ± 0.042	0.611 ± 0.023	0.948 ± 0.010	0.454 ± 0.026	0.314 ± 0.022	0.642 ± 0.046	0.487 ± 0.045	0.543 ± 0.021	0.552 ± 0.005
BET [55]	ABN [58]	DW [23]		KNN [80]	GCN [33]	0.786 ± 0.173	0.673 ± 0.180	0.577 ± 0.024	0.900 ± 0.037	0.385 ± 0.077	0.258 ± 0.063	0.486 ± 0.188	0.353 ± 0.155	0.553 ± 0.012	0.568 ± 0.002
Synth [20]	ABN [58]	DW [23]		KNN [80]	GCN [33]	0.909 ± 0.026	0.835 ± 0.042	0.612 ± 0.023	0.950 ± 0.010	0.457 ± 0.026	0.317 ± 0.023	0.661 ± 0.048	0.507 ± 0.047	0.565 ± 0.024	0.579 ± 0.005
ERNet [59]		DW [23]		KNN [80]	GCN [33]	0.927 ± 0.019	0.866 ± 0.032	0.627 ± 0.018	0.958 ± 0.009	0.483 ± 0.022	0.337 ± 0.019	0.652 ± 0.053	0.497 ± 0.054	0.554 ± 0.019	0.590 ± 0.008
BET [55]	DeepAtlas [70]	DW [23]		KNN [80]	GCN [33]	0.786 ± 0.173	0.673 ± 0.180	0.576 ± 0.024	0.898 ± 0.038	0.394 ± 0.077	0.268 ± 0.064	0.469 ± 0.182	0.337 ± 0.148	0.524 ± 0.015	0.562 ± 0.015
Synth [20]	DeepAtlas [70]	DW [23]		KNN [80]	GCN [33]	0.909 ± 0.026	0.835 ± 0.042	0.613 ± 0.023	0.949 ± 0.010	0.471 ± 0.028	0.334 ± 0.025	0.647 ± 0.051	0.492 ± 0.050	0.557 ± 0.032	0.572 ± 0.004
JERS [60]		DW [23]		KNN [80]	GCN [33]	0.929 ± 0.017	0.869 ± 0.030	0.628 ± 0.018	0.959 ± 0.009	0.508 ± 0.026	0.364 ± 0.024	0.655 ± 0.050	0.501 ± 0.051	0.551 ± 0.022	0.591 ± 0.006
JERS [60]		DW [23]		KNN [80]	BGN [37]	0.929 ± 0.017	0.869 ± 0.030	0.628 ± 0.018	0.959 ± 0.009	0.508 ± 0.026	0.364 ± 0.024	0.655 ± 0.050	0.501 ± 0.051	0.529 ± 0.072	0.534 ± 0.083
JERS [60]		DW [23]		KNN [80]	BNT [28]	0.929 ± 0.017	0.869 ± 0.030	0.628 ± 0.018	0.959 ± 0.009	0.508 ± 0.026	0.364 ± 0.024	0.655 ± 0.050	0.501 ± 0.051	0.571 ± 0.044	0.605 ± 0.022
UniBrain (ours)						**0.970 ± 0.012**	**0.942 ± 0.021**	**0.641 ± 0.015**	**0.965 ± 0.007**	**0.515 ± 0.024**	**0.367 ± 0.022**	**0.692 ± 0.035**	**0.540 ± 0.037**	**0.648 ± 0.043**	**0.781 ± 0.029**

outperforms existing methods in all tasks. 1) For the extraction task, we observed that joint-based extraction methods (ERNet, JERS and UniBrain) outperform single-stage extraction methods (BET and Synth), especially on the ABIDE dataset. Specifically, UniBrain achieves up to a 6.8% improvement in extraction dice scores over the best single-stage method Synth. 2) For the registration task, methods with strong extraction results typically yield better registration accuracy, highlighting the dependency of accurate registration on prior extraction quality. 3) In segmentation and parcellation tasks, again we observe that good registration directly translates to improved performance, given their reliance on accurate registration for mask generation. 4) Classification task results also reflect this trend, where methods with higher parcellation accuracy (like Synth-based, JERS-based, and UniBrain) yield better classification outcomes. This is likely due to the classification network leveraging parcellation masks for brain network construction. Overall, there's a clear interdependence among brain imaging analysis tasks, with strengths and errors propagating across them. Partially joint methods like ERNet, JERS, and DeepAtlas show improved performance in their joint tasks but are limited when combined with other separate models. In contrast, UniBrain, benefiting from full end-to-end joint learning, uniquely excels across all tasks.

4.2.3 Qualitative Analysis. Figure 3 visually compares the performance of our UniBrain with other methods on the ADHD test set. UniBrain is observed to excel in extraction, registration, segmentation, and parcellation tasks, corroborating the analysis presented in Section 4.2.2. Specifically, UniBrain's predicted brain extraction masks closely overlap with the ground truth, whereas other methods often include significant non-brain tissues in their predictions. For registration, UniBrain also achieves higher similarity between its final registered image and the target image. Crucially, inaccurate extraction methods result in irreversible errors that propagate into registration, where non-alignable non-brain tissues cannot be properly registered regardless of the method used. In segmentation and parcellation tasks, UniBrain continues to demonstrate superior performance, producing results that closely match the ground truth segmentation and parcellation masks. Overall, our findings highlight a clear trend: better extraction leads to better registration, which in turn enhances segmentation and parcellation outcomes.

4.2.4 Ablation Study. To study the effectiveness of each module in UniBrain, we design six variants of UniBrain for comparison, as shown in Table 5. We freeze the extraction module (Ext), registration module (Reg), segmentation module (Seg), parcellation module (Parc), brain network module (BN) and classification module (CLS) respectively, which means we bypass the network parameter updates and operations within this module. Our findings include: 1) We reverify the strong interdependence among tasks, where the performance of preceding tasks directly impacts subsequent ones. 2) All modules are essential, as achieving optimal in all tasks requires including all modules without exception. 3) UniBrain has the potential to handle tasks beyond classification (*e.g.*, regression), as evidenced by the negligible impact on the performance of preceding tasks when the classification module is removed.

5 RELATED WORK

Tasks of Brain Imaging Analysis. Brain Imaging Analysis has become standard practice for neuroscience, including multiple crucial processing steps and tasks, such as brain extraction, registration, segmentation, parcellation, network generation, and classification. Deep learning has greatly impacted the area of brain imaging analysis, advancing brain extraction [22, 34, 43, 79], registration [13, 15, 16, 39, 56], segmentation [1, 9, 26, 26], parcellation [40, 63], network generation [54, 81], and classification [28, 29, 36, 37]. However, these learning-based approaches often need extensive task-specific labels for effective training, which is a challenge considering that neuroimaging datasets are typically small and costly to annotate. To

