Contrastive Registration for Unsupervised Medical Image Segmentation

Lihao Liu
University of Cambridge
ll610@cam.ac.uk

Angelica I Aviles-Rivero
University of Cambridge
ai323@cam.ac.uk

Carola-Bibiane Schönlieb
University of Cambridge
cbs31@cam.ac.uk

Abstract

Medical image segmentation is a relevant task as it serves as the first step for several diagnosis processes, thus it is indispensable in clinical usage. Whilst major success has been reported using supervised techniques, they assume a large and well-representative labelled set. This is a strong assumption in the medical domain where annotations are expensive, time-consuming, and inherent to human bias. To address this problem, unsupervised techniques have been proposed in the literature yet it is still an open problem due to the difficulty of learning any transformation pattern. In this work, we present a novel optimisation model framed into a new CNN-based contrastive registration architecture for unsupervised medical image segmentation. The core of our approach is to exploit image-level registration and feature-level from a contrastive learning mechanism, to perform registration-based segmentation. Firstly, we propose an architecture to capture the image-to-image transformation pattern via registration for unsupervised medical image segmentation. Secondly, we embed a contrastive learning mechanism into the registration architecture to enhance the discriminating capacity of the network in the feature-level. We show that our proposed technique mitigates the major drawbacks of existing unsupervised techniques. We demonstrate, through numerical and visual experiments, that our technique substantially outperforms the current state-of-the-art unsupervised segmentation methods on two major medical image datasets.

1. Introduction

Medical image segmentation is a task of partitioning an image into multiple regions, which ideally reflect qualities such as well-defined structures guided by existing boundaries in the image domain. This task has been successfully applied in a range of medical applications using data coming from different parts of human anatomy including the heart [5], lungs [19] and brain [12]. This is a relevant task as medical image segmentation is the first step for several clinical applications such as tumor locating and neuropsychiatric disorders diagnosis.

The body of literature has reported several successful techniques for medical image segmentation, in which the most notorious progress has been done using supervised convolutional neural networks (CNNs) based methods; examples following this perspective include U-Net [32] and VoxResNet [6]. These techniques rely on learning prior patterns information from pair-wise data (medical images and their manual segmentations) to achieve astonishing performance at the same level as, and sometimes outperforming, radiologist. However, a major constraint of these methods is the assumption of having a well-representative set of annotations. This is not always possible and in the medical domain, and one also needs to deal with the inherent human bias. To deal with this constraint, the community has also explored techniques relying on less or non annotations using semi-supervised segmentation [38, 31, 9] and unsupervised segmentation [1, 27, 10, 11].

Notably, in recent years unsupervised techniques have become a great focus of attention as they do not require segmentation labels. In this context, clustering algorithms are often applied to perform unsupervised segmentation by grouping the image contents with similar intensities *e.g.* K-Means [21, 13] and DeepCluster [36]. However, these methods are still limited performance-wise as it is difficult to learn any image-to-mask pattern information when relying solely on images. To further embed the pattern information within an unsupervised architecture for better medical image segmentation performance, one feasible solution is to cast the segmentation task as an unsupervised deformable registration problem. The goal of this perspective is to find an optimal image-to-image transformation pattern to align a set of images into one coordinate system.

Formally, we aim to find an optimal image-to-image transformation pattern z between an unaligned image x and a reference image y, which is usually formulated by a metric function ϕ_z. Benefited from similar morphological attributes between the medical image and its corresponding segmentation mask (such as the shape of different organs), one can also transfer the segmentation mask of the reference image y_{seg} back to the coordinate system of the unaligned

image. This process is with the purpose of obtaining the segmentation mask of the unaligned image x_{seg} using the optimal pattern information z. Hence, when the image is correctly registered, the segmentation mask is automatically obtained, which we called *registration-based segmentation*.

Following this philosophy, we aim to develop an unsupervised segmentation model based on a registration architecture. A central observation of existing registration methods is that they only focus on capturing the pattern information from the image level, and fail to enhance the feature-level representation. Most recently, unsupervised feature representation learning has demonstrated promising results, in which contrastive based models, such as MoCo [17] and SimCLR [7], are leading solutions that have reached performance comparable to those produced by supervised techniques. The main idea is that by contrasting images to other ones, the differences between images are easily remembered within a network (i.e., learning distinctiveness); thus making the learned feature more robust to discriminate images from different groups. Therefore, to further improve the feature-level learning by contrasting the unaligned image x to the reference image y, we propose to embed the contrast feature learning in the registration architecture to extract feature maps with richer information to produce better segmentation.

In this paper, we present a novel contrastive registration architecture for unsupervised medical image segmentation. Our proposed technique is a simple yet effective unsupervised segmentation model. Unlike the existing technique, ours combines the image-level registration and feature-level contrastive representation learning. More specifically, our technique works as follows. We first propose two weight-shared feature encoders, where CNN features are extracted from the unaligned images and reference images respectively. Then, by contrasting the extracted CNN features from the unaligned images and reference images, one can obtain CNN features with richer information. Moreover, we use one single decoder to capture the pattern information z from the contrasted CNN features. Finally, we use the spatial transform network of that [20] along with the captured z to align the segmentation mask, of the reference image to the coordinate of unaligned images, to obtain the segmentation mask in an unsupervised manner. Our main contributions are as follow:

- We propose a simple yet effective contrastive registration architecture for unsupervised medical image segmentation, in which we highlight the combination of image-level registration architecture and feature-level contrastive learning. To our best acknowledge, this is the first work that embeds contrastive learning in a registration architecture for unsupervised segmentation.

