Abstract—Medical image segmentation is an important task in medical imaging, as it serves as the first step for clinical diagnosis and treatment planning. While major success has been reported using deep learning supervised techniques, they assume a large and well-representative labeled 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 segmentation techniques have been proposed in the literature. Yet, none of the existing unsupervised segmentation techniques reach accuracies that come even near to the state-of-the-art of supervised segmentation methods. In this work, we present a novel optimization model framed in a new convolutional neural network (CNN)-based contrastive registration architecture for unsupervised medical image segmentation called CLMorph. The core idea of our approach is to exploit image-level registration and feature-level contrastive learning, to perform registration-based segmentation. First, we propose an architecture to capture the image-to-image transformation mapping via registration for unsupervised medical image segmentation. Second, we embed a contrastive learning mechanism in the registration architecture to enhance the discriminative capacity of the network at the feature level. We show that our proposed CLMorph 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.

Index Terms—Brain segmentation, contrastive learning, deep learning, image registration, image segmentation.

I. INTRODUCTION

Medical image segmentation is the task of partitioning an image into multiple regions, which ideally reflect qualities such as well-defined structures guided by 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 [1], lungs [2], and brain [3]. Medical image segmentation is a relevant task, as it is the first step for several clinical applications such as tumor localization and neuropsychiatric disorder diagnosis.

The body of literature has reported several successful techniques for medical image segmentation, in which major progress has been reported using fully supervised convolutional neural networks (CNNs) [5], [6], [7], [8] or partially supervised CNN methods [9], [10], [11], [12]. These techniques rely on learning prior mapping information from pairwise data (mapping from medical images to their manual segmentation) to achieve astonishing performance at the same level as, and sometimes outperforming, radiologists. However, a major constraint of these methods is the assumption of having a well-representative set of annotations for training, which is not always possible in the medical domain. Moreover, obtaining such annotations is time-consuming, expensive, and requires expert knowledge. To deal with these constraints, a body of research has explored unsupervised segmentation techniques [13], [14], [15], [16], in which no annotations are needed.

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 content with similar intensities [17], [18], [19], [20]. However, these methods are still limited in performance-wise; as it is difficult to learn any image-to-mask transformation mapping when relying solely on images. To further embed the transformation mapping 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 [16], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30]. The goal of this perspective is to find an optimal image-to-image transformation mapping to align a set of images in one coordinate system.

In this work, we aim to find an optimal image-to-image transformation mapping $\mathbf{z}$ between an unaligned image $\mathbf{x}$ and a reference image $\mathbf{y}$, which is usually formulated by...

Manuscript received 25 October 2022; revised 21 August 2023; accepted 5 November 2023. Date of publication 20 November 2023; date of current version 8 January 2025. The work of Lihao Liu was supported by the University of Cambridge through the GlaxoSmithKline (GSK) Scholarship and the Girton College Graduate Research Fellowship. The work of Angelica I. Aviles-Rivero was supported by the University of Cambridge through Centre for Mathematical Imaging in Healthcare (CMIH), Cantab Capital Institute for the Mathematics of Information (C2I2M), and Cambridge Centre for Data-Driven Innovation (C2D3). The work of Carola-Bibiane Schönlieb was supported in part by the Philip Leverhulme Prize; in part by the Royal Society Wolfson Fellowship; in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/S026045/1, Grant EP/T003553/1, Grant EP/N014588/1, and Grant EP/T017961/1; in part by the Wellcome Innovator Award under Grant 209875; in part by the Leverhulme Trust Project Unveiling the Invisible; in part by the European Union Horizon 2020 Research and Innovation Programme under the Marie Skodowska-Curie Grant 777826 NoMADS; in part by the Cantab Capital Institute for the Mathematics of Information; and in part by the Alan Turing Institute. (Corresponding author: Lihao Liu.)

The authors are with the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, CB3 0WA Cambridge, U.K. (e-mail: l6160@cam.ac.uk).

Digital Object Identifier 10.1109/TNNLS.2023.3332003
a metric function $\phi$. Benefiting 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 $y_{\text{seg}}$ back to the coordinate system of the unaligned image. This process is with the purpose of obtaining the segmentation mask of the unaligned image $x_{\text{seg}}$ using the optimal mapping 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 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 mapping information on the image level but fail to enhance the feature-level representation. Most recently, unsupervised feature representation learning has demonstrated promising results. In particular, contrastive-based models, such as SimCLR [31] and BYOL [32], are reaching performance comparable to those produced by supervised techniques for different tasks. The main idea is that by contrasting images to others, the differences between images are easily remembered within a network (i.e., learning distinctiveness), thus making the learned feature more robust to discriminate images with different labels. Moreover, feature-level computation forms the foundation of contrastive learning, which calculates the similarity metrics for each element of the compared items. As comparisons at the image level could be prohibitively computationally expensive, feature-level computation is necessary for contrastive learning. Furthermore, mapping images into the feature level using CNNs allows for a more high-level extraction of representative information from images, rather than assessing the entire image. 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, producing better unsupervised segmentation results.

In this article, we present a novel Contrastive Learning registration architecture based on VoxelMorph for unsupervised medical image segmentation, which we named CLMorph. Our proposed CLMorph is a simple yet effective unsupervised segmentation model. Unlike existing techniques, our approach combines 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. 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 mapping information $z$ from the contrasted CNN features. Finally, we use the spatial transform network of [33] along with the captured $z$ to align the segmentation mask, of the reference image to the coordinate system of unaligned images, to obtain the segmentation mask in an unsupervised manner. Our main contributions are given as follows.

