

Unsupervised Domain Adaptation via Content Alignment for Hippocampus Segmentation

Hoda Kalabizadeh¹ Ludovica Griffanti¹ Pak-Hei Yeung²
Ana I. L. Namburete¹ Nicola K. Dinsdale¹ Konstantinos Kamnitsas¹

¹University of Oxford, United Kingdom

²Nanyang Technological University, Singapore

hoda.kalabizadeh@cs.ox.ac.uk

Abstract

Deep learning models for medical image segmentation often struggle when deployed across different datasets due to domain shifts - variations in both image appearance, known as style, and population-dependent anatomical characteristics, referred to as content. This paper presents a novel unsupervised domain adaptation framework that directly addresses domain shifts encountered in cross-domain hippocampus segmentation from MRI, with specific emphasis on content variations. Our approach combines efficient style harmonisation through z -normalisation with a bidirectional deformable image registration (DIR) strategy. The DIR network is jointly trained with segmentation and discriminator networks to guide the registration with respect to a region of interest and generate anatomically plausible transformations that align source images to the target domain. We validate our approach through comprehensive evaluations on both a synthetic dataset using Morpho-MNIST (for controlled validation of core principles) and three MRI hippocampus datasets representing populations with varying degrees of atrophy. Across all experiments, our method outperforms existing baselines. For hippocampus segmentation, when transferring from young, healthy populations to clinical dementia patients, our framework achieves up to 15% relative improvement in Dice score compared to standard augmentation methods, with the largest gains observed in scenarios with substantial content shift. These results highlight the efficacy of our approach for accurate hippocampus segmentation across diverse populations.

1. Introduction

Volumetric analysis of brain MRI, particularly of the hippocampus, is a common approach for diagnosing and monitoring dementia [1], with measurements typically obtained

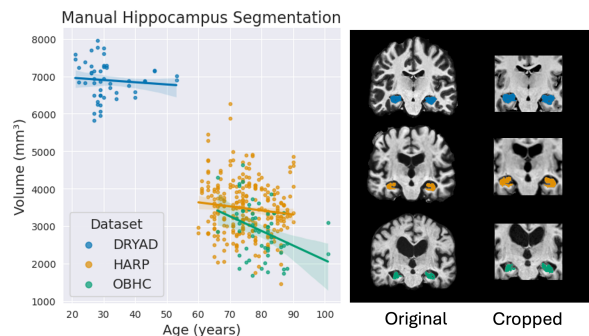


Figure 1. Comparison of hippocampal volumes between three populations. DRYAD: young and healthy; HarP: dementia research population; OBHC: dementia clinic population. Demonstrating the shrinking of the hippocampi across populations due to age and disease status/severity.

through manual or automated segmentation [2]. Manual segmentation is time-consuming and requires expert knowledge [3], creating a strong demand for accurate automated methods for hippocampus segmentation.

Deep learning (DL) models have demonstrated success in automatic segmentation across various medical imaging modalities [4]. However, their performance generally relies on two key prerequisites: large labelled training datasets and identical distributions between training and testing data. In medical applications, these prerequisites are often difficult to meet due to the high cost of expert annotations and distribution mismatch between training (source) and testing (target) domains, known as the *domain shift* [5]. Domain shifts between MRI datasets can be conceptualised as a combination of *style* variations—arising from scanner characteristics and acquisition parameters—and *content* variations, reflecting population differences across studies [6].

This paper aims to address distribution shifts between related domains in medical imaging, particularly in settings

where labels are scarce in the target domain. Specifically, we focus on hippocampus segmentation from MRI across different populations. These datasets exhibit noticeable domain shift in two distinct aspects: *style*, which arises from site and machine differences, and *content*, which is caused by substantial atrophy of the hippocampus and surrounding brain regions due to ageing and disease progression [7]. Figure 1 illustrates these variations in hippocampal volumes across datasets with different image acquisitions and population demographics, highlighting the need for domain adaptation methods that address both style and content shifts.

We aim to tackle the problem through unsupervised domain adaptation (UDA). Existing UDA approaches for medical image segmentation often employ either adversarial feature alignment or generative style transfer methods [8–10]. While these techniques effectively address style shifts such as scanner-related variations and imaging protocol differences, they fundamentally assume that the underlying anatomical structures remain consistent across domains. This assumption is violated in cross-population studies where content shifts—including disease progression, ageing, and demographic differences—lead to substantial morphological variations. Moreover, generative approaches often introduce computational overhead and training instability, while feature-level alignment methods may lose important spatial information crucial for dense prediction tasks like segmentation.

To address these limitations, we propose that content shifts should be explicitly modelled in image space through spatial transformations that can capture the underlying anatomical variations between populations. By leveraging deformable registration techniques, we can learn domain-specific anatomical transformations while preserving the spatial correspondence between images and their associated labels—a critical requirement for training segmentation models with transformed data.