Table 5: Ablation studies on holding out tasks of UniBrain on ADHD dataset

Methods	Extraction		Registration		Segmentation		Parcellation		Classification	
	Dice ↑	Jaccard ↑	MI ↑	CC ↑	Dice ↑	Jaccard ↑	Dice ↑	Jaccard ↑	ACC ↑	AUC-ROC ↑
UniBrain w/o Ext	0.140 ± 0.009	0.075 ± 0.005	0.593 ± 0.015	0.846 ± 0.029	0.260 ± 0.039	0.157 ± 0.028	0.196 ± 0.096	0.120 ± 0.063	0.525 ± 0.038	0.591 ± 0.023
UniBrain w/o Reg	0.968 ± 0.005	0.938 ± 0.009	0.362 ± 0.040	0.706 ± 0.043	0.286 ± 0.034	0.175 ± 0.025	0.140 ± 0.075	0.086 ± 0.050	0.571 ± 0.054	0.605 ± 0.045
UniBrain w/o Seg	0.968 ± 0.004	0.938 ± 0.007	0.650 ± 0.005	0.956 ± 0.008	0.044 ± 0.028	0.023 ± 0.015	0.706 ± 0.019	0.555 ± 0.021	0.650 ± 0.032	0.696 ± 0.029
UniBrain w/o Parc	0.969 ± 0.004	0.941 ± 0.007	0.651 ± 0.007	0.955 ± 0.006	0.514 ± 0.011	0.380 ± 0.010	0.140 ± 0.088	0.086 ± 0.058	0.569 ± 0.066	0.590 ± 0.031
UniBrain w/o BN	0.968 ± 0.003	0.939 ± 0.006	0.648 ± 0.007	0.956 ± 0.009	0.518 ± 0.015	0.375 ± 0.015	0.706 ± 0.021	0.555 ± 0.023	0.583 ± 0.061	0.631 ± 0.028
UniBrain w/o CLS	0.967 ± 0.004	0.937 ± 0.007	0.651 ± 0.009	0.956 ± 0.008	0.512 ± 0.012	0.367 ± 0.012	0.706 ± 0.020	0.555 ± 0.022	0.478 ± 0.000	0.500 ± 0.000
UniBrain	**0.970 ± 0.003**	**0.942 ± 0.006**	**0.652 ± 0.008**	**0.957 ± 0.008**	**0.520 ± 0.013**	**0.381 ± 0.013**	**0.708 ± 0.019**	**0.557 ± 0.022**	**0.652 ± 0.027**	**0.712 ± 0.030**

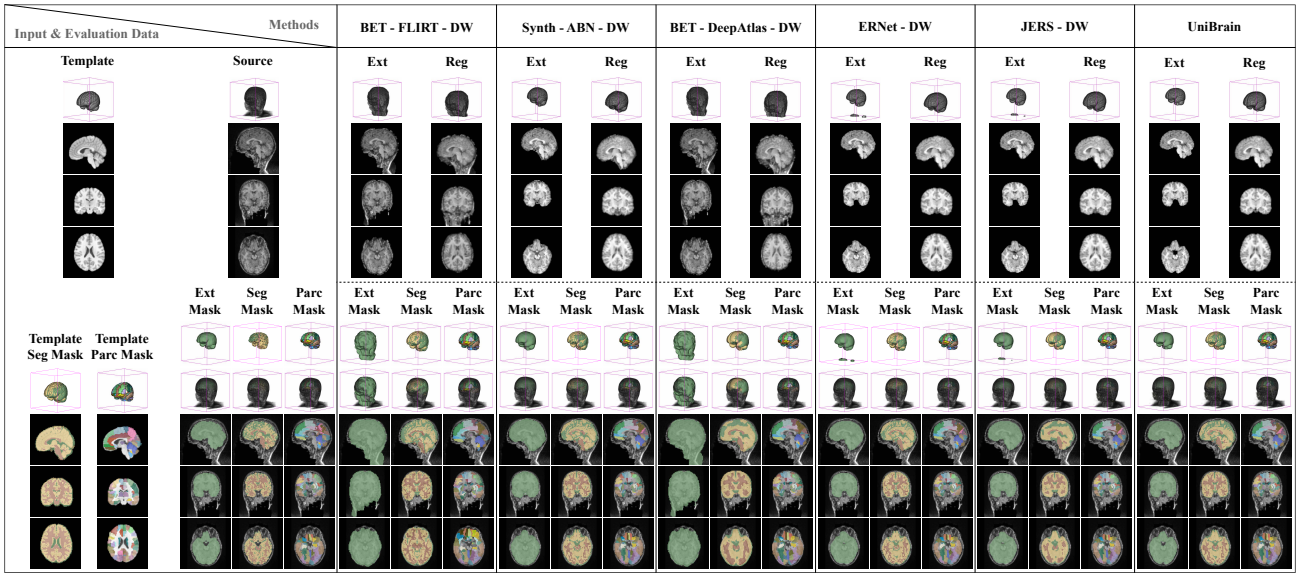


Figure 3: Visual comparisons for brain extraction, registration, segmentation, and parcellation tasks. We show 3D visualizations and middle slices in sagittal, axial, and coronal planes. On the left, source and target (template) images are displayed with their ground truth labels. The results and predicted labels for each method in extraction, registration, segmentation and parcellation are illustrated for performance assessment. Successful extraction, segmentation and parcellation imply a high overlap between predicted and ground truth masks. For registration, the higher similarity between the registered and template brains indicates better performance.

address this limitation, recent shifts towards unsupervised methods in extraction [12, 52, 53, 55], registration [5, 24, 58, 77], segmentation [23], parcellation [23], and network construction [38, 81]. These approaches reduce the dependency on extensive annotations. Nevertheless, combining these separate pipeline-based methods often requires detailed parameter tuning and manual quality control across tasks, which can be time-consuming and prone to inconsistency, affecting overall efficiency and effectiveness.

Joint Learning on Brain Imaging Analysis. Brain imaging analysis tasks are interrelated and can be jointly optimized. Specifically, effective extraction can enhance the stability of subsequent tasks. Registration can facilitate the segmentation and parcellation tasks by providing target mask information. In turn, the segmentation task can offer auxiliary tissue information beneficial for extraction and registration. Parcellation provides region information crucial for network generation. Finally, the robust network generation can lead to more precise classifications. To harness these interdependencies, various joint methods have been developed, including joint extraction-registration [59], joint registration-segmentation [19, 49, 70], joint extraction-registration-segmentation [60], and joint network generation-classification [27, 46, 74]. Though these partial joint learning solutions demonstrate great performance in their respective joint tasks, they overlook the potential relationships with other tasks, limiting the overall effectiveness.

Unified Model. Recent advancements in Unified Models have significantly impacted various AI domains. For instance, Unified-IO [42] exhibits multimodal capabilities in processing and generating images, text, audio, and actions, demonstrating its versatility across diverse tasks. In the field of anomaly detection, UniAD [73] effectively identifies anomalies across different categories through a unified approach. Additionally, MP3 [8] marks a significant advancement in integrating multiple functionalities crucial for autonomous driving. However, in the field of brain imaging analysis, there is a noticeable absence of such unified approaches. Our proposed UniBrain aims to address this gap by capturing relationships in various brain-related tasks and integrating them into a single model.

6 CONCLUSION

This paper introduces a novel unified framework, UniBrain, the first neural model to jointly perform a diverse set of brain imaging analysis tasks, including brain extraction, registration, segmentation, parcellation, network generation and classification. UniBrain integrates heterogeneous information into a single system, enabling efficient knowledge transfer across different modules, and avoiding the need for extensive task-specific labels. Experimental results show that UniBrain outperforms state-of-the-art methods in all tasks while also demonstrating robustness and time efficiency.

REFERENCES

[1] Zeynettin Akkus, Alfiia Galimzianova, Assaf Hoogi, Daniel L Rubin, and Bradley J Erickson. 2017. Deep learning for brain MRI segmentation: state of the art and future directions. *Journal of digital imaging* (2017).

[2] Salim Arslan, Sofia Ira Ktena, Antonios Makropoulos, Emma C Robinson, Daniel Rueckert, and Sarah Parisot. 2018. Human brain mapping: A systematic comparison of parcellation methods for the human cerebral cortex. *Neuroimage* (2018).

[3] 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. *MIA* (2008).

[4] Zilong Bai, Peter Walker, Anna Tschiffely, Fei Wang, and Ian Davidson. 2017. Unsupervised network discovery for brain imaging data. In *SIGKDD*.

[5] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. 2018. An unsupervised learning model for deformable medical image registration. In *CVPR*.

[6] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. 2019. VoxelMorph: a learning framework for deformable medical image registration. *IEEE TMI* (2019).

[7] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. Deep adversarial learning for multi-modality missing data completion. In *SIGKDD*.

[8] Sergio Casas, Abbas Sadat, and Raquel Urtasun. 2021. Mp3: A unified model to map, perceive, predict and plan. In *CVPR*.