1) 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, which we called CLMorph. To the best of our knowledge, this is the first work that embeds contrastive learning in a registration architecture for unsupervised segmentation.

2) 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 with feature-level contrast, we can largely improve the unsupervised segmentation performance and reduce the performance gap between supervised and unsupervised segmentation techniques.

3) 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.

II. RELATED WORK

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

A. Supervised Medical Image Segmentation

In this section, we revise exiting techniques that heavily rely on annotations (i.e., supervised techniques) and the close related paradigm of one-shot learning where labels are still needed for the technique to work.

1) Fully 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 handcrafted features, including thresholding [34], statistical model [35], and Bayesian model [36]. However, a major drawback is that they are not capable of capturing high-level semantic information. Thereby, they tend to fail in segmenting those organs accurately.

More recently, 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 [5], nnU-Net [6], VoxResNet [7], and MedicalNet [8]. Moreover, Transformer [37], with its self-attention mechanism and position encoding capabilities, has been widely applied in medical image segmentation tasks, such as SwinPA-Net [38] and TCSM [39]. Although effective in biomedical image segmentation, these supervised methods rely heavily on a well-representative set of annotations. This is a strong assumption in the medical domain as annotations are expensive to obtain (in many medical tasks, there is a need for at least two readers), highly uncertain, and require expert knowledge. Hence, when applying the trained model to another dataset, such approaches often fail to segment correctly.

2) One-Shot Medical Image Segmentation: Recently, the community has moved toward the perspective of developing techniques that require a tiny labeled set. For example, the principles of one-shot learning have been reported in the literature for medical image segmentation. For example, Zhao et al. [9] proposed DataAug, which uses the learned spatial and intensity transformation to synthesize labeled
images for one-shot medical image segmentation. Another work using this philosophy is LT-Net [10]. In that work, the authors embedded a cycle consistency loss into a registration architecture to perform one-shot medical image segmentation. However, in the training process, one still needs to provide a set of labels that are well-representative for the task at hand. An alternative to those sets of techniques is unsupervised learning, which has recently been a great focus of attention in the medical imaging community. This paradigm is the focus of this work and existing techniques are discussed next.

B. Unsupervised Medical Image Segmentation

1) Clustering: For unsupervised segmentation, clustering algorithms [17], [18], [19], [20] have been extensively explored. The idea is to divide the image into different groups of pixels/voxels according to the similarity of image intensities (i.e., each pixel value). Besides clustering techniques, works based on Autoencoders [40] and GANs [41] have also been explored for unsupervised medical image segmentation. However, due to the lack of segmentation masks, which determines whether it is impossible to learn any image-to-mask mapping information, these algorithms are still inefficient for the task of segmentation.

2) Registration Algorithms: Registration is the process of aligning multiple datasets into one common coordinate system. In this area, the most famous work is VoxelMorph. Dalca et al. [25] proposed VoxelMorph, which is a fundamental learning-based framework for deformable medical image registration, while in another work [16], the same research team introduces a probabilistic generative model using CNNs for unsupervised learning-based deformable image registration, overcoming the computational intensity of traditional methods and the limitations of recent learning-based techniques, achieving state-of-the-art accuracy and fast runtimes while ensuring diffeomorphic registration, and providing uncertainty estimates.

Influenced by VoxelMorph, some works try to enable different attributions based on this type of registration framework. Zhang [24] proposed an inverse-consistent deep network (ICNet) for unsupervised deformable image registration, incorporating inverse-consistent and antifolding constraints to encourage symmetric deformation between images and prevent registration errors, demonstrating superior performance over existing approaches on T1-weighted brain MRI scans for tissue segmentation and anatomical landmark detection. Zhao et al. [22] introduced an unsupervised learning method, volume tweening network (VTN), for 3-D medical image registration, featuring an end-to-end cascading scheme, efficient affine registration network integration, and an invertibility loss for backward consistency, achieving state-of-the-art performance and significant speed improvements over traditional methods. In the work of [23], the deep learning image registration (DLIR) framework for unsupervised affine and deformable image registration was introduced, which utilize CNN trained on image similarity, offering flexible designs for both affine and deformable registration, and achieving performance comparable to conventional methods but with a substantial increase in speed. The work of [21] proposed a method for learning a low-dimensional probabilistic deformation using a conditional variational autoencoder (CVAE), which enables the comparison, generation, and transportation of deformations between images.

More recently, Mok and Chung [28] proposed a deep Laplacian pyramid image registration network, addressing the limitations of existing deep learning-based methods in deformable image registration by handling larger deformations and maintaining desirable diffeomorphic properties, demonstrating significant outperformance over existing methods in both accuracy and speed on MR brain scan datasets. Zhao et al. [29] introduced recursive cascaded networks, a flexible and general architecture for unsupervised deformable image registration, where each cascade learns to perform progressive deformation, allowing iterative application for better fit between image pairs, demonstrating consistent gains and outperformance over state-of-the-art methods in 3-D medical images. Heinrich [30] proposed a network leveraging ideas from probabilistic dense displacement optimization, employing meaningful and theoretically proven constraints for differentiable displacement regularization in a weakly supervised setting, and addressing shortcomings in handling large deformations in medical images.

3) Registration-Based Segmentation: To enable the mapping information in the training process, another feasible solution is to cast the segmentation task an unsupervised deformable registration process [16], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30]. Instead of seeking the image-to-mask transformation mapping, the goal of unsupervised registration methods is to find an optimal image-to-image transformation mapping. 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 mapping information [26], [27]. 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 \( y_{\text{seg}} \) back to the coordinate system of the unaligned image to obtain the segmentation mask of the unaligned image \( x_{\text{seg}} \) using the optimal transformation mapping \( z \). Hence, as long as the image is correctly registered, the segmentation mask is automatically obtained, which we call registered segmentation.