To evaluate our approach comprehensively, we conduct experiments on multiple datasets spanning different complexity levels. First, we create a controlled synthetic dataset using a variant of MNIST (Morpho-MNIST [11]) to isolate and validate our method’s ability to handle content shifts in a simplified setting, where thick and thin digit variants simulate the structural changes observed in hippocampal atrophy. Second, we demonstrate clinical relevance using three real brain MRI datasets with progressively challenging domain shifts: a young healthy population (DRYAD [12]), a dementia research cohort (HarP [13]), and a clinical dementia population (OBHC [14]). As illustrated in Figure 1, these datasets exhibit substantial hippocampal volume variations across populations, with progressive atrophy from healthy young brains to clinical dementia cases, enabling comprehensive evaluation of our content shift

adaptation method across realistic clinical scenarios.

In this study, we develop a UDA framework in image space that *explicitly* addresses the domain shifts inherent in cross-domain hippocampus segmentation, with particular attention to content variations. Our contributions include:

- We address *style* shift through a robust preprocessing pipeline, eliminating the need for a computationally intensive generative style transfer method.
- We propose a DIR-based method for addressing *content* shift, which preserves spatial correspondence between images and their associated labels and can be used to train a downstream segmentation model.
- Extensive experiments across multiple datasets demonstrate noticeable registration and segmentation performance improvements over baselines.

2. Related Work

2.1. Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) is a method that aligns source and target domain distributions at either the feature or image level, using only labelled source data and unlabelled target data during training [6].

In feature space, feature alignment methods focus on reducing the disparity between source and target domains by learning domain-invariant features. The goal is to map images from both domains into a shared latent space to minimise discrepancies [10]. For medical image segmentation tasks, previous studies have explored implicit discrepancy minimisation approaches using adversarial learning to maximise domain confusion while minimising segmentation loss [8, 15]. These adversarial methods have demonstrated significant improvements in multi-site segmentation tasks, effectively addressing *style* shifts that arise from differences in imaging protocols, scanner manufacturers, and acquisition parameters across different medical institutions.

In image space, image translation methods perform domain alignment in pixel space rather than feature space. In previous studies, these methods focused on converting source domain images to match the *style* of the target domain. These approaches commonly employ Generative Adversarial Networks (GANs), particularly CycleGAN [9, 16], which enables unpaired image-to-image translation using cycle consistency loss. Other studies have shown that for segmentation tasks, robust preprocessing methods, such as *z*-normalisation, perform comparably to GAN-based style transfer approaches [17].

Moreover, despite most studies introducing domain shifts occurring due to style and content shifts, many assumed the content/anatomical information to be domain-invariant. This assumption also forms the underlying basis for studies investigating style-content disentanglement [18–

20], which posit that style is domain-specific while content remains domain-invariant.

Consequently, UDA methods that explicitly address content shift by treating it as domain-specific have remained largely unexplored. To fill this gap, our framework proposes to model content shifts explicitly in image space through learnable transformations rather than treating them as invariant features, while handling style shifts through pre-processing, thereby avoiding the computational overhead of generative approaches.

2.2. Deformable Image Registration

Deformable Image Registration (DIR) is a technique for estimating non-rigid voxel correspondences between fixed and moving image pairs. It relies on two core loss functions: a dissimilarity function between the deformed and fixed images, and the regularisation penalty of the registration field. To preserve topology, additional constraints are commonly imposed on the registration field to ensure the mapping is differentiable and invertible, referred to as diffeomorphic registration [21]. Given the advancements of DL methods, neural network architectures have been developed that take fixed and moving images as input to estimate a registration field [22–24]. A popular example of an unsupervised DL-based registration model is VoxelMorph [23, 25]. Often proposed for 3D MR brain scans, VoxelMorph learns a registration function for aligning input image pairs.

Inspired by DIR, new methods have emerged for data augmentation via learning deformable transformations. Deformation-based transformations are able to capture more general spatial variations, making them an ideal approach for directly modelling the content shift between source and target data. Recently, studies have investigated learning shape-based data augmentations via deformation-based transformations for medical image segmentation tasks [26, 27]. For example, Zhao et al., [26] proposed a model for independently learning the spatial transformations between a pair of labelled and unlabelled MR datasets. They then used the model to generate augmented data based on the underlying spatial variations, which was subsequently used to train a segmentation model. Building on these insights, our framework leverages DIR techniques to explicitly model content shifts through learnable spatial transformations, while ensuring that the spatial correspondence between images and labels is preserved—a critical requirement for effective cross-domain segmentation in the presence of anatomical variations.