[9] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. 2018. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage* (2018).

[10] Yongjun Chen, Hongyang Gao, Lei Cai, Min Shi, Dinggang Shen, and Shuiwang Ji. 2018. Voxel deconvolutional networks for 3D brain image labeling. In *SIGKDD*.

[11] ADHD-200 consortium. 2012. The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience* (2012).

[12] Robert W Cox. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research* (1996).

[13] Xin Dai, Xiangnan Kong, Xinyue Liu, John Boaz Lee, and Constance Moore. 2020. Dual-Attention Recurrent Networks for Affine Registration of Neuroimaging Data. In *SDM*.

[14] Yuhang Ding, Xin Yu, and Yi Yang. 2021. Modeling the probabilistic distribution of unlabeled data for one-shot medical image segmentation. In *AAAI*.

[15] Koen AJ Eppenhof and Josien PW Pluim. 2018. Pulmonary CT registration through supervised learning with convolutional neural networks. *IEEE TMI* 38, 5 (2018), 1097–1105.

[16] Jingfan Fan, Xiaohuan Cao, Pew-Thian Yap, and Dinggang Shen. 2019. BIRNet: Brain image registration using dual-supervised fully convolutional networks. *MIA* 54 (2019), 193–206.

[17] Aditya Grover, Aaron Zweig, and Stefano Ermon. 2019. Graphite: Iterative generative modeling of graphs. In *ICML*.

[18] M Mehmet Haznedar, Monte S Buchsbaum, Tse-Chung Wei, Patrick R Hof, Charles Cartwright, Carol A Bienstock, and Eric Hollander. 2000. Limbic circuitry in patients with autism spectrum disorders studied with positron emission tomography and magnetic resonance imaging. *American Journal of Psychiatry* (2000).

[19] Yuting He, Tiantian Li, Guanyu Yang, Youyong Kong, Yang Chen, Huazhong Shu, Jean-Louis Coatrieux, Jean-Louis Dillenseger, and Shuo Li. 2020. Deep complementary joint model for complex scene registration and few-shot segmentation on medical images. In *ECCV*.

[20] Andrew Hoopes, Jocelyn S Mora, Adrian V Dalca, Bruce Fischl, and Malte Hoffmann. 2022. SynthStrip: Skull-stripping for any brain image. *NeuroImage* (2022).

[21] Shuai Huang, Jing Li, Jieping Ye, Adam Fleisher, Kewei Chen, Teresa Wu, and Eric Reiman. 2011. Brain effective connectivity modeling for Alzheimer's disease by sparse Gaussian Bayesian network. In *SIGKDD*.

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

[23] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. 2015. Spatial Transformer Networks. In *NeurIPS*.

[24] Mark Jenkinson and Stephen Smith. 2001. A global optimisation method for robust affine registration of brain images. *MIA* (2001).

[25] Minyoung Jung, Hirotaka Kosaka, Daisuke N Saito, Makoto Ishitobi, Tomoyo Morita, Keisuke Inohara, Mizuki Asano, Sumiyoshi Arai, Toshio Munesue, Akemi Tomoda, et al. 2014. Default mode network in young male adults with autism spectrum disorder: relationship with autism spectrum traits. *Molecular autism* (2014).

[26] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *MIA* (2017).

[27] Xuan Kan, Hejie Cui, Joshua Lukemire, Ying Guo, and Carl Yang. 2022. Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation. In *MIDL*.

[28] Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. 2022. Brain network transformer. In *NeurIPS*.

[29] Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. 2017. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage* (2017).

[30] Valeria Kebets, Avram J Holmes, Csaba Orban, Siyi Tang, Jingwei Li, Nanbo Sun, Ru Kong, Russell A Poldrack, and BT Thomas Yeo. 2019. Somatosensory-motor dysconnectivity spans multiple transdiagnostic dimensions of psychopathology. *Biological psychiatry* (2019).

[31] Byung-Hoon Kim, Jong Chul Ye, and Jae-Jin Kim. 2021. Learning dynamic graph representation of brain connectome with spatio-temporal attention. In *NeurIPS*.

[32] Thomas N Kipf and Max Welling. 2016. Variational Graph Auto-Encoders. *NeurIPS Workshop* (2016).

[33] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

[34] 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* (2016).

[35] Xiangnan Kong, Philip S Yu, Xue Wang, and Ann B Ragin. 2013. Discriminative feature selection for uncertain graph classification. In *SDM*.

[36] Gaotang Li, Marlena Duda, Xiang Zhang, Danai Koutra, and Yujun Yan. 2023. Interpretable Sparsification of Brain Graphs: Better Practices and Effective Designs for Graph Neural Networks. In *SIGKDD*.

[37] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. 2021. Braingnn: Interpretable brain graph neural network for fmri analysis. *MIA* (2021).

[38] Xia Liang, Jinhui Wang, Chaogan Yan, Ni Shu, Ke Xu, Gaolang Gong, and Yong He. 2012. Effects of different correlation metrics and preprocessing factors on small-world brain functional networks: a resting-state functional MRI study. *PloS one* (2012).

[39] Haofu Liao, Wei-An Lin, Jiarui Zhang, Jingdan Zhang, Jiebo Luo, and S Kevin Zhou. 2019. Multiview 2D/3D rigid registration via a point-of-interest network for tracking and triangulation. In *CVPR*.

[40] Eun-Cheon Lim, Uk-Su Choi, Kyu Yeong Choi, Jang Jae Lee, Yul-Wan Sung, Seiji Ogawa, Byeong Chae Kim, Kun Ho Lee, Jungsoo Gim, Alzheimer's Disease Neuroimaging Initiative, et al. 2022. DeepParcellation: a novel deep learning method for robust brain magnetic resonance imaging parcellation in older East Asians. *FAN* (2022).

[41] Sikun Lin, Shuyun Tang, Scott T. Grafton, and Ambuj K. Singh. 2022. Deep Representations for Time-Varying Brain Datasets. In *SIGKDD*.

[42] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. UNIFIED-IO: A Unified Model for Vision, Language, and Multi-modal Tasks. In *ICLR*.

[43] 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. *AIM* (2019).

[44] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. 1997. Multimodality image registration by maximization of mutual information. *TMI* (1997).

[45] Frederik Maes, Dirk Vandermeulen, and Paul Suetens. 2003. Medical image registration using mutual information. *Proc. IEEE* (2003).

[46] Usman Mahmood, Zening Fu, Vince D Calhoun, and Sergey Plis. 2021. A deep learning model for data-driven discovery of functional connectivity. *Algorithms* (2021).

[47] Evangelos E Papalexakis, Alona Fyshe, Nicholas D Sidiropoulos, Partha Pratim Talukdar, Tom M Mitchell, and Christos Faloutsos. 2014. Good-enough brain model: Challenges, algorithms and discoveries in multi-subject experiments. In *SIGKDD*.

[48] Josien PW Pluim, JB Antoine Maintz, and Max A Viergever. 2000. Image registration by maximization of combined mutual information and gradient information. In *MICCAI*.

[49] Liang Qiu and Hongliang Ren. 2021. U-RSNet: An unsupervised probabilistic model for joint registration and segmentation. *Neurocomputing* (2021).

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

[51] Anwar Said, Roza G Bayrak, Tyler Derr, Mudassir Shabbir, Daniel Moyer, Catie Chang, and Xenofon Koutsoukos. 2023. NeuroGraph: Benchmarks for Graph Machine Learning in Brain Connectomics. In *NeurIPS*.

[52] 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* (2004).

[53] David W Shattuck and Richard M Leahy. 2002. BrainSuite: an automated cortical surface identification tool. *MIA* (2002).