4) Contrastive Learning: The current unsupervised registration methods only focus on capturing the mapping 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 unsupervised learning [42], [43] and contrastive representation learning including momentum contrast (MoCo) [44], SimCLR [31], contrastive multiview coding (CMC) [45], and BYOL [32]. The main idea of contrastive representation learning is to maximize the differences between images from different groups as well as to maximize the agreements between images and their different augmented views, such as SimCLR [31]. Another research line relies on teacher–student learning mechanisms to learn from each other without group difference.
We display the overall workflow of our unsupervised segmentation technique in Fig. 1. Our technique takes as input two 3-D images, the unaligned image $x$ and the reference image $y$, which are fed into a Siamese (two-symmetric) weight-shared 3-D CNNs to extract the highly semantic feature maps from the unaligned and reference images—these parts are shown in blue and yellow in Fig. 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 backpropagating 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 mapping $z$—see the green part in Fig. 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 mapping $z$, which is based on the recovered image-resolution feature map. After training and finding 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 [33].
B. Transformation Mapping Estimation

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

To compute the transformation mapping, we estimate $z$ by maximizing 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 (MAP) estimation). However, solving 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 optimization 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., [48], [49], [50]. We use the most commonly used divergence: the Kullback–Leibler (KL) divergence [51], [52].

With this purpose, we seek to minimize the KL divergence from $q_\psi(z|y; x)$ to $p(z|y; x)$, which expression reads

$$
\psi^* = \min_\psi \text{KL}[q_\psi(z|y; x) \mid p(z|y; x)]
$$

where $\text{KL}[q, p]$ is the evidence lower bound (ELBO), and const is a normalization constant. $q_\psi(z|y; x)$ comes from a multivariate normal distribution $\mathcal{N}$

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

where $\mu_\psi(y; x)$ and $\sigma^2_\psi(y; x)$ denote the mean and variance of the distribution, respectively, which can be directly learned through the convolutional layers, as shown in Fig. 1, while $p(z)$ and $p(y|z; x)$ follow the multivariate normal distribution, which are modeled as

$$
p(z) = \mathcal{N}(z; 0, \sigma^2_z)
$$

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

where $\sigma_z$ is the variance (a diagonal matrix) of this distribution and $x \circ \phi_z$ is the noisy observed image in which $\sigma^2_z$ is the noise of the noisy term. Whilst several works in the literature have addressed the problem of image registration from a probabilistic perspective, for example, the works of that [16], [21], we highlight that our framework presents a new interpretation, and in fact, our novel function is based on other principles such as contrastive term, smooth and reconstruction terms. They are explained in Section III-C.

C. Full Optimization Model

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

1) Reconstruction and Smooth Loss: According to the derivation from (1), there are two terms to be optimized. 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) to (4), the derivation is written as

$$
\mathcal{L}_{cs}(\psi; x, y) = \text{KL}[q_\psi(z|y; x) \mid p(z)] - E_q[\log p(y|z; x)]
$$

$$
= \frac{1}{2} \left[ \text{tr}(\sigma^2_\psi(y; x)) + ||\mu_\psi(y; x)||^2 - \log \det(\sigma^2_\psi(y; x)) \right]
$$

$$
+ \frac{1}{2\sigma^2_z}||y - x \circ \phi_z||^2.
$$

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

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

where $\log \det(\sigma^2_\psi(y; x))$ is a Jacobian determinant, which spatially smooths the means. 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

$$
\mathcal{L}_{recon} = \frac{1}{2\sigma^2_z}||y - x \circ \phi_z||^2.
$$

2) Contrastive Loss: Contrastive learning is a learning paradigm that seeks to learn distinctiveness. It aims to maximize 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 3-D 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 structure sizes. 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 is done 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 [53] and [31].

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
From our four-component process, we can now formalize the loss we used in our framework. It is based on a contrastive loss [31], [44], [53], 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 define \( \text{sim}(u, v) = (u^T v) / (\|u\| \cdot \|v\|) \) as the cosine similarity between \( u \) and \( v \). We then define the contrastive loss function as

\[
L_{\text{contrast}} = -\log \frac{\exp(\text{sim}(f(x), f(y))/\tau)}{\sum_{i \in I} \mathbb{1}_{i \neq x} \exp(\text{sim}(f(x), f(i))/\tau)}
\]  

(8)

where \( f() \) denotes the CNN encoder, \( f(x) \) and \( f(y) \) are the generated features from the CNN encoder, and \( \mathbb{1}_{i \neq x} \in [0, 1] \) is an indicator, which values 1 only when \( i \neq x \). We also define \( \tau \) as a temperature hyperparameter [54].

3) Optimization Model: Our unsupervised framework is composed of three loss functions. The first two ones are directly derived from the optimization model described in (5), in which the image-to-image reconstruction loss and the transformation mapping (deformation field) smooth loss are introduced, while the third 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

\[
L_{\text{total}} = L_{\text{recon}} + \alpha L_{\text{smooth}} + \beta L_{\text{contrast}}
\]

\[
= \frac{1}{2\sigma^2} \|y - x \circ \phi_y\|^2 + \alpha \left( \frac{1}{2} \text{tr}(\Sigma^2_{\text{y}, x} z) + \|\mu_{\text{y}, x}\|^2 \right)
\]

\[
- \log \text{det}(\Sigma^2_{\text{y}, x} z)
\]

\[
+ \beta \left( -\log \frac{\exp(\text{sim}(f(x), f(y))/\tau)}{\sum_{i \in I} \mathbb{1}_{i \neq x} \exp(\text{sim}(f(x), f(i))/\tau)} \right)
\]

(9)

where \( \alpha \) and \( \beta \) are the hyperparameters balancing \( L_{\text{recon}} \), \( L_{\text{smooth}} \), and \( L_{\text{contrast}} \), which we empirically set as 1 and 0.01, respectively. The full training and test pseudocode logic are presented in Algorithm 1.