- We evaluate our technique using a range of numerical and visual results on two major benchmark datasets. We show that, by casting the unsupervised segmentation task via registration as the registration task with feature-level contrast, we can largely improve the unsupervised segmentation performance and narrow the performance difference with supervised techniques.

- We demonstrate that our contrastive registration architecture, as a by-product, can also lead to a better registration performance than the state-of-the-art unsupervised registration techniques.

2. Related Work

The body of literature has reported impressive results for image segmentation. In this section, we review the existing techniques in turn.

2.1. Supervised Medical Image Segmentation

Medical image segmentation has been extensively investigated in the literature, in which supervised methods have been most successful. Early works learn to segment the different organs based on hand-crafted features including thresholding [35], statistical model [15], and Bayesian model [40]. For example, the work of [35] presented an iterative thresholding algorithm. This technique recursively selected an optimum threshold value for segmenting different brain structures based on a hand-crafted shape prior. Fischl *et al.* [15] designed a statistic model to segment brain structures based on probabilistic information automatically estimated from the pair-wise date. The authors of [40] employed a Bayesian model to segment coronary arteries by exploiting the probabilistic relationships between the surface and boundary of the coronaries. These methods seek to differentiate the different organs based on hand-crafted features (i.e., shape and boundary). However, a major drawback is that they are not capable of capturing the high-level semantic information. Thereby, they tend to fail in segmenting those organs accurately.

More recently, convolutional neural networks (CNN) based methods, which train with annotations to capture the high-level semantic information, have demonstrated remarkable performance beyond radiologist execution, such as U-Net [32] and VoxResNet [6]. Ronneberger *et al.* [32] proposed a U-shape architecture, which recursively combined the high-level semantic features and low-level features for better biomedical image segmentation. Similarly, Chen *et al.* [6] presented a VoxResNet, which seamlessly integrated the low-level features with high-level semantic features for multi-modality medical image segmentation. Although effective in biomedical image segmentation, these supervised methods rely heavily on a well-representative set of annotations, which is not always available. Hence, when

applying the trained model to another dataset, it often fails to segment the medical images correctly.

2.2. Unsupervised Medical Image Segmentation

Recently, to deal with the constraint of requiring annotated data, unsupervised segmentation techniques (no labels required) have a focus of great attention. In this direction, clustering algorithms have been extensively explored to perform unsupervised segmentation, in which the idea is to divide the image into different groups according to the similarity of image intensities (i.e., each pixel value). Jose *et al.* [21] applied the well-known K-means algorithm to segment the abnormal region in brains. Likewise, Tian *et al.* [36] embedded the K-means technique into a CNN architecture to cluster the CNN features for better segmentation results. However, due to the lack of segmentation masks, which determines that it is impossible to learn any image-to-mask pattern information, clustering algorithms are still inefficient in segmentation tasks.

To enable the pattern information in the training process, another feasible solution is to cast the segmentation task as an unsupervised deformable registration process. Instead of seeking the image-to-mask transformation pattern, the goal of unsupervised registration methods is to find an optimal image-to-image transformation pattern. Specifically, given an unaligned image x and a reference image y, the main goal of unsupervised registration methods is to calculate the latent variable z, which contains the image-to-image transformation pattern information. Benefited from the morphological similarity between the medical image and its corresponding segmentation mask, unsupervised registration architecture can transfer the segmentation mask of the reference image y_{seg} back to the coordinate system of the unaligned image to obtain the segmentation mask of the unaligned image x_{seg}, using the optimal transformation pattern z. Hence, as long as the image is correctly registered, the segmentation mask is automatically obtained, which we call registered segmentation.

2.3. Contrastive Learning

Current unsupervised registration methods only focus on capturing the pattern information from the image level and ignore the feature-level representation learning. To further enhance the feature representation learning without annotations, a body of researchers has demonstrated promising results using contrastive representation learning. The main idea of contrastive representation learning is to maximise the differences between images from different groups as well as to maximise the agreements between images and their different augmented views, which makes the learned feature more robust to discriminate images from different groups. He *et al.* [17] presented Momentum Contrast (MoCo), which built a momentum encoder to dynamically queue the extracted CNN feature for a fast contrastive feature learning. Chen *et al.* [7] proposed a simple framework for contrastive learning (SimCLR) by introducing a simple non-linear transformation between the feature representation before calculating the contrastive loss and adding multiple commonly-used data augmentations methods. Current contrastive learning methods mainly focus on improving the discriminating capacity of CNN-based models on image classification tasks. To further transfer the robust feature-level learning of contrastive learning mechanism to the downstream tasks (i.e., segmentation), we propose to embed the contrast feature learning in the registration architecture to extract feature maps with richer information for better segmentation.

3. Proposed Technique

This section contains two key parts: i) the registration architecture which employs a CNN-based probabilistic model to calculate image-to-image transformation pattern z, and ii) our new full optimisation model that is composed of a reconstruction and a smooth loss and combined with a contrastive learning mechanism.

3.1. Our Technique Overview

In this section, we describe the workflow of our technique, in which we highlight their main components that are described in detail in the following subsections.