[54] Antonín Škoch, Barbora Rehák Bučková, Jan Mareš, Jaroslav Tintěra, Pavel Sanda, Lucia Jajcay, Jiří Horáček, Filip Španiel, and Jaroslav Hlinka. 2022. Human brain

structural connectivity matrices–ready for modelling. *Scientific Data* (2022).
[55] Stephen M Smith. 2002. Fast robust automated brain extraction. *Human brain mapping* (2002).
[56] 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 *MICCAI*.
[57] Olaf Sporns. 2013. Structure and function of complex brain networks. *Dialogues in clinical neuroscience* (2013).
[58] Yao Su, , Xin Dai, Lifang He, and Xiangnan Kong. 2022. ABN: Anti-Blur Neural Networks for Multi-Stage Deformable Image Registration. In *ICDM*.
[59] Yao Su, Zhentian Qian, Lifang He, and Xiangnan Kong. 2022. Ernet: Unsupervised collective extraction and registration in neuroimaging data. In *SIGKDD*.
[60] Yao Su, Zhentian Qian, Lei Ma, Lifang He, and Xiangnan Kong. 2023. One-shot Joint Extraction, Registration and Segmentation of Neuroimaging Data. In *SIGKDD*.
[61] 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 *SIGKDD*.
[62] Philippe Thévenaz and Michael Unser. 2000. Optimization of mutual information for multiresolution image registration. *TIP* (2000).
[63] Benjamin Thyreau and Yasuyuki Taki. 2020. Learning a cortical parcellation of the brain robust to the MRI segmentation with convolutional neural networks. *MIA* (2020).
[64] J Michael Tyszka, Daniel P Kennedy, Lynn K Paul, and Ralph Adolphs. 2014. Largely typical patterns of resting-state functional connectivity in high-functioning adults with autism. *Cerebral cortex* (2014).
[65] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* (2002).
[66] Lucina Q Uddin, AM Clare Kelly, Bharat B Biswal, Daniel S Margulies, Zarrar Shehzad, David Shaw, Manely Ghaffari, John Rotrosen, Lenard A Adler, F Xavier Castellanos, et al. 2008. Network homogeneity reveals decreased integrity of default-mode network in ADHD. *Neuroscience methods* (2008).
[67] Bernadette CM Van Wijk, Cornelis J Stam, and Andreas Daffertshofer. 2010. Comparing brain networks of different size and connectivity density using graph theory. *PloS one* (2010).
[68] Shuxin Wang, Shilei Cao, Dong Wei, Renzhen Wang, Kai Ma, Liansheng Wang, Deyu Meng, and Yefeng Zheng. 2020. LT-Net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation. In *CVPR*.
[69] Shen Wang, Lifang He, Bokai Cao, Chun-Ta Lu, Philip S Yu, and Ann B Ragin. 2017. Structural deep brain network mining. In *SIGKDD*.
[70] Zhenlin Xu and Marc Niethammer. 2019. DeepAtlas: Joint semi-supervised learning of image registration and segmentation. In *MICCAI*.
[71] Yujun Yan, Jiong Zhu, Marlena Duda, Eric Solarz, Chandra Sripada, and Danai Koutra. 2019. Groupinn: Grouping-based interpretable neural network for classification of limited, noisy brain data. In *SIGKDD*.
[72] Sen Yang, Qian Sun, Shuiwang Ji, Peter Wonka, Ian Davidson, and Jieping Ye. 2015. Structural graphical lasso for learning mouse brain connectivity. In *SIGKDD*.
[73] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. 2022. A unified model for multi-class anomaly detection. *NeurIPS* (2022).
[74] Yue Yu, Xuan Kan, Hejie Cui, Ran Xu, Yujia Zheng, Xiangchen Song, Yanqiao Zhu, Kun Zhang, Razieh Nabi, Ying Guo, et al. 2023. Deep dag learning of effective brain connectivity for fmri analysis. In *ISBI*.
[75] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. 2019. Data augmentation using learned transformations for one-shot medical image segmentation. In *CVPR*.
[76] Shijie Zhao, Yan Cui, Linwei Huang, Li Xie, Yaowu Chen, Junwei Han, Lei Guo, Shu Zhang, Tianming Liu, and Jinglei Lv. 2020. Supervised brain network learning based on deep recurrent neural networks. *Access* (2020).
[77] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. 2019. Recursive cascaded networks for unsupervised medical image registration. In *ICCV*.
[78] Shijie Zhao, Junwei Han, Jinglei Lv, Xi Jiang, Xintao Hu, Yu Zhao, Bao Ge, Lei Guo, and Tianming Liu. 2015. Supervised dictionary learning for inferring concurrent brain networks. *TMI* (2015).
[79] Tao Zhong, Fenqiang Zhao, Yuchen Pei, Zhenyuan Ning, Lufan Liao, Zhengwang Wu, Yuyu Niu, Li Wang, Dinggang Shen, Yu Zhang, et al. 2021. DIKA-Nets: Domain-invariant knowledge-guided attention networks for brain skull stripping of early developing macaques. *NeuroImage* 227 (2021), 117649.
[80] Houliang Zhou, Yu Zhang, Brian Y Chen, Li Shen, and Lifang He. 2022. Sparse Interpretation of Graph Convolutional Networks for Multi-modal Diagnosis of Alzheimer's Disease. In *MICCAI*.
[81] Zhen Zhou, Xiaobo Chen, Yu Zhang, Dan Hu, Lishan Qiao, Renping Yu, Pew-Thian Yap, Gang Pan, Han Zhang, and Dinggang Shen. 2020. A toolbox for brain network construction and classification (BrainNetClass). *HBM* (2020).
[82] Dongmian Zou and Gilad Lerman. 2019. Encoding robust representation for graph generation. In *IJCNN*.

A APPENDIX

This section first provides a detailed analysis of the evaluation, then describes the experiment settings to support the reproducibility of the results in this paper. Our code and data have been made publicly available at https://github.com/Anonymous7852/UniBrain.

A.1 Additional Experiment

A.1.1 Influence of Parameters. We study four crucial hyperparameters of our UniBrain: the extraction loss weight α; the registration loss weight β; the segmentation loss weight γ; and the number of registration stage M. As mentioned in Section 3.6 and Section 3.2, we introduced α, β, and γ to balance the impact of each loss term, and M to boost the registration performance. We vary these hyperparameters to study their influence on their corresponding task and the final classification task. Given the high costs of training deep learning models, we employed a practical approach for hyperparameter tuning. Initially, we set all weights to 1, and turned the hyperparameter one by one. Based on the performance on the validation set, we find the optimal values of $\alpha = 1$, $\beta = 0.1$, $\gamma = 1$, and $M = 5$. For the extraction loss weight α, we observed that as the weight increased, both extraction and classification accuracies improved. However, beyond a certain point, while extraction accuracy plateaued, classification accuracy began to decline, as depicted in Figure 4(a). This suggests that the extraction loss not only boosts extraction performance but also enhances final classification. Similar observations were made for registration loss weight β, segmentation loss weight γ and the number of registration stages M, as shown in Figure 4(b,c,d). Each contributed to improving the performance of their respective tasks and the overall classification outcome, confirming their effectiveness.

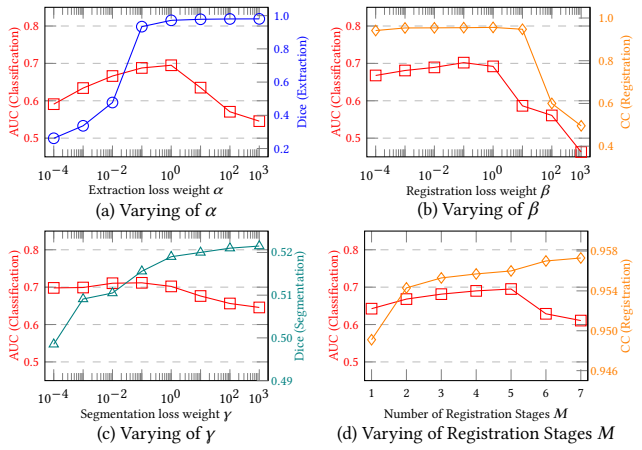


Figure 4: Effect of varying the extraction loss weight α, registration loss weight β, segmentation loss weight γ and registration stages M.