IV. EXPERIMENTAL RESULTS

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

A. Dataset Description and Evaluation Protocol

1) Benchmark Datasets: We evaluate our technique using two benchmarking datasets: the LONI Probabilistic Brain Atlas (LPBA40) dataset [55] and the MindBoggle101 dataset [56]. The characteristics are detailed next.

The LPBA40 dataset [55] is composed of a series of maps from the regions of the brain. It contains 40 T1-weighted 3-D brain images, from 40 human healthy volunteers, of size 181 \( \times \) 217 \( \times \) 181 with a uniform space of 1 \( \times \) 1 \( \times \) 1 mm\(^3\). The 3-D brain volumes were manually segmented to identify 56 structures, while the MindBoggle101 dataset [56] is composed of a collection of 101 T1-weighted 3-D brain MRIs from several public datasets: OASIS-TRT-20, NKI-TRT-20, NKI-RS-22, MMRR-21, and Extra-18. It contains 101 skull-stripped T1-weighted 3-D brain MRI volumes, from healthy subjects, of size 182 \( \times \) 218 \( \times \) 182 that are evenly spaced by 1 \( \times \) 1 \( \times \) 1 mm. 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 MN152 space and have manual segmentation masks (50 anatomical labels).

2) Evaluation Metrics: To evaluate our technique against the state-of-the-art unsupervised brain image segmentation, we use three widely used metrics: the Dice similarity coefficient [57], Hausdorff distance (HD) [58], and average symmetric surface distance (ASSD) [59]. Dice measures the region-based similarity between the predicted segmentation

---

Algorithm 1 Training and Testing Scheme

**Training Process**

Input: unaligned image \( x \), reference image \( y \), segmentation mask for.

Output: deformation field \( z \), registered result \( x \circ \phi_z \).

for each training pair do

1. Feed both \( x \) and \( y \) into the same-weight 3D CNN to obtain highly semantic features \( f(x) \) and \( f(y) \).
2. Combine features \( f(x) \) and \( f(y) \), and input them into a decoder to restore values \( \mu_{\text{z}, y} \) and \( \sigma_{\text{z}, y} \).
3. Use \( \mu_{\text{z}, y} \) and \( \sigma_{\text{z}, y} \) to reconstruct the deformation field \( z \) using a probabilistic model (Transformation Mapping Estimation).
4. Apply the calculated deformation field \( z \) to \( x \) via a Spatial Transformer Network (STN) to obtain the registered result \( x \circ \phi_z \).
5. Ensure quality registration by calculating the total loss, comprising the reconstruction loss between the registered image \( x \circ \phi_z \) and reference image \( y \), smooth loss on the deformation field \( z \), and the contrastive loss on \( f(x) \) and \( f(y) \); optimize the total loss through the back-propagation process.

end

**Testing Process**

Input: unaligned image \( x \), reference image \( y \), segmentation of reference image \( \hat{y} \).

Output: deformation field \( z \), segmentation result of unaligned image \( \hat{y} \circ \phi_z \).

for each testing pair do

1. Feed both \( x \) and \( y \) into the 3D CNN using the saved weights from above Training Process to obtain highly semantic features \( f(x) \) and \( f(y) \).
2. Combine features \( f(x) \) and \( f(y) \), and input them into a decoder to restore values \( \mu_{\text{z}, y} \) and \( \sigma_{\text{z}, y} \).
3. Use \( \mu_{\text{z}, y} \) and \( \sigma_{\text{z}, y} \) to reconstruct the deformation field \( z \) using a probabilistic model.
4. Apply the calculated deformation field \( z \) to \( \hat{y} \) via a STN to get the segmentation result \( x \circ \phi_z \).

end
TABLE I
NUMERICAL COMPARISON OF OUR TECHNIQUE VERSUS SOTA TECHNIQUES FOR THE LPBA40 AND MINDBOGGLE101 DATASETS. THE NUMERICAL VALUES DISPLAY THE AVERAGE OF DICE, HD, AND ASSD OVER ALL REGIONS. THE HIGHER THE DICE IS, THE BETTER THE MODEL PERFORMS, WHILE THE LOWER THE HD AND ASSD ARE, THE BETTER THE MODEL PERFORMS. THE BEST PERFORMANCE IS DENOTED IN BOLD FONT. THE PER-REGION RESULTS ARE SHOWN IN FIGS. 2 AND 3

