DEEP LEARNING METHODS TO PROCESS AND ANALYSE MRI IMAGES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Evan Ma Yu

May 2022

DEEP LEARNING METHODS TO PROCESS AND ANALYSE MRI IMAGES

Evan Ma Yu, Ph.D.

Cornell University 2022

Our understanding of brain anatomy and physiology has advanced greatly thanks to the introduction of magnetic resonance imaging (MRI). As the technology develops and the amount of data multiplies, it is essential to develop methods to effectively and efficiently extract useful information from our data. However, scans are often collected under varying conditions, making analysis difficult. For this reason it is important to properly prepare the MRI for a quantitative and qualitative study. In this thesis, we investigate the use of deep learning models to prepare, process and analyse structural brain MRI scans. More specifically, we first introduce an unsupervised and interpretable method that register brain with high accuracy, especially in the context of large displacements. Then we show a segmentation strategy that requires only a single labeled example to train, while leveraging all the available unlabeled scans. Next, we present a novel method that warps brain template given a subject attributes. Finally, we discuss a scientific application of processed MRI and how our strategy can be useful to study neuroanatomical shape.

BIOGRAPHICAL SKETCH

Healthcare data plays a crucial role in the research, diagnosis, and treatment of diseases. Recent developments in hardware have led to an advancement in data collection. However, a huge gap exists between data acquisition and analysis. I am passionate about the development of novel technologies that will drive the analysis of data, closing this gap.

During my PhD, my research has focused on MRI scans of the brain. Contrasting natural images, obtaining MRI labels to train algorithms is very costly and time consuming. Additionally, there are many irregularities and confounding factors that make analysis challenging. Nevertheless, I enjoyed developing new and robust unsupervised algorithms to tackle each challenge. In addition, in order to hone my skill set, I participated in internships that provided relevant experience. Working at New York-Presbytarian Hospital, allowed me to experience intricacies involved in acquiring healthcare data, interacting with patients, and collaborating with doctors. Finally, while interning at HP, I spearheaded the development of a non-contact sensing method to monitor physiological measurements. These experiences provided me with a well-rounded background in academic, clinical and commercial research.

ACKNOWLEDGEMENTS

I am deeply thankful to my advisor, Mert R. Sabuncu. This thesis would not have existed without his guidance and support. I still remember his eagerness and enthusiasm to teach the first day I joined his lab. His unwavering mentorship then continued throughout every step of my graduate school journey, helping me overcome hardships and roadblocks. Words cannot express my gratitude for the experience I had in this lab.

I am lucky to have amazing labmates and friends who supported me throughout these years. I very grateful to Zhilu Zhang, Hubert Lin and Meenakshi Khosla for their friendship, insight and encouragement in every aspect of my life. I owe a lot of my success to the wonderful company in our lab: Heejong Kim, Alan Wang, Batuhan Karaman, Gia H. Ngo, Tianyu Ma, Matthew Pool, Carmen Khoo, Victor Butoi, Cagla Bahadir, Zijin Gu and Jinwei Zhang.

I am forever grateful for the unconditional support of my family, especially my parents, Maliang and Chunhao, and my sister, Gloria. Their endless love drives me forward in every step of my life.

To my family,

for their endless love and support

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

Medical imaging has become one of the essential components in healthcare. Advancements in imaging modalities such as magnetic resonance imaging (MRI) have given us an unprecedented look at the human anatomy and physiology. In the present, they play a key role in the diagnosis, treatment and research of diseases. The ubiquity of imaging modalities in clinical setting has led to an exponential increase of patients data. As a result, development of methods to effectively and efficiently extract useful information from clinical data is essential. However, it is not an easy task. There are many challenges that are often encountered in medical scans. Although large dataset of MRI exists, they are often obtained under different settings with varying degree of conditions and quality. Existing algorithms can easily break if the data is not preprocess accordingly. Alternatively, training your own model can be also difficult due to lack of image and labels pairs.

Recent resurgence of neural network or deep learning based models has enabled state-of-the-art performance in many tasks, particularly in computer vision. This breakthrough has enable super-human performance in problems such as classification and segmentation [101, 106, 127]. In this thesis, we have developed new methodologies that leverage the use of deep learning models to prepare, process and analyze MRI scans. Our proposed methods cover many different aspect within the pipeline of MRI image analysis. As we progress through the thesis, we will highlight how our contributions excel at each given task.

1.2 MRI Processing

In clinical and research settings, computational tools are often used to provide us insight into MRI images. However, scans are often collected under varying conditions, making analysis difficult. For this reason it is important to properly prepare the data for a quantitative and qualitative study. We will focus on structural brain MRI images in this thesis. In this section, some classical steps to process our data are highlighted.

1.2.1 Intensity Correction

The first step in many MRI protocol often involves defining a new range of values for the voxels in order to address contrast difference arising from different acquisition protocols. This can be simply accomplish through histogram matching, linear or piece-wise linear transformation. However, this step does not remove noise that are present in the MRI. Image acquisition inherently introduce noise to the MRI. Therefore, depending on the level of noise, it might be desirable to implement denoising methods. Common technique involves the use of specialized filters or exploiting sparseness and self-similarity properties within the image [60, 120, 121]. Other types of intensity artifact arises from inhomogeneities in the MRI. They are the result from inconsistencies in the radio-frequency coils and object-dependent interaction during scan. This is known as the bias field which is a low-frequency artifact that corrupts the signal intensity across the image. There are two common approaches to overcome this issue, namely a prospective and a retrospective strategy [162]. The former aims to reduce the artifact during the acquisition process. These can range from improvement of the acquisition hardware, development

of specialized MRI sequences and usage of calibration protocols [28, 30, 102]. As with the retrospective approach, the aim is to reduce the bias field after the image acquisition. These methods includes the use of special filter, surface fitting, segmentation and histogram [2, 61, 158, 173, 203].

1.2.2 Skull Stripping

Skull stripping involves removing non-brain brain tissues from the scan. Manual extraction of the brain can performed, but it is a time consuming and expensive task [143]. It is common to rely on automatic or semi-automatic approaches to help us remove non-brain tissues [92, 143]. Many types of approches have been developed. For example, intensity-based methods relies in the the distribution of intensity values to differentiate between the brain tissue and non-brain tissue [32, 68]. Morphological-based methods uses filters, morphological operations and/or thresholding to remove the skull [22, 24, 156]. Deformable surface-based method iteratively mold a surface to the boundary of the brain through some optimization constraint [32, 87, 159]. Atlas-based method requires fitting a template to the subject's MRI in order to determine different regions in the head [39, 54]. Finally, there exist many methods that combines idea from the aforementioned techniques [15, 25, 94, 153].

1.2.3 Registration

Head movement introduces motion artifacts to the scan. It is also becoming common for patients to have multiple scans across different modalities and time points [175]. This allows clinician and researchers to track anatomical changes over

time. At the same time, different MRI sequences help highlight different region within the brain due to differences in fluid, muscle and fat content. As a result, there is a inherent need to register subjects across time and space. This entails finding the appropriate transformation to align corresponding anatomical structures. Registered images then provides common reference frame to study structural changes due to variation of attributes such as difference of age or presence of disease. Not to mention, many analysis pipelines also depends on the assumption of registered images in order to perform tasks like segmentation and motion tracking [4, 196]. There has been extensive review of classical approaches for medical image registration [76, 134]. A wide range of transformation can be applied to the misaligned image. Parametric transformation such as rigid, affine, and thin-spline have relatively lower computational complexity due to the smaller number of parameters that has to be learned [9, 152]. More general types of transformations are achieved by regularizing the deformation field. Common constraints includes diffusive regularisation, curvature regularization, elastic model, fluid model, topological preservation and diffeomorphism [7, 12, 17, 19, 53, 165, 168, 180].

1.2.4 Segmentation

Quantitative study of the MRI brain often requires us to identify the regions of interest (ROI). For instance, many downstream task such as morphological studies, disease identification and guided intervention requires proper delineation of the brain. During this step, we assign labels to groups of voxels that share characteristics of interest. It is common to segment the brain manually. Unfortunately, it is a time consuming task and requires expertise in the anatomy. In addition, poor reproducibility is often encountered during manual delineation due to inter- and

intra- user variability [63]. Therefore, there has been significant strides to develop automatic approaches.

The simplest method relies label propagation or fusion using brain atlas [83, 151,186]. On the other hand, learning-based method uses a training set containing images and labels [55, 170, 171]. While deformable-based method often optimize contours in an iterative manner given an objective [31, 47, 137]. Finally, region-based method groups voxels according to a given specified criteria [66,189,192,193]. A vast amount of literature also exist where a hybrid combination of the different techniques are used [70, 151, 172].

1.3 Deep Learning in Medical Imaging

Recent resurgence of neural network or deep learning based models enabled us to achieve the state of the art performance in most computer vision task [5, 65, 101, 106, 127]. Once trained, inference time is significantly faster than many classical methods since we don't have to repeat the optimization for each new test set. In addition, their execution time is significantly faster because they are highly parallelizable on GPUs. In the present, convolutional neural networks (CNNs) are the most successful type of of architecture in the domain of image analysis. Research on CNNs dates back decades [58, 103, 104], however they did not gain significant momentum until the breakthrough of AlexNet [101]. They are now prevalent in application and research, including in the field of medical imaging. We will briefly discuss relevant work used in MRI processing. For a more thorough review, we refer to the readers to [95].

For both skull stripping and brain segmentation, it is common to pass the MRI

volume or slices through a CNN to make the corresponding prediction. Working with 3D volume can provide better contextual information at each voxel, but it is more computationally expensive compared to working with 2D slices of the brain. Neural networks are typically trained end-to-end using image and label pairs. The models then output labels for each voxel or pixel within the scan [3,93,98,146]. Due to the difficulty of obtaining labeled dataset, there is also extensive work on weakly-supervised and unsupervised methods. These range from using anatomical prior, adversarial strategy, semantic constraint, and augmentation method [35,36,91,200]

In the realm of registration, the literature can be classified into supervised, weakly supervised or unsupervised method. Supervision can be provided through ground truth deformation, synthesized warping or deformation produced by classical methods [23,46,49,107,177,194]. On the other hand, weakly supervised models uses landmark or labels in the anatomy to guide registration [51, 81, 82]. Finally, unsupervised models can use similarity metric like that of classical method or losses on the representation of the network [13, 40, 50, 99, 140, 188].

1.4 Challenges

The aforementioned network-based approaches have achieved state-of-the-art results at their respective tasks. However, there are many challenges and direction yet to be explored. In the area of brain registration, most proposed approaches are not suitable for the registration of 3D images with large misalignment, and they often assume that the images are roughly in the same orientation and position. In addition, current neural network models directly output the deformation field or transformation parameters in a black-box fashion. On the other hand, there

is a constant need to develop better weakly supervised or unsupervised segmentation techniques. Supervised segmentation typically yield tools that are sensitive to changes in image characteristics. For instance, modifications of the imaging protocol, software, or hardware [88]. Finally, many registration and segmentation models relies on a fixed atlas or template. However, human brains are very heterogeneous. Demographic, clinical, or other con-founding factors can influence the shapes and sizes of brain regions. As a result, single and fixed templates can struggle to accommodate complex structural differences across our population.

1.5 Contributions

In this thesis, we propose novel methodologies to tackle problems that are present in MRI processing. Chapter 1 provides an overview of MRI processing pipeline, along with a review of existing methodologies and challenges.

Chapter 2 introduces the background on deep learning techniques relevant to our models. Each subsequent chapters then provides a detailed look at each of our contributions.

Chapter 3 presents an unsupervised end-to-end learning-based image registration framework that relies on automatically detecting corresponding keypoints. This leads to substantially more robust registration and yields better interpretability since the keypoints reveal which parts of the image are driving the final alignment. We demonstrate registration accuracy that surpasses current state-of-the-art methods, especially in the context of large displacements.

Chapter 4 presents the Segmentation Autoencoder (SAE), a weakly supervised

segmentation model. It is a novel perspective of segmentation as a discrete representation learning problem. It is also a variational autoencoder segmentation strategy that is flexible and adaptive. Our method, leverages all available unlabeled scans and merely requires a segmentation prior, which can be *a single unpaired* segmentation image.

Chapter 5 tackles the limitations of fixed templates. We developed a novel neural network model that captures morphometric variability across clinical and demographic groups. Our model learns to compute an attribute-specific spatial deformation that warps a brain template. We demonstrate the ability of our model to deform a brain template given a wide range of ages, presence of disease and different sexes.

Chapter 6 shows how processed MRI can be used to study anatomical shape. In this work, a novel machine learning strategy for studying neuroanatomical shape variation is introduced. Our model works with volumetric binary segmentation images, and does not require extraction of surface points or a mesh. The learned shape descriptor is invariant to affine transformations, including shifts, rotations and scaling.

Chapter 7 concludes this thesis. A summary of our contribution is listed along with future extension of our work.

CHAPTER 2

MACHINE LEARNING IN MEDICAL IMAGING

Deep networks, also known as neural networks and multiplayer perceptrons, originated as mathematical models for biological learning system [147]. In the present, they are not constrained to have biological realism. Instead, they are use as an efficient function approximator. Recent developments have allowed neural networks to achieve state of the art accuracy in many task including classification, segmentation, and detection. This chapter introduces the fundamental idea behind some of these models. Section 2.1 starts by defining the functional form of neural networks along with common architectures. Section 2.2 introduces strategies to regularize the models to allow better training and generalization. In Section 2.3, optimization strategies are presented. Finally, Section 2.4 showcases application of these models in the context of image processing and analysis.

2.1 Feed Forward Network

Many task consist of mapping an input $\boldsymbol{x}$ to an output $\boldsymbol{y}$. For example, $\boldsymbol{x}$ could be a brain MRI of a patient and we want to determine whether a tumor is present. Our goal is to approximate such mapping $\boldsymbol{y} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$ by optimizing the parameters $\boldsymbol{\theta}$ during training. Feed forward networks define a specific functional form by stacking multiple layers and forming different representations along the way. For a simple layer construction, we start with a linear combination with weights $\boldsymbol{W}^0$, bias $\boldsymbol{b}^0$ and a nonlinearity σ,

$$\boldsymbol{z}^0 = \sigma\left(\boldsymbol{W}^0\boldsymbol{x} + \boldsymbol{b}^0\right). \tag{2.1}$$

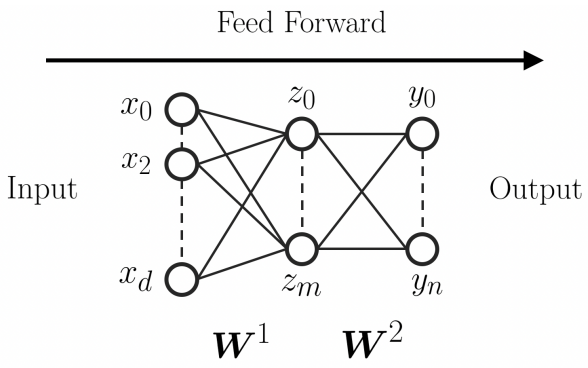


Figure 2.1: An example of fully connected neural network. The dimension of input x, representation z, and output y is given by d, m, n, respectively.

This process can be repeated using the output of each layer, forming a fully connected and feed foward network

$$\boldsymbol{y} = \boldsymbol{W}^L \left(\boldsymbol{W}^{L-1} \cdots \sigma \left(\boldsymbol{W}^0 \boldsymbol{x} + \boldsymbol{b}^0 \right) + \boldsymbol{b}^{L-1} \right) + \boldsymbol{b}^L. \tag{2.2}$$

A simple depiction of this process is shown in Fig. 2.1. The presence of nonlinear or activation function is important, without σ the resulting function will be a linear mapping from $\boldsymbol{x}$ to $\boldsymbol{y}$. Many type of activation are used in practice, common ones include ReLU, softmax, sigmoid, and tanh [64]. An important aspect of the neural network is its ability to approximate arbitrary functions given the appropriate weights and architecture. Under some settings, neural network can learn any function provided that it has enough width and depth [33, 80, 113].

2.2 Regularization

In practice, we are interested in the performance of the network beyond the training dataset. Therefore, method that mitigate overfitting is essential. In this section we discuss some of the strategies that are frequently employed.

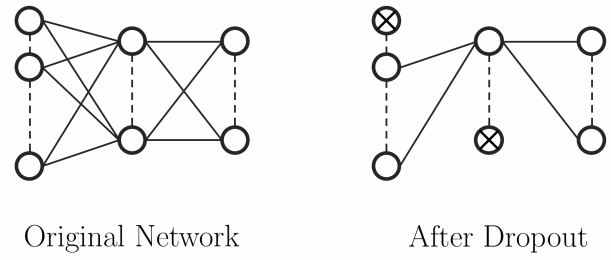


Figure 2.2: Random subset of the network nodes are temporarily removed during dropout.

2.2.1 Norm Penalty

In my traditional machine learning approaches, we often have a penalty term for the parameters of our model. This term limit the capacity of our algorithm by constraining the magnitude of the parameters. The same idea can be applied to the neural network. For an arbitrary loss $\mathcal{L}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta})$, we can introduce penalty $R(\boldsymbol{\theta})$ through

$$\mathcal{L}'(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}) + \alpha R(\boldsymbol{\theta}). \tag{2.3}$$

The weight α controls the significance of the regularizer. Commonly used functions of R includes L^1- and L^2-norm.

2.2.2 Dropout

Combing models through bagging improves the performance of machine learning models. However, it can be prohibitively expensive to train an array of neutral networks. Dropout provides a fast way to approximate combination of models by temporarily removing nodes of the network [164]. Note that unlike bagging, dropout networks are not independent because they share a subset of their parameters. Removal of nodes is done randomly and independently. All its incoming

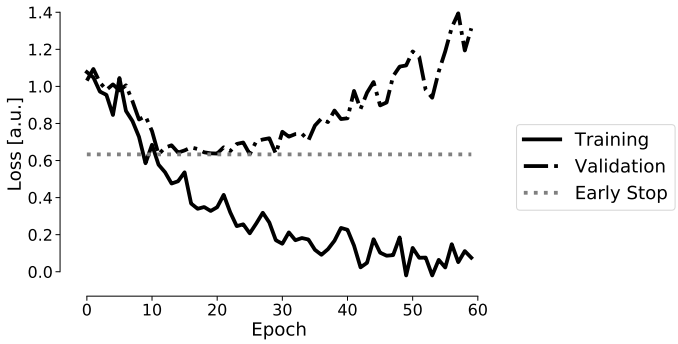


Figure 2.3: Loss on training and validation dataset. Early stop line denotes the performance if the model is stopped at the lowest point of the validation curve.

and outgoing connections are dropped, as depicted in Fig. 2.2. The probability of an element being masked out is controlled by a hyperparameter p.

2.2.3 Early Stopping

In the setting of supervised learning, it is common to have a training, validation and test set. Neural networks are capable of memorizing random noise [198]. Therefore, it can an be easy for the model to obtain very good accuracy or low loss in the training dataset. By keeping track of the validation performance, we are able to tune hyperparameters and detect signs of overfitting. Fig. 2.3 depict this occurrence. As the network memorize the training dataset, the validation performance gets worse overtime. However, if we stop the model at the lowest point of validation performance, we can obtain better generalization error.

2.2.4 Data Augmentation

Neural networks benefit from a large training dataset [100, 118]. However, in some studies, it can be very hard to obtain additional data. In order to make the network

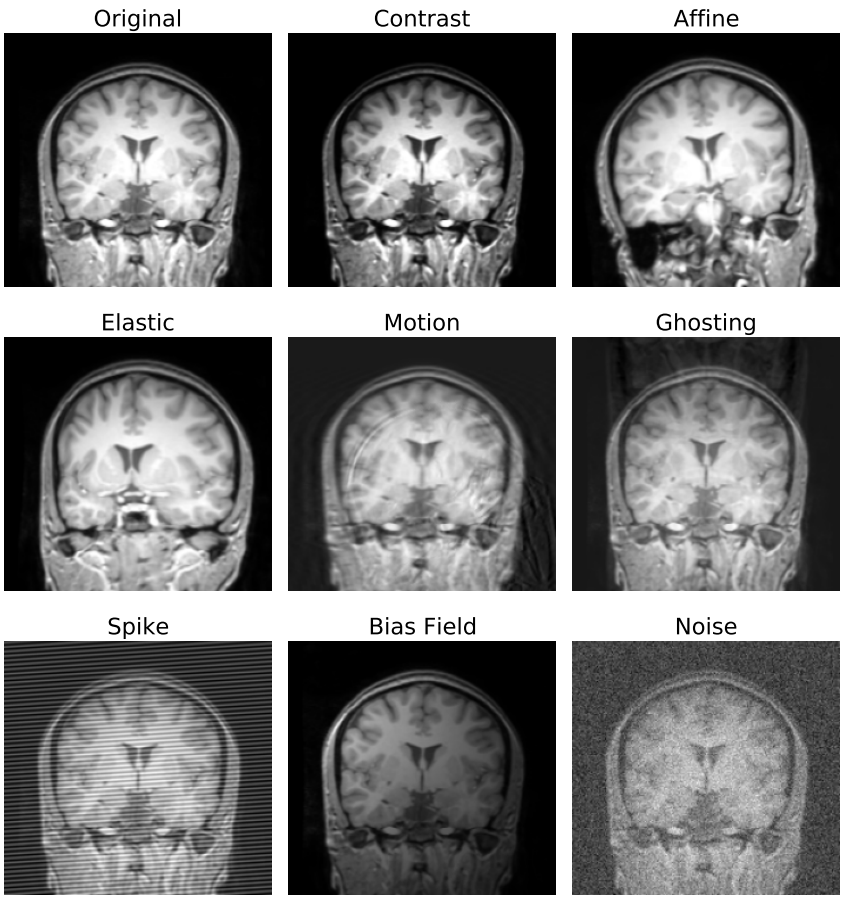


Figure 2.4: Example of data augmentation. Different transformations were introduce to the original brain to simulate variations that can be seen during model deployment

more robust against variations on testing data, we can artifically introduce artifacts and transformation during training. This enables us to create additional data point to present to our model. It is important to decide on which transformation are relevant during model deployment. For example, in MRI imaging, we might introduce affine and elastic deformation to accentuate differences across subject brains and placements within the scan. At the same time, we can also reproduce common artifacts due to motion, ghosting, spike and bias field. Fig. 2.4 shows some of the possible transformation that can introduce during training.

2.3 Optimization

For a loss function $\mathcal{L}$, we are interested in finding the parameters of our model such that it minimizes our cost $\boldsymbol{\theta}^* = \text{argmin}\, \mathcal{L}(\boldsymbol{\theta})$. Gradient descent usually used to accomplish this task especially for machine learning models where the close solution does not exist or it is very expensive to compute. The gradient $\nabla\mathcal{L}(\boldsymbol{\theta})$ allow us to investigate how small changes in $\boldsymbol{\theta}$ affects our loss. By moving in the direction negative of the gradient [26], we can iteratively minimize the loss

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla\mathcal{L}_i(\boldsymbol{\theta}_t). \tag{2.4}$$

The learning rate η control the size of the step size at each iteration t. The loss function is computed with each individual sample of the training data with size n. It is often expensive to compute the full gradient with all the data points. Therefore, we randomly choose m subsamples where $m < n$. This is known as batch or stochastic gradient descent (SGD). The remaining of this section briefly introduces popular variants of gradient descent algorithms.

2.3.1 Momentum

Gradient descent with momentum add the history of past update in each iteration. Loss landscape can be steeper in one dimension with respect to another. Momentum favors the direction in where there are smaller gradient oscillation. For this algorithm, the update rule is:

$$\begin{aligned} \boldsymbol{m}_t &= \mu\boldsymbol{m}_{t-1} + \eta\nabla\mathcal{L}(\boldsymbol{\theta}_t) \\ \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \boldsymbol{m}_t, \end{aligned} \tag{2.5}$$

where μ is the momentum's hyperparameter.