A.1.2 Running Efficiency. We measure the efficiency of UniBrain by comparing its running time with other baselines. The measurement is made on the same device with an AMD EPYC 7543 CPU and an NVIDIA Tesla A100 GPU. The running time is reported as the average processing time for each image in its corresponding task. As indicated in Table 6, fully separate methods are the slowest due to the need for individual optimization of each task. Partially joint learning methods demonstrate increased speed in their joined tasks but still require combination with other methods, limiting overall time efficiency. UniBrain is the fastest method, which efficiently performs all tasks in an end-to-end manner on the same device, enhancing overall speed.

Table 6: Running Time of compared methods on ADHD dataset.

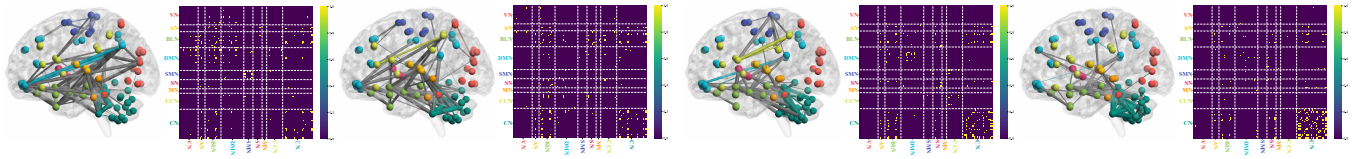
Methods						Time (Sec) ↓					
Ext	Reg	Seg	Parc	NG	Cls	Ext	Reg	Seg	Parc	NG	Cls
BET [55]	FLIRT [24]	DW [23]		KNN [80]	GCN [33]	1.2	4.1	1.2×10^{-1}	1.5×10^{-1}	8.2×10^{-1}	2.0×10^{-5}
Synth [20]	FLIRT [24]	DW [23]		KNN [80]	GCN [33]	9.8	5.2	1.3×10^{-1}	1.6×10^{-1}	7.9×10^{-1}	3.6×10^{-5}
BET [55]	VM [5]	DW [23]		KNN [80]	GCN [33]	1.2	5.7×10^{-3}	1.0×10^{-4}	1.1×10^{-4}	7.9×10^{-1}	4.0×10^{-5}
Synth [20]	ABN [58]	DW [23]		KNN [80]	GCN [33]	9.8	9.2×10^{-3}	1.0×10^{-4}	1.1×10^{-4}	8.2×10^{-1}	3.4×10^{-5}
ERNet [59]		DW [23]		KNN [80]	GCN [33]		4.0×10^{-2}	1.0×10^{-4}	1.1×10^{-4}	8.0×10^{-1}	3.1×10^{-5}
Synth [20]	DeepAtlas [70]	DW [23]		KNN [80]	GCN [33]	9.8	7.5×10^{-3}		1.1×10^{-4}	8.1×10^{-1}	4.4×10^{-5}
JERS [60]		DW [23]		KNN [80]	GCN [33]		4.9×10^{-2}		1.1×10^{-4}	8.1×10^{-1}	7.4×10^{-5}
JERS [60]		DW [23]		KNN [80]	BGN [37]		4.9×10^{-2}		1.1×10^{-4}	8.1×10^{-1}	2.4×10^{-3}
JERS [60]		DW [23]		KNN [80]	BNT [28]		4.9×10^{-2}		1.1×10^{-4}	8.0×10^{-1}	4.2×10^{-4}
UniBrain (ours)							2.2×10^{-1}				

A.1.3 Voxel-based End-to-End Learning. In this section, we compare UniBrain with another voxel-based end-to-end brain imaging analysis solution. In this set of experiments, we disregard graph-based models, relying only on voxel information from images for final classification predictions. We devised three groups: 1) Direct use of raw MRI images as input (including non-brain tissues, images in different coordinate spaces) for label classification. 2) Use of extracted brain images as input (still in different coordinate spaces) for label classification. 3) Use of the brain been extracted and registered to a standard space as input for classification. As shown in Table 7, we observed that the performance is worse when using raw images as input due to the inclusion of non-brain tissues and spatial transformation noise. Images processed through extraction and registration yielded higher accuracy. UniBrain, integrating preprocessing and classification in a joint learning approach, outperformed all other models.

Table 7: Voxel-based End-to-End Learning on ADHD dataset

Methods			Extraction		Registration		Classification	
Ext	Reg	Cls	Dice ↑	Jaccard ↑	MI ↑	CC ↑	ACC ↑	AUC-ROC ↑
-	-	3D-CNN	-	-	-	-	0.539 ± 0.048	0.623 ± 0.014
BET [55]	-	3D-CNN	0.830 ± 0.058	0.713 ± 0.079	-	-	0.539 ± 0.021	0.587 ± 0.019
Synth [20]	-	3D-CNN	0.920 ± 0.012	0.853 ± 0.021	-	-	0.547 ± 0.034	0.634 ± 0.018
ERNet_{ext} [59]	-	3D-CNN	0.935 ± 0.016	0.879 ± 0.028	-	-	0.582 ± 0.044	0.656 ± 0.025
JERS_{ext} [60]	-	3D-CNN	0.938 ± 0.014	0.883 ± 0.025	-	-	0.573 ± 0.037	0.638 ± 0.020
Synth [20]	FLIRT [24]	3D-CNN	0.920 ± 0.012	0.853 ± 0.021	0.621 ± 0.018	0.942 ± 0.006	0.647 ± 0.056	0.656 ± 0.051
Synth [20]	VM [5]	3D-CNN	0.920 ± 0.012	0.853 ± 0.021	0.632 ± 0.020	0.940 ± 0.007	0.617 ± 0.060	0.651 ± 0.029
Synth [20]	ABN [58]	3D-CNN	0.920 ± 0.012	0.853 ± 0.021	0.635 ± 0.021	0.943 ± 0.009	0.634 ± 0.028	0.622 ± 0.019
Synth [20]	DeepAtlas [70]	3D-CNN	0.920 ± 0.012	0.853 ± 0.021	0.632 ± 0.021	0.940 ± 0.007	0.645 ± 0.039	0.642 ± 0.031
ERNet [59]		3D-CNN	0.935 ± 0.016	0.879 ± 0.028	0.636 ± 0.014	0.952 ± 0.009	0.573 ± 0.042	0.570 ± 0.022
JERS		3D-CNN	0.938 ± 0.014	0.883 ± 0.025	0.637 ± 0.014	0.952 ± 0.009	0.613 ± 0.049	0.596 ± 0.025
UniBrain (ours)			**0.970 ± 0.003**	**0.942 ± 0.006**	**0.652 ± 0.008**	**0.957 ± 0.008**	**0.652 ± 0.027**	**0.712 ± 0.030**

A.1.4 Visualization of Generated Brain Network. In Figure 5, we show the predicted brain network connections for healthy controls (HC) and patients. For the ADHD dataset, we observed that patients show significantly fewer interactions within the DMN and SMN systems compared to HC. This aligns with earlier research [30, 66] findings, suggesting that alterations in DMN and SMN network connectivity may be associated with ADHD. For the ABIDE dataset, we observed that patients exhibit significantly fewer interactions within the BLN and DMN systems, as well as with other systems, compared to HC. This is consistent with findings [18, 25] related to ASD, suggesting that alterations in DMN and SMN network connectivity are associated with ASD.