We display the overall workflow of our unsupervised segmentation architecture in Figure 1. Our technique takes as input two 3D images, the unaligned image x and the reference image y, which are fed into a two-symmetric weight-shared 3D CNNs to extract the highly semantic feature maps from the unaligned and reference images – these parts are illustrated in blue and yellow colours in Figure 1. We then employ a contrastive loss on the projection of the two extracted CNN feature maps, which forces the network to contrast the difference between the two CNN feature maps. The two symmetric weight-shared CNNs can generate more robust CNN feature maps via back-propagating the contrastive loss during training.

Based on the contrasted feature maps of the two CNNs, we use a single decoder to integrate all the feature maps, with different resolutions of the two CNNs, and estimate the transformation pattern z – see the green part in Figure 1. We first concatenate the feature maps with the same resolutions of the two CNNs. We then adopt a decoder to recursively use the feature maps, with high-level semantic information (low resolutions) to refine the feature maps with low-level detailed information (high resolutions), until we obtain a feature map with the same resolution as the input images. We introduce a probabilistic model to estimate the optimal transformation pattern z, which is based on the recovered image-resolution feature map. After training and finding

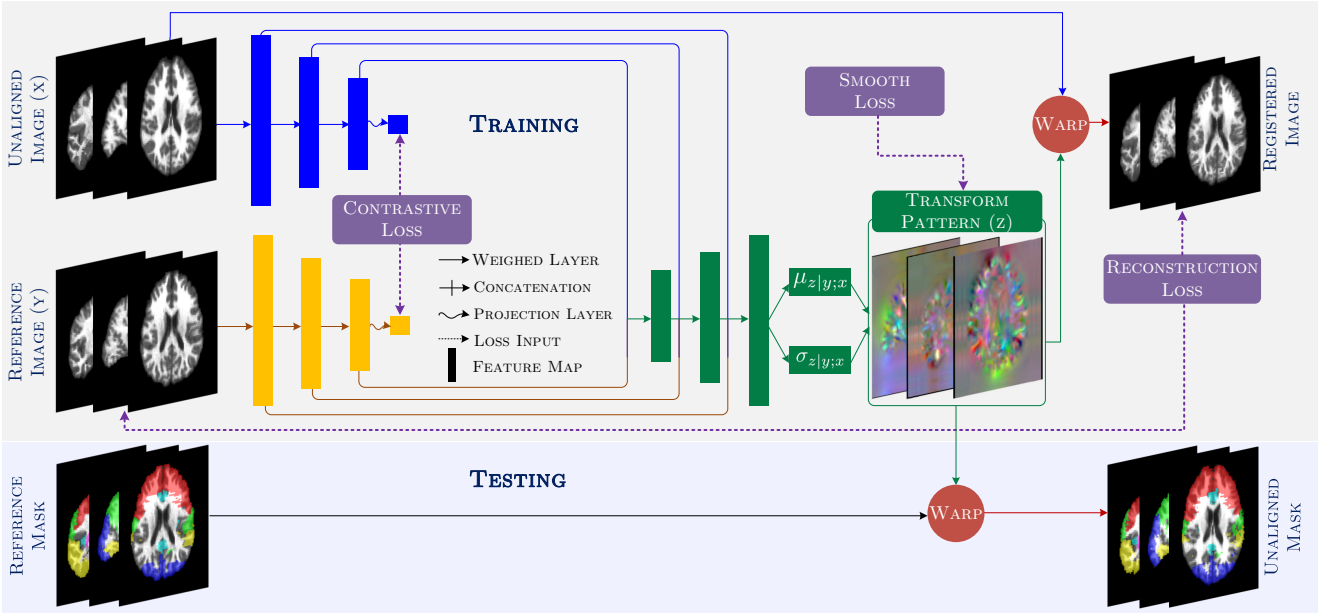

Figure 1. The overall workflow of the proposed network architecture. It takes two images as input for the two CNNs, which generates two sets of feature maps. Two projection layers are applied on the feature maps to produce two vector presentations of the input images. Based on these vector presentations, a contrastive loss is employed. It then recursively concatenates and enlarges the contrasted CNN feature maps, from the two CNNs, using a single decoder until the transformation pattern is obtained. Based on the transformation pattern, the reconstruction and smooth loss are adopted to ensure that the registration is well-performed. After the above training process is completed, we adopt the learned transformation pattern to transfer the reference mask to obtain the segmentation mask of the unaligned image.

the optimal z, we obtain the segmentation output by aligning the segmentation mask of the reference images to the coordinate system of the unaligned image. We do this using the optimal z via a spatial transform network [20].

3.2. Transformation Pattern Estimation

Given a pair images called unaligned image $x : \mathcal{X} \rightarrow \Omega$ and a reference image $y : \mathcal{X} \rightarrow \Omega$, where $\mathcal{X} = [w] \times [h] \times [d] \subset \mathbb{Z}^3$ is the input domain and Ω is the value data domain (i.e., gray scale), we seek to determine an optimal transformation pattern z, which parametrises a spatial transformation function denoted as ψ_z, such that the transformed unaligned image $x \circ \phi_z$ is aligned with y.