2.3.2 RMSprop

Root mean square propagation (RMSprop) [77] is an adaptie learning rate algorithm. The squared of the gradient is accumulated through a running average

$$\boldsymbol{v}_t = \gamma \boldsymbol{v}_{t-1} + (1-\gamma)[\nabla \mathcal{L}(\boldsymbol{\theta}_t)]^2. \tag{2.6}$$

The learning rate is then scaled by the squared of the resulting term

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{\sqrt{\boldsymbol{v}_t + \epsilon}} \nabla \mathcal{L}(\boldsymbol{\theta}_t), \tag{2.7}$$

for some hyperparameter γ and noise ϵ to prevent division by zero.

2.3.3 Adam

Adaptive moment optimizer or Adam [96] is another method that adapts the learning rate. A running average (with hyperameter β_1 and β_2) of the first moment and second moment of the gradient are kept:

$$\begin{aligned} \boldsymbol{m}_t &= \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1) \nabla \mathcal{L}(\boldsymbol{\theta}_t) \\ \boldsymbol{v}_t &= \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)[\nabla \mathcal{L}(\boldsymbol{\theta}_t)]^2. \end{aligned} \tag{2.8}$$

These estimates are then corrected through:

$$\begin{aligned} \hat{\boldsymbol{m}}_t &= \frac{\boldsymbol{m}_t}{1-\beta_1^t} \\ \hat{\boldsymbol{v}}_t &= \frac{\boldsymbol{v}_t}{1-\beta_2^t}\ , \end{aligned} \tag{2.9}$$

where β^t denotes β to the power of t. The update rule is given by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{\sqrt{\hat{\boldsymbol{v}}_t + \epsilon}} \hat{\boldsymbol{m}}_t, \tag{2.10}$$

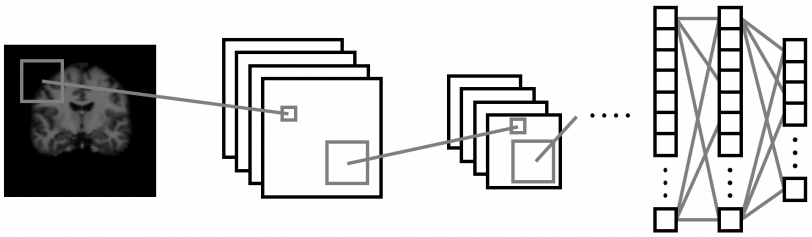


Figure 2.5: An example of a simple convolutional neural network architecture for classification problems. Input image is passed through a series of convolutional layers before it is flatten and passed through a fully connected network.

2.4 Applications

In the present, neural networks have become the most prevalent method for many tasks. In computer vision and medical imaging, the most common type of architecture is the convolutional neural networks or CNNs. Recently, transformers have been increasingly use for image data [45, 110, 197]. However, CNNs remain state of the art for many tasks. In this section, we briefly discussed some popular applications of these models.

2.4.1 Convolutional Neural Networks

Within a convolutional layer, we take the dot product between a kernel K (with learnable weights) and a location within the image I. The kernel is then moved to another location of the image. We repeat this process until the image is covered

$$h(x, y, z) = (K * I)(x, y, z) = \sum_i \sum_j \sum_k I(x + i, y + j, z + k)K(i, j, k). \quad (2.11)$$

Fig. 2.5 shows a simple architecture, based on VGG [157], that can be used for classification. The input image is passed through a series of convolutional layers along

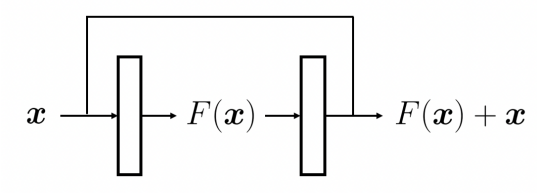


Figure 2.6: Skip connection building block. After a set number of layers F, the original input is added to the output.

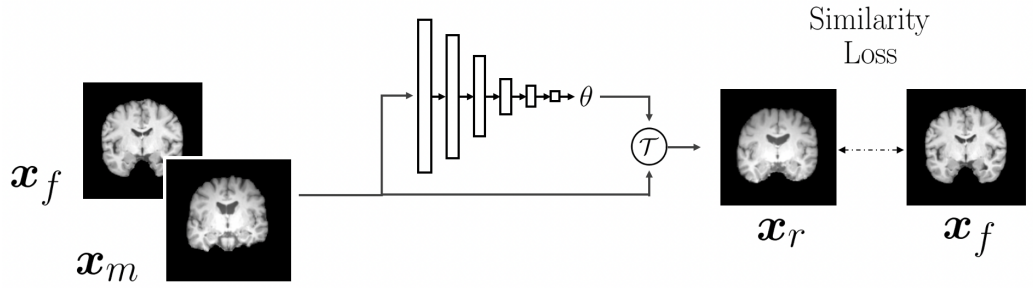


Figure 2.7: An example of a registration network. The moving and fixed image pair are passed to a neural network that predicts the transformation of interest. This is applied to the moving image to produce a registered image.

with activation functions and pooling layers. During this process, the network's field of view increases and it learns representation that is useful for classification. The final representation from the fully convolutional network is then flatten and passed through fully connected layers to produce the class prediction. Usually, we can achieve higher accuracy if we make the network deeper [131]. However, a larger network is often harder to optimize due to problems such as the vanishing gradient. Skip connections [72] are commonly used to alleviate this issue. With this simple modification, the input or representation is passed through a set of layers as usual, but we also add the original input to the output, as shown in Fig. 2.6.

2.4.2 Registration

Traditionally, pairwise iterative-based approaches have been extensively used in medical image registration [76, 134]. A similarity criteria such as mean-squared error (MSE), normalized cross correlation (NCC) or mutual information [8, 9, 74, 75, 75, 78, 125, 182] is used to measure the quality of alignment and optimize for the parameters of the transformation. However, this is expensive since it is repeated for each new image pairs.

On the other hand, popular unsupervised neural network approaches often takes the moving and fixed image pair as input, as shown in Fig. 2.7. The model then predicts the deformation of interest. This can range from parametric transformation to nonlinear deformation [13, 40, 50, 85, 99, 140, 188]. The transformation is then applied to the moving image and the quality of the registration is measured by a similarity function. The network can be trained end-to-end through this approach. During inference, this type of models is significantly faster than iterative approaches since it does not need to optimize each new moving and fixed image pair.

2.4.3 Segmentation

Many different CNN architectures have been proposed for segmentation [10, 111, 133, 201]. One of the most popular method used in medical imaging is the U-Net [146]. The U-Net has a contracting and expanding path. The former consist of a series of convolutional and down-sampling layers which helps increase the field of view of the network and capture large context within the input image. The expanding path then up-sample the resulting representation to enable localization

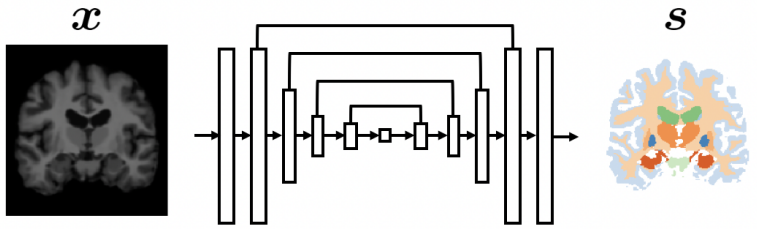


Figure 2.8: Illustration of U-Net [146] to segment an image. First half of the network is the contracting path, whereas the latter half is the expanding path. Skip connections connects both side of the network to help propagate contextual information.

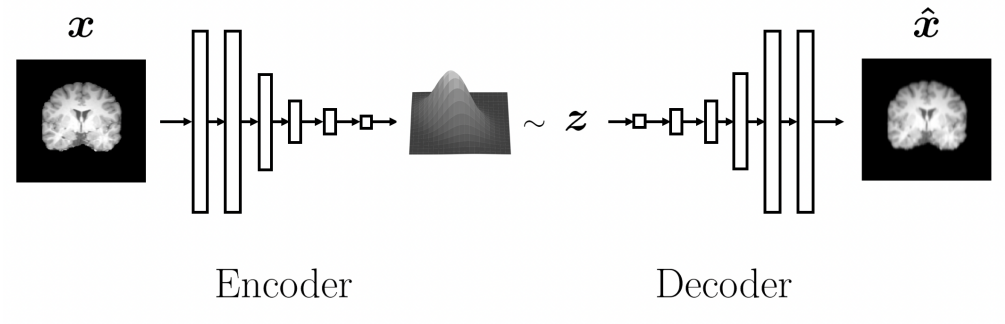


Figure 2.9: Illustration of the variational autoencoder (VAE). The encoder approximates the approximate posterior $q(\boldsymbol{s}|\boldsymbol{x})$ and the decoder $p(\boldsymbol{x}|\boldsymbol{s})$ approximate the image likelihood.

of the regions of interest. Skips connections between the contracting and expanding path is also used to help propagate contextual information. An illustration of this idea is shown in Fig. 2.8.

2.4.4 Generative Model

Neural networks are currently the state of the art method for generative modeling. The most prominent methods are generative adversarial network (GAN) [65] and variational autoencoder (VAE) [97]. In this section, we will focus on VAE.

Variational inference provides a powerful tool to approximate probability distri-

butions. Generally, our goal is to find a model that maximizes the log probability of the observed data $\boldsymbol{x}$. We often associate $\boldsymbol{x}$ with a latent representation $\boldsymbol{s}$

$$\log p(\boldsymbol{x}) = \log \int p(\boldsymbol{x}|\boldsymbol{s})p(\boldsymbol{s})d\boldsymbol{s}. \tag{2.12}$$

However, exact computation of Eq. 2.12 is often intractable due to the integration over the latent variables $\boldsymbol{s}$. Instead, we maximizes the evidence lower bound or ELBO, which can be computed at least approximately

$$\log p(\boldsymbol{x}) \geq -\mathrm{KL}(q(\boldsymbol{s}|\boldsymbol{x})||p(\boldsymbol{s})) + \mathop{\mathbb{E}}_{\boldsymbol{s}\sim q(\boldsymbol{s}|\boldsymbol{x})} \left[\log p(\boldsymbol{x}|\boldsymbol{s})\right]. \tag{2.13}$$

In VAE, the ELBO is optimize through the use of neural networks. In this setup, the approximate posterior $q(\boldsymbol{s}|\boldsymbol{x})$ is computed through a network called encoder and the image likelihood $p(\boldsymbol{x}|\boldsymbol{s})$ is computed through another network known as decoder. Fig. 2.9 shows a illustration of this model.

CHAPTER 3

ROBUST AFFINE IMAGE REGISTRATION

Registration is a fundamental task in medical imaging, and recent machine learning methods have become the state-of-the-art. However, these approaches are often not interpretable, lack robustness to large misalignments, and do not incorporate symmetries of the problem. In this work, we propose KeypointMorph, an unsupervised end-to-end learning-based image registration framework that relies on automatically detecting corresponding keypoints. Our core insight is straightforward: matching keypoints between images can be used to obtain the optimal transformation via a differentiable closed-form expression. We use this observation to drive the unsupervised learning of anatomically-consistent keypoints from images. This not only leads to substantially more robust registration but also yields better interpretability, since the keypoints reveal which parts of the image are driving the final alignment. Moreover, KeypointMorph can be designed to be equivariant under image translations and/or symmetric with respect to the input image ordering. We demonstrate the proposed framework in solving 3D affine registration of multimodal brain MRI scans. Remarkably, we show that this strategy leads to consistent keypoints, even across modalities. We demonstrate registration accuracy that surpasses current state-of-the-art methods, especially in the context of large displacements. Our code is available at https://github.com/evanmy/KeypointMorph

3.1 Introduction

Registration is a core problem in biomedical imaging applications. Multiple images, often encompassing a variety of contrasts, are commonly acquired [175]. Classical (i.e. non-learning-based) registration methods involve an iterative optimization of

a similarity metric over a space of transformations [134,163]. Recent deep learning-based strategies leverage large datasets of images to solve registration. Given a pair of images $(\boldsymbol{x}_f, \boldsymbol{x}_m)$, these strategies use neural network architectures that either output transformation parameters (e.g. affine or spline) [40,107] or a dense deformation field [13] which aligns them. Since the registration step is achieved via a single feed-forward pass, it is substantially faster than iterative methods.

However, prior learning-based methods often fail when the given image pair has a large misalignment. Existing systems typically require that image pairs are roughly aligned [13, 35, 40, 140], or at least in the same orientation [128]. In addition, today's learning-based registration methods lack interpretability, as they are essentially "black-box models" that output either transformation parameters or a deformation field and provide little insight into what drives the alignment. Finally, the machine learning models used in registration tasks generally do not exploit the symmetries and equivariances present in the problem. For example, if one of the images is translated by a fixed amount, this has a pre-determined effect on the optimal registration solution, and this property is not built into any of the architectures used today for image registration.

Contribution. We propose KeypointMorph, an unsupervised end-to-end deep-learning-based framework that aims to address all aforementioned issues. Our main insight is that *matched keypoints* can be used to derive the optimal transformation in closed-form. This keypoint-based formulation is robust against misalignments as the closed-form solution does not depend on the initial position of the keypoints. Additionally, the model is interpretable since the keypoints that drive the alignment can be visualized. Finally, we also show how to incorporate symmetries into the model design. For example, the architecture we describe

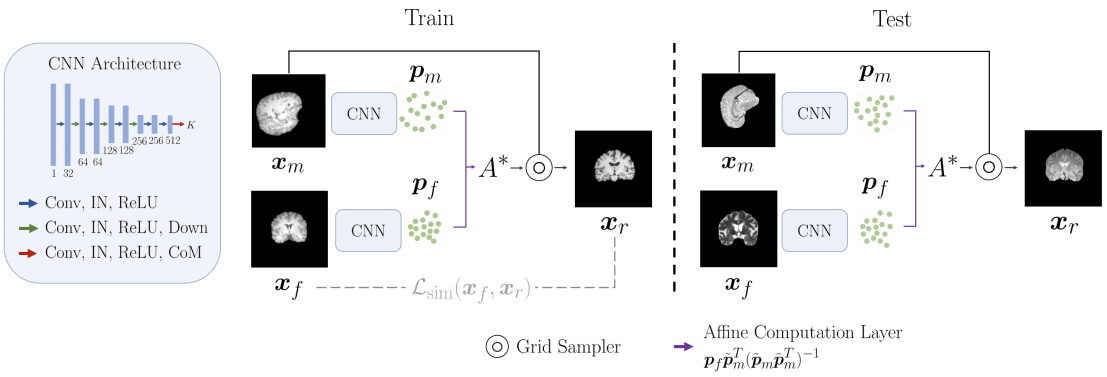


Figure 3.1: Proposed framework. Fixed and moving 3D images are passed through the same keypoint detection network composed of convolutional layers and a final center-of-mass (CoM) layer that predicts keypoints useful for registration. The affine matrix is then computed using Eq. 3.2 and is used to transform the moving image.

in this paper is translation equivariant and leverages a recently-proposed center-of-mass layer [116, 160]. Rather than treating matched keypoint detection as a supervised learning problem requiring human-annotated keypoints, we propose to use an end-to-end unsupervised strategy tailored toward registration. By unifying image registration and keypoint detection, we can train a model that finds matching keypoints useful for aligning images. We demonstrate this framework in the context of affine registration of 3D multi-modal brain MR scans.

3.2 Related Works

Classical Image Registration Methods. Pairwise iterative, optimization-based approaches have been extensively studied in medical image registration [76, 134]. These methods employ a variety of similarity functions, types of deformation, transformation constraints or regularization strategies, and optimization techniques. Intensity-based similarity criteria are most often used, such as mean-squared error (MSE) or normalized cross correlation for registering images

of the same modality [8, 9, 75]. For registering image pairs from different modalities, statistical measures like mutual information or contrast-invariant features like MIND are popular [74, 75, 78, 125, 182].

Another registration paradigm first detects features or keypoints in the images, and then establishes their correspondence. This approach often involves handcrafted features [174], features extracted from curvature of contours [148], image intensity [56, 71], color information [129, 178], or segmented regions [124, 185]. Features can be also obtained so that they are invariant to viewpoints [16, 21, 112, 169]. These algorithms then optimize similarity functions based on these features over the space of transformations [29, 76]. This strategy is sensitive to the quality of the keypoints and often suffer in the presence of substantial contrast and/or color variation [181].

Learning-based Methods. In learning-based image registration, supervision can be provided through ground-truth transformations, either synthesized or computed by classical methods [23, 46, 49, 107, 177, 194]. Unsupervised strategies use loss functions similar to those employed in classical methods [13, 35, 40, 50, 99, 140, 188]. Weakly supervised models employ (additional) landmarks or labels to guide training [13, 51, 81, 82].

Recent learning-based methods compute image features or keypoints [114] that can be used for image recognition, retrieval, or registration. Learning the keypoints can be done with supervision [181, 195], self-supervision [43] or without supervision [14, 108, 135]. In contrast, our focus is on robust image registration via corresponding keypoints; these prior works aim explicitly to obtain keypoints that are repeatable under different viewpoints and/or image acquisition conditions. Nevertheless, we build on previous ideas and introduce a framework that output

matched keypoints in 3D space regardless of the initial position of the image.

3.3 Proposed Method

Let $\boldsymbol{x}_m$ and $\boldsymbol{x}_f$ be moving (source) and fixed (target) volumes, which may vary in modality and orientation.[1] In this paper, we focus on 3D affine transformations, where the goal is to find the optimal affine transformation matrix $A^* \in \mathbb{R}^{3\times4}$ such that the moved (registered) image $\boldsymbol{x}_r = \boldsymbol{x}_m \circ A^*$ matches the fixed image $\boldsymbol{x}_f$, where $\circ$ denotes the spatial transformation of an image. To efficiently compute A^*, we employ *corresponding* keypoints from $\boldsymbol{x}_m$ and $\boldsymbol{x}_f$. The matching keypoints are computed using a convolutional neural network, and A^* is estimated using a non-learnable computation layer.

3.3.1 Keypoint Detector Network

To compute K matching keypoints, we use a single neural network g_ϕ with parameters ϕ to produce $\boldsymbol{p}_m = g_\phi(\boldsymbol{x}_m)$ and $\boldsymbol{p}_f = g_\phi(\boldsymbol{x}_f)$, where $\boldsymbol{p}_m$ and $\boldsymbol{p}_f$ are matrices of shape $d \times K$ (i.e. each column is a corresponding keypoint pair of d dimensions).

In our implementation, the backbone architecture of the keypoint detector consists of convolutional layers, followed by instance normalization [176], ReLU activation, and 2x downsampling via strided convolution, as shown in Fig. 3.1. The output from the backbone is followed by a center-of-mass (CoM) layer [116, 160], which computes the center-of-mass for each of the K activation maps. This spe-

[1] Although we consider 3D volumes in this work, we stress that our method is agnostic to the number of dimensions. The terms "image" and "volume" are used interchangeably.

cialized layer is (approximately) translation equivariant and enables precise localization. We provide more details and compare CoM to fully connected layers in Appendix A.2.

3.3.2 Affine Computation Layer

Given K corresponding keypoint pairs, we can derive a differentiable closed-form expression for an affine transformation that aligns the keypoints. Let $\tilde{\boldsymbol{p}}_m = [\boldsymbol{p}_m \ \mathbb{1}]^T \in \mathbb{R}^{(d+1)\times K}$, where $\mathbb{1} \in \mathbb{R}^{1\times K}$ is a vector of ones and $K > d$. We aim to find the optimal affine transformation

$$A^* = \arg\min_A \|A\tilde{\boldsymbol{p}}_m - \boldsymbol{p}_f\|_F \,, \tag{3.1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This leads to the closed-form solution

$$A^* = \mathcal{A}(\boldsymbol{p}_f, \boldsymbol{p}_m) = \boldsymbol{p}_f \tilde{\boldsymbol{p}}_m^T (\tilde{\boldsymbol{p}}_m \tilde{\boldsymbol{p}}_m^T)^{-1}. \tag{3.2}$$

We provide the derivation in the Appendix A.1.

3.3.3 Training

We found the following self-supervised pre-training strategy to be effective for initializing the keypoint detector backbone. We first pick a set of initial *keypoints* $\boldsymbol{p}$ chosen uniformly at random over the image coordinate grid. For a single subject, we assume that we have access to aligned multi-modal (e.g. T1, T2, PD-weighted MRI) volumes $\{\boldsymbol{x}_i\}$. During pre-training, in each mini-batch, we apply random affine transformations $\mathcal{T}$ (drawn from a uniform distribution over the parameter

space) to $\boldsymbol{x}_i$ and $\boldsymbol{p}$, and minimize:

$$\arg\min_{\phi} \sum_i \mathbb{E}_{\mathcal{T}} \left\| \mathcal{T}\boldsymbol{p} - g_\phi(\boldsymbol{x}_i \circ \mathcal{T}) \right\|_2^2, \tag{3.3}$$

where $\mathbb{E}$ denotes expectation and $\mathcal{T}\boldsymbol{p}$ transforms the list of coordinates in $\boldsymbol{p}$ by the affine transformation $\mathcal{T}$.

Following the pre-training strategy, we train KeypointMorph on random image pairs using the entire training dataset. We consider the typical real-world scenario of multi-parametric MRI, where scans of potentially different contrasts (e.g. T1, T2, PD-weighted) need to be registered. We present only within-modality image pairs to the network during training, where a simple loss like MSE may be used. Thus, given a dataset $\mathcal{D}$ composed of same-modality image pairs $(\boldsymbol{x}_f, \boldsymbol{x}_m)$, the overall objective is:

$$\arg\min_{\phi} \mathop{\mathbb{E}}_{(\boldsymbol{x}_f, \boldsymbol{x}_m) \sim \mathcal{D}} \left\| \boldsymbol{x}_m \circ \mathcal{A}\left(g_\phi(\boldsymbol{x}_f), g_\phi(\boldsymbol{x}_m)\right) - \boldsymbol{x}_f \right\|_2^2. \tag{3.4}$$

During training, we apply random affine transformations to the images as an augmentation strategy. Remarkably, we found that KeypointMorph learns to detect anatomically consistent keypoints *across modalities*, even though it is trained on same-modality pairs. In our experiments, we demonstrate how this can be used to perform multi-modal registration at test-time.

3.3.4 KeypointMorph Variants

We use `KeypointMorph_mse` to refer to the main unsupervised variant of our model, trained with Equation 3.4. In addition, we employ a *supervised* variant, `KeypointMorph_dice`, that exploits segmentations during training. During pre-training of this variant, we use the center-of-mass of segmentation labels as

ground truth keypoints and do not restrict to a single subject. During subsequent end-to-end training, we use soft-Dice, a loss function that measures volume overlap of the moving and registered label maps [13, 78, 81].

3.4 Experiments

3.4.1 Dataset

We used the IXI brain MRI dataset[2] for evaluation. Each subject has T1, T2, and PD-weighted 3D MRI scans in spatial alignment. We partitioned the 577 total subjects into sets of 427, 50, and 100 for training, validation, and testing, respectively. We performed standard skull stripping [98] on all images.

We used a pre-trained and validated SynthSeg model [18] to automatically delineate 23 regions of interest (ROIs)[3]. We used these segmentations for training a subset of models, as described below. Furthermore, all performance evaluations were based on examining the overlap of ROIs in the test images.

3.4.2 Test-time Performance Evaluation

We used each testing subject as a moving volume $\boldsymbol{x}_m$, paired with a different test volume treated as fixed image $\boldsymbol{x}_f$. For all test volumes, we use the 23-label segmentation maps to quantify alignment. We simulated different degrees of misalignment by transforming $\boldsymbol{x}_m$ using rotation, scaling, or translation. For a given

[2] `https://brain-development.org/ixi-dataset/`

[3] ROIs were pallidum, amygdala, caudate, cerebral cortex, hippocampus, thalamus, putamen, white matter, cerebellar cortex, ventricle, cerebral white matter, and brainstem.

transformation type, we choose 1-3 axes randomly and apply a uniform random transformation (e.g. rotation) up to a given amount (e.g. degree). We use the predicted transformation to resample the moved segmentation labels on the fixed image grid, and compute the Dice score to quantify alignment quality.

We performed registration across all combinations of available modalities (registering T1 to T1, T1 to T2, etc). All pairings and amount of transformations was kept the same across the registration experiments.