2.3. Joint Registration and Segmentation

Registration and segmentation tasks have great complementarity, which can bring mutual improvement in complex scenes with limited annotations [28]. The registration model generates augmented training data through im-

age warping (creating warped image-label pairs) to enhance segmentation training [26]. This approach reduces annotation requirements while improving segmentation generalisation. Conversely, the segmentation model provides region-based constraints [29] that focus registration attention on regions of interest, enabling more precise alignment in complex anatomical structures.

Previous studies have utilised DIR-based learning for one-shot or few-shot segmentation [26, 30]. However, their approach is not optimised for the segmentation task, allowing background structures, unrelated to the task, to have a substantial influence on the overall alignment. To address this limitation, He et al. proposed a complementary joint model for registration and segmentation, suggesting that region constraints fed back via the segmentation model can provide attention on ROIs, resulting in finer registration in complex settings [28]. Moreover, they include a discriminator network for adversarial training to further guide alignment by evaluating the similarity between transformed and target images. In another study, Wang et al. propose learning reversible voxel-wise correspondence for one-shot segmentation, investigating bidirectional warping to make the learning process more robust [30].

While these studies have made significant contributions to one-shot and few-shot segmentation, they do not inherently address the problem of domain shift, as they focus on learning from limited data within the same domain. Our framework bridges this gap by extending the proven complementarity of registration and segmentation to cross-domain scenarios. Here, the registration component learns population-specific anatomical transformations that enable effective domain adaptation rather than merely addressing data scarcity, while segmentation provides anatomically-informed constraints to enable systematic handling of cross-population morphological variations.

3. Method

3.1. Overview

Figure 2 visualises the proposed framework. We assume access to a source dataset, $\mathcal{D}_s = \{\mathcal{X}_s, \mathcal{Y}_s\}$, and an unlabelled target dataset, $\mathcal{D}_t = \{\mathcal{X}_t\}$. The framework consists of three core components: preprocessing for style alignment, content alignment, and downstream supervised segmentation. Source and target images are initially preprocessed to address the style shift between domains. Following this, the style-aligned images are used to train the content alignment module, which outputs source images aligned to the target domain, with attention focused on the region of interest (ROI) to be segmented, aiding the training process. Lastly, the transformed images and corresponding labels are used to train a segmentation network, which is then evaluated on unseen target images.

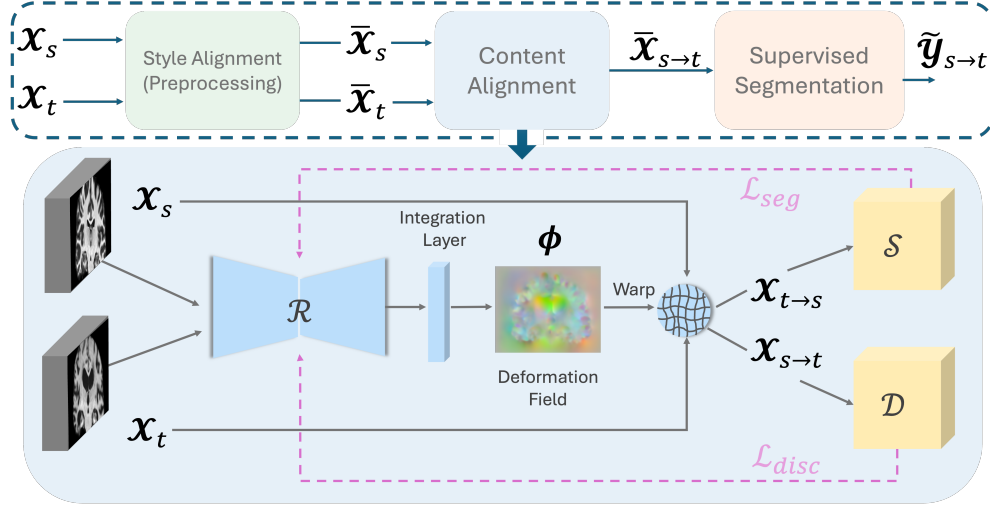


Figure 2. Top: Overview of the proposed framework, showing the sequence of steps for input source ($\mathcal{X}_s$) and target ($\mathcal{X}_t$) images to output style and content aligned images that are used to train a downstream segmentation network. Bottom: Proposed model architecture for learning content alignment, formed of registration ($\mathcal{R}$), segmentation ($\mathcal{S}$) and discriminator ($\mathcal{D}$) networks for joint training.

3.2. Data Preprocessing and Style Alignment

All MR scans were affine registered to a standard space (MNI) to ensure spatial correspondence across participants. The scans also underwent brain extraction to isolate brain tissue from surrounding non-brain structures. N4 bias field correction was then applied to address low-frequency intensity non-uniformities.