To compute the transformation pattern, we estimate z by maximising the posterior registration probability $p(z|y;x)$ given x and y – that is, to estimate the central tendency of the posterior probability (maximum a posteriori estimation MAP). However, to solve the partition function is intractable and cannot be solved analytically. To address this problem, one can use variational inference and approximate the solution through an optimisation problem over variational parameters. Following this principle, we adopt a CNN-based variational approach to compute $p(z|y;x)$. We first introduce an approximate posterior probability $q_\psi(z|y;x)$ which we assume is normally distributed. To measure the similarity between these two distributions, a divergence $D(q_\psi(z|y;x)||p(z|y;x))$ measure can be applied, e.g., [2, 34, 29]. We use the most commonly used divergence: the Kullback-Leibler (KL) divergence [26, 8]. With this purpose, we seek to minimise the KL divergence from $q_\psi(z|y;x)$ to $p(z|y;x)$ which expression reads:

$$\begin{aligned}
\psi^* &= \min_\psi \mathrm{KL}[q_\psi(z|y;x) \,||\, p(z|y;x)] \\
&= \min_\psi \mathbb{E}_q[log\ q_\psi(z|y;x) \ - log\ p(z|y;x)] \\
&= \min_\psi \mathbb{E}_q[log\ q_\psi(z|y;x) \ - log\ p(z,y;x)] + log\ p(y;x) \\
&= \min_\psi \mathrm{KL}[q_\psi(z|y;x) \,||\, p(z)] - \mathbb{E}_q[log\ p(y|z;x)] + const,
\end{aligned} \tag{1}$$

where $const$ is a normalisation constant and $q_\psi(z|y;x)$ comes from a multivariate normal distribution $\mathcal{N}$:

$$q_\psi(z|y;x) = \mathcal{N}(z; \mu_{z|y;x}, \sigma^2_{z|y;x}) \ , \tag{2}$$

where $\mu_{z|y;x}$ and $\sigma_{z|y;x}$ denote the mean and standard variance of the distribution respectively, which can be directly learned through the convolutional layers; as shown in Figure 1. Whilst $p(z)$ and $p(y|z;x)$ follow the multivariate normal distribution, which are modelled as:

$$p(z) = \mathcal{N}(z; 0, \sigma_z^2) \ , \tag{3}$$

$$p(y|z;x) = \mathcal{N}(y; x \circ \phi_z, \sigma^2) \ , \tag{4}$$

where σ_z is the variance (a diagonal matrix) of this distribution and $x \circ \phi_z$ is the noisy observed registered image in which σ^2 is the variance of the noisy term.

3.3. Full Optimisation Model

In this section, we detail our optimisation model that is composed of a reconstruction and smooth loss and a contrastive mechanism.

Reconstruction & Smooth Loss. According to the derivation from (1), there are two terms to be optimised. The first term is the KL divergence between the approximate posterior probability $q_\psi(z|y;x)$ and the prior probability $p(z)$, and the second term is the expected log-likelihood $E_q[log\ p(y|z;x)]$. Based on our assumptions from (2)-(4), the derivation is written as:

$$\begin{aligned}\mathcal{L}_{cs}(\psi;x,y) =& \mathrm{KL}[q_\psi(z|y;x)\ ||\ p(z)] - \mathbb{E}_q[log\ p(y|z;x)] \\ =& \frac{1}{2}[tr(\sigma^2_{z|y;x}) + ||\mu_{z|y;x}|| - log\ det(\sigma^2_{z|y;x})] \\ &+ \frac{1}{2\sigma^2_{z|y;x}}||y - x \circ \phi_z||^2.\end{aligned} \tag{5}$$

For sake of clarity in the notation, we decouple $\mathcal{L}_{cs}$ to detail the terms. The first term (the second line) is a close form of $KL[q_\psi(z|y;x)$ It enforces the two distribution $q_\psi(z|y;x)$ and $p(z)$ to be as close as possible. We called this term the smooth loss:

$$\mathcal{L}_{smooth} = \frac{1}{2}[tr(\sigma^2_{z|y;x}) + ||\mu_{z|y;x}|| - log\ det(\sigma^2_{z|y;x})] \tag{6}$$

The second term (the third line) is the expected log-likelihood $E_q[log\ p(y|z;x)]$. It enforces the registered image $x \circ \phi_Z$ to be similar to reference image y, which we refer as our reconstruction loss:

$$\mathcal{L}_{recon} = \frac{1}{2\sigma^2_{z|y;x}}||y - x \circ \phi_z||^2 \tag{7}$$

Contrastive Loss. Contrastive learning is a learning paradigm that seeks to learn distinctiveness. It aims to maximise agreements between images and their augmented views, and from different groups via a contrastive loss in a latent space. In our work, we follow a four components principle for the contrastive learning process, which is described next.

The first component in our contrastive learning process is the 3D images. The image contents are basically the same including the number of brain structures and the relative locations of each structure. They are mainly different in the structure size. Therefore, we view the unaligned and the reference images as images sampled from different augmented views. As our second component, we set the two CNN encoders as weight-shared. This with the purpose of ensuring that the CNN-based encoder can extract unified CNN features from the unaligned and reference images. For the third component, we adopt a fully-connected layer, as the projection layer, to map the CNN-features to a latent space where contrastive loss is applied. Finally, our fourth component is our contrastive loss following the standard definition presented in [16, 7].

From our four-component process, we can now formalise the loss we used in our framework. It is based on a contrastive loss [16, 7, 17] which is a function whose value is low when the image is similar to its augmented sample and dissimilar to all other samples. Formally, given a set of images I, we view our unaligned image x and reference image y as an augmented image pair, and any other images in I as negative samples. Moreover, we denote $sim(u,v) = \frac{u^\intercal v}{||u||\cdot||v||}$ as the cosine similarity between u and v. We then define the contrastive loss function as:

$$\mathcal{L}_{contrast} = -log \frac{\exp\left(sim(x,y)/\tau\right)}{\sum\limits_{i \in I} \mathbb{1}_{i \neq y} \exp\left(sim(x,i)/\tau\right)} \tag{8}$$

where $\mathbb{1}_{i \neq y} \in \{0,1\}$ is an indicator, which values 1 only when $i \neq y$. We also define τ as a temperature hyperparameter [39].