3.4.3 Baselines

Advanced Normalizing Tools (ANTs) is a widely used software package for medical image registration [9]. We use the "TRSAA" affine implementation, which consist of translation, rigid, similarity and two affine transformation steps. The volumes are registered successively at three different resolutions: 0.25x, 0.5x and finally at full resolution. At 0.25x and 0.5x resolution, Gaussian smoothing with σ of two and one voxels is applied, respectively. In addition, we use the ANTs affine initializer, which conducts a grid search over a range of rotations and translations to find a good initialization. We used mutual information as the similarity metric, which is suitable for registering images with different contrasts.

Deep Learning for Image Registration (DLIR) is a recent learning-based method that includes affine image registration [40]. DLIR implements a Spatial Transformer Network [85], which was originally used to improve class prediction accuracy and has since become the backbone of many subsequent works in image registration [13,41,107]. For a direct comparison, we used the same backbone archi-

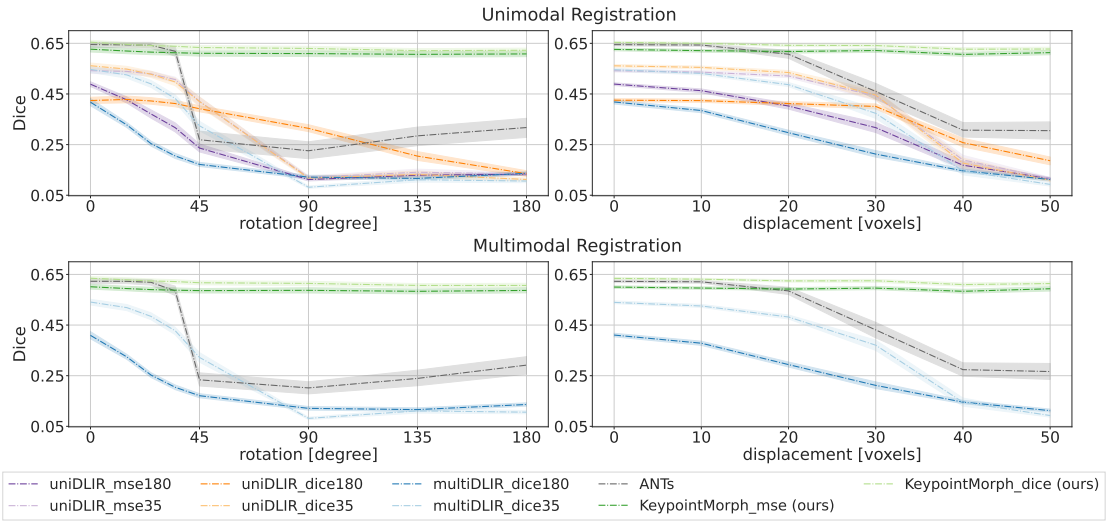
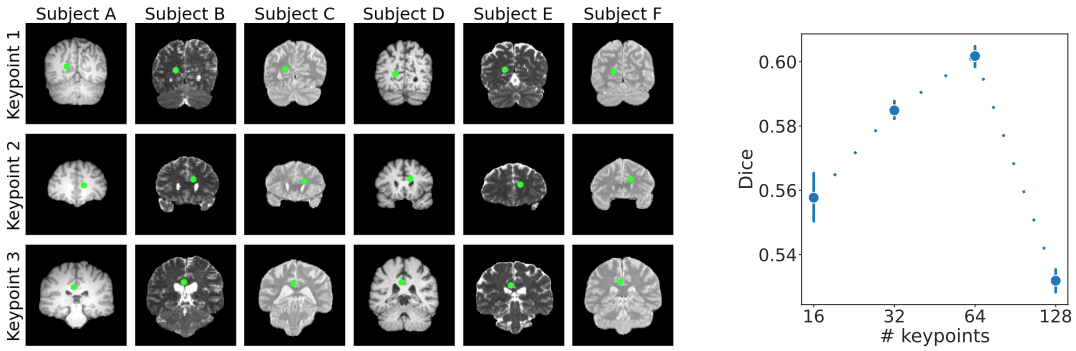


Figure 3.2: **Registration results.** The x-axis of the second column shows the average absolute displacement in the moving volume after rotation, scaling, or translation. The Dice score is averaged for all test subjects and brain anatomical regions. See Section 3.4.3 for details on the naming scheme.

tecture as KeypointMorph replacing the center-of-mass layer with a fully-connected (FC) layer which outputs 12 parameters for the 3D affine transformation.[4] In Appendix A.2, we also investigated a KeypointMorph implementation with the same FC layer as DLIR.

We find that DLIR often cannot register image pairs with large misalignments. We alleviate this by using more aggressive augmentation during training, where the images are randomly transformed. We consider two different amounts of rotation for the training of DLIR: maximum $\pm 35°$ or $\pm 180°$. We use the same loss function and training scheme as we used for KeypointMorph. We trained separate DLIR models for each modality as it produces better results than training DLIR across modalities with mutual information. We also trained *supervised* modality-specific and multi-modal DLIR models using a soft-Dice loss computed on the aligned segmentation maps [107].

[4]We used instance Norms for `multiDLIR` and Batch Norms for `uniDLIR`, which we found to work well in practice.



(a) (b)

Figure 3.3: (a) Sample keypoints learned without supervision by KeypointMorph. Each row shows a different keypoint and each column shows a different subject and/or modality. (b) Mean registration performance for rotation, scaling and translation under different number of keypoints.

Altogether, we implement six different DLIR variants. The naming scheme follows the convention `<mod>DLIR_<loss><degree>`, where `<mod>` denotes whether the model was trained on a single (`uni`) or multiple (`multi`) modalities, `<loss>` denotes the training loss function, and `<degree>` denotes the maximum angle of rotation for augmentation.

3.5 Results

Fig. 3.2 and Table A.2 summarize the results. We find that all DLIR models suffer substantially as the rotation angle increases. Training with aggressive augmentation increases performance for test pairs with large misalignments, but reduces the accuracy for those with smaller misalignments. Using supervision (`uniDLIR_dice35`) leads to improved accuracy. For unimodal registration, the DLIR model that was trained with all modalities (`multiDLIR_dice`) did not produce better accuracy than a model that was trained with each modality separately. ANTs yields excellent results when the initial misalignment is small (e.g., less than

45 degrees rotation). However, the accuracy drops substantially when the misalignment exceeds this range.

In contrast to these models, KeypointMorph variants performed well across all types of transformations and ranges, with only marginal drops in accuracy in large misalignments. In the case where we have access to ROIs during training, we find that KeypointMorph trained with Dice outperforms all models in nearly all the tasks. However, the unsupervised variant trained with MSE still yields excellent accuracy across all settings and is only minimally suboptimal compared to its supervised counterpart. We provide qualitative results in Appendix A.4, and compare the computational time across different models in Appendix A.3. Overall, the KeypointMorph variants substantially outperform other learning-based baselines, and KeypointMorph performs comparably or often substantially better (at large misalignments) than state-of-the-art ANTs registration, while requiring substantially less runtime.

3.5.1 Keypoint Analysis

Number of Keypoints. We trained five unsupervised KeypointMorphmodel variants with 16, 32, 64, and 128 keypoints, respectively. Fig. 3.3b illustrates that increasing the number of keypoints leads to improved Dice scores (and lower variability), up to 64 keypoints. However, we find that further increasing the keypoints can lead to a drop in performance (at 128 keypoints). We hypothesize that it is harder to find a relatively large number of anatomical landmarks that are consistent across individuals.

Keypoint Visualization. In contrast to existing models that compute the transformation parameters using a "black-box" neural network, we can investigate the keypoints that KeypointMorph learns to drive the alignment. Fig. 3.3a illustrates a set of representative keypoint examples for a learned unsupervised KeypointMorph model. We find that the final learned keypoints correspond to the same anatomical region in different subjects and modalities. In appendix A.6, we provide a quantitative study on keypoint consistency across the scans of each subject, and in Appendix A.7 we show an expanded version of Fig. 3.3a.

CHAPTER 4

WEAKLY SUPERVISED IMAGE SEGMENTATION

Deep neural networks are powerful tools for biomedical image segmentation. These models are often trained with heavy supervision, relying on pairs of images and corresponding voxel-level labels. However, obtaining segmentations of anatomical regions on a large number of cases can be prohibitively expensive. Thus there is a strong need for deep learning-based segmentation tools that do not require heavy supervision and can continuously adapt. In this paper, we propose a novel perspective of segmentation as a discrete representation learning problem, and present a variational autoencoder segmentation strategy that is flexible and adaptive. Our method, called Segmentation Auto-Encoder (SAE), leverages all available unlabeled scans and merely requires a segmentation prior, which can be *a single unpaired* segmentation image. In experiments, we apply SAE to brain MRI scans. Our results show that SAE can produce good quality segmentations, particularly when the prior is good. We demonstrate that a Markov Random Field prior can yield significantly better results than a spatially independent prior.

4.1 Introduction

Quantitative biomedical image analysis often builds on a segmentation of the anatomy into regions of interest (ROIs). Recently, deep learning techniques have been increasingly used in a range of segmentation applications [3, 93, 109, 146]. These methods often rely on a large number of *paired* scans and segmentations (voxel-level labels) to train a neural network. Training labels are either generated by human experts, which can be costly and/or hard to scale, or automatic soft-

ware [44], which can constrain performance. Furthermore, supervised techniques typically yield tools that are sensitive to changes in image characteristics, for instance, due to a modification of the imaging protocol [88]. This is a significant obstacle for the widespread clinical adoption of these technologies.

One approach to improve robustness and performance is to relax the dependency on paired training data and simply use unpaired examples of segmentations, sometimes called "atlases." Building on unpaired atlases, a segmentation model can then be trained continuously on new sets of unlabeled images [36, 38, 91]. For example, recently Dalca *et al.* [36] proposed an approach where an autoencoder is pre-trained on *thousands* of unpaired atlases. For a new set of unlabeled images, the encoder is then re-trained via an unsupervised strategy. Another widely-used approach to improve generalizability is data augmentation on labeled training data [27, 200]. For example, zhao2019data demonstrated an adaptive approach that learns an augmentation model on a dataset of unlabeled images. This model was then applied to augment a single paired atlas to perform one-shot segmentation within a supervised learning framework. Another popular approach is to use registration to propagate atlas labels to a test image [105, 151].

In this paper, we present a novel perspective for *minimally supervised* image segmentation. Instead of viewing segmentation from the lens of supervised learning or inverse inference, we regard it as a discrete representation learning problem, which we solve with a variational autoencoder (VAE) like strategy [97]. We call our framework Segmentation Auto-encoder, or SAE. As we demonstrate below, SAE is flexible and can leverage *all* available data, including unpaired atlases and unlabeled images. We show that we can train a good segmentation model using SAE with as little as a *single unpaired* atlas. In conventional representation learning,

e.g., VAE [97], an encoder maps an input to a continuous latent representation, which often lacks interpretability. In contrast, in SAE, the encoder computes a discrete representation that is a segmentation image, which is guided by an atlas prior. Finally, we employ the Gumbel-softmax relaxation [86] to train the SAE network. The Gumbel-softmax approximates the non-differentiable argmax (tresholding) operation with a softmax in order to make the function differentiable. It provides us with a simple and efficient way to perform the reparameterization trick for a categorical distribution, allowing the network to be trained via backpropagation. In our experiments, we demonstrate that SAE produces high quality segmentation maps, even with a single unpaired atlas. We also quantify the boost in performance as we exploit richer prior models. For example, a Markov Random Field model yields significantly better results than a spatially independent prior.

4.2 Method

We consider a dataset of N observed images (e.g. MRI scans) $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$, which we model as independent samples from the same distribution. Let $\boldsymbol{s}$ denote the (latent) segmentation, where each voxel is assigned a unique discrete anatomical label. Using Bayes' rule:

$$\log p(\boldsymbol{x}^{(i)}) = \log \sum_{\boldsymbol{s}} p(\boldsymbol{x}^{(i)}|\boldsymbol{s})p(\boldsymbol{s}), \tag{4.1}$$

where $p(\boldsymbol{s})$ denotes a prior distribution on the segmentation, $p(\boldsymbol{x}^{(i)}|\boldsymbol{s})$ is the posterior probability of the observed image conditioned on the latent segmentation, often called the image likelihood, and the sum is over all possible values of $\boldsymbol{s}$. We assume the prior $p(\boldsymbol{s})$ is provided and "learning" involves finding the parameters that describe the image likelihood $p(\boldsymbol{x}^{(i)}|\boldsymbol{s})$. Since Eq. 4.1 is computationally in-

tractable for most practical scenarios, we follow the classical variational strategy and maximize the evidence lower bound objective (ELBO):

$$\log p(\boldsymbol{x}^{(i)}) \geq -\mathrm{KL}(q(\boldsymbol{s}|\boldsymbol{x}^{(i)})||p(\boldsymbol{s})) + \mathop{\mathbb{E}}_{\boldsymbol{s}\sim q(\boldsymbol{s}|\boldsymbol{x}^{(i)})} \log p(\boldsymbol{x}^{(i)}|\boldsymbol{s}), \tag{4.2}$$

where $\mathrm{KL}(\cdot||\cdot)$ denotes the KL-divergence and $q(\boldsymbol{s}|\boldsymbol{x}^{(i)})$ is an efficient-to-manipulate distribution that approximates the true posterior $p(\boldsymbol{s}|\boldsymbol{x}^{(i)})$.

Following the VAE [97] framework, we use two neural networks to compute the approximate posterior $q(\cdot|\cdot)$ and the image likelihood $p(\cdot|\cdot)$. A so-called encoder network computes the approximate posterior $q_\phi(\boldsymbol{s}|\boldsymbol{x})$, where ϕ denotes the parameters of the encoder. The image likelihood $p_\theta(\boldsymbol{x}|\boldsymbol{s})$ is computed by the a decoder network, parameterized by θ. In our formulation, the encoder can be viewed as a segmentation network. The decoder corresponds to a generative or "reconstruction" model that describes the process of creating an observed image from an underlying segmentation.

A natural choice for the approximate posterior is a voxel-wise independent model:

$$q_\phi(\boldsymbol{s}|\boldsymbol{x}^{(i)}) = \prod_{j=1}^{V} \mathrm{Cat}(s_j|\boldsymbol{x}^{(i)}, \phi), \tag{4.3}$$

where $\mathrm{Cat}(s_j|\boldsymbol{x}^{(i)}, \phi)$ is a categorical distribution computed as the soft-max output of the encoder network at the j^{th} voxel evaluated for label s_j. Assuming an additive Gaussian noise likelihood model:

$$p_\theta(\boldsymbol{x}|\boldsymbol{s}) = \prod_{j=1}^{V} \mathcal{N}(\boldsymbol{x}; \hat{\boldsymbol{x}}_j(\boldsymbol{s};\theta), \sigma^2), \tag{4.4}$$

where $\hat{\boldsymbol{x}}(\boldsymbol{s};\theta)$ is a "reconstruction" image computed by the decoder network, subscript j is the voxel index, and $\mathcal{N}(\cdot;\mu,\sigma^2)$ denotes a Gaussian with mean μ and variance σ^2.

Putting together Eq. 4.2 and 4.4 and relying on Monte Carlo sampling to approximate the expectation, we obtain the following loss function to be minimized over θ and ϕ:

$$\mathcal{L} = \sum_{i=1}^{N} \mathrm{KL}(q_\phi(\boldsymbol{s}|\boldsymbol{x}^{(i)})||p(\boldsymbol{s})) + \frac{V}{2}\log\sigma^2 + \frac{1}{2\sigma^2 K}\sum_{k=1}^{K}||\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}(\boldsymbol{s}_{ik};\theta)||_2^2, \quad (4.5)$$

where $\boldsymbol{s}_{ik}$ is an independent sample segmentation image drawn from $q_\phi(\boldsymbol{s}|\boldsymbol{x}^{(i)})$. Following the convention in the field, in practice we set $K = 1$, which yields an unbiased yet noisy estimate of the loss and its gradient. Eq. 4.5 does not explicitly require paired images and segmentations $\{\boldsymbol{x}^{(i)}, \boldsymbol{s}^{(i)}\}$. Instead, it merely needs a prior $p(\boldsymbol{s})$. There are many ways to define a prior, but in our experiments we use a classical construction: a probabilistic atlas that describes the probability of labels at each location, which can be coupled with a Markov random field component that encourages certain topological arrangements.

4.2.1 Spatial Prior

The first prior we consider is a probabilistic atlas that assigns an independent label probability vector at each voxel, p_j. We call this a spatial prior:

$$p_{spatial}(\boldsymbol{s}) = \prod_{j=1}^{V} p_j(s_j). \quad (4.6)$$

There are many ways to construct this type of prior. For example, we can aggregate segmentations of different subjects and compute the frequency of anatomical labels at each voxel. If instead we only have a single segmentation image, we can apply a spatial blur to this segmentation in order to account for inter-subject variation. With the spatial prior, the first term in Eq. 4.5 reduces to:

$$\mathrm{KL}(q_\phi(\boldsymbol{s}|\boldsymbol{x}^{(i)})||p_{spatial}(\boldsymbol{s})) = \sum_{j=1}^{V} H(\mathrm{Cat}(s_j|\boldsymbol{x}^{(i)}), p_j(s_j)) - H(\mathrm{Cat}(s_j|\boldsymbol{x}^{(i)})) \quad (4.7)$$

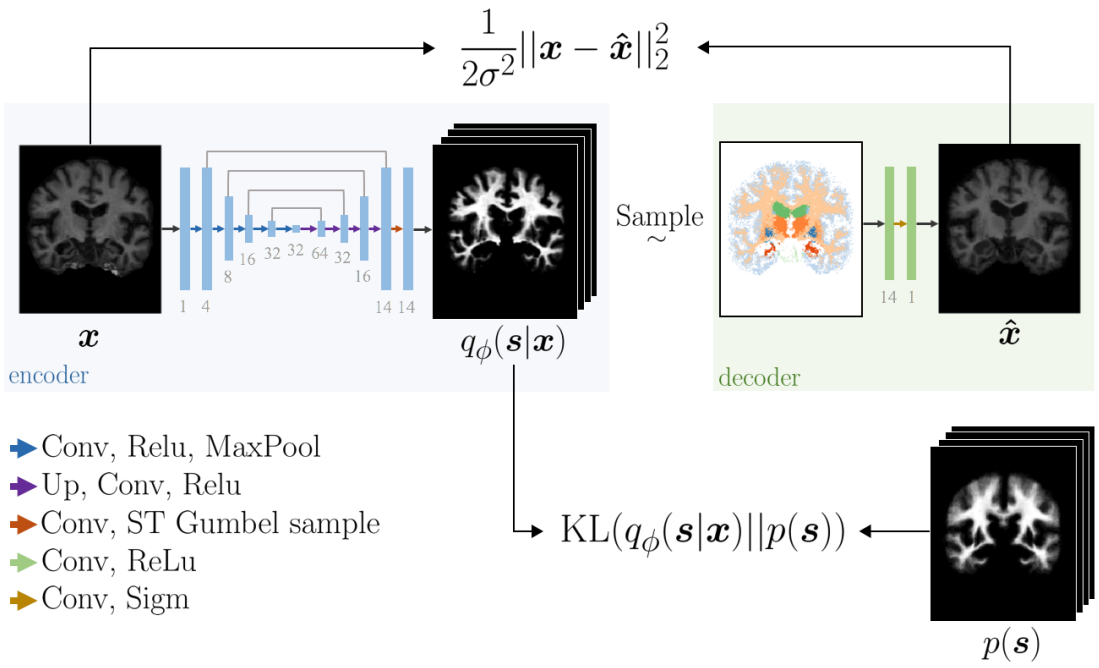


Figure 4.1: **Proposed architecture**. The encoder (blue) is a U-Net and decoder (green) is a simple CNN. (Conv) 3x3x3 convolution (Relu) rectified linear unit (Maxpool) 2x downsample (Up) 2x upsample (ST Gumbel) straight through Gumbel softmax (Sigm) sigmoid. The number of channels are displayed below each layer.

where the first term denotes cross-entropy and second term is marginal entropy.

4.2.2 Markov Random Field Prior

The spatial prior can be modified using a Markov Random Field (MRF) to capture neighborhood relationships in a segmentation image. Following [55, 199], we define the MRF prior as:

$$p_{MRF}(\mathbf{s}) = \frac{1}{Z} \exp\left[\sum_{j=0}^{V} V_j(s_j) + \sum_{j=0}^{V} \sum_{k \in N_j} V(s_k, s_j)\right] \tag{4.8}$$

where N_j is the $3 \times 3 \times 3$-neighborhood around voxel j, $V_j(\cdot)$ is the unitary potential at voxel j, $V(\cdot, \cdot)$ is the pairwise clique potential, and Z is a normalization constant.

Similar to [55], we define these potential functions based on a provided probabilistic atlas. Specifically, V_j is the voxelwise log frequency of each label: $\log p_j$; and $V(\cdot,\cdot)$ is the log normalized counts of label co-occurrences in neighboring voxels. E.g., $V(l_1, l_2)$ is computed as the logarithm of the count of neighboring voxel pairs with labels l_1 and l_2 divided by the count of voxels with label l_2. If the pairwise potential is set to zero, the MRF prior reduces to the spatial prior. With the MRF prior, the first term in Eq. 4.5 becomes:

$$\mathrm{KL}(q_\phi(\boldsymbol{s}|\boldsymbol{x}^{(i)})||p_{MRF}(\boldsymbol{s})) = \mathrm{KL}(q_\phi(\boldsymbol{s}|\boldsymbol{x}^{(i)})||p_{spatial}(\boldsymbol{s})) + \mathcal{L}_{MRF} + \text{const.}, \tag{4.9}$$

where the first term is from Eq. 4.7, and the second term can be expressed as:

$$\mathcal{L}_{MRF} = -\sum_{j=0}^{V}\left(\sum_{l_j=0}^{L-1} q_j(l_j|\mathbf{x}^{(i)}) \sum_{l_k=0}^{L-1}\sum_{k\in\mathcal{N}_j} q_y(l_k|\mathbf{x}^{(i)}) V(s_k = l_k, s_j = l_j)\right). \tag{4.10}$$

The MRF loss term quantifies the dissimilarity between the label topology of the prior and the approximate posterior $q(\cdot|\cdot)$. Appendix B provides more details about this constraint

4.2.3 Implementation Details

Our SAE architecture is shown in Fig. 4.1. The encoder is a 3D U-Net [146] and the decoder is a simple fully convolutional network. Training involves optimizing Eq. 4.5 with back-propagation. To implement the sampling layer, we employed the straight-through Gumbel-softmax relaxation scheme [86, 117], with the recommended setting for the temperature τ to 2/3. We estimated σ^2 by using the the global mean square error (MSE) between the reconstructed scan $\hat{\boldsymbol{x}}$ and the input scan $\boldsymbol{x}^{(i)}$. To initialize σ^2, we set the weight on the the reconstruction loss to be zero for the first 16 subjects (effectively setting σ^2 to infinity) so that the

segmentation (encoder) network was trained only based on the prior. In subsequent batches, σ^2 was updated as the average MSE over the latest 16 subjects and rounded to the nearest power of 10 in order to reduce fluctuation. Our complete model is trained end-to-end with the ADAM optimizer [96], with a learning rate of 10^{-4} and default parameter for its first and second moments. At test time, segmentation involves a computationally efficient single forward pass through the encoder and we output the `argmax` label at each voxel. Our code in PyThorch is available at https://github.com/evanmy/sae.

4.3 Experiments

4.3.1 Dataset

We evaluated SAE on T1-weighted 3D brain MRI scans, which we preprocessed with FreeSurfer, including skull stripping, bias-field correction, intensity normalization, affine registration to Talairach space, and resampling to 1 mm^3 isotropic resolution [54]. We focused on 12 brain regions (listed below) that were manually segmented and visually inspected for quality assurance. These manual segmentations were only used to quantify performance. The total number of subjects was 38: 30 subjects were used for training and 8 subjects for testing. Although we call our sets training and testing, we emphasize that SAE did not have access to the segmentation images during training, as we are proposing an unsupervised paradigm. We repeated the experiment 5 times with different random subject assignments to the train/test partitioning.