(a) ADHD (b) ABIDE

Figure 5: We show the top 100 positive connections within the generated brain network for visualization. Nodes (ROIs) are categorized into different neural systems, including Visual Network (VN), Auditory Network (AN), Bilateral Limbic Network (BLN), Default Mode Network (DMN), Somato-Motor Network (SMN), Subcortical Network (SN), Memory Network (MN), Cognitive Control Network (CCN) and Cerebellum Network (CN). Edges that connect nodes within the same neural system are colored. The width of the edge represents its weight in the graph.

A.2 Evaluation Metrics

The end-to-end brain imaging analysis problem aims to perform brain extraction, registration, segmentation, parcellation, network generation and classification tasks simultaneously. We quantitatively evaluated all tasks except network generation, due to ground-truth brain network connections not included in the dataset.

A.2.1 Extraction Performance. The brain MRI datasets contain the ground truth brain extraction mask, outlining the brain in the source image. As in previous work [59, 60], to assess the accuracy of extraction, we evaluate the volume overlap of extraction masks using the Dice score and Jaccard score. The Dice score is expressed as:

$$\text{Dice} = 2 \cdot \frac{|\hat{\mathbf{M}} \cap \mathbf{M}|}{|\hat{\mathbf{M}}| + |\mathbf{M}|}, \tag{12}$$

where $\hat{\mathbf{M}}$ denotes the predicted brain mask and $\mathbf{M}$ is the corresponding ground truth. If $\hat{\mathbf{M}}$ represents a precise extraction, a high degree of overlap is expected between the non-zero regions of $\hat{\mathbf{M}}$ and $\mathbf{M}$. The Jaccard score is expressed as:

$$\text{Jaccard} = \frac{|\hat{\mathbf{M}} \cap \mathbf{M}|}{|\hat{\mathbf{M}} \cup \mathbf{M}|}, \tag{13}$$

If $\hat{\mathbf{M}}$ accurately represents extraction, we expect it to largely overlap with $\mathbf{M}$, yielding a high Jaccard score.

A.2.2 Registration Performance. Same to [58, 60], We use mutual information [44, 45, 48, 62] and normalized cross-correlation [3] to evaluate the registration performance. The mutual information between the warped image (*i.e.*, registered) $\mathbf{W}$ and the target image $\mathbf{T}$ is expressed as:

$$\text{MI}(\mathbf{W}, \mathbf{T}) = \sum_{w,t} p_{\mathbf{WT}}(w,t) \log \frac{p_{\mathbf{WT}}(w,t)}{p_{\mathbf{W}}(w) \cdot p_{\mathbf{T}}(t)} \tag{14}$$

where $p_{\mathbf{W}}(w)$ and $p_{\mathbf{T}}(t)$ are the marginal probability distributions of image $\mathbf{W}$ and $\mathbf{T}$, respectively. $p_{\mathbf{WT}}(w,t)$ is the joint probability distribution. The mutual information measures the mutual dependence between $\mathbf{W}$ and $\mathbf{T}$. If the warped image $\mathbf{W}$ and the target image $\mathbf{T}$ are geometrically aligned, we expect the mutual information to be maximal. The normalized cross-correlation between the warped image (*i.e.*, registered) $\mathbf{W}$ and the target image $\mathbf{T}$ is expressed as:

$$\text{NCC}(\mathbf{W}, \mathbf{T}) = \frac{\sum_{x,y,z}(\mathbf{W}_{xyz} - \overline{\mathbf{W}})(\mathbf{T}_{xyz} - \overline{\mathbf{T}})}{\sqrt{\sum_{x,y,z}(\mathbf{W}_{xyz} - \overline{\mathbf{W}})^2 \sum_{x,y,z}(\mathbf{T}_{xyz} - \overline{\mathbf{T}})^2}} \tag{15}$$

where $\mathbf{W}_{xyz}$ and $\mathbf{T}_{xyz}$ represent the voxel values of the images $\mathbf{W}$ and $\mathbf{T}$ at the position (x, y, z). $\overline{\mathbf{W}}$ and $\overline{\mathbf{T}}$ are the mean voxel values of these images, respectively. The summation runs over all the coordinates (x, y, z) in the three-dimensional space. A higher NCC value indicates a greater similarity between the two images.

A.2.3 Segmentation and Parcellation Performance. We evaluate the segmentation by measuring the volume overlap between the predicted mask and the ground-truth mask. If the segmentation task performs well, the predicted segmentation mask should overlap with the ground truth segmentation mask. Similar to the extraction evaluation, we use the Dice score and the Jaccard score to evaluate the overlap of the segmentation masks. If the mask includes multiple labeled tissue types, the final score is calculated as the average of the scores for each tissue type. Parcellation follows the same standard and metrics for evaluation.

A.2.4 Classification Performance. We evaluate the classification task using accuracy and AUC-ROC score since the dataset contains ground truth label y.

A.2.5 Evaluation details. The extraction, registration, segmentation, and parcellation tasks contain voxel-level labels for evaluation (*e.g.*, whether each voxel in the mask is predicted correctly). Consistent with [6, 59, 60, 77], we report the mean and standard deviation (std) at the subject level, *i.e.*, the average performance across subjects on the test set is presented. The classification labels are at the instance level. To more effectively showcase the robustness of the classification task, we present average and std results from 10 runs. In each run, the highest AUC-ROC performance on the validation set is tested for performance comparison.

A.3 Details of Data Preprocessing

The proposed method and baselines are evaluated on two different public brain MRI datasets: ADHD and ABIDE.

• *Attention Deficit Hyperactivity Disorder (ADHD)* [11]: The dataset is collected from ADHD-200 global competition dataset, which contains records of sMRI images for 776 subjects, labeled as real patients (positive) and normal controls (negative). Each subject includes a T1-weighted 3D brain sMRI scan, along with the corresponding ground truth brain extraction and segmentation masks. The extraction mask indicates the brain and non-brain tissues, and the segmentation mask indicates the 3 brain tissue types: gray matter, white matter, and cerebrospinal fluid. Each subject contains the ground truth transformation from the subject's image space to the standard MNI image space. The transformation is used to collect the ground truth parcellation from the MNI atlas (*e.g.*, AAL). The

original dataset is unbalanced, following [35], we randomly sampled 100 ADHD patients and 100 normal controls from the dataset for performance evaluation. Out of the 200 scans, 160 are used for training, 20 for validation, and 20 for testing. All scans are cropped and resized to $96 \times 96 \times 96$ dimensions.

• *Autism Brain Imaging Data Exchange (ABIDE)* [64]: The dataset is collected from Autism Brain Imaging Data Exchange I (ABIDE I), which contains 1112 subjects, labeled as real patients and normal controls. Each subject has a T1-weight 3D brain sMRI scan, with the corresponding ground truth brain extraction and segmentation masks. Similar to the ADHD dataset, each subject in ABIDE contains the ground truth extraction mask, segmentation mask and transformation to MNI space. We randomly sampled 500 ASD patients and 500 normal controls from the dataset for performance evaluation. Out of the 1000 scans, 800 are used for training, 100 for validation, and 100 for testing. All scans are cropped and resized to $96 \times 96 \times 96$ dimensions.

• *MNI 152 and AAL Atlas* [65]: We use MNI 152 as the template image, a standard brain commonly used in neuroimaging studies. The MNI 152 image contains a segmentation mask indicating gray matter, white matter, and cerebrospinal fluid. We employ the AAL-116 atlas for parcellation, featuring 116 ROIs within the MNI 152 image. Using this atlas and each subject's ground truth transformation, we collect parcellation labels for evaluation. All scans are cropped and resized to $96 \times 96 \times 96$ dimensions.