| U-Net (Upper Bound) | Dice | HD | ASSD | Dice | HD | ASSD |
|---------------------|------|----|------|------|----|------|
|                      | 0.832| 9.54| 1.028| 0.811| 10.25| 0.887|
| VoxelMorph [61]     | 0.665±0.06| 14.16±2.34| 1.906±0.09| 0.439±0.17| 18.21±3.04| 0.972±0.09|
| Syn [62]            | 0.701±0.04| 13.61±2.56| 1.986±0.08| 0.543±0.16| 16.83±3.68| 0.926±0.07|
| VoxelMorph [26]     | 0.716±0.08| 11.38±3.46| 1.874±0.41| 0.559±0.12| 15.03±2.92| 1.171±0.10|
| FAIM [63]           | 0.729±0.07| 11.04±4.11| 1.809±0.40| 0.583±0.04| 14.50±2.71| 1.108±0.08|
| Lap[IRN [28]        | 0.739±0.07| 10.77±3.85| 1.727±0.60| 0.607±0.04| 14.54±2.58| 1.041±0.08|
| VTN [29]           | 0.745±0.08| 11.11±3.41| 1.601±0.31| 0.626±0.04| 14.64±2.57| 1.007±0.09|
| PDD-Net [30]       | 0.749±0.06| 10.75±3.77| 1.714±0.42| 0.632±0.05| 15.21±2.92| 0.990±0.10|
| ViT-V-Net [64]     | 0.750±0.20| 11.05±4.23| 1.701±0.35| 0.622±0.01| 14.35±3.16| 1.101±0.16|
| C2FVIT [65]       | 0.757±0.03| 10.40±2.66| 1.555±0.39| 0.639±0.03| 13.17±3.88| 0.991±0.11|
| CLMorph (Ours)     | 0.763±0.07| 10.38±2.59| 1.458±0.31| 0.646±0.04| 12.76±2.52| 0.892±0.05|

Fig. 2. Numerical comparisons between our framework and existing SOTA techniques on the LPBA40 dataset. The results are reported in terms of the Dice, HD, and ASSD metrics reflecting a per region, of the brain, evaluation.

Fig. 3. Comparisons on the MindBoggle101 dataset. We compare our technique against existing SOTA methods. The comparisons displayed are in terms of the Dice, HD, and ASSD metrics for each region on the brain.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fig. 4. Visual comparisons between our technique and unsupervised SOTA techniques for segmentation. The rows show the three views from the 3-D data. The third column displays the ground truth, while the last three samples show segmentation from our technique and the compared ones. Zoomed-in views are displayed to highlight interesting regions where our technique performs better than the other models.

Table II

|                | LPBA40   | MindBoggle101 |
|----------------|----------|---------------|
| UtilizReg [61] | 0.665±0.06 | 0.440±0.16    |
| SyN [62]       | 0.700±0.04 | 0.540±0.15    |
| VoxelMorph [26]| 0.709±0.07 | 0.558±0.10    |
| FAIM [63]      | 0.720±0.08 | 0.566±0.09    |
| LapIRN [28]    | 0.727±0.09 | 0.608±0.08    |
| VTN [29]       | 0.722±0.06 | 0.598±0.04    |
| PDD-Net [30]   | 0.733±0.05 | 0.612±0.08    |
| CLMorph (Ours) | 0.750±0.08 | 0.631±0.05    |

1) Implementation Details: In our experiments, we follow the training and testing splitting setting from [66] for the LPBA40 dataset, where the first 30 images are used as training datasets and the last ten images are used as testing datasets. We group the 56 structures into seven large regions such that we can show the segmentation results more intuitively. Moreover, for the MindBoggle101 dataset, we follow the protocol from [63] and use 42 images from NIKI-TRT-20 and NIKI-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. As for the reference image, we select the image that is the most similar to the anatomical average as the atlas. In particular, we selected image #30 as the reference image for the LPBA40 dataset and image #39 for the Mindboggle101. Our proposed technique has been implemented in Pytorch [67].

We follow the standard preprocessing protocol to normalize the images to have zero mean and unit variance. We randomly select two images as the pairwise input data (unaligned image and reference image). Moreover, we enrich our input data by...
applying three different transformations (data augmentation) in the following order: random flips in the y-coordinate, random rotation with an angle of fewer than $10^\circ$, and random crops of size $160 \times 192 \times 160$.

2) Training Scheme: In our training scheme, we initialize the parameters of all convolutional layers following the initialization protocol of that [68]. We also initialize the parameters of all batch normalization layers using random variables drawn from Gaussian distributions with zero mean and 0.1 derivation [69]. Moreover, we use the Adam optimizer [70] with a batch size of 8. We set the initial learning rate to $3 \times 10^{-3}$ and then decrease it by multiplying 0.1 every 20 epochs and terminate the training process after 200 epochs. Due to the small number of the dataset, which is a common issue in the 3-D medical image area, we did not split the data into validation sets from the training or testing sets for model selection. Hence, we use the above learning rate decay policy to decrease the learning rate to a small value in the last few epochs to ensure that the model does not diverge and save the last epoch model as our final model. This is a standard protocol followed in the area. Our technique took $\sim 20$ h to train on a single Tesla P100 GPU.

3) 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 [33] 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 s to process one whole MRI image on a single GPU (Tesla P100 GPU).

C. Results and Discussion

In this section, we present the numerical and visual results outlined in Section IV-A2 and discuss our findings and how they are compared with current existing techniques.