4.3.2 Variants of SAE

We employed two atlases. The first one (Atlas1) was based on a single unpaired segmentation image that we obtained from [122], which was automatically segmented using FreeSurfer [54]. We applied spatial blurring (Gaussian with 3 mm isotropic standard deviation) to the one-hot encoded segmentation image to obtain a probabilistic prior. As a second prior (Atlas2), we used a publicly available probabilistic atlas [138], which was computed based on 20 manually labeled subjects. Both priors and all input MRI scans were affine registered to Talairach space. For both of these priors, we implemented two versions: including and excluding the MRF loss of Eq. 4.10. Specifically, SAE1 (w/o MRF) uses the spatial prior derived by smoothing the single OASIS segmentation. SAE1 (w/ MRF) adds the MRF term of Eq. 4.10, where the pairwise potential function is computed based on the neighborhood statistics in the OASIS segmentation image. Finally, SAE2 uses the probabilistic atlas prior [138], instantiated with and without the MRF loss.

4.3.3 Benchmark Methods

As naive baselines, we used the most probable label at each voxel in the two priors. ***Baseline1*** corresponds to Atlas1 and ***Baseline2*** corresponds to Atlas2. As a strong baseline, we used an implementation of a widely-used atlas-based brain MRI segmentation tool [179], which uses Expectation-Maximization (EM) [42] to invert a probabilistic generative model. This EM baseline was run with the two atlases, which we refer to as ***EM1*** and ***EM2***. For each image, the EM baseline numerically solves an optimization problem and is thus relatively slow. Finally, all the data in the EM baseline has been pre-processed the exact way as we did for

our model.

As an effective upper bound on performance, we also implemented a supervised model, where a 3D U-Net [146] with the same settings as our encoder was trained with the paired manual segmentations in the training data. Negative generalized (soft) Dice [167] was used as the loss function and 6 of the 30 training subjects were reserved for validation. Training was terminated when validation loss stopped improving. As with our previous setup, we repeated this experiment 5 times with different train (N=24), validation (N=6), and test (N=8) splits[1].

4.3.4 Metrics

All presented results are computed on the test images of each round. For quantitative evaluation, we rely on two metrics: the Dice score that measures the volumetric overlap between the automatic segmentation and the ground truth manual segmentation; and the 95%-Hausdorff distance (HD) that quantifies the distance between the boundaries of the automatic and manual segmentations. When the two segmentation maps are exactly the same, Dice score will achieve its maximum value of 1 and HD will be equal to zero.

4.3.5 Experimental Results

Table 4.1 lists the global average Dice and HD values for the baselines and SAE variants. Regional and subject-level results are also presented in Fig. 4.3. We observe that in every single case and region, SAE produces segmentations that

[1]Test subjects are always the same for all methods

	Performance Measure	
Model	Haussdorff (mm)	Dice Overlap (%)
Baseline1	4.11±0.07	62.82±0.53
EM1 Baseline	4.25±0.09	71.24±0.71
Baseline2	3.50±0.06	71.45±0.65
SAE1 (w/o MRF)	3.88±0.05	74.64±0.30
SAE1 (w MRF)	3.81±0.05	75.36±0.32
EM2 Baseline	2.65±0.05	79.70±0.54
SAE2 (w/o MRF)	2.73±0.04	79.94±0.34
SAE2 (w MRF)	2.68±0.05	80.54±0.36
Supervised	2.23±0.07	84.60±0.26

Model	Test Time (s)
EM	61.07
SAE (CPU)	6.58
SAE (GPU)	1.58

Table 4.1: Mean performance of all methods with their standard errors and computational time per volume at testing.

are better than the naive baselines. SAE Dice scores, overall, were 8-12 points higher than the naive atlas based baselines and slightly better than the strong EM baselines. On a modern CPU, the EM baseline had a run-time of around 60 seconds, whereas SAE took less than 7 seconds per single volume at test time (less than 2 sec on a GPU). This represents more than a 10x speed-up over a popular brain MRI segmentation tool, with no discernible reduction in the quality of results.

For SAE, we observe that the adopted prior has a significant impact on the results. With a superior prior, SAE2 (derived from multiple subjects) yields substantially better results than SAE1. In addition, adding spatial consistency via the MRF loss improves the accuracy in all model variants (paired t-test $p < 10^{-6}$, for both atlases). This result highlights the importance of having a sophisticated prior. The best unsupervised model, SAE2 (w/ MRF), yielded a Dice score that was about 4 points below the fully supervised model, which is a strong upper bound in our experiment.

A qualitative visualization of SAE2 (w/ MRF) results is provided in Fig. 4.4. We can see that despite having a fixed prior $p(\mathbf{s})$, our model is able to capture

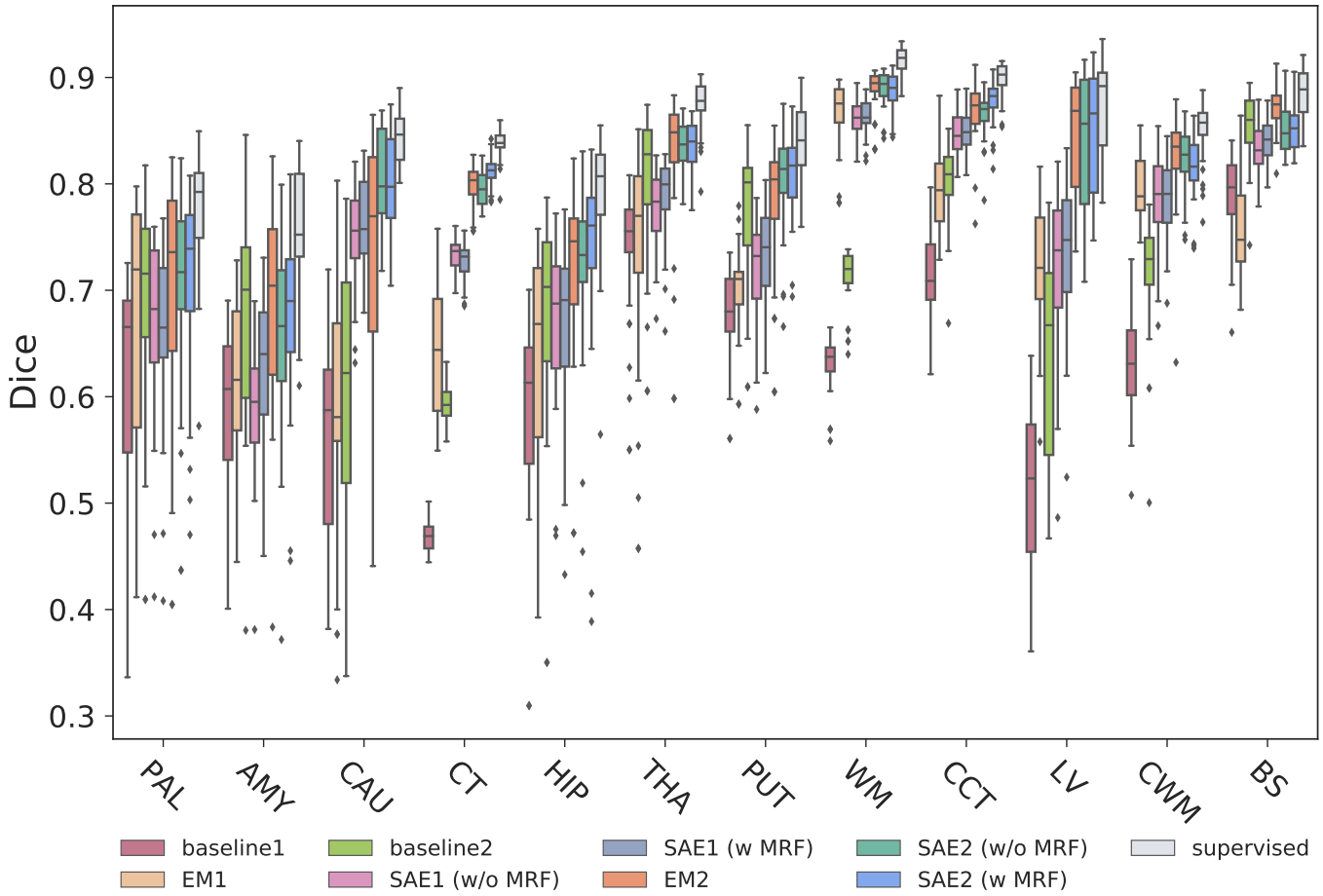


Figure 4.2: Boxplot of dice scores. Legend: (PAL) pallidum (AMY) amygdala (CAU) caudate (CT) cerebral cortex (HIP) hippocampus (THA) thalamus (PUT) putamen (WM) white matter (CCT) cerebellar cortex (LV) left ventricle (CMW) cerebral white matter (BS) brainstem.

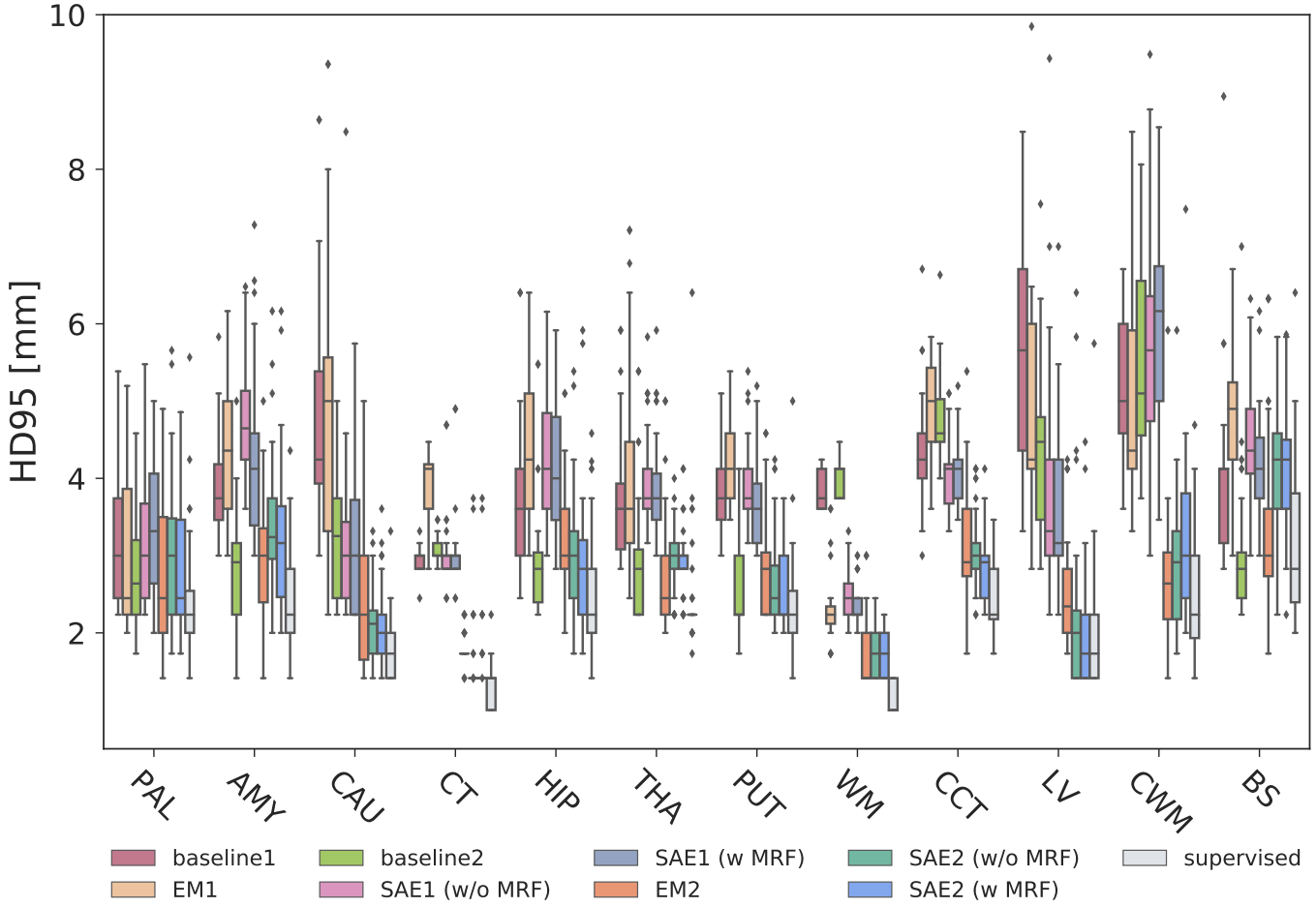


Figure 4.3: Boxplot of Hausdorff distance. Legend: (PAL) pallidum (AMY) amygdala (CAU) caudate (CT) cerebral cortex (HIP) hippocampus (THA) thalamus (PUT) putamen (WM) white matter (CCT) cerebellar cortex (LV) left ventricle (CMW) cerebral white matter (BS) brainstem.

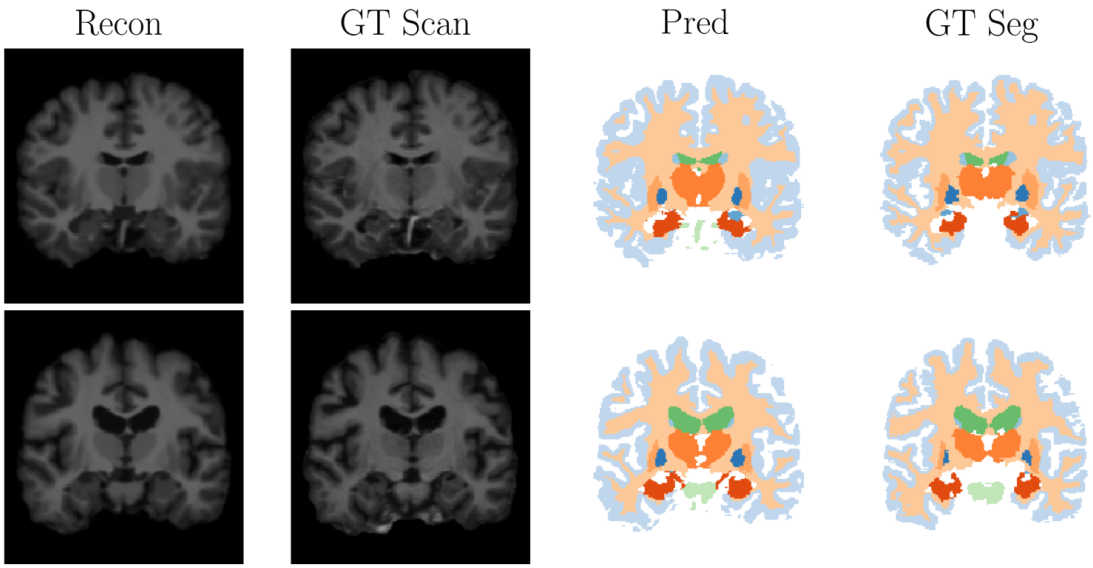


Figure 4.4: Representative segmentation results obtained with SAE2 (w/ MRF) on two subjects. Recon is the output of the decoder. GT scan and segmentation are the input MRI and manual segmentation, respectively. Pred is the segmentation obtained through `argmax` of the one-hot encoding $q_\phi(\boldsymbol{s}|\boldsymbol{x}^{(i)})$.

inter-subject neuroanatomical variation. This is mainly due to the decoder, which enforces the latent representation to be useful for reconstruction.

CHAPTER 5

CONDITIONAL DEFORMABLE TEMPLATES

Registration and segmentation often use brain templates. They describes the anatomical layout of an "average" brain is an essential building block of neuroimage analysis pipelines. However, a single template is often not sufficient to fully capture the variability in a heterogeneous population. Brain structures have very different shapes and sizes in different clinical and demographic groups. In this paper, we develop a novel neural network model that captures this morphometric variability. Our model learns to compute an attribute-specific spatial deformation that warps a brain template. We train this model on individual brain MRI segmentations in an end-to-end fashion, allowing for fast inference during testing. We demonstrate the ability of our model to deform a brain template given a wide range of ages, presence of disease and different sexes. Detailed qualitative and quantitative experiments are provided in order to demonstrate the flexibility of our model. Finally, we study the surface of the deformed template's hippocampus to show how our model can be used for shape analysis.

5.1 Introduction

Advances in neuroimaging have enabled examination of the brain at an unprecedented scale. The shapes and sizes of brain regions are an important area of neuroscientific study. Changes associated with factors such as aging, sexual dimorphism, and neurological diseases have been analyzed in many studies [57, 73, 136, 141, 142, 154, 183].

Today, conventional brain analysis pipelines rely on a probabilistic template

(or atlas), which assigns anatomical label probabilities at each voxel. Once constructed, imaging data from different individuals are spatially registered to the template for statistical analysis. However, demographic, clinical, or other confounding factors can influence the shapes and sizes of brain regions. As a result, a single and fixed template can struggle to accommodate complex structural differences across a heterogeneous group of individuals, which can complicate downstream statistical analyses. One approach to address this issue is to explicitly endow the template with more flexibility that might account for subject-specific characteristics [37, 150].

In this paper, we consider modeling attribute-specific neuroanatomical variability via a deformable template model, where the deformation is an explicit function of a given attribute vector. We present an end-to-end learning strategy to train the proposed neural network model and present empirical results that demonstrate utility and reveal interesting neuroanatomical shape variability associated with aging, sex, and Alzheimer's disease (AD).

5.2 Background and Related Work

There are many ways to construct a template or atlas that assigns labels on a voxel grid. The segmentation of a representative brain MRI from a dataset can be used as a naive reference by applying spatial blurring to account for inter-subject variability at the boundaries. However, today, most atlases are constructed from multi-subject data [89, 115]. A common approach involves co-registering the subjects and computing the frequency of anatomical labels at each voxel. It is widely recognized that a single atlas has difficulty accounting for the morphological

variability across a heterogeneous group of individuals [37, 145, 150]. Multiple atlases can be constructed for different subgroups. However, this would demand a significant amount of time, funds, and expertise.

Given a template, the morphological variability in the population is largely captured via deformations [134]. Deformation models can include global affine transformations or more flexible non-linear transformations. A popular parameterization employs B-slines [149]. Non-parametric deformation strategies can build on an elastic model [12] or diffusion model [79, 168]. A popular approach is to use diffeomorphic transformations, which ensures that the underlying topology is preserved through the use of continuous, invertible, and differentiable deformations [7, 8, 17, 90, 180].

Recently, deep learning based approaches to image registration have showed a lot of promise. Instead of solving an optimization problem for a pair of images, these methods use a neural network that learns to directly compute the transformation that aligns two input images [13, 41, 161]. By obviating the need of optimization during inference, these methods are significantly faster than non-learning-based approaches.

A recent paper closely related to ours, proposed a method to construct image templates in a learning-based framework using convolutional neural networks (CNN) [37]. The network synthesizes a conditional template for a given attribute value and produces a deformation field that aligns the conditional template with an input image. Our paper builds on this prior work but introduces a different approach. Unlike [37], we learn a function that computes a deformation field that warps a universal (unconditional) population template. This way, we are explicitly modeling morphological changes associated with the attributes as a diffeomorphic

deformation (i.e. shape and size variation). In contrast, previous works involving the estimation of multiple (conditional) templates, including [37] allowed these templates to have different appearances.

5.3 Proposed Method

We adopt a template $\boldsymbol{t}$ that assigns probabilistic labels at each voxel. This can describe an unconditional prior on an individual segmentation image $\boldsymbol{s}$:

$$p(\boldsymbol{s};\boldsymbol{t}) = \frac{1}{Z}\exp\Big(\text{SoftDice}(\boldsymbol{s},\boldsymbol{t})\Big), \tag{5.1}$$

where SoftDice is the soft Dice between the two segmentation maps [126] and Z is the partition function. We are interested in modifying the prior as a function of demographic and/or clinical variables that are collected in an attribute vector $\boldsymbol{a}$. Thus, our objective is to learn the conditional distribution of $\boldsymbol{s}$ given a set of attributes and a template $p(\boldsymbol{s}|\boldsymbol{a};\boldsymbol{t})$:

$$p(\boldsymbol{s}|\boldsymbol{a};\boldsymbol{t}) = \int p(\boldsymbol{s}|\boldsymbol{z};\boldsymbol{t})p(\boldsymbol{z}|\boldsymbol{a})d\boldsymbol{z}. \tag{5.2}$$

where $\boldsymbol{z}$ is a latent embedding vector that parameterizes the attribute-specific deformation of the template, i.e., $\phi_{\boldsymbol{z}}$. In our implementation, $\boldsymbol{z}$ is of size $5 \times 6 \times 7 \times 128$ and follows a Gaussian distribution

$$p(\boldsymbol{z}|\boldsymbol{a}) = \mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_{\boldsymbol{a}},\boldsymbol{\Sigma}_{\boldsymbol{a}}), \tag{5.3}$$

where $\boldsymbol{\mu}_{\boldsymbol{a}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{a}}$ are the mean and diagonal covariance matrix, respectively. $\phi_{\boldsymbol{z}}$ is computed via a series of up-sampling and convolution layers (details below).

For the likelihood term $p(\boldsymbol{s}|\boldsymbol{z};\boldsymbol{t})$, we assume that $\boldsymbol{s}$ is generated by warping the

template $\boldsymbol{t}$ with the deformation field $\phi_{\boldsymbol{z}}$:

$$p(\boldsymbol{s}|\boldsymbol{z};\boldsymbol{t}) \propto \exp\Big(\text{SoftDice}(\boldsymbol{s}, \boldsymbol{t} \circ \phi_{\boldsymbol{z}})\Big). \tag{5.4}$$

The conditional prior $p(\boldsymbol{z}|\boldsymbol{a})$ captures the dependency between the template and attributes. Eq. 5.2 is computationally intractable, and we rely on Monte Carlo samples to approximate the expectation:

$$p(\boldsymbol{s}|\boldsymbol{a};\boldsymbol{t}) = \mathop{\mathbb{E}}_{\boldsymbol{z}\sim p(\boldsymbol{z}|\boldsymbol{a})} p(\boldsymbol{s}|\boldsymbol{z};\boldsymbol{t}) \approx \frac{1}{K}\sum_{k=1}^{K} p(\boldsymbol{s}|\boldsymbol{z}_k;\boldsymbol{t}), \text{ where } \boldsymbol{z}_k \sim p(\boldsymbol{z}|\boldsymbol{a}). \tag{5.5}$$

We implemented parts of the model with neural networks. Details of our architecture can be found in Figure 5.1. We use a fully-connected layer to parameterize $p(\boldsymbol{z}|\boldsymbol{a})$. For a given set of attributes, the network outputs a mean $\boldsymbol{\mu}_{\boldsymbol{a}}$, and variance $\boldsymbol{\Sigma}_{\boldsymbol{a}}$. As is common in Monte Carlo based deep learning techniques, a single instance of $\boldsymbol{z}$ is sampled using the reparametrization trick [97]. The sample is reshaped to a small cube. Then, a series of convolutional layers with kernel 2x2x2 and stride of 2 are used to upsample the latent space $\boldsymbol{z}$. We repeat this process until we have a tensor with the same size as the template $\boldsymbol{t}$. This tensor is the "stationary" velocity field $\boldsymbol{u}$, which parameterizes a diffeomorphic deformation, as in [35]. Thus the final deformation can be computed via applying the following ordinary differential equation:

$$\frac{\partial \phi_{\boldsymbol{z}}}{\partial \tau} = \boldsymbol{u}\left(\phi_{\boldsymbol{z}}^{\tau}\right), \tag{5.6}$$

which describes a particle flowing according to a stationary velocity field. As we integrate over time, we start from an identity transformation at $\tau = 0$ to the final deformation $\phi_{\boldsymbol{z}}$ at $\tau = 1$. This integration can be approximated in a neural network with a scaling and squaring layer [6, 35].