Inspired by findings from [17] demonstrating that robust intensity normalisation techniques sufficiently address style variations without requiring the computational complexity of GANs, we applied z -score normalisation along with histogram standardisation to harmonise intensity distributions, thereby reducing the impact of style differences. While pre-processing steps mitigate style-related domain shifts, they do not address the more challenging content shift. To overcome this, we developed a DIR-based model detailed in the following section.

3.3. Deformable Image Registration for Content Alignment

3.3.1. Registration Network:

To train a DIR network, $\mathcal{R}$, that learns to perform non-linear transformations from source to target images, we utilise the diffeomorphic VoxelMorph [31] backbone. $\mathcal{R}$ takes as input the source and target images, $\{\mathcal{X}_s, \mathcal{X}_t\}$, and outputs a velocity field which is passed through an integration layer to output a diffeomorphic deformation field, ϕ . We implement bidirectional warping by also computing the inverse deformation field ϕ^{-1} . Subsequently, $\{\mathcal{X}_s, \mathcal{Y}_s, \mathcal{X}_t\}$ are passed into a spatial transform model, with the deformation fields to compute: $\mathcal{X}_{s \rightarrow t} = \mathcal{X}_s \circ \phi$; $\mathcal{Y}_{s \rightarrow t} = \mathcal{Y}_s \circ \phi$, and $\mathcal{X}_{t \rightarrow s} = \mathcal{X}_t \circ \phi^{-1}$. The registration network is optimised

using two key loss components: $\mathcal{L}_{sim}$, a similarity loss that uses mean squared error (MSE) to ensure the deformed images closely match their targets in both forward and inverse directions; and $\mathcal{L}_{smooth}$, a smoothness constraint implementing $L2$ -regularisation on both the forward and inverse deformation fields to prevent physiologically implausible transformations and ensure spatial consistency of the registration.

$$\mathcal{L}_{sim} = \mathcal{L}_{mse}(\mathcal{X}_{s \rightarrow t}, \mathcal{X}_t) + \mathcal{L}_{mse}(\mathcal{X}_{t \rightarrow s}, \mathcal{X}_s) \quad (1)$$

$$\mathcal{L}_{smooth} = \mathcal{L}_{L2}(\phi) + \mathcal{L}_{L2}(\phi^{-1}) \quad (2)$$

3.3.2. Joint Learning Strategy:

We incorporated a discriminator network, $\mathcal{D}$, that was trained concurrently with the registration network in an adversarial framework. The discriminator learned to differentiate between the deformed source images ($\mathcal{X}_{s \rightarrow t}$) and real target images ($\mathcal{X}_t$) while simultaneously providing adversarial loss feedback to guide the registration process. This feedback was implemented as a binary cross-entropy loss: $\mathcal{L}_{disc} = \mathcal{L}_{bce}(\mathcal{X}_{s \rightarrow t})$. The adversarial training encourages the registration network to generate transformations that produce more target-like appearances, overcoming the limitations of conventional similarity metrics by providing additional constraints for content alignment.

We also included a 3D U-Net segmentation network, $\mathcal{S}$, which primarily provides feedback to the registration network while undergoing training concurrently on the labelled source data. For each training iteration, $\mathcal{S}$ processes two images. First, it takes the source image ($\mathcal{X}_s$) as input and outputs the predicted segmentation mask ($\hat{\mathcal{Y}}_s$). A dice loss, $\mathcal{L}_{dice}(\hat{\mathcal{Y}}_s, \mathcal{Y}_s)$, is computed between the predicted and ground truth mask, which is used to train the

segmenter $\mathcal{S}$. Subsequently, the segmenter processes the inverse-deformed target image ($\mathcal{X}_{t \rightarrow s}$), output by the registration network, predicting the segmentation mask ($\hat{\mathcal{Y}}_{t \rightarrow s}$). This approach allows utilisation of the ground truth source labels, resulting in a Dice loss that is backpropagated to the registration network: $\mathcal{L}_{seg} = \mathcal{L}_{dice}(\hat{\mathcal{Y}}_{t \rightarrow s}, \mathcal{Y}_s)$.

3.3.3. Combined Loss Function:

The network is trained using a combined loss, formed using a similarity loss $\mathcal{L}_{sim}$, smoothing loss $\mathcal{L}_{smooth}$, discriminator loss $\mathcal{L}_{disc}$ and segmentation loss $\mathcal{L}_{seg}$, weighted by λ_{sim} , λ_{smooth} , λ_{disc} and λ_{seg} respectively, resulting in the following objective function:

$$\mathcal{L}(\mathcal{R}, \mathcal{S}, \mathcal{D}) = \lambda_{sim}\mathcal{L}_{sim} + \lambda_{smooth}\mathcal{L}_{smooth} + \lambda_{disc}\mathcal{L}_{disc} + \lambda_{seg}\mathcal{L}_{seg} \quad (3)$$

Once $\mathcal{R}$ was trained, we used the transformed images to train a segmentation model $\mathcal{F}(\mathcal{X}_{s \rightarrow t}, \mathcal{Y}_{s \rightarrow t})$, such that the performance for $\mathcal{D}_t$ was maximised.