Optimisation Model. Our unsupervised framework is composed of three loss functions. The first two ones are directly derived from the optimisation model described in (5), in which the image-to-image reconstruction loss and the transformation pattern (deformation field) smooth loss are introduced. Whilst the third loss is the contrastive loss as in (8) which forces the network to contrast the difference between unaligned images and reference images. The total loss is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \alpha\mathcal{L}_{smooth} + \beta\mathcal{L}_{contrast} \tag{9}$$

where α and β are the hyper-parameters balancing $\mathcal{L}_{recon}$, $\mathcal{L}_{smooth}$ and $\mathcal{L}_{contrast}$, which we empirically set as 1 and 0.01, respectively.

4. Experimental Results

In this section, we describe the range of experiments that we conducted to validate our proposed technique.

4.1. Dataset Description & Evaluation Protocol

We evaluate our technique using two benchmarking datasets: the LONI Probabilistic Brain Atlas (LPBA40) dataset [33] and the MindBoggle101 dataset [23]. The characteristics are detailed next.

Dataset Description. The LPBA40 dataset [33] is composed of a series of maps from the regions of the brain. It contains 40 T1-weighted 3D brain images, from 40 human healthy volunteers, of size 181×217×181 with uniformly space of 1×1×1 mm^3. The 3D brain volumes were manually segmented to identify 56 structures. Whilst the MindBoggle101 dataset [23] is composed of a collection of 101 T1-weighted 3D brain MRIs from several public datasets: OASIS-TRT-20, NKI-TRT-20, NKI-RS-22,

TECHNIQUE	FRONTAL	PARIETAL	OCCIPITAL	TEMPORAL	CINGULATE	PUTAMEN	HIPPO	AVG
UtilzReg [37]	0.691	0.617	0.612	0.665	0.665	0.710	0.692	0.665
VoxelMorph [4]	0.875	0.726	0.737	0.794	0.630	0.618	0.630	0.716
FAIM [25]	0.879	0.720	0.740	0.781	0.643	0.710	0.634	0.729
Our Method	**0.906**	**0.757**	**0.752**	**0.828**	**0.681**	**0.713**	**0.704**	**0.763**

Table 1. Numerical comparison of our technique vs SOTA techniques for the the LPBA40 dataset. The numerical values display per region $Dice$ similarity metric whilst the last column refers to the average of the $Dice$ metric over all regions.

TECHNIQUE	FRONTAL	PARIETAL	OCCIPITAL	TEMPORAL	CINGULATE	AVG
UtilzReg [37]	0.482	0.456	0.425	0.385	0.446	0.439
VoxelMorph [4]	0.585	0.549	0.480	0.611	0.572	0.559
FAIM [25]	0.621	0.589	0.490	0.623	0.591	0.583
Our Method	**0.644**	**0.620**	**0.537**	**0.703**	**0.640**	**0.629**

Table 2. Numerical comparison of our technique vs SOTA techniques on the MindBoggle dataset. The values displayed reflect the computed $Dice$ metric per each brain region. 'AVG' denotes the average of the Dice metric over all regions.

MMRR-21, and Extra-18. It contains 101 skull-stripped T1-weighted 3D brain MRI volumes, from healthy subjects, of size $182\times218\times182$ that are evenly spaced by $1\times1\times1$ mm^3. From this dataset, we use 62 MRI images from OASIS-TRT-20, NKI-TRT-20 and NKI-RS-22, since they are already wrapped to MNI152 space and have manual segmentation masks (50 anatomical labels). To evaluate our technique against the state-of-the-art unsupervised brain image segmentation, we use the widely-used metric the Dice similarity coefficient [14].

4.2. Implementation Details & Running Scheme

In this section, we provide the implementation details and the training & testing scheme that we followed to produce the reported results.

Implementation Details. In our experiments, we follow the training and testing splitting setting from [28] for the LPBA40 dataset. We group the 56 structures into seven large regions such that we can show the segmentation results more intuitively. Moreover and for the MindBoggle101 dataset, we follow the protocol from [25] and use 42 images from NKI-TRT-20 and NKI-RS-22 for training and 20 images from OASIS-TRT-20 for testing. For evaluation purposes, we also group the 50 small regions into five larger regions of interest. Our proposed technique has been implemented in Pytorch [30].

To ensure a large batch size and facilitate the training process, we first resize the 3D data (volume) to $91\times109\times91$ with an evenly spacing of $2\times2\times2$ mm^3. We then follow the standard pre-processing protocol to normalise the images to have zero mean and unit variance. Moreover, we enrich our input data by applying three different transformations (data augmentation) in the following order: random flips in y coordinate, random rotation with an angle of less than 10 degrees, and random crops of size $80\times106\times80$.

Training Scheme. In our training scheme, we initialise the parameters of all convolutional layers following the initialisation protocol of that [18]. We also initialise the parameters of all batch normalisation layers using random variables drawn from Gaussian distributions with zero mean and 0.1 derivation [24]. Moreover, we use Adam optimiser [22] with a batch size of 8. We set the initial learning rate to 3×10^{-3} and then decrease it by multiplying 0.1 every 20 epochs, and terminate the training process after 200 epochs. Our technique took $\sim$ 20 hours to train on a single Tesla P100 GPU.