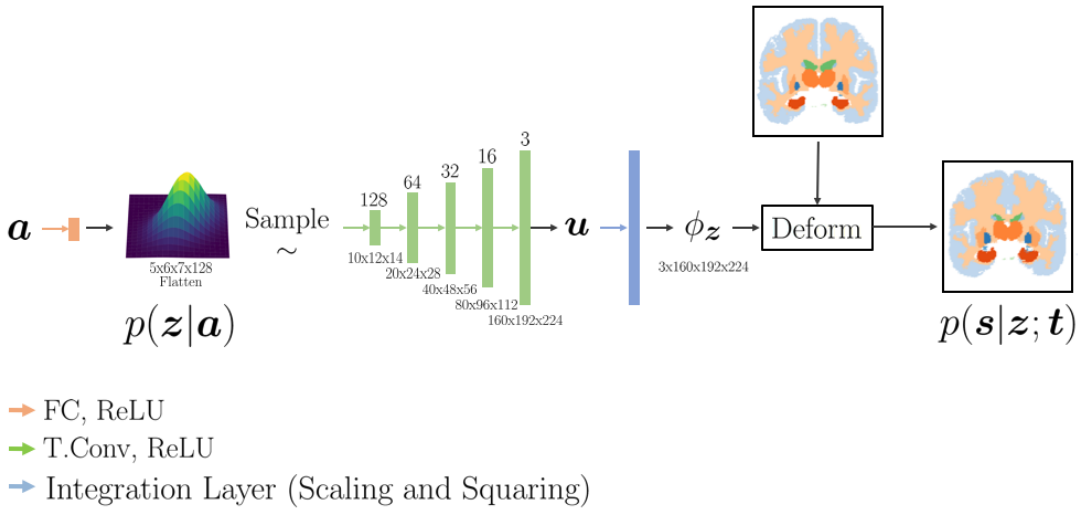


Figure 5.1: **Proposed architecture**. Attributes are passed to a fully-connected network (FC) and rectified linear unit (ReLU) to obtain the mean and variance of a multivariate Gaussian. Samples from $\boldsymbol{z}$ are up-sampled using 2x2x2 transpose convolution (T.Conv) with a stride of 2. The last layer before the velocity field $\boldsymbol{u}$ does not have a ReLU. Scaling and squaring are used to integrate $\boldsymbol{u}$ [35]. The deformed template is obtained by $\boldsymbol{t} \circ \phi_{\boldsymbol{z}}$

5.4 Experiments

5.4.1 Dataset

To demonstrate the ability of our model to deform a brain template given a wide range of attributes, we used 3D brain segmentations of T1-weighted MRI scans from the OASIS-1 dataset[1] [123]. Using FreeSurfer [54], we performed skull stripping, bias-field correction, intensity normalization, affine registration to Talairach space, resampling to 1mm^3 isotropic resolution, and segmentation into 12 regions of interest (listed in caption of Figure 5.3). Each segmentation was visually inspected for quality assurance. We used 415 subjects (255 females), with ages ranging from 18 to 96 years old (52.8 ± 25 years). The dataset includes 100 subjects diagnosed

[1] https://www.oasis-brains.org/

Models	Dice %
Template 1	62.25±3.07
Template 2	69.42±4.56
Model 1	71.64±3.50
Model 2	73.25±3.59

Table 5.1: Overall Dice score for different models. Mean ± standard deviation.

with probable Alzheimer's disease (AD). We randomly picked 375 subjects for training and the remaining 40 were reserved for testing. We ensured that the age distribution was consistent between the training the test samples.

5.4.2 Experimental Setup

We wanted to evaluate how well our model is able to deform a template given a set of attributes $\boldsymbol{a}$, namely: sex, age and diagnosis of AD. We experimented with two templates $\boldsymbol{t}$. For the first template (**Template 1**), we used the one-hot encoded segmentation from an independent healthy subject in the MCIC dataset [62], which was processed in the same manner as the OASIS data. The second template (**Template 2**) was derived from a publicly available probabilistic atlas [138], computed based on 20 manually labeled subjects.

Optimizing the logarithm of Eq. 5.5 (with $K = 1$ and ignoring the variation in the partition function) is equivalent to maximizing the soft Dice between the deformed template $\boldsymbol{t} \circ \phi_{\boldsymbol{z}}$ and the individual input subjects. We used the ADAM optimizer [96] to optimize soft Dice in the training data, with a learning rate of 10^{-4} and default moments parameters. Once trained, templates can be efficiently deformed with a single forward pass, allowing for fast inference of the conditional template. Our model is freely available https://github.com/evanmy/

conditional_deformation.

5.4.3 Evaluation

We emphasize that the input to our network is only a subject's attribute $\boldsymbol{a}$ and a probabilistic template. To evaluate our model, we provide the attribute of each test subject for our model to deform the probabilistic template $\boldsymbol{t}$. We repeated all experiments 5 times with different train/test subject splits. Only the templates remained the same during repeated experiments. We computed Dice scores between the deformed template and the subject's actual segmentation. In Table 5.1, we can see that our model is able individualize the template in order to account attribute specific characteristics. The performance of the model depends on the quality of the template. The model that learns to deform a single subject template (**Model 1**) does not perform as well as a model using the probabilistic template that was constructed from multiple subjects (**Model 2**). In Fig. 5.2a-c, we show the boxplots of grouped Dice scores according to different attribute configurations. Regardless of the grouping, the deformed templates achieved better alignment of regions than the corresponding templates.

5.4.4 Visualization of Attribute-specific Templates

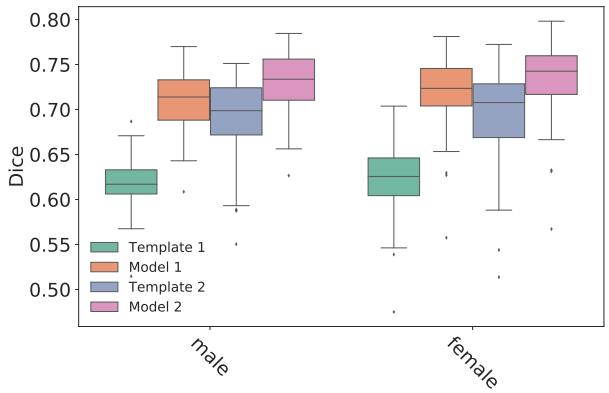
Our framework enables us to efficiently create a template for any configuration of attribute values, which we think is a unique way to interrogate associations between brain anatomy and demographic or clinical variables. In Fig. 5.3a, **Model 2** is used to deform the template for a range of attributes. For example, ventricular enlargement (green regions) is substantial in a healthy female, going from 20 to 70

years old. Similarly, AD-associated deformation of ventricles in a 70y old female is evident. Other brain regions' shapes and sizes vary with age and AD too. In Fig. 5.3b, we visualize the change in total normalized[2] grey matter (GM) or white matter (WM) volume as a function of age and AD status. Associated monotonic and non-linear atrophy patterns are apparent.

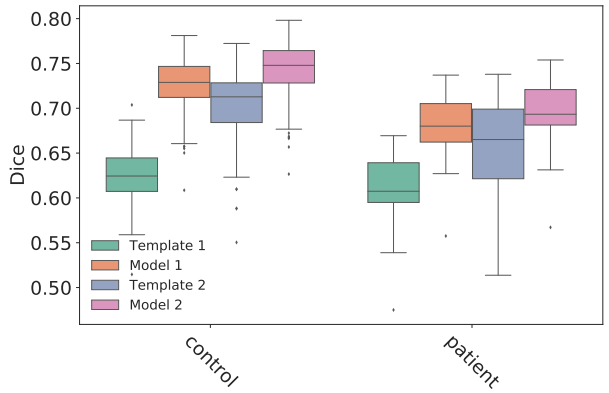
5.4.5 Shape Analysis

We can use the learned model to further investigate and visualize shape changes associated with specific variables. In this experiment, we wanted to capture the effect of aging and AD on the shape of the hippocampus, an anatomical structure that is strongly associated with both attributes [1]. In Fig. 5.4a, we visualize the difference in hippocampal shape between 18 and 90 years, for a healthy male. We used **Model 2** to create a whole-brain, volumetric template for these attribute values. We isolated the hippocampi and applied morphological opening to remove noise. The binary masks were then converted to a mesh and was smoothed with Laplacian smoothing. We then visualized the (reference) mesh for the 90-year old hippocampus. At each reference mesh vertex, we showed the signed distance to the closest point on the (target) 18-year old hippocampus mesh. If the closest target mesh point was outside of the reference mesh, the sign was negative; and otherwise positive. We employed the same visualization (Fig. 5.4b) to compare the hippocampal shape of a 65 year old Alzheimer's patient (reference) to a healthy one of the same age (target). These results support prior evidence that there is regional atrophy linked to aging and Alzheimer's, which probably differentially involve hippocampal sub-fields and thus lead to shape differences [1].

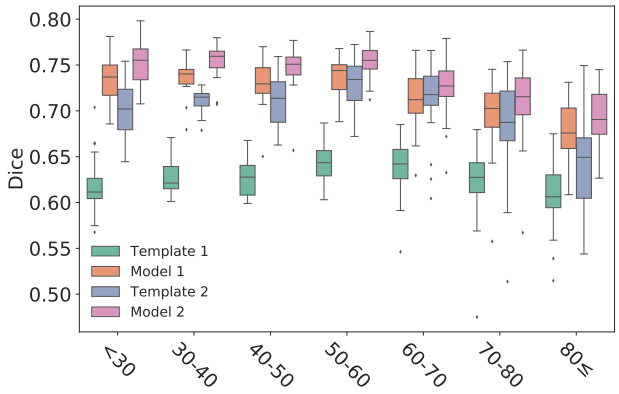
[2]Divided by the total WM/GM volume of a 20 year old male or female



(a) Sex

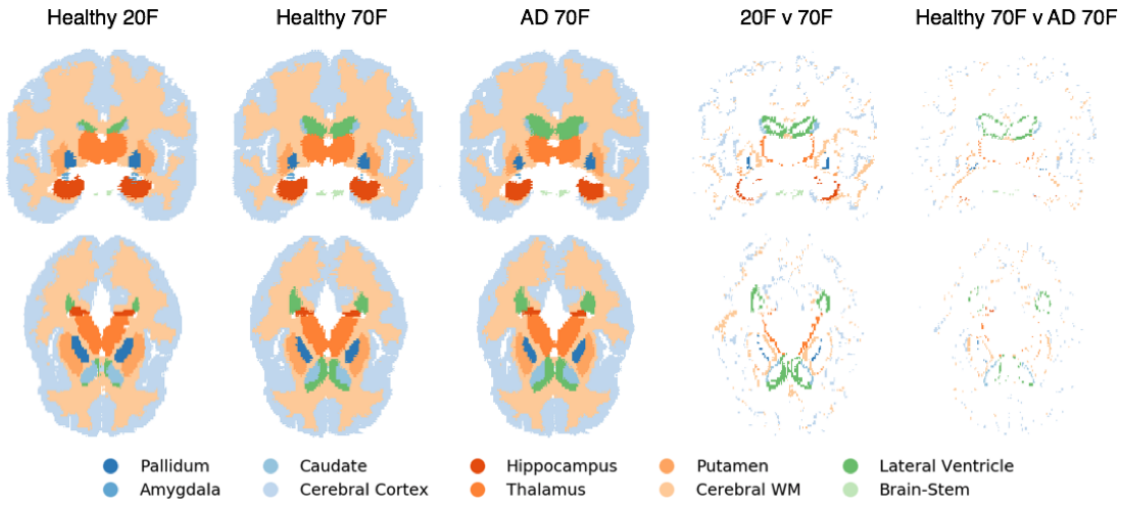


(b) Alzheimer

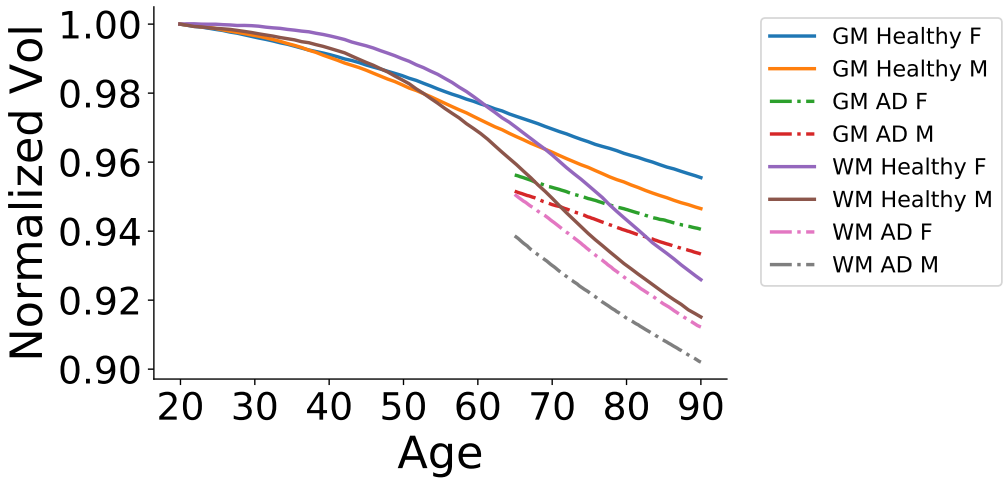


(c) Age

Figure 5.2: Dice scores between individual segmentations and templates. **Template 1** is the un-deformed template from a single subject. **Template 2** uses un-deformed template from a multi-subject probabilitistic atlas. **Model 1** or **2** learns to apply an attribute-specific deformation to Template 1 or 2, respectively. Similarly, (a-c) Box-plot of Dice score across different groups of attributes.

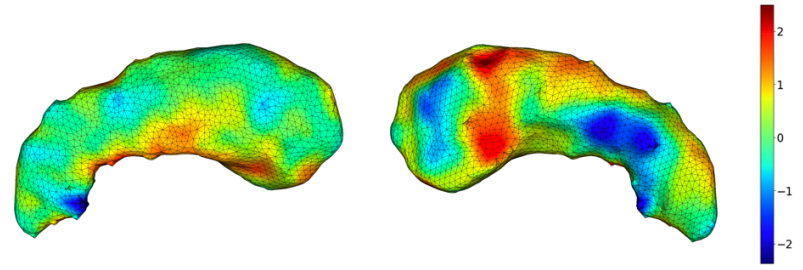


(a)

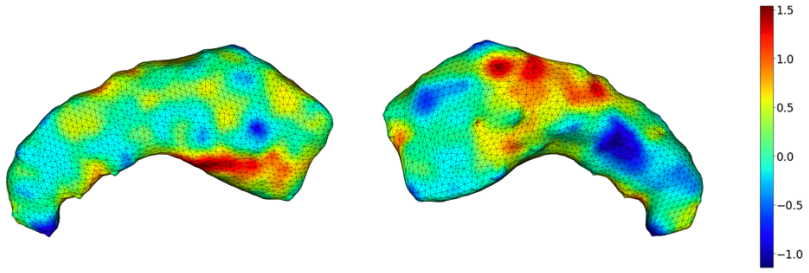


(b)

Figure 5.3: (a) Coronal and axial views of **Model 2** deformed templates for 20 and 70 year old healthy female, and 70 year old female AD patient. Difference maps between pairs of deformed templates are shown in last two columns. White pixels indicate that the compared templates have same label. Colored pixels show the label for the second template. (b) Changes of deformed template's grey matter (GM) and white matter (WM) volumes over age for different attributes. Each regional volume was normalized with respect to a 20 year old healthy male or female.



(a)



(b)

Figure 5.4: Signed distance visualization of the difference between hippcocampal surface meshes derived from **Model 2** deformed templates. Distance to the closest point on the target mesh are visualized on the reference mesh. Closest target mesh points that fall outside the reference mesh have negative value. (a) 90 yo healthy male (reference) vs 18 yo healthy male (target); (b) 65 yo male AD patient (reference) vs 65 yo healthy male (target).

CHAPTER 6

APPLICATION OF MRI IMAGING

We propose a novel machine learning strategy for studying neuroanatomical shape variation. Our model works with volumetric binary segmentation images, and requires no pre-processing such as the extraction of surface points or a mesh. The learned shape descriptor is invariant to affine transformations, including shifts, rotations and scaling. Thanks to the adopted autoencoder framework, inter-subject differences are automatically enhanced in the learned representation, while intra-subject variances are minimized. Our experimental results on a shape retrieval task showed that the proposed representation outperforms a state-of-the-art benchmark for brain structures extracted from MRI scans.

6.1 Introduction

Over the last two decades, neuroimaging has revolutionized our understanding of brain anatomy by allowing us to examine population variation at an unprecedented scale. One aspect of brain morphology that has received considerable attention is *shape*. Shape, in general, refers to the geometric properties of an object (e.g., a brain structure or a region of interest) that are independent of size or volume.

A broad range of techniques have been developed for the study of shapes of brain structures, which are under significant genetic influence [59] and can yield sensitive biomarkers of disease. A thorough review is provided by Ng *et al.* [132]. Following their convention, shape analysis techniques can be broadly grouped into five distinct types. One group covers techniques that work on point-based or local features [67], whereas another category includes methods that are surface based [144, 184]. A third category utilizes basis functions such as spherical harmonics [166] to represent the geometry, while a fourth group includes skeleton based schemes, such as medial profiles [69]. Finally, there is a category of methods that rely on characterizing deformations [48].

Today, most aforementioned techniques would be considered hand-crafted, as they heavily rely on arbitrary modeling choices in order to extract representations. Recently, deep neural networks have revitalized so-called end-to-end learning approaches that discover optimal representations in a data-driven fashion. In this work, we present an unsupervised approach to learn shape representations of different brain regions. In our framework, rotation, scaling and translation are normalized via a spatial transformer network [84] that aligns an input shape to a population template. Our shape template is not arbitrary, but also learned during training. Finally, the aligned structure is fed to an autoencoder that learns

to encode the input into a shape descriptor. The network is trained in an end-to-end fashion with binary segmentation volumes as input, and requires no other preprossessing. We report results for shape retrieval experiments in the OASIS dataset [122].

The rest of the paper is organized as follows. Section 2 discusses machine learning based approaches closely related to our approach. Section 3 introduces the proposed unsupervised learning strategy. Section 4 presents and discusses the empirical results. Finally, Section 5 concludes our paper.

6.2 Machine Learning based Shape Analysis

There has been a recent surge in the use of machine learning techniques to derive shape features. Some basic approaches include projecting the 3D objects onto different views before submitting these to a convolutional neural network (CNN) [11,190,202]. In this framework, the user needs to choose an arbitrary set of views, which can be sub-optimal for characterizing a 3D object.

An alternative approach is to represent objects as 3D point clouds, which are then processed using a discriminative neural network such as the PointNet [67,139]. However, this strategy yields representations that are optimal for a specific task, as in predicting Alzheimer's disease.

Another set of techniques rely on extracted surface meshes. For example, one can compute a heat kernel signature (HKS) of a mesh, which is then fed to an autoencoder [52, 191]. The HKS features are not scale invariant by design. In addition, techniques used in [52,191] are for object classification, but not instance

retrieval, which is our focus in this study. A different mesh-based approach was recently presented by Shakeri *et. al* [155], who used a spectral matching method to establish pointwise correspondence across samples and a hybrid auto-encoder and discriminator strategy to learn representations that are optimal for classifying subjects. More recently, there has been a growing effort in generalizing deep learning algorithms to non-Euclidean data such as those on mesh graphs or manifolds. Some of these algorithms have been applied to shape analysis [20, 130], which are collectively referred to as Geometric Deep Learning. The main drawback of mesh-based techniques is that the quality of the representation strongly depends on the quality of the surface mesh, which can suffer from topological errors.

Another approach related to our work is the 3D ShapeNet [187], which was originally developed to handle 2.5D depth data, and yield shape representations optimized for object class recognition. To our knowledge, ShapeNet has not been applied to the shape analysis of neuroanatomical structures yet. Furthermore, this strategy does not have isometry or affine invariance built into the model.

6.3 Proposed Method

Our method obtains a 3D shape descriptor without the need to extract a point cloud or mesh representation of the structure boundary. The learned shape descriptor is invariant to rotation, translation, and axis-independent scaling. Our proposed architecture consist of two components, namely a spatial transformer network (STN) [84] and a convolutional autoencoder (CAE), as illustrated in Figure 6.1. We achieve invariance against affine transformations (excluding shear) via the STN, which aligns the input 3D segmentation x_{in} to a structure-specific

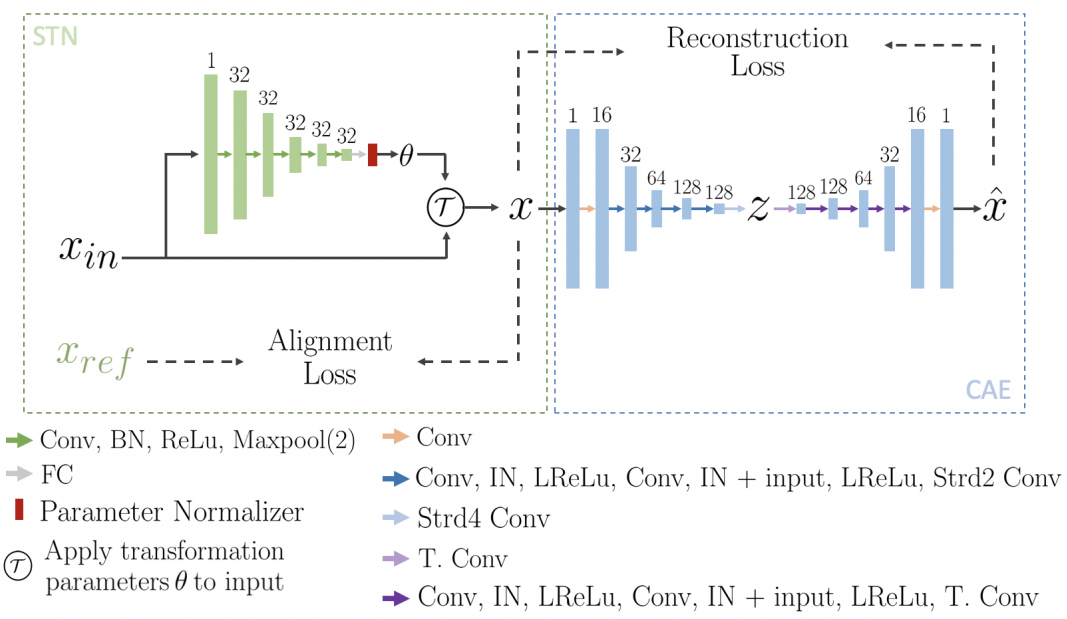


Figure 6.1: Proposed architecture. The network consists of a spatial transformer network (STN) and a convolutional autoendoer (CAE). The STN takes input x_{in}, a binary segmentation volume, and computes a set of affine transformation parameters θ, which are used to align to the learned reference template x_{ref} using the affine transformation $\mathcal{T}$. The template-aligned scan x is passed through a CAE in order to obtain a shape descriptor z from its bottleneck. The CAE has several residual blocks, where "+input" in the legend indicates a skip connection. Conv:for a 3×3 convolution, IN: Instance normalization, LReLU: Leaky Rectified Linear Unit, T.Conv: Transposed convolution, Strd2 Conv: convolution with stride 2. The number of channels is indicated above each layer.

reference template x_{ref}. Note that this template is not pre-set, but learned during training via minimizing the loss function described below. In the STN component, a convolutional neural network computes 9 transformation parameters (collectively denoted as θ) that make up an affine transformation matrix. These parameters are three rotation angles, three translations (axis aligned shifts) and three (axis specific) scales. Note that we did not include shearing in our transformation model as we considered this to affect the "shape." Next, the output of the STN is converted into a transformation matrix $\mathcal{T}$, which is then applied to the image grid, producing a deformed sampling grid. The sampling grid defines where on the input image to

sample in order to produce a template-aligned output x. Finally, the model passes x to an CAE, which goes through a bottleneck representation, namely the shape descriptor z, before decoding it into a reconstruction $\hat{x}$. The entire model (STN and CAE) is trained end-to-end, minimizing the following loss function:

$$\mathcal{L} = -\text{Dice}(x_{in}, x_{ref}) - \zeta(t)\text{Dice}(x, \hat{x}). \tag{6.1}$$

Dice is a commonly used metric that quantifies the similarity between two segmentation maps (e.g., binary volumes). In our implementation, we treated x_{in}, x_{ref}, and $\hat{x}$ as probabilistic segmentations, where each voxel took a value between 0 and 1, indicating the probability of the structure of interest. Hence, we defined the Dice metric as two times the sum of the voxel-wise product of the two input volumes divided by the sum of squared norm of the individual volumes. $\zeta(t)$ is an epoch dependent weighting function that follows a pre-determined schedule. At the beginning of training, we want the network to focus on alignment so $\zeta(t)$ was initialized with a small value. However, at later epochs, $\zeta(t)$ was gradually increased, so toward the end of training reconstruction quality was emphasized more in order to obtain a good shape descriptor. The first term in Eq. 6.1 is the alignment loss, whereas the second term is the reconstruction loss.