4. Experimental Setup

4.1. Datasets

Morpho-MNIST Dataset: We created a synthetic content shift scenario using the Morpho-MNIST dataset [11] to develop and validate our approach before applying it to complex brain MRI data. We obtained thick (source) and thin (target) images of digit zero, where the thinning creates a *content* shift that simulates hippocampal atrophy patterns observed in dementia progression. Additionally, to introduce *style* shift mimicking the variability in real brain scans, we added random image patches from CIFAR [32] as background (patch size=5×5) and foreground (patch size=15×15), along with Gaussian noise (mean=0, std=0.05). This approach enables rigorous method validation before transitioning to the more computationally intensive brain MRI applications. Examples of simulated thick and thin image pairs are shown in Figure 3.

Hippocampus Dataset: To demonstrate the content shift that occurs due to hippocampus atrophy between healthy and dementia populations, we focused on three datasets: Young-Healthy (DRYADS [12]), Dementia Research (HarP [13]) and Dementia Clinical (OBHC [14]).

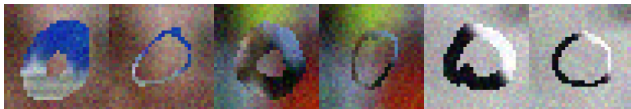


Figure 3. Example of simulated Morpho-MNIST thick and thin pairs.

These datasets present both style and content shifts. Specifically, DRYAD $\rightarrow$ HarP represents a transition from young healthy brains to an older population with mixed cognitive status, including both healthy individuals and those with dementia. DRYAD $\rightarrow$ OBHC represents a more extreme clinical scenario, transitioning directly from young healthy participants to elderly patients in a real-world clinical setting. Our experiments with these domain pairs allow us to evaluate our method under increasingly challenging and clinically relevant conditions. The three datasets are characterised as follows:

- **Young Healthy Dataset (DRYADS):** 25 young, healthy participants scanned using a 3T Siemens scanner with manually annotated hippocampus labels [12].
- **Dementia Research Dataset (HarP):** 130 T1-weighted MRI volumes with manually labelled hippocampi, including cognitively healthy controls, mild cognitive impairment (MCI), and Alzheimer’s disease (AD) patients acquired using various scanners [13, 33].
- **Dementia Clinical Dataset (OBHC):** 29 consented memory clinic patients from the Oxford Brain Health Clinic, representing real-world clinical diversity. Images acquired on a 3T Siemens scanner with hippocampi manually annotated by an experienced researcher [14].

Figure 1 illustrates the substantial anatomical variations across our three datasets and the progressive changes between populations. The left panel presents a scatter plot of age versus manually annotated hippocampal volumes, clearly demonstrating the demographic and volumetric differences between datasets. Notably, there is a pronounced decline in hippocampal volume from the young-healthy cohort (DRYAD) to the dementia populations (HarP and OBHC), reflecting the combined effects of ageing and disease progression. The right panel displays representative brain scans with corresponding hippocampal segmentation masks from each population, visualising the characteristic neuroanatomical changes: enlarged ventricular spaces and reduced hippocampal volumes that accompany advancing age and dementia severity.

4.2. Implementation Details

To speed up training time, the brains were cropped to a size of (96, 128, 80), as visualised in Figure 1, removing as much of the background and outer areas of the brain as possible. The content alignment and downstream segmentation tasks were evaluated using the following evaluation metrics: Dice score (DSC), Surface Dice score (SDSC), Hausdorff distance (HD), and Relative Absolute Volume Difference (RAVD). For the downstream segmentation task, images were split into left and right hemispheres. We used the Adam optimiser with a learning rate of 1×10^{-4} , 3×10^{-4} and 1×10^{-3} for training $\mathcal{R}$, $\mathcal{D}$ and $\mathcal{S}$, respectively. For Equation 3, we performed ablation studies on

Method	Training Data	Morpho-MNIST Thick	Morpho-MNIST Thin
U-Net	Target	0.981 ± 0.001	0.938 ± 0.001
U-Net	Src (MNIST-Thick)	-	0.669 ± 0.037
Image-Space Aug	Src (MNIST-Thick) + Intensity Aug	-	0.606 ± 0.014
Image-Space Aug	Src (MNIST-Thick) + Affine Aug	-	0.528 ± 0.025
Image-Space Aug	Src (MNIST-Thick) + Deformable Aug	-	0.647 ± 0.020
Feature-Space UDA [15]	Src (DRYAD) + Target	-	0.648 ± 0.163

Table 1. DSC for segmentation performance using common augmentation methods in image space, and a representative UDA method for segmentation in feature space. Tested on Morpho-MNIST Thick and Morpho-MNIST Thin.