TABLE III

|                     | LPBA40       | MindBoggle101 |
|---------------------|--------------|---------------|
|                     | Dice  | HD  | ASSD | Dice  | HD  | ASSD |
| U-Net (Upper Bound) | 0.832 | 9.54 | 1.028 | 0.811 | 10.25 | 0.887 |

|                     | Fully Supervised Baseline | Our Unsupervised Technique |
|---------------------|---------------------------|-----------------------------|
| $L_{\text{recon}}$ | 0.702±0.07 | 12.82±3.33 | 1.842±0.58 | 0.552±0.20 | 15.11±2.26 | 1.210±0.13 |
| $L_{\text{recon}} + L_{\text{smooth}}$ | 0.716±0.08 | 11.38±3.46 | 1.874±0.41 | 0.559±0.12 | 15.03±2.92 | 1.171±0.10 |
| $L_{\text{recon}} + L_{\text{contrast}}$ | 0.751±0.10 | 10.92±2.28 | 1.801±0.47 | 0.604±0.14 | 14.85±2.77 | 0.993±0.21 |
| $L_{\text{recon}} + L_{\text{smooth}} + L_{\text{contrast}}$ (ours) | 0.763±0.07 | 10.38±2.59 | 1.458±0.31 | 0.646±0.04 | 12.76±2.52 | 0.892±0.05 |
| $L_{\text{recon}} + L_{\text{byol}}$ | 0.744±0.18 | 11.41±3.29 | 1.693±0.77 | 0.619±0.13 | 14.73±2.39 | 1.153±0.89 |
| $L_{\text{recon}} + L_{\text{smooth}} + L_{\text{byol}}$ | 0.761±0.08 | 10.41±2.12 | 1.462±0.53 | 0.649±0.08 | 12.88±2.44 | 0.894±0.10 |

Fig. 6. Statistical analysis. Multiple comparisons are performed followed by a paired Wilcoxon test and $p$-values adjusted using the Bonferroni method. Our technique reported significant statistically different among all compared techniques.
We also present visual results shown in Fig. 7. We observe contrastive learning, our method yields promising results. Table II makes it evident that even when focusing on registration performance with the integration of approach. To further support our numerical 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 Fig. 4. 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 zoomed-in views in Fig. 4 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. In contrast, our technique was able to perform better in this regard by capturing fine details in 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.

We also included the visualization of the deformation field generated from the contrasted feature (w/ contrastive loss) and noncontrast feature (w/o contrastive loss). According to Fig. 5, when the deformation between the moving and fixed images is large, the deformation field generated from the contrasted feature is smoother because the contrasted feature maps have high likeness, while the deformation field generated from noncontrast features tends to be less smooth because there is a huge difference between the feature maps extracted from the moving and fixed images.

3) Statistical Analysis: We support statistically our visual and numerical findings. To do this, we run a Friedman test for multiple comparisons, $\chi^2(5) = 29.78$, $p < 0.0001$, followed by a Wilcoxon test for pairwise comparisons. As shown in Fig. 6, the pairwise test between groups revealed a statistically significant difference metricwise of our technique against the rest of compared techniques.

4) By-Product Registration Accuracy: As a by-product of our methods, we also provide the numerical and the visual registration results. Based on the learned deformation field, we register the segmentation mask of the unaligned image to the coordinate system of the reference image to get the aligned segmentation mask. Then, compare it with the reference image’s mask to get the dice scores, see Table II. The registration results demonstrate the effectiveness of our approach. Table II makes it evident that even when focusing solely on registration performance with the integration of contrastive learning, our method yields promising results. We also present visual results shown in Fig. 7. We observe that our technique was able to better preserve the fine details in the registration outputs than the compared techniques. Most notably, this can be observed in the zoomed-in views. For example, one can observe that our technique is able to preserve better brain structures, while 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 while keeping fine details. We now underline the main message of this article: our optimization model fulfills the intended purposes, and at this point in time, our technique outperforms the SOTA unsupervised segmentation results.

5) Hyperparameter Searching: We also run a set of new experiments for the hyperparameter analysis. Specifically, we set the weight for the smooth loss ($\alpha$) and contrastive loss ($\beta$) as 0.001, 0.01, 0.1, 1, and 10. As shown in Fig. 8, the best combination is our current setting ($\alpha = 1$ and $\beta = 0.01$), which gives the highest dice result of 0.763.

6) Ablation Study: To demonstrate that each component of our technique fulfills a purpose, we include an ablation study to evaluate their influence on performance. We consider our three major components in our model whose results are shown in Table III. Our ablation study is performed on the two benchmark datasets, in order to understand the general behavior of our technique. The results are displayed in terms of the Dice metric and we progressively evaluate our method with different loss combinations. From the results in Table III, we can observe that while 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 [5]. From the results, in Table III, one can observe that our technique opens the door for further investigation in unsupervised techniques, as the performance shows potential toward the one reported by supervised techniques.

7) Additional Experiments on a Cardiac Dataset: We also provide a set of new results using an additional cardiac dataset called ACDC [71]. The dataset is for medical image registration. It is composed of 150 labeled images. We selected 100 images for training and the remaining 50 images for testing. The results are reported in Table IV. From the results, we observe that our model outperforms the compared techniques in terms of Dice and ASSD metrics by a large margin while readily competing against LVM-S3 in terms of HD metric. These results further support the generalization and performance of our CLMorph technique.

8) Additional Experiments on a Lung Dataset: We now present findings on lung CT scans using the DirLab dataset [72], which is composed of lung CT images.
Fig. 7. Visual comparison of our technique and the SOTA techniques for image registration. The rows display the $x$, $y$, and $z$ views from the 3-D medical data. The columns display outputs samples from FAIM, VoxelMorph, and our technique. The zoomed-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 mapping $z$ between the unaligned image and the reference image produced by our method.

Fig. 8. Hyperparameter searching. We display different combination values for $\alpha$ and $\beta$ used in (9). The parameter values are evaluated using the Dice metric. The best combination is $\alpha = 1$ and $\beta = 0.01$, which gives the highest Dice result of 0.763.

characterizing the inspiration-toExpiration motion of the lung from ten patients. Each image has 300 anatomical landmarks as ground truth. Furthermore, we evaluate the target registration error (TRE). The TRE is defined as the distance between a set of manually identified corresponding points, typically referred to as landmarks, in the registered image and the corresponding points in the target image. As shown in the new results of Table V, the TRE score confirms that our CLMorph consistently performs well across these three diverse datasets, reinforcing the method’s applicability to a wider range of medical image segmentation tasks.