Although the user can pick an arbitrary template x_{ref} (such as some training sample), in our implementation we optimized it via minimizing the loss function. The learned template volume was passed through a sigmoid layer in order to ensure that the voxel values lie between zero and one. Similar to the widely used batch-norm layer, we introduced a parameter normalizer layer at the output of the STN that ensures that the (mini-batch) average value of each of the nine parameters (3 rotation angles, 3 translations and 3 log scales) is equal to zero. This way, the learned template does not experience drift in rotation, scaling and translation over the training epochs. Without a parameter normalizer, there's nothing in the

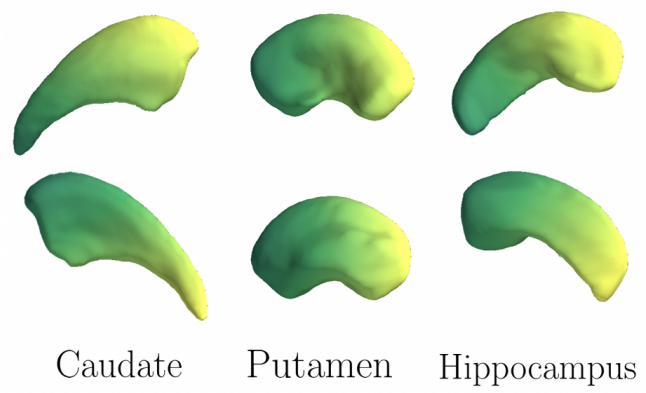


Figure 6.2: Lateral and medial views of the learned templates

learning dynamics that would prevent the learned template to continuously rotate, for example.

6.4 Experiments and Results

6.4.1 Dataset

To showcase and validate the proposed algorithm, we used healthy subjects from the OASIS-1 dataset [122] https://www.oasis-brains.org/, spanning the ages of 18 through 96. The total sample size was 315. We split the data into three non-overlapping groups containing 165, 50, and 100 subjects. These were used for training, validation, and testing, respectively. For each MRI scan, we extracted 6 brain regions: caudate, putamen and hippocampus in the two hemispheres. These structures were automatically segmented using FreeSurfer (v 5.1), which were visually inspected for quality assurance. The left hemisphere ROIs were mirrored and combined with the right hemishpere data.

There were 20 healthy subjects who obtained a second (repeat) scan on a subse-

Retrieval Experiment				
Setting	Type		ShapeDNA	Ours
Same Scan	Similarity	Top 1	**94.17**	91.58
		Top 5	97.17	**98.66**
	Affine	Top 1	56.75	**93.67**
		Top 5	70.92	**98.61**
Repeated Subject	No	Top 1	55.83	**99.17**
		Top 5	81.67	**100**
	Similarity	Top 1	40.00	**73.06**
		Top 5	70.83	**90.83**
	Affine	Top 1	30.83	**78.33**
		Top 5	60.00	**93.89**
Expanded Look-Up	No	Top 1	42.50	**99.17**
		Top 5	65.00	**100**
	Similarity	Top 1	28.33	**68.89**
		Top 5	55.83	**86.11**
	Affine	Top 1	20.00	**72.78**
		Top 5	40.83	**90.83**

Table 6.1: Retrieval accuracy under different settings

quent visit within 90 days of their initial session. These subjects were all included in our test dataset and the repeat scans were used for our retrieval experiment, as described below.

6.4.2 Implementation Details of Proposed Approach

We augmented our training data by randomly rotating up to 35° around all three axes. Similarly, we applied an axis-independent random scale up to ±50%. We trained a single model for the three different structure types we considered in our experiments: caudate, putamen, and hippocampus. However, we used a separate template for each structure, thus learning three templates. The templates were initialized using a single, average 43-year old training subject. The learned templates are shown in Figure 6.2.

We optimized our loss function using ADAM [96], with stochastic gradients computed on a mini batch size of 12 (4 examples of each structure). $\zeta(t)$ was set to 10^{-10} during the first epoch, and increased up to 10^{-3} linearly with each epoch. The implementation is in PyTorch and the code is freely available at https://github.com/evanmy/voxel_shape_analysis.

6.4.3 Benchmark Method

We compared our method to ShapeDNA [144], which uses a surface-based strategy to derive shape descriptors that are invariant to isometric transformations. ShapeDNA uses the Laplace-Beltrami spectrum of the surface mesh and has been successfully applied to the study of brain structures [184].

6.4.4 Retrieval Experiments

A good shape descriptor should not only be invariant to specific transformations, but also capture meaningful differences across subjects while remaining stable for a given subject. To this end, we performed two retrieval experiments.

In first experiment, we applied random transformations to test subjects (up to 15° rotation and 20% scale on each axis, respectively) to create query images. We considered two scenarios for the transformations: Similarity and Affine. In the Similarity case, we applied random rotations coupled with global scales. In the Affine case, each axis was randomly scaled independently, in addition to the random rotations.

The L2-norm was computed between the representation derived for the ran-

domly transformed query image and the representations from all original images of the test subjects. We then ranked these distances and report the top-1 and top-5 accuracy values in Table 6.1. Top-X refers to the fraction of query instances where the query subject ranked among the top X smallest L2-distances in shape space.

In the second experiment, we used the repeat scans from the 20 subjects in the test set. Similar to above, the shape descriptor derived from the repeat scans should lie close to the first scans of the corresponding subjects. We considered three transformation scenarios. No transformation, random similarity and random affine, where the random transformations were implemented as in the first experiment. We also considered two retrieval scenarios. In first case (Repeat Subjects Only), the look-up dataset consisted only of the 20 test subjects with repeat scans. In the second case (Expanded Look-Up), the look-up dataset included all 100 test subjects. We computed top-1 and top-5 accuracy values for these different scenarios, reported in Table 6.1.

In first experiment, ShapeDNA yielded high accuracy for similarity transformations, which is unsurprising since it is isometry invariant. However, ShapeDNA performed poorly with affine transformations, whereas the proposed method achieved high accuracy for both types of transformations. In the second experiment, each brain regions from the repeated MRI will not be exactly the same as the ones from the initial scan. Since our method works directly in voxel space, it is not as sensitive to the mesh surface that is required for the ShapeDNA benchmark. As a result, our algorithm vastly outperformed the benchmark, under all considered scenarios. We believe that this makes our method more stable across repeat scans of the same subject.

CHAPTER 7

CONCLUSION

MRI processing has many challenges. In this thesis, we delved into the use of deep learning models to prepare, process and analyse structural brain MRI scans. More specifically we focused on the registration and segmentation aspect of the pipeline and introduce methods related to its application.

In Chapter 3, we introduced KeypointMorph, a robust deep learning-based affine registration framework that employs unsupervised keypoint extraction. Our key insight is that matched keypoints yield a closed-form solution for affine registration, even in the case of large misalignments, and this in turn can be used to drive unsupervised keypoint detection. We showed that state-of-the-art optimization and learning-based methods for image registration struggle to register image pairs that have large misalignment. In contrast to many "black-box" machine learning-based registration methods, KeypointMorph also offers the ability to investigate what drives the registration by visualizing the keypoints. We envision that KeypointMorph can be used in a variety of applications that exhibit large misalignments, and can be extended to compute non-linear deformations.

In Chapter 4, we presented SAE, a flexible deep learning framework that can be used to train image segmentation models with minimal supervision. We applied SAE to segment brain MRI scans, relying on an unpaired atlas prior. Importantly, SAE does not need manual segmentations paired with the images, which opens up to possibility to deploy it on new imaging techniques, e.g., with high resolution or different contrast. Empirically, we presented the change in segmentation accuracy as we use different types of priors. Current implementation of SAE assumes that the input MRI is affine normalized with the prior by working in Talairach space.

However, SAE can be implemented with very different types of priors, which we would like to explore in the future. For example, in the present paper, we did not experiment with a spatial deformation model that would warp the atlas to better align with the input image. We envision that we can integrate a "spatial transformer" type neural networks, such as VoxelMorph [34], to relax our assumption. By adding a deformation model to the prior, we believe that we can handle complications like moving organs. Alternatively, we can implement more sophisticated priors, such as those that exploit an adversarial strategy, as in adversarial autoencoders [119].

In Chapter 5, we showed a novel framework to learn a deformable template, where the deformation is a function of attributes such as age, sex or diseases status. We believe our modeling approach has at least three different use cases. First, this model can be used as a subject-specific prior in a segmentation framework. Our results suggest that a subject-specific prior can yield improved quality segmentations than a model based on a single global prior. Second, our framework can be used to interrogate morphological changes associated with certain variables of interest. In our experiments, we demonstrated how we can visualize shape changes linked to aging and Alzheimer's disease. Finally, we believe that the proposed framework can be useful to normalize for confounding variables. The conventional approach in computational anatomy is to spatially register with a single template and then control for confounding variables such as aging by including them as regresors in subsequent statistical analyses. This approach cannot account for non-linear effects. Instead, one can use the proposed framework that would allow us to directly normalize for nuisance variation in shape and size. We will explore this direction in future research.

In Chapter 6, we demonstrated a data driven method to learn a 3D shape descriptor. Geometric transformations are normalized for through the spatial transformer by aligning the input to a learned template. Furthermore, a concise shape descriptor is obtained through an autoencoder. Our method outperforms an existing benchmark on retrieval experiments of subjects with longitudinal scans. Future work will include visualizing the learned shape descriptor and its application to examining associations between genetic/clinical variables and neuroanatomical shape.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR "KEYPOINTMORPH: ROBUST MULTI-MODAL AFFINE REGISTRATION VIA UNSUPERVISED KEYPOINT DETECTION"

A.1 Derivation of Closed-form Expression

Let $\boldsymbol{p}_f \in \mathbb{R}^{d\times K}$ and $\tilde{\boldsymbol{p}}_m = [\boldsymbol{p}_m \ \mathbb{1}]^T \in \mathbb{R}^{(d+1)\times K}$, where $\mathbb{1} \in \mathbb{R}^{1\times K}$ is a vector of ones and $K > d$. We wish to find the optimal affine transformation $A \in \mathbb{R}^{d\times(d+1)}$ which minimizes:

$$\mathcal{L} = \left\| A\tilde{\boldsymbol{p}}_m - \boldsymbol{p}_f \right\|_F ,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Taking the derivative with respect to A and setting the result to zero, we obtain:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial A} &= (A\tilde{\boldsymbol{p}}_m - \boldsymbol{p}_f)\tilde{\boldsymbol{p}}_m^T = \mathbf{0} \\ &\implies A\tilde{\boldsymbol{p}}_m\tilde{\boldsymbol{p}}_m^T = \boldsymbol{p}_f\tilde{\boldsymbol{p}}_m^T \\ &\implies A = \boldsymbol{p}_f\tilde{\boldsymbol{p}}_m^T(\tilde{\boldsymbol{p}}_m\tilde{\boldsymbol{p}}_m^T)^{-1}. \end{aligned}$$

A.2 Center-of-Mass Layer vs Fully Connected Layer

The backbone of our architecture is composed of CNN layers and a center-of-mass (CoM) layer at the end of the network [116, 160]. The CoM layer computes the center-of-mass of the activation map for each channel of the CNN output. In other words, weighted average between the voxel values and the grid coordinates. These center-of-masses are then used as keypoints in KeypointMorph. CoM layer

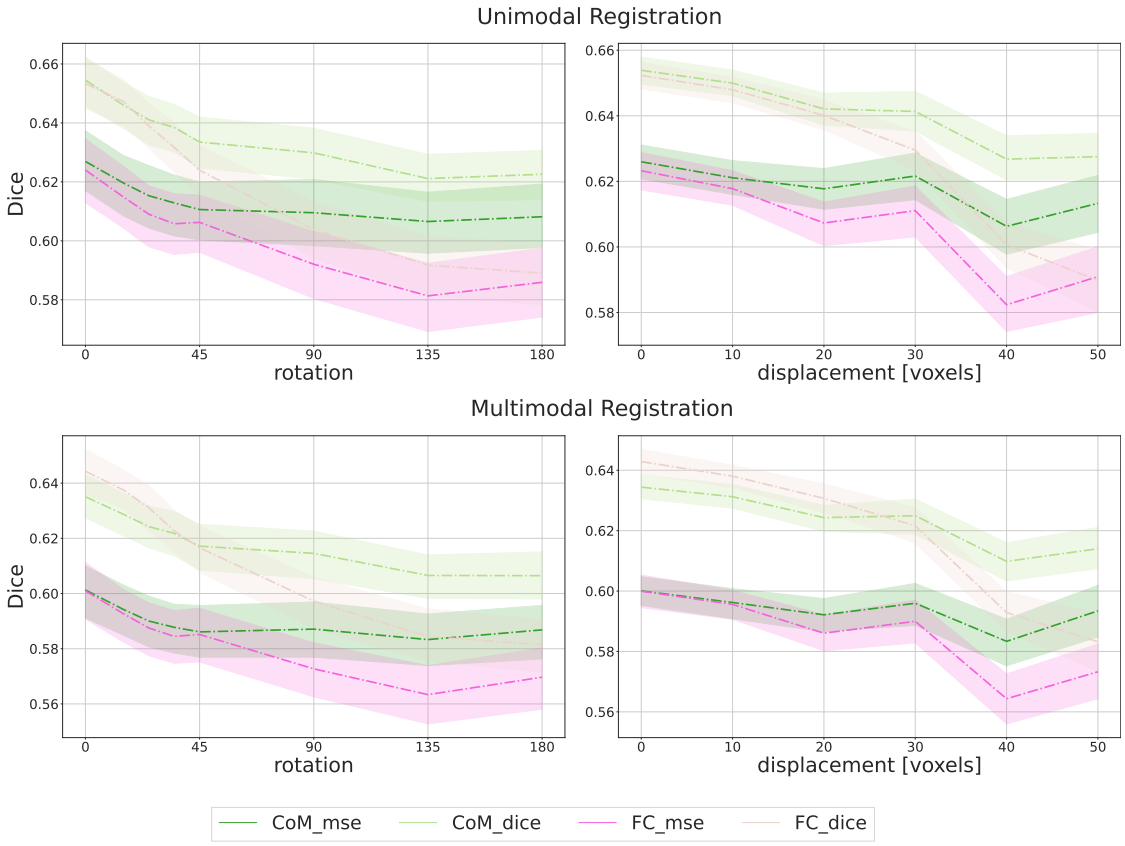


Figure A.1: Performance comparison bewtween KeypointMorph models that uses center-of-mass (CoM) and fully connected (FC) layers to predict keypoints. The suffix `mse` and `dice` represent the unsupervised and supervised version of KeypointMorph, respectively.

is shift equivariant as compared to fully connected (FC) layer. As an ablation study, we compared the performance of using a CoM layer and FC layer, which is commonly used in registration [40, 41, 85, 107]. In Fig. A.1, the model using FC layers follows the same architecture as DLIR model that was used in the baseline. However, instead of outputting 12 affine parameters, it outputs 64 keypoints. We repeated the experiments found in Section 3.4. We can see that regardless of the type of layer used to compute the keypoints, KeypointMorph provide robustness to large deformation. Models that uses CoM have comparable performance at low deformation compared to models that uses FC layers. However, CoM models provide the best performance at higher degrees of misalignment.

A.3 Computation Time

Table A.1 summarizes runtime at inference of all methods. On a modern CPU, the ANTs baseline, which does not have GPU support, requires more than 60 seconds, whereas KeypointMorph requires roughly 2.7 seconds per subject at test time (about 0.2 sec on GPU). Comparing ANTs and KeypointMorph (CPU) and KeypointMorph (GPU), this represent more than 20x and 300x speed-up, respectively.

A.4 Qualitative Results

We present some qualitative results of KeypointMorph trained without supervision. Moving and fixed subjects were picked randomly from the test set. We introduced random affine transformation with ±180 degrees of rotation, ±20% of scaling, and ±25 voxels of translation to the moving image. Fig. A.2 presents the registration result of the moved brains in the third column. We overlayed the aligned image (green) to the fixed/target brain (red).

Model	*CPU Time (s)*	*GPU Time (s)*
ANTs	66.95±1.56	-
DLIR	1.49±0.09	0.02±0.001
KeypointMorph	2.68±0.44	0.21±0.012

Table A.1: Average computation time across different models

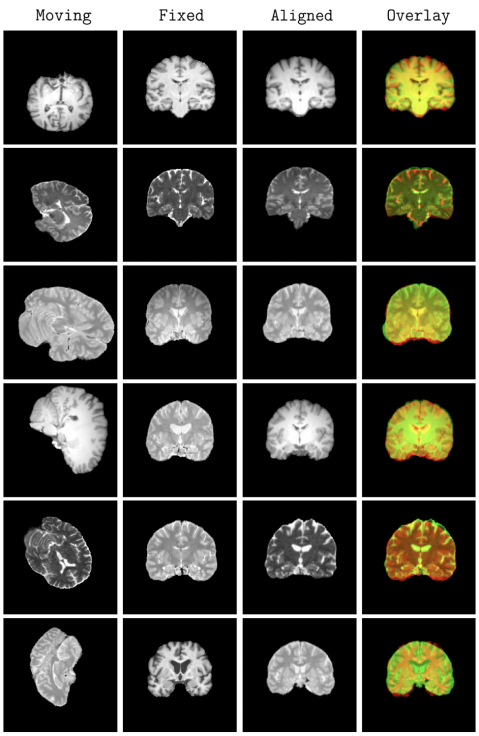


Figure A.2: Sample registration results obtained with KeypointMorph trained with no supervision (`KeypointMorph_mse`). Each row shows a different moving and fixed image pair. Red is the fixed image and green is resampled (moved) image.

A.5 Quantitative of Results

This section provides more details on the quantitative results for Fig.2 from the main paper. We randomly a picked a different fixed volume $\boldsymbol{x}_f$ on the test set for each moving subjects. We introduce random amount of misaligntment to $\boldsymbol{x}_m$. Each of the 3 axes of $\boldsymbol{x}_m$ had 50% chance of being deformed with a given transformation with at least one axis being perturbed. Since each subjects in the IXI dataset has a corresponding T1, T2 and PD weighted MRI scan, we performed registration across all combination of modalities (e.g. registering T1 to T1, T1 to T2 etc..). All pairings and random transformations were kept the same across the registration experiments.

T	Model	Dice Score								
		T1→T1	T2→T1	PD→T1	T1→T2	T2→T2	PD→T2	T1→PD	T2→PD	PD→PD
rotation	uniDLIR_mse180	26.8±14.84	-	-	-	26.86±15.01	-	-	-	29.36±17.01
	uniDLIR_mse35	51.33±7.95	-	-	-	53.88±7.2	-	-	-	53.57±6.35
	uniDLIR_dice180	38.64±14.65	-	-	-	28.59±12.62	-	-	-	35.24±13.98
	uniDLIR_dice35	52.57±7.81	-	-	-	53.41±9.33	-	-	-	54.14±8.03
	multiDLIR_dice180	21.68±11.06	21.78±11.42	22.0±11.46	21.32±10.6	21.88±11.48	21.99±11.37	21.22±10.67	21.78±11.5	22.09±11.62
	multiDLIR_dice35	49.84±8.48	48.98±8.16	49.37±8.41	49.27±8.33	49.42±8.21	49.41±8.36	49.5±8.3	49.15±8.11	50.28±8.6
	ANTs	48.07±26.14	44.32±25.34	42.27±25.0	43.48±25.23	45.06±25.59	42.69±25.04	40.56±25.01	43.51±24.89	44.66±25.84
	KeypointMorph_mse	62.24±6.56	58.75±5.98	58.8±6.59	58.62±6.07	60.79±6.28	59.38±5.88	58.83±6.01	59.37±5.68	61.79±6.21
	KeypointMorph_dice	**65.17 ± 5.11**	**61.94 ± 5.04**	**62.1 ± 4.82**	**62.54 ± 4.96**	**62.96 ± 5.3**	**61.61 ± 4.83**	**62.15 ± 4.87**	**61.35 ± 5.11**	**63.24 ± 5.26**
scaling	uniDLIR_mse180	45.97±8.07	-	-	-	45.96±8.46	-	-	-	53.47±7.32
	uniDLIR_mse35	53.32±7.26	-	-	-	54.46±7.37	-	-	-	53.37±5.98
	uniDLIR_dice180	44.23±9.13	-	-	-	39.09±7.48	-	-	-	42.91±7.74
	uniDLIR_dice35	55.19±6.95	-	-	-	54.92±7.96	-	-	-	57.57±6.8
	multiDLIR_dice180	39.12±9.49	41.08±7.93	41.32±7.81	38.12±9.45	41.74±8.43	41.7±8.1	37.96±9.93	41.48±8.85	42.02±8.81
	multiDLIR_dice35	53.43±6.35	52.68±6.33	52.86±6.54	52.57±6.56	52.84±6.98	52.49±6.71	52.69±6.37	52.38±6.76	53.32±6.74
	ANTs	66.29±6.19	**63.4 ± 5.7**	62.51±5.93	62.7±5.81	63.84±6.33	62.42±5.74	61.49±5.94	62.28±5.7	64.29±6.48
	KeypointMorph_mse	63.0±6.79	59.23±6.02	59.7±6.65	59.35±6.13	61.73±6.49	60.27±6.01	59.66±6.28	60.16±5.62	62.8±6.64
	KeypointMorph_dice	**66.51 ± 5.24**	63.25±4.86	**63.35 ± 4.6**	**63.33 ± 4.98**	**64.16 ± 5.35**	**62.7 ± 4.55**	**63.37 ± 5.0**	**62.77 ± 4.81**	**64.85 ± 5.42**
translation	uniDLIR_mse180	45.88±8.12	-	-	-	45.57±8.25	-	-	-	53.14±7.73
	uniDLIR_mse35	52.8±7.84	-	-	-	54.43±7.56	-	-	-	53.32±6.33
	uniDLIR_dice180	45.05±8.84	-	-	-	38.59±7.28	-	-	-	43.09±8.07
	uniDLIR_dice35	55.26±6.97	-	-	-	54.77±7.97	-	-	-	57.65±6.69
	multiDLIR_dice180	37.52±9.69	38.78±8.52	39.25±8.35	36.32±9.57	39.1±9.18	39.3±8.79	36.18±10.03	38.92±9.5	39.7±9.38
	multiDLIR_dice35	53.41±6.17	52.38±6.07	52.67±6.5	52.84±6.27	52.78±6.58	52.6±6.68	52.88±6.49	52.33±6.62	53.33±6.93
	ANTs	66.34±6.37	63.45±5.7	62.49±5.9	62.79±5.81	63.94±6.46	62.44±5.73	61.5±5.97	62.34±5.72	64.3±6.62
	KeypointMorph_mse	63.5±7.09	59.5±5.92	59.84±6.64	59.86±6.02	62.33±6.63	60.58±5.9	60.19±6.19	60.64±5.5	63.28±6.78
	KeypointMorph_dice	**66.99 ± 5.41**	**63.55 ± 4.82**	**63.75 ± 4.64**	**63.94 ± 4.71**	**64.79 ± 5.62**	**63.34 ± 4.55**	**63.83 ± 4.74**	**63.16 ± 4.86**	**65.42 ± 5.66**

Table A.2: Mean performance of all method with their standard deviation. The average Dice score is computed across test subject pairs, brain regions, and modalities. The notation $A \rightarrow B$ refers to registering moving volumes of modality A to fixed volumes of modality B. Bold numbers highlight the highest Dice score of a task given a transformation shown in the first column T.