Method	Training Data	DRYAD	HarP	OBHC
U-Net	Target	0.880 ± 0.012	0.854 ± 0.025	0.811 ± 0.027
U-Net	Src (DRYAD)	-	0.705 ± 0.135	0.550 ± 0.153
Image-Space Aug	Src (DRYAD) + Intensity Aug	-	0.719 ± 0.083	0.552 ± 0.172
Image-Space Aug	Src (DRYAD) + Affine Aug	-	0.717 ± 0.113	0.592 ± 0.150
Image-Space Aug	Src (DRYAD) + Deformable Aug	-	0.710 ± 0.126	0.562 ± 0.217
Feature-Space UDA [15]	Src (DRYAD) + Target	-	0.686 ± 0.123	0.477 ± 0.170

Table 2. DSC for segmentation performance using common augmentation methods in image space, and a representative UDA method for segmentation in feature space. Models tested on DRYAD, HarP and OBHC.

both the Morpho-MNIST and MRI datasets and obtained the best performance with the following loss weights:

Morpho-MNIST: $\lambda_{sim} = 1$, $\lambda_{smooth} = 0.001$, $\lambda_{disc} = 0.0001$, $\lambda_{seg} = 0.01$

MRI: $\lambda_{sim} = 0.1$, $\lambda_{smooth} = 1$, $\lambda_{disc} = 0.0001$, $\lambda_{seg} = 1$
Training was conducted with a batch size of 1 using 3-fold cross-validation. On an Nvidia A10 GPU, content alignment took approximately 40 minutes on Morpho-MNIST and 1.5 hours on MRI, while downstream segmentation required 30 minutes and 1 hour, respectively.

5. Results & Discussion

5.1. Domain Shift

To quantify domain shift between datasets and evaluate existing adaptation methods, we conducted comprehensive experiments summarised in Tables 1 and 2.

We first established baseline performance by training and testing segmentation networks within each domain, which served as the oracle performance. These within-domain models achieved average Dice scores exceeding 0.9 for Morpho-MNIST and above 0.8 across all MRI datasets (Tables 1 and 2, row 1). When evaluating out-of-domain (OoD) performance, we observed substantial degradation in segmentation accuracy across all target domains: Morpho-MNIST thick→thin yielded DSC=0.669 (Table 1, row 2), while DRYAD→HarP achieved DSC=0.705 and DRYAD→OBHC showed an even more pronounced decline to DSC=0.550 (Table 2, row 2). These results confirm that larger domain gaps correlate with greater performance deterioration. We subsequently evaluated common image-space augmentation strategies, specifically intensity and geometric augmentations, to address style and content shift, respectively. For Morpho-MNIST, random augmen-

tations failed to improve upon baseline OoD performance. In contrast, for MRI datasets, both intensity augmentations and affine transformations provided marginal improvements over baseline OoD performance for DRYAD→HarP, with more substantial gains observed for DRYAD→OBHC (Table 2, rows 3-5 vs. row 2).

Since existing image-space UDA approaches are not explicitly designed to handle content shift, we evaluated a representative feature-based UDA method for segmentation [15] as a comparative baseline (Tables 1 and 2, row 6). In the case of Morpho-MNIST, the feature-space UDA performed comparably to random augmentations in image space. Alternatively, for the MRI datasets, this approach yielded inferior Dice scores compared to the OoD baseline, with random augmentations demonstrating superior performance. These findings establish a comprehensive benchmark for evaluating our proposed image-space UDA framework, which specifically targets content shift alignment.

5.2. Content Alignment & Downstream Segmentation

Tables 3, 4 and 5 present quantitative evaluation of our method across content alignment (CA) and downstream segmentation (DS) tasks for three experimental scenarios: Morpho-MNIST thick→thin, DRYAD→HarP and DRYAD→OBHC, respectively. We evaluate the contribution of each proposed component through ablation studies that incrementally incorporate the discriminator loss $\mathcal{L}_{disc}$ and segmenter loss $\mathcal{L}_{seg}$ into the training framework.