Testing Scheme. After training, we performed unsupervised segmentation based on the learned parameters. We first fed the unaligned image x and the reference image y into the trained network to calculate the transformation relation z. Then, we used a spatial transform network (STN) [20], to align the segmentation mask of the reference image according to the calculated z, to obtain the segmentation result of the unaligned image. On average, our method takes less than 10 seconds to process one whole MRI image on a single GPU (Tesla P100 GPU).

4.3. Results & Discussion

In this section, we present the numerical and visual results outlined in previous subsections, and discuss our findings and how they are compared with current existing techniques.

Comparison with the State-of-the-arts. We begin by evaluating our method against the state-of-the-art meth-

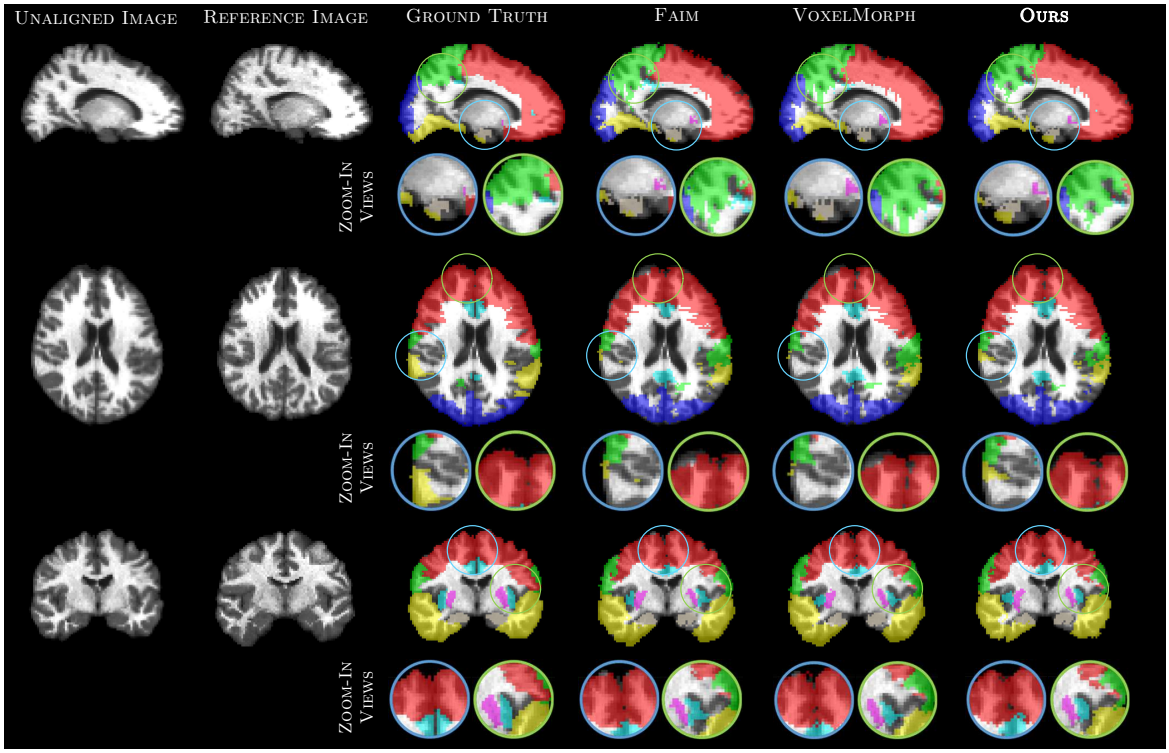

Figure 2. Visual comparisons between our technique and unsupervised SOTA techniques for segmentation. The rows show the three views from the 3D data. The third column displays the ground truth whilst the last three samples segmentation from our technique and the compared ones. Zoom-in views are displayed to highlight interesting regions where our technique performs better than the other models.

ods unsupervised brain image registration methods on the LPBA40 dataset: UtilzReg [37], VoxelMorph [3], and FAIM [25]. We remark that we use the registration architecture from other methods to predict the transformation pattern z and then adopt the same testing scheme as our network for image segmentation. We report the global results in Table 1 for the LPBA40 dataset, in order to understand the general behaviour and performance of our technique over the SOTA methods. The displayed numbers are the per regions and average of the image metrics across this dataset. In a close look at the results, we observe that the compared techniques perform similarly close to each other, whilst our technique outperforms all other techniques for all regions by a significant margin. This behaviour is consistent on the Mindboggle101 dataset whose results are reported in Table 2. From these numerical results, we can observe that the performance gain of our technique is even more significant for all regions. These comparisons, therefore, show that the combination of contrastive learning mechanism and registration architecture can better discriminate brain structures and generate more accurate segmentation outputs.

To further support our experimental results, we present a set of visual comparisons of a selection of images for our technique and the compared ones. We start by displaying the unsupervised segmentation outputs in Figure 2. By visual inspection, we observe that the segmentation outputs, generated from FAIM and VoxelMorph, tend to fail to segment correctly several relevant regions of the brain, and they do not adapt correctly to the contour of the brain structure. The zoom-in views in Figure 2 highlight these effects – for example the red region on the brain that the compared techniques fail to correctly segment; and the yellow region that is not well-captured by FAIM and VoxelMorph. By contrast, our technique was able to perform better in this regard by capturing fine details on both cases. Overall, our technique is able to accommodate better with the fine details in the brain regions, producing segmentation closer to the ground truth.