A.6 Keypoint Consistency

Keypoint consistency plays an important role for multimodal registration. In this section, we show that KeypointMorph learns consistent keypoints across different modalities in an unsupervised manner. In Fig. A.3, we plotted the mean absolute error between the keypoints of each test subject across different modalities, as a function of training iteration. We can observe that as the model trains, the keypoints become more consistent across modalities. In other words, the keypoints lie in almost the same location across modalities for a given subject, allowing KeypointMorph to perform accurate multimodal registration.

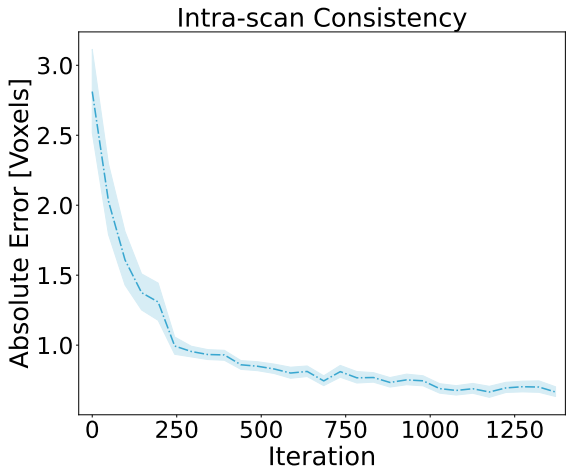


Figure A.3: Mean absolute error between keypoint locations of test subject across different modalities. Each training iteration represent a model update after 32 training subject.

A.7 Keypoint Visualization

In this section, we provide an extended visualization of keypoints. In the figures below, we investigated which regions KeypointMorph uses to register the volume pairs. We picked 12 random subjects and showed 12 of the 64 learned keypoints in each row. These keypoints were learned without supervision of any labeled regions. We can observe that the final keypoints lie within similar region of the brain across different subjects and scans.

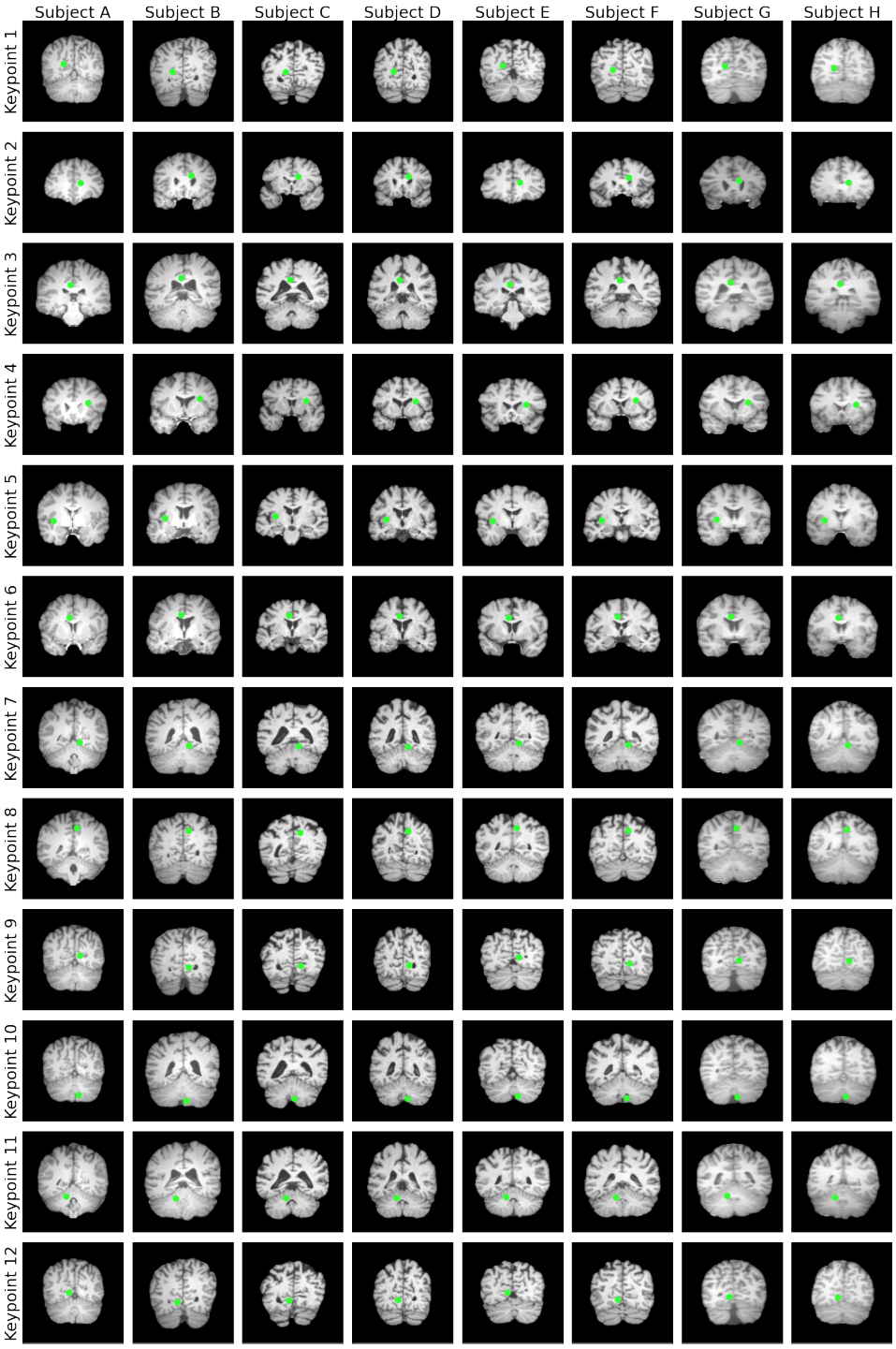


Figure A.4: Keypoints for different T1 scans

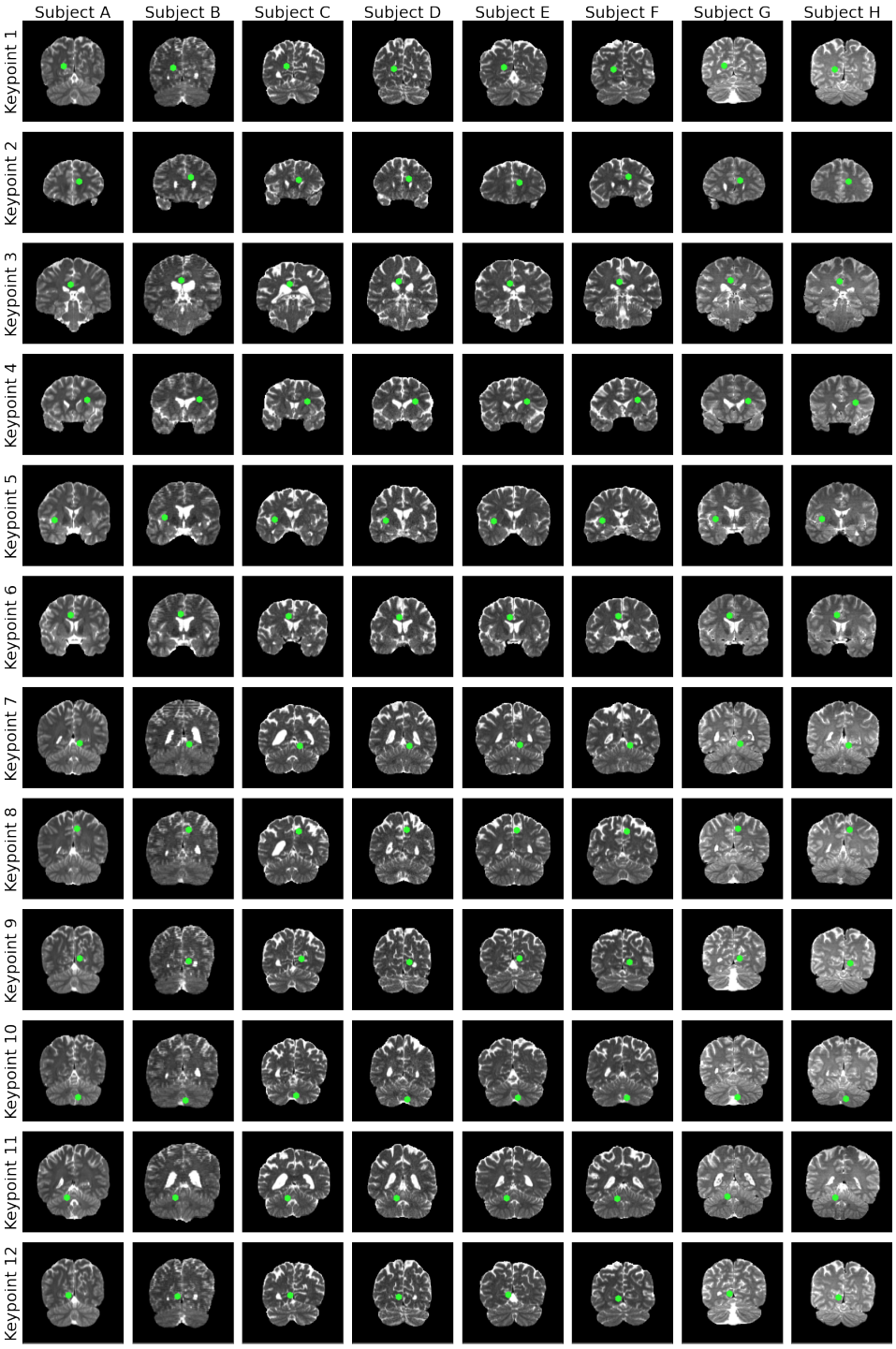


Figure A.5: Keypoints for different T2 scans

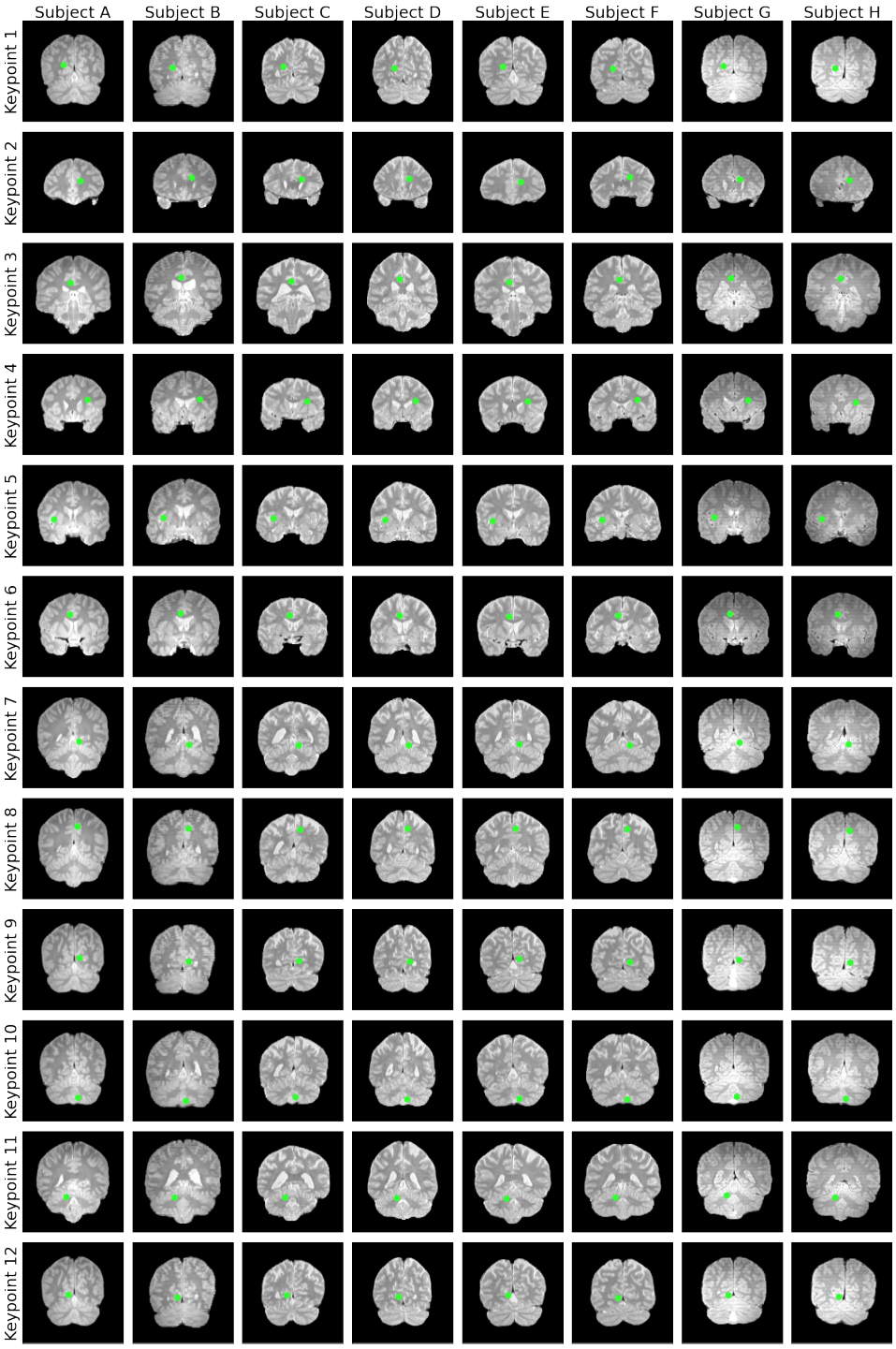


Figure A.6: Keypoints for different PD scans

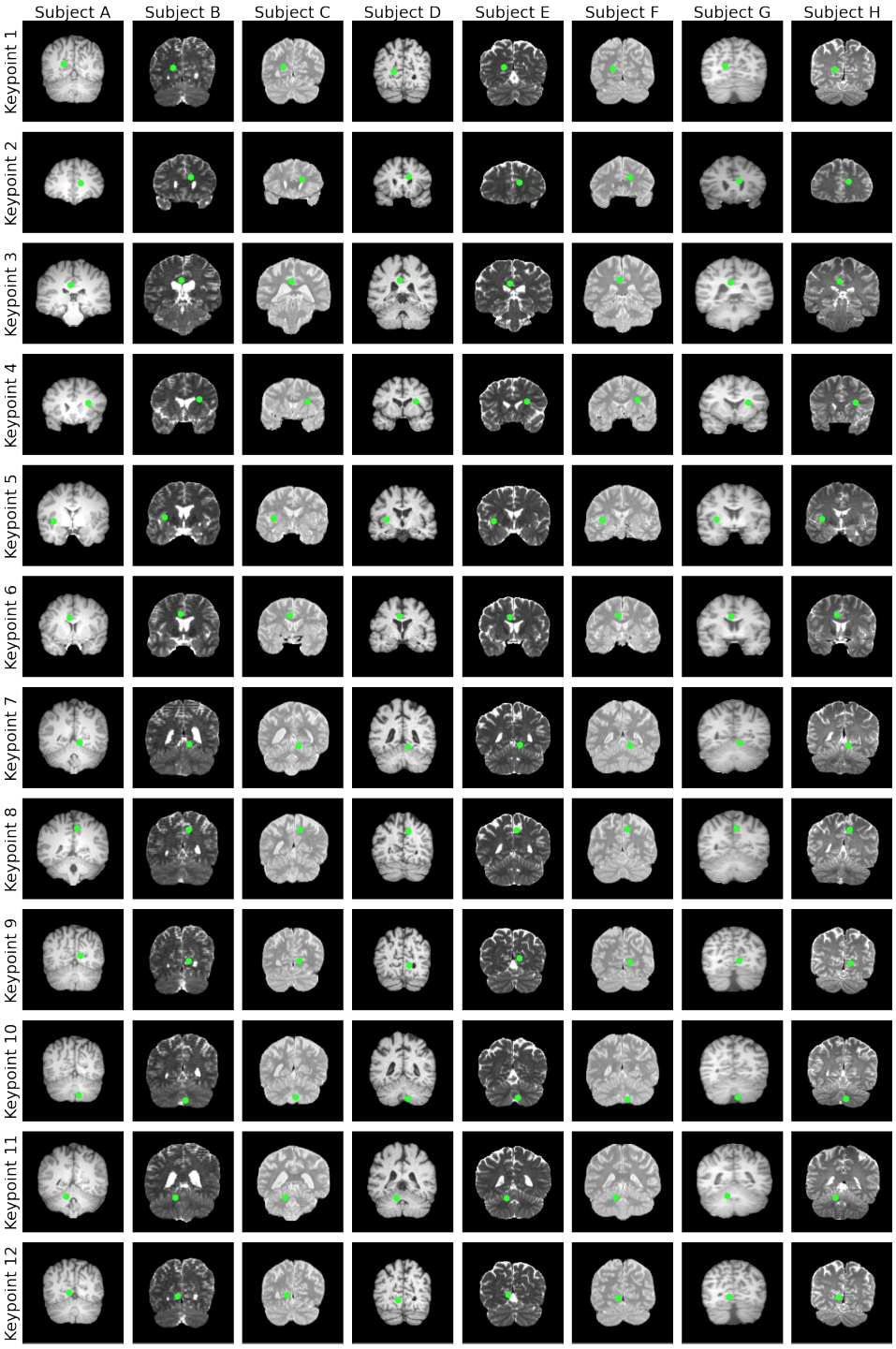


Figure A.7: Keypoints for different multimodal scans

APPENDIX B

SUPPLEMENTARY MATERIAL FOR "AN AUTO-ENCODER STRATEGY FOR ADAPTIVE IMAGE SEGMENTATION"

B.1 Close form solution of KL-Divergence with MRF prior

Let $p(\mathbf{s})$ be a prior with spatial information:

$$p(\mathbf{s}) = \frac{1}{z} \exp\left[\sum_{i=0}^{V} \log p(s_i) + \sum_{i=0}^{V} \sum_{y \in \mathcal{N}_i} \log p(s_y|s_i)\right]. \tag{B.1}$$

If approximate posterior is $q(\mathbf{s}|\mathbf{x}) = \prod_{i=0}^{V} q(s_i|\mathbf{x})$. Then the cross-entropy term from the KL-divergence $D_{KL}(q(\mathbf{s}|\mathbf{x})||p(\mathbf{s}))$ is given by:

$$\mathop{\mathbb{E}}_{\mathbf{s}\sim q(\mathbf{s}|\mathbf{x})}\left[c + \sum_{i=0}^{V} \log p(s_i) + \sum_{i=0}^{V} \sum_{y \in \mathcal{N}_i} \log p(s_y|s_i)\right]. \tag{B.2}$$

We are interested in the last term of the summation. By linearity of expectation, we define the spatial consistency loss as:

$$\begin{aligned}
\mathcal{L}_{mrf} &= -\sum_{i=0}^{V} \sum_{y \in \mathcal{N}_i} \mathop{\mathbb{E}}_{\mathbf{s}\sim q(\mathbf{s}|\mathbf{x})} \log p(s_y|s_i) \\
&= -\sum_{i=0}^{V} \sum_{y \in \mathcal{N}_i} \mathop{\mathbb{E}}_{\mathbf{s}\sim q_i q_j} \log p(s_y|s_i) \\
&= -\sum_{i=0}^{V} \sum_{y \in \mathcal{N}_i} \left(\sum_{l_i=0}^{L-1} \sum_{l_j=0}^{L-1} q_i(l_i|\mathbf{x}) q_y(l_j|\mathbf{x}) \log p(s_y = l_j|s_i = l_i)\right) \\
&= -\sum_{i=0}^{V} \left(\sum_{l_i=0}^{L-1} \sum_{l_j=0}^{L-1} q_i(l_i|\mathbf{x}) \sum_{y \in \mathcal{N}_i} q_y(l_j|\mathbf{x}) \log p(s_y = l_j|s_i = l_i)\right) \\
&= -\sum_{i=0}^{V} \left(\sum_{l_i=0}^{L-1} q_i(l_i|\mathbf{x}) \sum_{l_j=0}^{L-1} \sum_{y \in \mathcal{N}_i} q_y(l_j|\mathbf{x}) \log p(s_y = l_j|s_i = l_i)\right).
\end{aligned} \tag{B.3}$$

B.1.1 Intuition behind MRF prior

The intuition behind $p(s_y|s_i)$ is quite simple. For example, we know that certain brain region cannot be next to each other or very unlikely be next to each other. This can be captured using a lookup table. It is computed using the prior or single labeled example. We count the occurrence of each label in a 3x3x3 neighborhood around a given region of the brain. The result is displayed in in Fig. B.1 for the Buckner atlas. For a given label in the row, the probability of seen a label in the column is shown

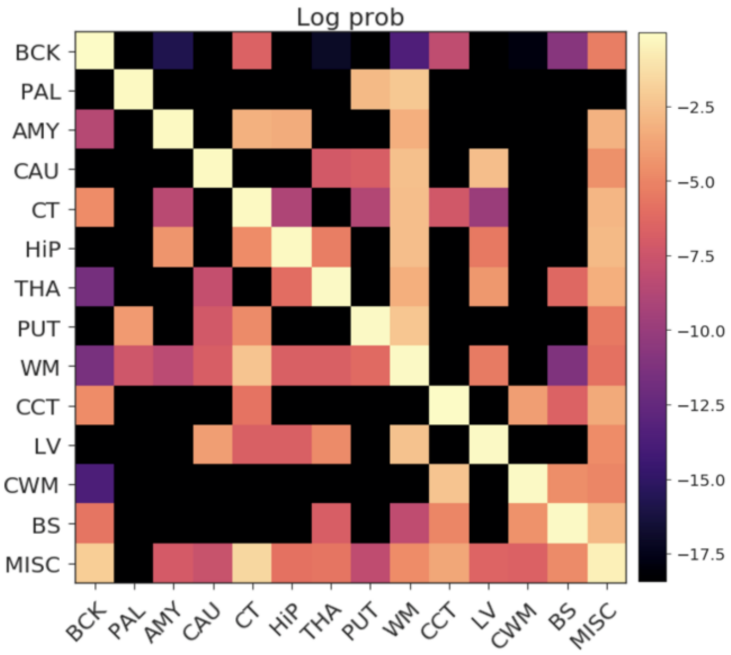


Figure B.1: Neighborhood probabilities from Buckner atlas. Given region in the column, the log probability of seeing a label in the row is shown.

BIBLIOGRAPHY

[1] Daniel H Adler, Laura EM Wisse, Ranjit Ittyerah, John B Pluta, Song-Lin Ding, Long Xie, Jiancong Wang, Salmon Kadivar, John L Robinson, Theresa Schuck, et al. Characterizing the human hippocampus in aging and alzheimer's disease using a computational atlas derived from ex vivo mri and histology. *Proceedings of the National Academy of Sciences*, 115(16):4252–4257, 2018.

[2] Tarun Kumar Agarwal, Mayank Tiwari, and Subir Singh Lamba. Modified histogram based contrast enhancement using homomorphic filtering for medical images. In *2014 IEEE International Advance Computing Conference (IACC)*, pages 964–968. IEEE, 2014.

[3] Zeynettin Akkus, Alfiia Galimzianova, Assaf Hoogi, Daniel L Rubin, and Bradley J Erickson. Deep learning for brain mri segmentation: state of the art and future directions. *Journal of digital imaging*, 30(4):449–459, 2017.

[4] Paul Aljabar, Rolf A Heckemann, Alexander Hammers, Joseph V Hajnal, and Daniel Rueckert. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *Neuroimage*, 46(3):726–738, 2009.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[6] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer, 2006.

[7] John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007.

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

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

[10] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[11] Bai et al. Gift: A real-time and scalable 3d shape search engine. In *Proc of CVPR*, 2016.

[12] Ruzena Bajcsy and Stane Kovačič. Multiresolution elastic matching. *Computer vision, graphics, and image processing*, 46(1):1–21, 1989.

[13] 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.

[14] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key. net: Keypoint detection by handcrafted and learned cnn filters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5836–5844, 2019.

[15] Stefan Bauer, Thomas Fejes, and Mauricio Reyes. A skull-stripping filter for itk. *Insight Journal*, 2012, 2013.

[16] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[17] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005.