Our results demonstrate consistent performance improvements across all datasets when both discriminator and segmenter losses are integrated during training. For Morpho-MNIST (Table 3), our content alignment network improved Dice scores from 0.758 (standard DIR) to 0.818

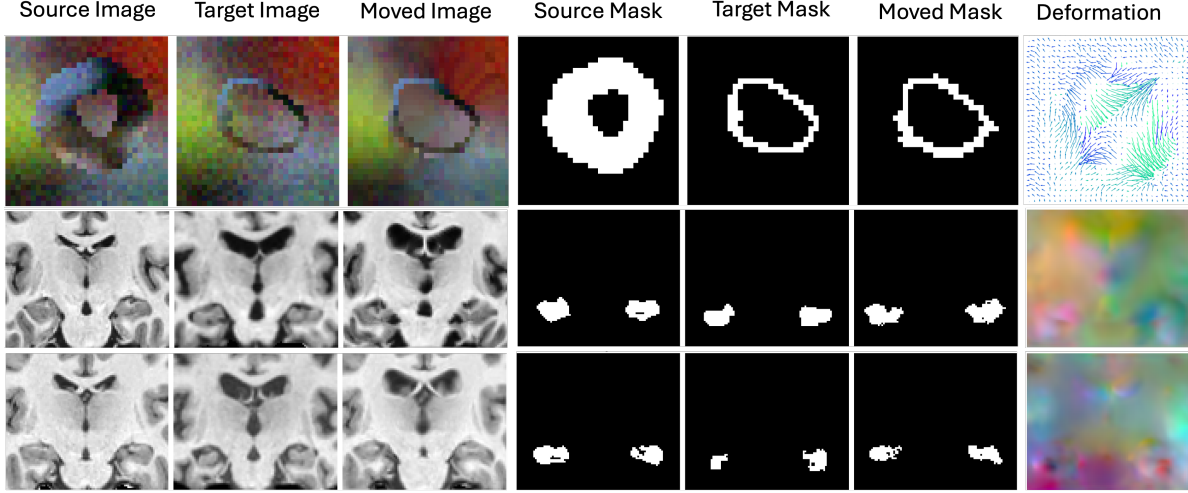


Figure 4. Examples of content alignment for Morpho-MNIST thick→thin (top row), DRYAD→HarP (middle row), DRYAD→OBHC (bottom row).

Task	Model	Avg DSC $\uparrow$	Avg SDSC $\uparrow$	95 % HD $\downarrow$	RAVD $\downarrow$
CA	DIR	0.758 $\pm$ 0.001	0.910 $\pm$ 0.000	2.037 $\pm$ 0.012	58.290 $\pm$ 0.101
CA	+ $\mathcal{L}_{disc}$	0.777 $\pm$ 0.004	0.923 $\pm$ 0.002	1.952 $\pm$ 0.031	44.749 $\pm$ 1.083
CA	+ $\mathcal{L}_{seg}$	0.791 $\pm$ 0.003	0.928 $\pm$ 0.003	1.742 $\pm$ 0.023	49.943 $\pm$ 1.401
CA	+ $\mathcal{L}_{disc}$ + $\mathcal{L}_{seg}$	0.818 $\pm$ 0.002	0.947 $\pm$ 0.003	1.540 $\pm$ 0.029	35.153 $\pm$ 2.018
DS	Trg (in-domain)	0.938 $\pm$ 0.001	0.988 $\pm$ 0.001	0.632 $\pm$ 0.018	5.456 $\pm$ 0.268
DS	Src (out-of-domain)	0.669 $\pm$ 0.037	0.909 $\pm$ 0.007	2.109 $\pm$ 0.137	64.833 $\pm$ 11.277
DS	DIR	0.725 $\pm$ 0.053	0.843 $\pm$ 0.054	3.161 $\pm$ 1.256	100.279 $\pm$ 33.005
DS	+ $\mathcal{L}_{disc}$	0.802 $\pm$ 0.036	0.908 $\pm$ 0.028	1.846 $\pm$ 0.313	59.118 $\pm$ 17.084
DS	+ $\mathcal{L}_{seg}$	0.797 $\pm$ 0.044	0.909 $\pm$ 0.038	2.217 $\pm$ 0.734	58.571 $\pm$ 21.941
DS	+ $\mathcal{L}_{disc}$ + $\mathcal{L}_{seg}$	0.830$\pm$0.026	0.936$\pm$0.018	1.640$\pm$0.285	43.588$\pm$12.746

Table 3. Content alignment (CA) and downstream segmentation (DS) performance for Morpho-MNIST thick $\rightarrow$ thin, tested on thin. Bold: best performance.

(row 1 vs. row 4). Correspondingly, we achieved substantial improvements in downstream segmentation performance, with Dice scores increasing from 0.669 (OoD baseline) to 0.830 using our complete framework (row 6 vs. row 10).

For DRYAD $\rightarrow$ HarP (Table 4), our content alignment network enhanced Dice scores from 0.566 (standard DIR) to 0.619 (row 1 vs. row 4), while downstream segmentation performance improved from 0.705 (without CA) to 0.771 (row 6 vs. row 10), surpassing random augmentation-based approaches (0.717) reported in Table 2.