A.4 Detailed Settings of UniBrain

Training settings of UniBrain. Our experiments are conducted on Ubuntu 20.04 LTS, utilizing an AMD EPYC 7543 CPU and an NVIDIA Tesla A100 GPU. The code is implemented in Python 3.7.6, and the neural networks are built using PyTorch 1.7.1. The implementation also makes use of Numpy 1.21.6, SimpleITK 2.0.2, and Nibabel 3.1.1. To overcome GPU memory limitations, we employ batch gradient descent, with each training batch consisting of one sample. The models are optimized using the Adam optimizer, with a learning rate of 1×10^{-5} and a weight decay rate of 1×10^{-6}. We also apply image augmentation techniques, including random translation, rotation, and scaling, to the source images during training. Also, for graph learning, we introduce feature and edge perturbation to provide robust graph generation and classification. Random noise is added to the node feature, and edges are randomly added/removed. We detail the settings in Table 8.

Table 8: Range of augmentation.

Transformation			Perturbation	
Translation (Voxels)	Rotation (Degree)	Scale (Times)	Feature (Rate)	Edge (Rate)
± 2	± 2	0.98 ∼ 1.02	0.01	0.01

Parameters settings of UniBrain. We use the validation set to tune the hyperparameters. The number of registration stages is set to 5. The extraction loss weight α, registration loss weight β and segmentation loss weight γ in Eq. (11) are 1, 0.1 and 1, respectively. The extraction network contains 10 convolutional layers with 16, 32, 32, 64, 64, 64, 32, 32, 32 and 16 filters. The registration network contains 6 convolutional layers with 16, 32, 64, 128, 256 and 512 filters. The segmentation network contains 10 convolutional layers with 128, 256, 256, 512, 512, 512, 256, 256, 256 and 128 filters. The feature extraction network contains 2 MLP layers and 1 output layer with a feature size of 256, 256 and 256. The graph classification network contains 2 convolutional layers and one MLP layer with feature sizes of 128, 128 and 128.

A.5 Settings of Baselines

Brain Extraction Tool (BET) [55]: This skull-stripping method is a component of the FSL (FMRIB Software Library) package. It employs a deformable approach to accurately fit the brain surface by utilizing 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.

SynthStrip [20]: This is the state-of-the-art skull-stripping tool, which leverages a deep learning strategy to synthesize arbitrary training images for segmentation maps, yielding a robust performance. It is included in the FreeSurfer package. We use the default command to execute SynthStrip: `mri_synthstrip -i <input> -o <output> -m <output brain mask>`.

FMRIB's Linear Image Registration Tool (FLIRT) [24]: This is a fully automated affine brain image registration tool included in the FSL (FMRIB Software Library) package. It performs the registration process without requiring manual intervention, allowing for the alignment of brain images based on affine transformations. 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`.

VoxelMorph (VM) [5]: This unsupervised image registration method utilizes a neural network to predict the transformation between images. In order to ensure a fair comparison, we re-implemented the method using an affine transformation. The network architecture consists of 6 convolutional layers with filter sizes of 16, 32, 64, 128, 256, and 512.

Anti-Blur Registration Networks (ABN) [58]: This is an unsupervised anti-blur multi-stage registration method that involves iteratively transforming the source image to align with a target image. Same to UniBrain, the number of stages is set to 5 for fair comparison. Within each stage, we configure the network architecture with 6 convolutional layers, featuring filter sizes of 16, 32, 64, 128, 256, and 512.

Directly Warping (DW) [23]: This differentiable operation refers to inversely generating the segmentation and parcellation masks via the process of registration. Once the registration is completed, the segmentation and parcellation masks of the target image can be directly warped and transformed into the source image space.

DeepAtlas [70]: This is a joint registration and segmentation method. For a fair comparison, we configure 6 convolutional layers with 16, 32, 64, 128, 256 and 512 filters for the registration module, and the segmentation network contains 10 convolutional layers with 128, 256, 256, 512, 512, 512, 256, 256, 256 and 128 filters.

ERNet [59]: This is a joint extraction and registration method. In line with the original configuration, the number of extraction and registration stages is set to 5. The extraction network contains 10 convolutional layers with 16, 32, 32, 64, 64, 64, 32, 32, 32 and 16 filters. The registration network contains 6 convolutional layers with 16, 32, 64, 128, 256 and 512 filters.

JERS [60]: This is a joint extraction, registration and segmentation method. In line with the original configuration, the extraction and registration stages are set to 5. The extraction network contains 10 convolutional layers with 16, 32, 32, 64, 64, 64, 32, 32, 32 and 16 filters. The registration network contains 6 convolutional layers with 16, 32, 64, 128, 256 and 512 filters. The segmentation network contains 10 convolutional layers with 128, 256, 256, 512, 512, 512, 256, 256, 256 and 128 filters.

K-Nearest Neighbor (KNN) [80]: For the brain network construction, we follow the recent work [80] using K-Nearest Neighbor (KNN) to measure the similarity between each ROI, and generating weighted adjacency matrix. In line with the original configuration, the number of K is set to 10.

Graph Convolutional Networks (GCN) [33]: This is a type of neural network that operates directly on graphs, enabling it to capture the complex relationships and structures within networked data. The network contains 2 convolutional layers and one MLP layer with feature sizes of 128, 128 and 128.

BrainGNN (BGN) [37]: This method contains ROI-aware graph convolutional layers and ROI-selection pooling layers (R-pool) that highlight salient ROIs (nodes in the graph) from brain networks. The default model setting contains 2 convolutional layers and two MLP layers with feature sizes of 32, 32, 32 and 512.

Brain Network Transformer (BNT) [28]: This is a transformer-based model, which uses connection profiles as node features and learns pairwise connection strengths among ROIs with attention weights. The default model setting contains 2 multi-head self-attention layers with the number of heads of 2 and feature sizes of 128.

A.6 Details of Definitions and Notations

In this section, we detail the relevant concepts and notations.

Definition 1 (Source image, target image and labels). Given a training dataset $\mathcal{D} = \{(\mathbf{S}_i, \mathbf{M}_i, y_i)\}_{i=1}^{Z}$, along with a template $(\mathbf{T}, \mathbf{B}, \mathbf{P})$. The dataset contains Z source images $\mathbf{S}_i \in \mathbb{R}^{W\times H\times D}$ (*i.e.*, raw MRI scan), each paired with a brain extraction mask $\mathbf{M}_i \in \{0,1\}^{W\times H\times D}$ and a classification label $y_i \in \mathcal{Y}$. The template contains a target image $\mathbf{T} \in \mathbb{R}^{W\times H\times D}$, with its segmentation mask $\mathbf{B} \in \{0,1\}^{C\times W\times H\times D}$ and parcellation mask $\mathbf{P} \in \{0,1\}^{K\times W\times H\times D}$. Here, W, H, and D denote the width, height and depth dimensions of the 3D images, C denotes the number of segmentation labels (*i.e.*, the number of labeled brain tissue types), K denotes the number of labeled brain regions (*i.e.*, the number of ROIs). $\mathcal{Y}$ is the classification label space (*e.g.*, $\{0,1\}$ for binary classification).

Definition 2 (Extraction and extracted image). The predicted brain extraction mask $\hat{\mathbf{M}}$ is a binary tensor of identical dimensions to the source image $\mathbf{S}$. It represents cerebral tissues in $\mathbf{S}$ with a value of 1 and non-cerebral tissues with 0. The extracted image $\mathbf{E} = \mathbf{S} \circ \hat{\mathbf{M}}$ is obtained by applying the $\hat{\mathbf{M}}$ on $\mathbf{S}$ via a element-wise product $\circ$.