V. CONCLUSION

This article presents a novel CNN-based registration architecture for unsupervised medical image segmentation. First, we proposed to use unsupervised registration-based segmentation by capturing the image-to-image transformation mapping. Second, 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 optimization model mitigates some major drawbacks of existing unsupervised techniques. We demonstrate, through several experiments, that our technique is able to report state-of-the-art results for unsupervised medical image segmentation. While supervised techniques still report better performance than unsupervised ones, in this work, we show the potential in terms of performance when no labels are available. This is of great interest,
particularly in domains such as the medical area where annotations require expert knowledge and are expensive.

REFERENCES

[1] C. Chen et al., “Deep learning for cardiac image segmentation: A review,” *Frontiers Cardiovascular Med.*, vol. 7, p. 25, Mar. 2020.

[2] J. Liu, A. I. Aviles-Rivero, H. Ji, and C.-B. Schönlieb, “Rethinking medical image reconstruction via shape prior, going deeper and faster: Deep joint indirect registration and reconstruction,” *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101930.

[3] S. Hu, E. A. Hoffman, and J. M. Reinhardt, “Automatic lung segmentation for accurate quantification of volumetric X-ray CT images,” *IEEE Trans. Med. Imag.*, vol. 20, no. 6, pp. 490–498, Jun. 2001.

[4] A. de Brébisson and G. Montana, “Deep neural networks for anatomical brain segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 20–28.

[5] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Oper. (MICCAI)*, Cham, Switzerland: Springer, 2015, pp. 234–241.

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

[7] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, “VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images,” *NeuroImage*, vol. 170, pp. 446–455, Apr. 2018.

[8] S. Chen, K. Ma, and Y. Zheng, “Med3D: Transfer learning for 3D medical image analysis,” *IEEE J. Biomed. Health Inform.*, 2019, arXiv:2006.00625.

[9] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, “Data augmentation using learned transformations for one-shot medical image segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8535–8545.

[10] S. Wang et al., “LT-net: Label transfer by learning reversible voxelwise correspondence for one-shot medical image segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9159–9168.

[11] N. M. Portela, G. D. C. Cavalcanti, and T. I. Ren, “Semi-supervised clustering for MR brain image segmentation,” *Exp. Syst. Appl.*, vol. 41, no. 4, pp. 1492–1497, Mar. 2014.

[12] W. Cui et al., “Semi-supervised brain lesion segmentation with an adapted mean teacher model,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Oper. (MICCAI)*, Cham, Switzerland: Springer, 2019, pp. 554–565.

[13] B. Alfano et al., “Unsupervised, automated segmentation of the normal brain using a multispectral relaxometric magnetic resonance approach,” *Magn. Reson. Med.*, vol. 37, no. 1, pp. 84–93, Jan. 1997.

[14] C. Lee, S. Huh, T. A. Ketter, and M. Unser, “Unsupervised connectivity-based thresholding segmentation of midsagittal brain MR images,” *Comput. Biol. Med.*, vol. 28, no. 3, pp. 309–338, May 1998.

[15] A. V. Dalca, J. Guttag, and M. R. Sabuncu, “Anatomical priors in convolutional networks for unsupervised biomedical segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9290–9299.

[16] A. V. Dalca, E. Yu, P. Golland, B. Fischl, M. R. Sabuncu, and J. E. Iglesias, “Unsupervised deep learning for Bayesian brain MRI segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Oper. (MICCAI)*, Cham, Switzerland: Springer, 2019, pp. 356–365.

[17] H. P. Ng, S. H. Ong, K. C. W. Foong, P. S. Goh, and W. L. Nowinski, “Medical image segmentation using K-means clustering and improved watershed algorithm,” in *Proc. IEEE Southwest Symp. Image Anal. Interpreta.* 2006, pp. 61–65.

[18] B. N. Li, C. K. Chu, S. Chang, and S. H. Ong, “Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation,” *Comput. Biol. Med.*, vol. 41, no. 1, pp. 1–10, Jan. 2011.

[19] A. Jose, S. Ravi, and M. Sambath, “Brain tumor segmentation using K-means clustering and fuzzy C-means algorithms and its area calculation,” *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 2, no. 3, pp. 1–6, 2014.

[20] K. Tian, S. Zhou, and J. Guan, “DeepCluster: A general clustering framework based on deep learning,” in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2017, pp. 809–825.
E. Osipov et al., “Hyperseed: Unsupervised learning with vector symbolic architectures,” IEEE Trans. Neural Netw. Learn. Syst., early access, Nov. 16, 2022, doi: 10.1109/TNNLS.2022.3211274.

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9726–9735.

Y. Tian, D. Krishnan, and P. Isola, “Contrastive multiview coding,” 2019, arXiv:1906.05849.

X. Zhao et al., “Contrastive learning for label-efficient semantic segmentation,” 2020, arXiv:2012.06985.

K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, “Contrastive learning of global and local features for medical image segmentation with limited annotations,” 2020, arXiv:2006.10511.

S.-I. Amari, “α-divergence is unique, belonging to both f-divergence and Bregman divergence classes,” IEEE Trans. Inf. Theory, vol. 55, no. 11, pp. 4079–4106, Nov. 2009.

C. Stein et al., “A bound for the error in the normal approximation to the distribution of a sum of dependent random variables,” in Proc. 6th Berkeley Symp. Math. Statist. Probab., vol. 2, San Francisco, CA, USA: The Regents of the University of California, 1972, pp. 583–602.