As a by-product of our methods, we also provide the registration results. We observe that our technique was able to better preserve the fine details in the registration outputs than the compared techniques. Most notable, this can be observed in the zoom-in views. For example, one can observe that our technique is able to preserve better the brain structures whilst the compared techniques fail in these terms. Moreover, we can see that our registration outputs are closer to the reference image displaying less blurry effects whilst keeping fine details. We now underline the main message of our paper: *our optimisation model fulfills the intended purposes, and at this point in time our technique outperforms the SOTA unsupervised segmentation results.*

Ablation Study. To demonstrate that each component

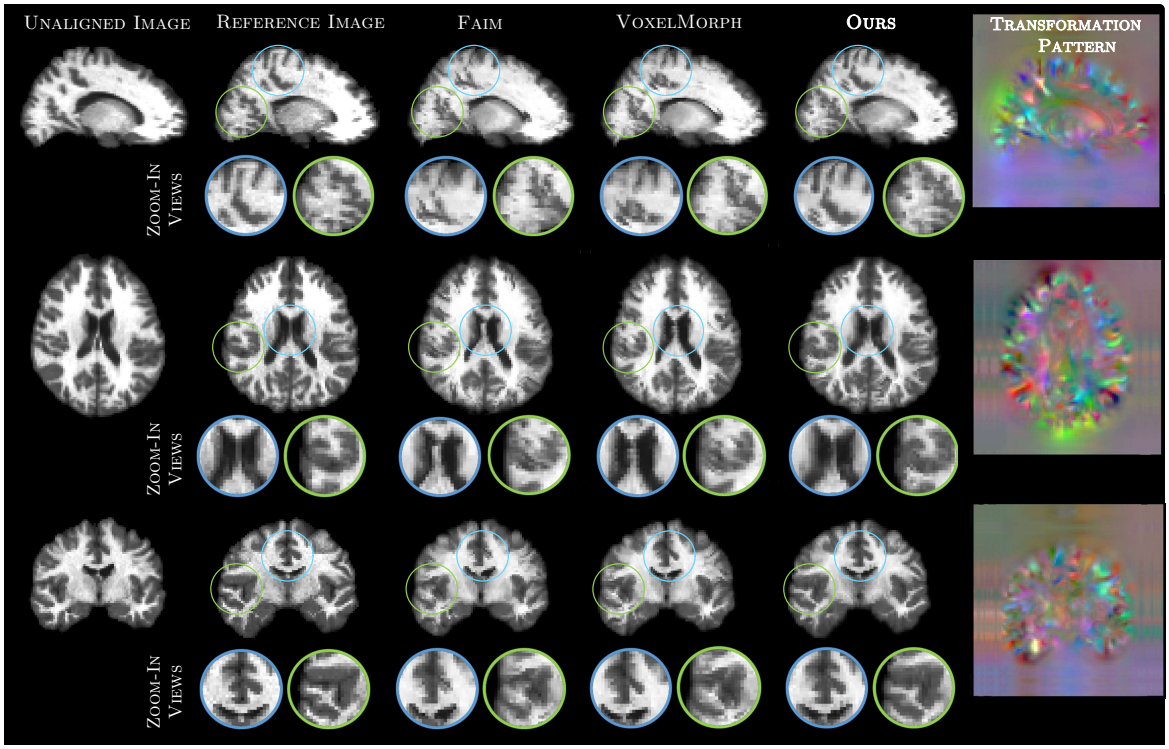

Figure 3. Visual comparison of our technique and the SOTA techniques for image registration. The rows display the x, y, and z views from the 3D medical data. The columns display outputs samples from FAIM, VoxelMorph, and our technique. The zoom-in views show interesting structures that clearly show the improvement in terms of preserving the brain structures and fine details. The last column (deformation field) presents the transformation pattern z between unaligned image and reference image produced by our method.

	LPBA40	MindBoggle101
Fully Supervised Baseline		
U-Net (Upper Bound)	0.832	0.811
Our Unsupervised Technique		
w/ $\mathcal{L}_{recon}$	0.702	0.552
w/ $\mathcal{L}_{recon}$ + $\mathcal{L}_{smooth}$	0.716	0.559
w/ $\mathcal{L}_{recon}$ + $\mathcal{L}_{contrast}$	0.751	0.604
w/ $\mathcal{L}_{recon}$ + $\mathcal{L}_{smooth}$ + $\mathcal{L}_{contrast}$	**0.763**	**0.629**

Table 3. Ablation Study, in terms of the mean Dice metric, of our proposed technique. Top part displays a numerical comparison of our technique vs fully supervised segmentation. Bottom part displays a comparison of our technique with its different components.

of our technique fulfills a purpose, we include an ablation study to evaluate their influence on the performance. We consider our three major components in our model whose results are displayed in Table 3. Our ablation study is performed on the two benchmark datasets, in order to understand the general behaviour of our technique. The results are displayed in terms of the Dice metric and we progressively evaluate our method with different losses combination. From the results in Table 3, we can observe that whilst the contrastive mechanism, in our technique, indeed provides a positive effect in terms of performance, it benefits our carefully designed components. In this work, we also pose the question of – at this point in time, what is the performance gap between supervised and unsupervised segmentation techniques? To respond to this question, we use as baseline U-Net [32]. From the results in the table, one can observe that our unsupervised technique is reaching an unprecedented performance narrowing the gap between supervised methods.