Definition 3 (Transformation and warped image). Without loss of generality, 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 3D affine transformation matrix $\mathbf{A} \in \mathbb{R}^{4\times 4}$. The warped image $\mathbf{W} = \mathcal{T}(\mathbf{E}, \mathbf{a})$ results from applying the affine transformation on the extracted image $\mathbf{E}$, where $\mathcal{T}(\cdot,\cdot)$ denotes the affine transformation operator. The following relationship holds for $\mathbf{W}$ and $\mathbf{E}$ on the voxel level:

$$\mathbf{W}_{xyz} = \mathbf{E}_{x'y'z'}, \tag{16}$$

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

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \mathbf{A} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ a_5 & a_6 & a_7 & a_8 \\ a_9 & a_{10} & a_{11} & a_{12} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}. \tag{17}$$

Definition 4 (Source segmentation and parcellation masks). Source segmentation mask $\mathbf{R} \in \{0,1\}^{C\times W\times H\times D}$ and parcellation mask $\mathbf{U} \in \{0,1\}^{K\times W\times H\times D}$ are binary tensor with the first dimension being the tissue numbers (*i.e.*, C) and region numbers (*i.e.*, K). The rest dimensions (*i.e.*, W, H, D) are same to the source image $\mathbf{S}$.

Definition 5 (ROI extraction and ROI feature). For ROI extraction, we first expand $\mathbf{S}$ to $\tilde{\mathbf{S}} \in \mathbb{R}^{K\times W\times H\times D}$, where each k^{th} slice of $\tilde{\mathbf{S}}$ is a copy of $\mathbf{S}$. Then, the ROI extracted image (*i.e.*, parcellated image) $\mathbf{F} \in \mathbb{R}^{K\times W\times H\times D} = \tilde{\mathbf{S}} \circ \mathbf{U}$ is obtained by applying the $\tilde{\mathbf{S}}$ on $\mathbf{U}$ via a element-wise product $\circ$. We learn features from each ROI in $\mathbf{F}$, converting them into vectors. These vectors are concatenated to form the matrix $\mathbf{H} \in \mathbb{R}^{K\times N}$, where K is the number of ROIs and N is the feature vector length.

Definition 6 (Brain Network). The brain network connectivity matrix $\mathbf{C} \in \mathbb{R}^{K\times K} = \mathbf{H}\mathbf{H}^{\top}$ is constructed by performing the Gram matrix computation on ROI feature $\mathbf{H}$. Then, we define brain network $G = (\mathcal{V}, \mathbf{C}, \mathbf{H})$, where $\mathcal{V} = \{v_j\}_{j=1}^{K}$ is the set of nodes representing K different brain regions. Each node corresponds to an ROI within brain parcellation.

Table 9: Basic Notation.

Notation	Description
W, H, D	width, height and depth dimensions of the 3D images
$\mathbf{S}_i$	source image (*i.e.*, raw MRI scan), $\mathbf{S}_i \in \mathbb{R}^{W\times H\times D}$
$\mathbf{M}_i$	ground truth brain extraction mask of $\mathbf{S}_i$, $\mathbf{M}_i \in \{0,1\}^{W\times H\times D}$
y_i	ground truth classification label of $\mathbf{S}_i$, $y_i \in \mathcal{Y}$
$\mathcal{Y}$	classification label space (*e.g.*, $\{0,1\}$ for binary classification)
$\mathcal{D} = \{(\mathbf{S}_i, \mathbf{M}_i, y_i)\}_{i=1}^{Z}$	dataset contains Z source images, each paired with a brain extraction mask $\mathbf{M}_i$ and classification label y_i
$\mathbf{T}$	target image, $\mathbf{T} \in \mathbb{R}^{W\times H\times D}$
$\mathbf{B}$	segmentation mask of $\mathbf{T}$, $\mathbf{B} \in \{0,1\}^{C\times W\times H\times D}$
$\mathbf{P}$	parcellation mask of $\mathbf{T}$, $\mathbf{P} \in \{0,1\}^{K\times W\times H\times D}$
C	number of segmentation labels (*i.e.*, the number of labeled brain tissue types)
K	number of labeled brain regions (*i.e.*, the number of ROIs)
$\hat{\mathbf{M}}$	predicted brain extraction mask, $\hat{\mathbf{M}} \in \{0,1\}^{W\times H\times D}$
$\hat{y}$	predicted classification label, $\hat{y} \in \mathcal{Y}$
$\mathbf{E}$	extracted image, obtained by $\mathbf{E} = \mathbf{S} \circ \hat{\mathbf{M}}$, $\mathbf{E} \in \mathbb{R}^{W\times H\times D}$
$\mathbf{a}$	vector of the affine transformation, $\mathbf{a} \in \mathbb{R}^{12}$
$\mathbf{A}$	matrix of an 3D affine transformation, $\mathbf{A} \in \mathbb{R}^{4\times 4}$
$\mathcal{T}(\cdot,\cdot)$	affine transformation operator
$\mathbf{W}$	warped image, obtained by $\mathbf{W} = \mathcal{T}(\mathbf{E}, \mathbf{A})$
$\mathbf{R}$	source segmentation mask, $\mathbf{R} \in \{0,1\}^{C\times W\times H\times D}$
$\mathbf{V}$	warped segmentation mask, $\mathbf{V} \in \{0,1\}^{C\times W\times H\times D}$
$\mathbf{U}$	warped parcellation mask, $\mathbf{U} \in \{0,1\}^{K\times W\times H\times D}$
$\tilde{\mathbf{S}}$	source images expanded from $\mathbf{S}$ for ROI extraction, $\tilde{\mathbf{S}} \in \mathbb{R}^{K\times W\times H\times D}$
$\mathbf{F}$	ROI extracted image (*i.e.*, parcellated image), obtained by $\mathbf{F} = \tilde{\mathbf{S}} \circ \mathbf{U}$, $\mathbf{F} \in \mathbb{R}^{K\times W\times H\times D}$
$\mathbf{H}$	matrix of ROI feature (*i.e.*, node feature), $\mathbf{H} \in \mathbb{R}^{K\times N}$
N	feature vector length of $\mathbf{H}$
$\mathbf{C}$	adjacency matrix, measures brain network connectivity, $\mathbf{C} \in \mathbb{R}^{K\times K}$
$\mathcal{V}$	set of nodes (*i.e.*, ROIs), $\mathcal{V} = \{v_j\}_{j=1}^{K}$
G	brain network, $G = (\mathcal{V}, \mathbf{C}, \mathbf{H})$
$f_\theta(\cdot)$	extraction function, $f_\theta : \mathbb{R}^{W\times H\times D} \to \mathbb{R}^{W\times H\times D}; \mathbf{S} \mapsto \hat{\mathbf{M}}$
$g_\phi(\cdot,\cdot)$	registration function, $g_\phi : \mathbb{R}^{W\times H\times D} \times \mathbb{R}^{W\times H\times D} \to \mathbb{R}^{12}; (\mathbf{E}, \mathbf{T}) \mapsto \mathbf{A}$
$h_\psi(\cdot)$	segmentation function $h_\psi : \mathbb{R}^{W\times H\times D} \to \mathbb{R}^{C\times W\times H\times D}; \mathbf{S} \mapsto \mathbf{R}$
$n_\xi(\cdot)$	brain network generation function, $n_\xi : \mathbb{R}^{K\times W\times H\times D} \to \mathbb{R}^{K\times N}; \mathbf{F} \mapsto \mathbf{H}$
$c_\eta(\cdot,\cdot)$	classification function $c_\eta : (\mathbb{R}^{K\times K}, \mathbb{R}^{K\times N}) \to \mathcal{Y}; (\mathbf{C}, \mathbf{H}) \mapsto \hat{y}$
$\mathcal{P}^*$	optimal parameter set, $\mathcal{P}^* = \{\theta^*, \phi^*, \psi^*, \xi^*, \eta^*\}$
$\mathcal{L}_{cls}(\cdot,\cdot)$	classification loss term, *i.e.*, cross-entropy loss
$\mathcal{L}_{ext}(\cdot,\cdot)$	extraction loss term, *i.e.*, cross-entropy loss
$\mathcal{L}_{sim}(\cdot,\cdot)$	image dissimilarity loss term, *i.e.*, negative cross-correlation
$\mathcal{L}_{seg}(\cdot,\cdot)$	segmentation loss term, *i.e.*, cross-entropy loss