[18] Benjamin Billot, Douglas Greve, Koen Van Leemput, Bruce Fischl, Juan Eugenio Iglesias, and Adrian V Dalca. A learning strategy for contrast-agnostic mri segmentation. *arXiv preprint arXiv:2003.01995*, 2020.

[19] Morten Bro-Nielsen and Claus Gramkow. Fast fluid registration of medical images. In *International Conference on Visualization in Biomedical Computing*, pages 265–276. Springer, 1996.

[20] Bronstein et al. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Proc Mag*, 2017.

[21] Matthew Brown, Richard Szeliski, and Simon Winder. Multi-image matching using multi-scale oriented patches. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 510–517. IEEE, 2005.

[22] Marijn E Brummer, Russell M Mersereau, Robert L Eisner, and Richard RJ Lewine. Automatic detection of brain contours in mri data sets. *IEEE Transactions on medical imaging*, 12(2):153–166, 1993.

[23] Xiaohuan Cao, Jianhua Yang, Jun Zhang, Qian Wang, Pew-Thian Yap, and Dinggang Shen. Deformable image registration using a cue-aware deep regression network. *IEEE Transactions on Biomedical Engineering*, 65(9):1900–1911, 2018.

[24] Aaron Carass, Jennifer Cuzzocreo, M Bryan Wheeler, Pierre-Louis Bazin, Susan M Resnick, and Jerry L Prince. Simple paradigm for extra-cerebral tissue removal: algorithm and analysis. *NeuroImage*, 56(4):1982–1992, 2011.

[25] Aaron Carass, M Bryan Wheeler, Jennifer Cuzzocreo, Pierre-Louis Bazin, Susan S Bassett, and Jerry L Prince. A joint registration and segmentation approach to skull stripping. In *2007 4th IEEE international symposium on biomedical imaging: from nano to macro*, pages 656–659. IEEE, 2007.

[26] Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

[27] Krishna Chaitanya, Neerav Karani, Christian F Baumgartner, Anton Becker, Olivio Donati, and Ender Konukoglu. Semi-supervised and task-driven data augmentation. In *International Conference on Information Processing in Medical Imaging*, pages 29–41. Springer, 2019.

[28] Jr-Yuan Chiou, Chang Beom Ahn, Lutfi Tugan Muftuler, and Orhan Nalcioglu. A simple simultaneous geometric and intensity correction method for echo-planar imaging by epi-based phase modulation. *IEEE transactions on medical imaging*, 22(2):200–205, 2003.

[29] Haili Chui and Anand Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2-3):114–141, 2003.

[30] G Collewet, A Davenel, C Toussaint, and Serge Akoka. Correction of intensity nonuniformity in spin-echo t1-weighted images. *Magnetic resonance imaging*, 20(4):365–373, 2002.

[31] Olivier Colliot, Oscar Camara, and Isabelle Bloch. Integration of fuzzy spatial relations in deformable models—application to brain mri segmentation. *Pattern recognition*, 39(8):1401–1414, 2006.

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

[33] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[34] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018.

[35] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236, 2019.

[36] 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*, pages 9290–9299, 2018.

[37] Adrian V Dalca, Marianne Rakic, John Guttag, and Mert R Sabuncu. Learning conditional deformable templates with convolutional networks. *arXiv preprint arXiv:1908.02738*, 2019.

[38] Adrian V Dalca, Evan Yu, Polina Golland, Bruce Fischl, Mert R Sabuncu, and Juan Eugenio Iglesias. Unsupervised deep learning for bayesian brain mri segmentation. *arXiv preprint arXiv:1904.11319*, 2019.

[39] Anders M Dale, Bruce Fischl, and Martin I Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194, 1999.

[40] Bob D de Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.

[41] Bob D de Vos, Floris F Berendsen, Max A Viergever, Marius Staring, and

Ivana Išgum. End-to-end unsupervised deformable image registration with a convolutional neural network. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 204–212. Springer, 2017.

[42] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[43] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[44] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. 3d fully convolutional networks for subcortical segmentation in mri: A large-scale study. *NeuroImage*, 170:456–470, 2018.

[45] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[46] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[47] Simon Duchesne, Jens C Pruessner, and D Louis Collins. Appearance-based segmentation of medial temporal lobe structures. *NeuroImage*, 17(2):515–531, 2002.

[48] Durrleman et al. Inferring brain variability from diffeomorphic deformations of currents: an integrative approach. *Medical image analysis*, 2008.

[49] Koen AJ Eppenhof and Josien PW Pluim. Pulmonary ct registration through supervised learning with convolutional neural networks. *IEEE transactions on medical imaging*, 38(5):1097–1105, 2018.

[50] Jingfan Fan, Xiaohuan Cao, Zhong Xue, Pew-Thian Yap, and Dinggang Shen. Adversarial similarity network for evaluating image alignment in deep learning based registration. pages 739–746, 2018.

[51] Jingfan Fan, Xiaohuan Cao, Pew-Thian Yap, and Dinggang Shen. Birnet: Brain image registration using dual-supervised fully convolutional networks. *Medical image analysis*, 54:193–206, 2019.

[52] Fang et al. 3D deep shape descriptor. In *Proc of CVPR*, 2015.

[53] Bernd Fischer and Jan Modersitzki. Curvature based image registration. *Journal of Mathematical Imaging and Vision*, 18(1):81–85, 2003.

[54] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[55] 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.

[56] Wolfgang Förstner and Eberhard Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *Proc. ISPRS intercommission conference on fast processing of photogrammetric data*, pages 281–305. Interlaken, 1987.

[57] Anthony F Fotenos, AZ Snyder, LE Girton, JC Morris, and RL Buckner. Normative estimates of cross-sectional and longitudinal brain volume decline in aging and ad. *Neurology*, 64(6):1032–1039, 2005.

[58] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

[59] Ge et al. Multidimensional heritability analysis of neuroanatomical shape. *Nature Communications*, 2016.

[60] Guido Gerig, Olaf Kubler, Ron Kikinis, and Ferenc A Jolesz. Nonlinear anisotropic filtering of mri data. *IEEE Transactions on medical imaging*, 11(2):221–232, 1992.

[61] Juan D Gispert, Santiago Reig, Javier Pascau, Juan J Vaquero, Pedro García-Barreno, and Manuel Desco. Method for bias field correction of brain t1-weighted magnetic resonance images minimizing segmentation error. *Human brain mapping*, 22(2):133–144, 2004.

[62] Randy L Gollub, Jody M Shoemaker, Margaret D King, Tonya White, Stefan Ehrlich, Scott R Sponheim, Vincent P Clark, Jessica A Turner, Bryon A Mueller, Vince Magnotta, et al. The mcic collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics*, 11(3):367–388, 2013.

[63] Sandra González-Villà, Arnau Oliver, Sergi Valverde, Liping Wang, Reyer Zwiggelaar, and Xavier Lladó. A review on brain structures segmentation in magnetic resonance imaging. *Artificial intelligence in medicine*, 73:45–69, 2016.

[64] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[65] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[66] Laura Gui, Radoslaw Lisowski, Tamara Faundez, Petra S Hüppi, François Lazeyras, and Michel Kocher. Morphology-driven automatic segmentation of mr images of the neonatal brain. *Medical image analysis*, 16(8):1565–1579, 2012.

[67] Gutierrez-Becker and Wachingger. Deep multi-structural shape analysis: Application to neuroanatomy. *In MICCAI*, 2018.

[68] Horst K Hahn and Heinz-Otto Peitgen. The skull stripping problem in mri solved by a single 3d watershed transform. In *International Conference on medical image computing and computer-assisted intervention*, pages 134–143. Springer, 2000.

[69] Hamarneh et al. Medial profiles for modeling deformation and statistical analysis of shape and their use in medical image segmentation. *Int J of Shape Mod*, 2004.

[70] Yongfu Hao, Tianyao Wang, Xinqing Zhang, Yunyun Duan, Chunshui Yu, Tianzi Jiang, Yong Fan, and Alzheimer's Disease Neuroimaging Initiative. Local label learning (lll) for subcortical structure segmentation: application to hippocampus segmentation. *Human brain mapping*, 35(6):2674–2697, 2014.

[71] Christopher G Harris, Mike Stephens, et al. A combined corner and edge

detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988.

[72] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[73] Trey Hedden and John DE Gabrieli. Insights into the ageing mind: a view from cognitive neuroscience. *Nature reviews neuroscience*, 5(2):87–96, 2004.

[74] Mattias P Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V Gleeson, Michael Brady, and Julia A Schnabel. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical image analysis*, 16(7):1423–1435, 2012.

[75] Gerardo Hermosillo, Christophe Chefd'Hotel, and Olivier Faugeras. Variational methods for multimodal image matching. *International Journal of Computer Vision*, 50(3):329–343, 2002.

[76] Derek LG Hill, Philipp G Batchelor, Mark Holden, and David J Hawkes. Medical image registration. *Physics in medicine & biology*, 46(3):R1, 2001.

[77] Geoffrey Hinton. Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude, 2012.

[78] Malte Hoffmann, Benjamin Billot, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. Learning multi-modal image registration without real data. *arXiv preprint arXiv:2004.10282*, 2020.

[79] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[80] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[81] Yipeng Hu, Marc Modat, Eli Gibson, Nooshin Ghavami, Ester Bonmati, Caroline M Moore, Mark Emberton, J Alison Noble, Dean C Barratt, and Tom Vercauteren. Label-driven weakly-supervised learning for multimodal deformable image registration. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1070–1074. IEEE, 2018.

[82] Yipeng Hu, Marc Modat, Eli Gibson, Wenqi Li, Nooshin Ghavami, Ester

Bonmati, Guotai Wang, Steven Bandula, Caroline M Moore, Mark Emberton, et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Medical image analysis*, 49:1–13, 2018.

[83] Juan Eugenio Iglesias and Mert R Sabuncu. Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis*, 24(1):205–219, 2015.

[84] Jaderberg et al. Spatial transformer networks. In *Advances in neural information processing systems*, 2015.

[85] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*, 2015.

[86] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[87] Mark Jenkinson, Mickael Pechaud, Stephen Smith, et al. Bet2: Mr-based estimation of brain, skull and scalp surfaces. In *Eleventh annual meeting of the organization for human brain mapping*, volume 17, page 167. Toronto., 2005.

[88] Amod Jog and Bruce Fischl. Pulse sequence resilient fast brain segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 654–662. Springer, 2018.

[89] Sarang Joshi, Brad Davis, Matthieu Jomier, and Guido Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151–S160, 2004.

[90] Sarang C Joshi and Michael I Miller. Landmark matching via large deformation diffeomorphisms. *IEEE transactions on image processing*, 9(8):1357–1370, 2000.

[91] Thomas Joyce, Agisilaos Chartsias, and Sotirios A Tsaftaris. Deep multi-class segmentation without ground-truth labels. 2018.

[92] P Kalavathi and VB Surya Prasath. Methods on skull stripping of mri head scan images—a review. *Journal of digital imaging*, 29(3):365–379, 2016.

[93] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben

Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[94] Tina Kapur, W Eric L Grimson, William M Wells III, and Ron Kikinis. Segmentation of brain tissue from magnetic resonance images. *Medical image analysis*, 1(2):109–127, 1996.

[95] Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. Deep learning applications in medical image analysis. *Ieee Access*, 6:9375–9389, 2017.

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

[97] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

[99] Julian Krebs, Hervé Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE transactions on medical imaging*, 38(9):2165–2176, 2019.

[100] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[101] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[102] Shang-Hong Lai and Ming Fang. A dual image approach for bias field correction in magnetic resonance imaging. *Magnetic resonance imaging*, 21(2):121–125, 2003.

[103] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113, 1997.

[104] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

[105] Hyeon Woo Lee, Mert R Sabuncu, and Adrian V Dalca. Few labeled atlases are necessary for deep-learning-based segmentation. *arXiv preprint arXiv:1908.04466*, 2019.

[106] Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H Sebastian Seung. Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*, 2017.

[107] Matthew CH Lee, Ozan Oktay, Andreas Schuh, Michiel Schaap, and Ben Glocker. Image-and-spatial transformer networks for structure-guided image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–345. Springer, 2019.

[108] Karel Lenc and Andrea Vedaldi. Learning covariant feature detectors. In *European conference on computer vision*, pages 100–117. Springer, 2016.

[109] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[110] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[111] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[112] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[113] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6232–6240, 2017.

[114] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1):23–79, 2021.

[115] Jun Ma, Michael I Miller, Alain Trouvé, and Laurent Younes. Bayesian template estimation in computational anatomy. *NeuroImage*, 42(1):252–261, 2008.

[116] Tianyu Ma, Ajay Gupta, and Mert R. Sabuncu. Volumetric landmark detection with a multi-scale shift equivariant neural network, 2020.

[117] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[118] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[119] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[120] José V Manjón. Mri preprocessing. In *Imaging Biomarkers*, pages 53–63. Springer, 2017.

[121] José V Manjón, Pierrick Coupé, Antonio Buades, D Louis Collins, and Montserrat Robles. New methods for mri denoising based on sparseness and self-similarity. *Medical image analysis*, 16(1):18–27, 2012.

[122] Marcus et al. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 2007.

[123] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

[124] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide-

baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004.

[125] David Mattes, David R Haynor, Hubert Vesselle, Thomas K Lewellen, and William Eubank. Pet-ct image registration in the chest using free-form deformations. *IEEE transactions on medical imaging*, 22(1):120–128, 2003.

[126] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[127] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[128] Sergiu Mocanu, Alan R Moody, and April Khademi. Flowreg: Fast deformable unsupervised medical image registration using optical flow. *arXiv preprint arXiv:2101.09639*, 2021.

[129] Philippe Montesinos, Valérie Gouet, and Rachid Deriche. Differential invariants for color images. In *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, volume 1, pages 838–840. IEEE, 1998.

[130] Monti et al. Geometric deep learning on graphs and manifolds using mixture model CNNs. In *Proc. of CVPR*, 2017.

[131] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

[132] Bernard Ng, Matthew Toews, Stanley Durrleman, and Yonggang Shi. Shape analysis for brain structures. *Shape Analysis in Medical Image Analysis*, pages 3–49, 2014.

[133] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[134] Francisco PM Oliveira and Joao Manuel RS Tavares. Medical image regis-

tration: a review. *Computer methods in biomechanics and biomedical engineering*, 17(2):73–93, 2014.

[135] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: Learning local features from images. *arXiv preprint arXiv:1805.09662*, 2018.

[136] Theodore J Passe, Pradeep Rajagopalan, Larry A Tupler, Christopher E Byrum, James R Macfall, and K Krishnan. Age and sex effects on brain morphology. *Progress in neuro-psychopharmacology & biological psychiatry*, 1997.

[137] Alain Pitiot, Hervé Delingette, Paul M Thompson, and Nicholas Ayache. Expert knowledge-guided segmentation system for brain mri. *NeuroImage*, 23:S85–S96, 2004.

[138] Oula Puonti, Juan Eugenio Iglesias, and Koen Van Leemput. Fast and sequence-adaptive whole-brain segmentation using parametric bayesian modeling. *NeuroImage*, 143:235–249, 2016.

[139] Qi et al. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc of CVPR*, 2017.

[140] Chen Qin, Bibo Shi, Rui Liao, Tommaso Mansi, Daniel Rueckert, and Ali Kamen. Unsupervised deformable registration for multi-modal images via disentangled representations. *Lecture Notes in Computer Science Information Processing in Medical Imaging*, page 249–261, 2019.

[141] Naftali Raz, Faith Gunning-Dixon, Denise Head, Karen M Rodrigue, Adrienne Williamson, and James D Acker. Aging, sexual dimorphism, and hemispheric asymmetry of the cerebral cortex: replicability of regional differences in volume. *Neurobiology of aging*, 25(3):377–396, 2004.

[142] Naftali Raz, Ulman Lindenberger, Karen M Rodrigue, Kristen M Kennedy, Denise Head, Adrienne Williamson, Cheryl Dahle, Denis Gerstorf, and James D Acker. Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. *Cerebral cortex*, 15(11):1676–1689, 2005.

[143] Hafiz Zia Ur Rehman, Hyunho Hwang, and Sungon Lee. Conventional and deep learning methods for skull stripping in brain mri. *Applied Sciences*, 10(5):1773, 2020.

[144] Reuter et al. Laplace–Beltrami spectra as Shape-DNA of surfaces and solids. *Computer-Aided Design*, 2006.

[145] Annemie Ribbens, Jeroen Hermans, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Unsupervised segmentation, clustering, and groupwise registration of heterogeneous populations of brain mr images. *IEEE transactions on medical imaging*, 33(2):201–224, 2013.

[146] 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*, pages 234–241. Springer, 2015.

[147] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[148] Azriel Rosenfeld and Mark Thurston. Edge and curve detection for visual scene analysis. *IEEE Transactions on computers*, 100(5):562–569, 1971.

[149] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999.

[150] Mert R Sabuncu, Serdar K Balci, and Polina Golland. Discovering modes of an image population through mixture modeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 381–389. Springer, 2008.

[151] Mert R Sabuncu, BT Thomas Yeo, Koen Van Leemput, Bruce Fischl, and Polina Golland. A generative model for image segmentation based on label fusion. *IEEE transactions on medical imaging*, 29(10):1714–1729, 2010.

[152] Julia A Schnabel, Daniel Rueckert, Marcel Quist, Jane M Blackall, Andy D Castellano-Smith, Thomas Hartkens, Graeme P Penney, Walter A Hall, Haiying Liu, Charles L Truwit, et al. A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 573–581. Springer, 2001.

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

[154] Alberto Serrano-Pozo, Matthew P Frosch, Eliezer Masliah, and Bradley T Hyman. Neuropathological alterations in alzheimer disease. *Cold Spring Harbor perspectives in medicine*, 1(1):a006189, 2011.

[155] Shakeri et al. Deep spectral-based shape features for Alzheimer's disease classification. In *Int Workshop on Spectral and Shape Analysis in Medical Imaging*, 2016.

[156] David W Shattuck, Stephanie R Sandor-Leahy, Kirt A Schaper, David A Rottenberg, and Richard M Leahy. Magnetic resonance image tissue classification using a partial volume model. *NeuroImage*, 13(5):856–876, 2001.

[157] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[158] John G Sled, Alex P Zijdenbos, and Alan C Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging*, 17(1):87–97, 1998.

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

[160] Michal Sofka, Fausto Milletari, Jimmy Jia, and Alex Rothberg. Fully convolutional regression network for accurate detection of measurement points. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 258–266. Springer, 2017.

[161] Hessam Sokooti, Bob De Vos, Floris Berendsen, Boudewijn PF Lelieveldt, Ivana Išgum, and Marius Staring. Nonrigid image registration using multi-scale 3d convolutional neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 232–239. Springer, 2017.

[162] Shuang Song, Yuanjie Zheng, and Yunlong He. A review of methods for bias correction in medical images. *Biomedical Engineering Review*, 1(1), 2017.

[163] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, 32(7):1153–1190, 2013.

[164] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks

from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[165] Marius Staring, Stefan Klein, and Josien PW Pluim. A rigidity penalty term for nonrigid registration. *Medical physics*, 34(11):4098–4108, 2007.

[166] Styner et al. Framework for the statistical shape analysis of brain structures using SPHARM-PDM. *The Insight Journal*, 2006.

[167] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.

[168] J-P Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical image analysis*, 2(3):243–260, 1998.

[169] Matthew Toews, Lilla Zöllei, and William M Wells. Feature-based alignment of volumetric multi-modal images. In *International Conference on Information Processing in Medical Imaging*, pages 25–36. Springer, 2013.

[170] Tong Tong, Robin Wolz, Pierrick Coupé, Joseph V Hajnal, Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. Segmentation of mr images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling. *NeuroImage*, 76:11–23, 2013.

[171] Zhuowen Tu and Xiang Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 32(10):1744–1757, 2009.

[172] Zhuowen Tu, Katherine L Narr, Piotr Dollár, Ivo Dinov, Paul M Thompson, and Arthur W Toga. Brain anatomical structure segmentation by hybrid discriminative/generative models. *IEEE transactions on medical imaging*, 27(4):495–508, 2008.

[173] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.

[174] Tinne Tuytelaars and Krystian Mikolajczyk. *Local invariant feature detectors: a survey*. Now Publishers Inc, 2008.

[175] Kâmil Uludağ and Alard Roebroeck. General overview on the merits of multimodal neuroimaging data fusion. *Neuroimage*, 102:3–10, 2014.

[176] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[177] Hristina Uzunova, Matthias Wilms, Heinz Handels, and Jan Ehrhardt. Training cnns for image registration from few samples with model-based data augmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 223–231. Springer, 2017.

[178] Joost Van de Weijer, Theo Gevers, and Andrew D Bagdanov. Boosting color saliency in image feature detection. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):150–156, 2005.

[179] Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Automated model-based tissue classification of mr images of the brain. *IEEE transactions on medical imaging*, 18(10):897–908, 1999.

[180] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009.

[181] Yannick Verdie, Kwang Yi, Pascal Fua, and Vincent Lepetit. Tilde: A temporally invariant learned detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5279–5288, 2015.

[182] Paul Viola and William M Wells III. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997.

[183] A Vita, L De Peri, C Silenzi, and M Dieci. Brain morphology in first-episode schizophrenia: a meta-analysis of quantitative magnetic resonance imaging studies. *Schizophrenia research*, 82(1):75–88, 2006.

[184] Wachinger et al. BrainPrint: A discriminative characterization of brain morphology. *NeuroImage*, 2015.

[185] Christian Wachinger, Matthew Toews, Georg Langs, William Wells, and Polina Golland. Keypoint transfer for fast whole-body segmentation. *IEEE transactions on medical imaging*, 39(2):273–282, 2018.

[186] Hongzhi Wang, Jung W Suh, Sandhitsu R Das, John B Pluta, Caryne Craige, and Paul A Yushkevich. Multi-atlas segmentation with joint label fusion. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):611–623, 2012.

[187] Wu et al. 3d shapenets: A deep representation for volumetric shapes. In *Proc of CVPR*, 2015.

[188] Guorong Wu, Minjeong Kim, Qian Wang, Brent C Munsell, and Dinggang Shen. Scalable high-performance image registration framework by unsupervised deep feature representations learning. *IEEE Transactions on Biomedical Engineering*, 63(7):1505–1516, 2015.

[189] Yan Xia, Keith Bettinger, Lin Shen, and Allan L Reiss. Automatic segmentation of the caudate nucleus from human brain mr images. *IEEE Transactions on Medical Imaging*, 26(4):509–517, 2007.

[190] Xie et al. Projective feature learning for 3d shapes with multi-view depth images. In *Computer Graphics Forum*, 2015.

[191] Xie et al. Deepshape: Deep-learned shape descriptor for 3d shape retrieval. *IEEE Tran on PAMI*, 2017.

[192] Jing-Hao Xue, Su Ruan, Bruno Moretti, Marinette Revenu, and Daniel Bloyet. Knowledge-based segmentation and labeling of brain structures from mri images. *Pattern recognition letters*, 22(3-4):395–405, 2001.

[193] Jing-Hao Xue, Su Ruan, Bruno Moretti, Marinette Revenu, Daniel Bloyet, and Wilfried Philips. Fuzzy modeling of knowledge for mri brain structure segmentation. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, volume 1, pages 617–620. IEEE, 2000.

[194] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage*, 158:378–396, 2017.

[195] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European conference on computer vision*, pages 467–483. Springer, 2016.

[196] Evan M Yu, Juan Eugenio Iglesias, Adrian V Dalca, and Mert R Sabuncu.

An auto-encoder strategy for adaptive image segmentation. *arXiv preprint arXiv:2004.13903*, 2020.

[197] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet, 2021.

[198] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[199] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.

[200] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8553, 2019.

[201] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[202] Zhu et al. Deep learning representation using autoencoder for 3d shape retrieval. *Neurocomputing*, 2016.

[203] Ying Zhuge, Jayaram K Udupa, Jiamin Liu, Punam K Saha, and Tad Iwanage. Scale-based method for correcting background intensity variation in acquired images. In *Medical Imaging 2002: Image Processing*, volume 4684, pages 1103–1111. International Society for Optics and Photonics, 2002.