T. Minka et al., “Divergence measures and message passing,” Microsoft Res., Cambridge, U.K., Tech. Rep. MSR-TR-2005-173, 2005.

R. Kullback and R. A. Leibler, “On information and sufficiency,” Ann. Math. Statist., vol. 22, no. 1, pp. 79–86, 1951.

T. M. Cover, Elements of Information Theory. Hoboken, NJ, USA: Wiley, 1999.

R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2, Jul. 2006, pp. 1735–1742.

Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3733–3742.

D. W. Shattuck et al., “Construction of a 3D probabilistic atlas of human cortical structures,” NeuroImage, vol. 39, no. 3, pp. 1064–1080, Feb. 2008.

A. Klein and J. Tourville, “101 labeled brain images and a consistent human cortical labeling protocol,” Frontiers Neurosci., vol. 6, p. 171, Dec. 2012.

L. R. Dice, “Measures of the amount of ecologic association between species,” Ecology, vol. 26, no. 3, pp. 297–302, Jul. 1945.

D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, “Comparing images using the Hausdorff distance,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 9, pp. 850–863, Jun. 1993.

V. Yeghiazaryan and I. Voiculescu, “Family of boundary overlap metrics for the evaluation of medical image segmentation,” J. Med. Imag., vol. 5, no. 1, Feb. 2018, Art. no. 015006.

L. Liu, X. Hu, L. Zhu, C.-W. Fu, J. Qin, and P.-A. Heng, “ϕ-net: Stacking densely convolutional LSTMs for sub-cortical brain structure segmentation,” IEEE Trans. Med. Imag., vol. 39, no. 9, pp. 2806–2817, Sep. 2020.

F.-X. Vialard, L. Risser, D. Rueckert, and C. J. Cotter, “Diffeomorphic 3D image registration via geodesic shooting using an efficient adjoint calculation,” Int. J. Comput. Vis., vol. 97, no. 2, pp. 229–241, Apr. 2012.

B. B. Avants et al., “Advanced normalization tools (ANTs),” Insight j, vol. 2, no. 365, pp. 1–35, 2009.

D. Kuang and T. Schmah, “Faim—A ConvNet method for unsupervised 3D medical image registration,” in Proc. Int. Workshop Mach. Learn. Med. Imag. Cham, Switzerland: Springer, 2019, pp. 646–654.

J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, “VIT-V-net: Vision transformer for unsupervised volumetric medical image registration,” 2021, arXiv:2104.06408.

T. C. Mok and A. Chung, “Affine medical image registration with coarse-to-fine vision transformer,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2022, pp. 20835–20844.

L. Liu, X. Hu, L. Zhu, and P.-A. Heng, “Probabilistic multilayer regularizer for unsupervised 3D brain image registration,” in Proc. Int. Conf. Med. Image Comput. Assist. Intervent. (MICCAI), Cham, Switzerland: Springer, 2019, pp. 346–354.

A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 8026–8037.

K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1026–1034.

A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

O. Bernard et al., “Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?” IEEE Trans. Med. Imag., vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

R. Castillo et al., “A reference dataset for deformable image registration spatial accuracy evaluation using the COPDgene study archive,” Phys. Med. Biol., vol. 58, no. 9, pp. 2861–2877, May 2013.

Lihao Liu received the B.Eng. degree from the Department of Software Engineering, Chongqing University, Chongqing, China, in 2016, and the M.Phil. degree from the Department of Computer Science and Engineering (CSE), The Chinese University of Hong Kong, Hong Kong, in 2020. He is currently pursuing the Ph.D. degree with the Cambridge Image Analysis Group (CIA) and the Centre for Mathematical Imaging in Healthcare (CMIH), Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, U.K.

His research interests include medical image analysis and video understanding. Specifically, he focuses on using machine learning techniques to solve unsupervised medical image registration and segmentation tasks, and surgical video processing and video content understanding tasks.

Angelica I. Aviles-Rivero received the Ph.D. degree from the Polytechnic University of Catalonia, Barcelona, Spain, in 2017.

She is currently a Senior Research Associate with the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, U.K. She has widely explored how to develop hybrid models, where explicit mathematical modeling and machine learning are considered. This line of research (Math + ML + X) has allowed her to gain expertise in a wide range of applications (X), including medical imaging, computational photography, computer graphics, and remote sensing, to name a few. Her research interests include the intersection of computational mathematics, computer vision, machine learning for applications to large-scale real-world problems, inverse problems, machine learning, and graph-based techniques for imaging and nonimaging application.

Carola-Bibiane Schönlieb received the B.S. degree from the Institute for Mathematics, University of Salzburg, Salzburg, Austria, in 2004, and the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 2009.

From 2004 to 2005, she held a teaching position in Salzburg. She held a post-doctoral position at the University of Göttingen, Göttingen, Germany, for one year. She became a Lecturer at the Department of Applied Mathematics and Theoretical Physics (DAMTP), University of Cambridge, in 2010, where she was promoted to a Reader in 2015 and a Professor in 2018. Since 2011, she has been a fellow of the Jesus College, University of Cambridge, where she is currently a Professor of applied mathematics with DAMTP, the Head of the Cambridge Image Analysis Group, the Director of the Cantab Capital Institute for Mathematics of Information, and the Director of the Engineering and Physical Sciences Research Council Center for Mathematical and Statistical Analysis of Multimodal Clinical Imaging. Her research interests include variational methods, partial differential equations, machine learning for image analysis, image processing, and inverse imaging problems.