5. Conclusion

This paper presents a novel CNN-based registration architecture for unsupervised medical image segmentation. Firstly, we proposed to use unsupervised registration-based segmentation by capturing the image-to-image transformation pattern. Secondly, to promote the image- and feature-level learning, for better segmentation results, we embed a contrastive feature learning mechanism into the registration architecture. Our network can learn to be more discriminative to the different images via contrasting unaligned image and reference images. We show that our carefully designed optimisation model mitigates some major drawbacks of existing unsupervised techniques. We demonstrate, through a several experiments, that at this point in time our technique is able to report state-of-the-art results for unsupervised medical image segmentation.

References

[1] Bruno Alfano, Arturo Brunetti, Eugenio M Covelli, Mario Quarantelli, Maria Rosaria Panico, Andrea Ciarmiello, and Marco Salvatore. Unsupervised, automated segmentation of the normal brain using a multispectral relaxometric magnetic resonance approach. *Magnetic resonance in medicine*, 37(1):84–93, 1997. 1

[2] Shun-Ichi Amari. \alpha -divergence is unique, belonging to both f-divergence and bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, 2009. 4

[3] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018. 7

[4] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019. 6

[5] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: A review. *Frontiers in Cardiovascular Medicine*, 7:25, 2020. 1

[6] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*, 170:446–455, 2018. 1, 2

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2, 3, 5

[8] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 4

[9] Wenhui Cui, Yanlin Liu, Yuxing Li, Menghao Guo, Yiming Li, Xiuli Li, Tianle Wang, Xiangzhu Zeng, and Chuyang Ye. Semi-supervised brain lesion segmentation with an adapted mean teacher model. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 554–565. Springer, 2019. 1

[10] Adrian V Dalca, John Guttag, and Mert R Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 9290–9299, 2018. 1

[11] Adrian V Dalca, Evan Yu, Polina Golland, Bruce Fischl, Mert R Sabuncu, and Juan Eugenio Iglesias. Unsupervised deep learning for bayesian brain MRI segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 356–365. Springer, 2019. 1

[12] Alexander de Brebisson and Giovanni Montana. Deep neural networks for anatomical brain segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 20–28, 2015. 1

[13] Nameirakpam Dhanachandra and et al. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771, 2015. 1

[14] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 6

[15] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002. 2

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 5

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 9729–9738, 2020. 2, 3, 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 6

[19] Shiying Hu, Eric A Hoffman, and Joseph M Reinhardt. Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. *IEEE transactions on medical imaging*, 20(6):490–498, 2001. 1

[20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 2, 4, 6

[21] Alan Jose, S Ravi, and M Sambath. Brain tumor segmentation using k-means clustering and fuzzy c-means algorithms and its area calculation. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(3), 2014. 1, 3

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] Arno Klein and Jason Tourville. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in neuroscience*, 6:171, 2012. 5

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 6

[25] Dongyang Kuang and Tanya Schmah. Faim–a convnet method for unsupervised 3d medical image registration. In *International Workshop on Machine Learning in Medical Imaging*, pages 646–654. Springer, 2019. 6, 7

[26] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 4

[27] Chulhee Lee, Shin Huh, Terence A Ketter, and Michael Unser. Unsupervised connectivity-based thresholding segmentation of midsagittal brain MR images. *Computers in biology and medicine*, 28(3):309–338, 1998. 1

[28] Lihao Liu, Xiaowei Hu, Lei Zhu, and Pheng-Ann Heng. Probabilistic multilayer regularization network for unsupervised 3d brain image registration. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 346–354. Springer, 2019. 6

[29] Tom Minka et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005. 4

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019. 6

[31] Nara M Portela, George DC Cavalcanti, and Tsang Ing Ren. Semi-supervised clustering for MR brain image segmentation. *Expert Systems with Applications*, 41(4):1492–1497, 2014. 1

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 234–241. Springer, 2015. 1, 2, 8

[33] David W Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L Narr, Russell A Poldrack, Robert M Bilder, and Arthur W Toga. Construction of a 3d probabilistic atlas of human cortical structures. *Neuroimage*, 39(3):1064–1080, 2008. 5

[34] Charles Stein et al. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972. 4

[35] Hidetomo Suzuki and Jun-ichiro Toriwaki. Automatic segmentation of head MRI images by knowledge guided thresholding. *Computerized medical imaging and graphics*, 15(4):233–240, 1991. 2

[36] Kai Tian, Shuigeng Zhou, and Jihong Guan. DeepCluster: A general clustering framework based on deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 809–825. Springer, 2017. 1, 3

[37] François-Xavier Vialard, Laurent Risser, Daniel Rueckert, and Colin J Cotter. Diffeomorphic 3D image registration via geodesic shooting using an efficient adjoint calculation. *International Journal of Computer Vision*, 97(2):229–241, 2012. 6, 7

[38] Shuxin Wang, Shilei Cao, Dong Wei, Renzhen Wang, Kai Ma, Liansheng Wang, Deyu Meng, and Yefeng Zheng. Lt-net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9162–9171, 2020. 1

[39] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 5

[40] Yan Yang, Allen Tannenbaum, Don Giddens, and Arthur Stillman. Automatic segmentation of coronary arteries using bayesian driven implicit surfaces. In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 189–192. IEEE, 2007. 2