For MRI, the most substantial improvements were observed in the clinically relevant DRYAD $\rightarrow$ OBHC scenario (Table 5), where our approach enhanced content alignment from 0.514 to 0.575 (row 1 vs. row 4) and downstream segmentation from 0.550 to 0.680 (row 6 vs. row 10). This represents a 24% relative improvement over the baseline without CA and a 15% relative improvement over the best random augmentation-based methods (0.592), demonstrating the effectiveness of our approach in addressing content shift in realistic clinical settings characterised by significant

hippocampal atrophy.

Examples of registration across each dataset are shown in Figure 4, along with their corresponding segmentation masks and deformation fields. The final column displays the learned deformation fields for DRYAD $\rightarrow$ HarP (middle row) and DRYAD $\rightarrow$ OBHC (bottom row), where the high-magnitude deformations (bright regions) are concentrated around the hippocampal structures. This spatial distribution of deformation validates our joint training approach, demonstrating that the segmentation-guided registration network effectively prioritises anatomical regions of interest. The concentration of deformation around the hippocampi reflects the network’s learned ability to focus computational resources on regions with the greatest anatomical variability, as enforced by the segmentation loss $\mathcal{L}_{seg}$ that backpropagates through the registration network during joint optimisation. Figure 5 presents representative segmentation results on target domain samples from HarP (top row) and OBHC (bottom row) datasets. The first column displays the manual ground truth annotations, while

Task	Model	Avg DSC $\uparrow$	Avg SDSC $\uparrow$	95 % HD $\downarrow$	RAVD $\downarrow$
CA	DIR	0.566 $\pm$ 0.013	0.646 $\pm$ 0.013	4.642 $\pm$ 0.140	52.007 $\pm$ 1.009
CA	+ $\mathcal{L}_{disc}$	0.573 $\pm$ 0.014	0.654 $\pm$ 0.014	4.615 $\pm$ 0.140	51.900 $\pm$ 1.007
CA	+ $\mathcal{L}_{seg}$	0.623 $\pm$ 0.013	0.742 $\pm$ 0.012	3.702 $\pm$ 0.157	29.248 $\pm$ 1.862
CA	+ $\mathcal{L}_{disc}$ + $\mathcal{L}_{seg}$	0.619 $\pm$ 0.008	0.745 $\pm$ 0.014	3.598 $\pm$ 0.154	24.188 $\pm$ 3.836
DS	Trg (in-domain)	0.854 $\pm$ 0.025	0.959 $\pm$ 0.020	1.298 $\pm$ 0.425	6.468 $\pm$ 4.635
DS	Src (Out-of-Domain)	0.705 $\pm$ 0.135	0.803 $\pm$ 0.136	8.740 $\pm$ 19.892	19.943 $\pm$ 18.797
DS	DIR	0.740 $\pm$ 0.166	0.842 $\pm$ 0.160	3.492 $\pm$ 6.605	17.194 $\pm$ 20.587
DS	+ $\mathcal{L}_{disc}$	0.758 $\pm$ 0.115	0.864 $\pm$ 0.087	4.489 $\pm$ 12.188	14.591 $\pm$ 14.402
DS	+ $\mathcal{L}_{seg}$	0.764 $\pm$ 0.142	0.875 $\pm$ 0.110	3.446 $\pm$ 8.376	12.863 $\pm$ 16.999
DS	+ $\mathcal{L}_{disc}$ + $\mathcal{L}_{seg}$	0.771$\pm$0.096	0.880$\pm$0.071	2.696$\pm$6.732	10.967$\pm$12.764

Table 4. Content alignment (CA) and downstream segmentation (DS) performance for DRYAD $\rightarrow$ HarP, tested on HarP. Bold: best performance.

Task	Model	Avg DSC $\uparrow$	Avg SDSC $\uparrow$	95 % HD $\downarrow$	RAVD $\downarrow$
CA	DIR	0.514 $\pm$ 0.003	0.566 $\pm$ 0.003	6.408 $\pm$ 0.02	89.627 $\pm$ 1.337
CA	+ $\mathcal{L}_{disc}$	0.510 $\pm$ 0.006	0.562 $\pm$ 0.006	6.395 $\pm$ 0.043	88.353 $\pm$ 0.157
CA	+ $\mathcal{L}_{seg}$	0.615 $\pm$ 0.006	0.727 $\pm$ 0.005	4.822 $\pm$ 0.161	32.546 $\pm$ 3.299
CA	+ $\mathcal{L}_{disc}$ + $\mathcal{L}_{seg}$	0.575 $\pm$ 0.048	0.673 $\pm$ 0.073	5.185 $\pm$ 0.716	48.283 $\pm$ 32.003
DS	Trg (in-domain)	0.811 $\pm$ 0.027	0.917 $\pm$ 0.028	4.462 $\pm$ 0.550	10.714 $\pm$ 7.566
DS	Src (Out-of-Domain)	0.550 $\pm$ 0.153	0.661 $\pm$ 0.133	8.929 $\pm$ 4.211	37.311 $\pm$ 24.647
DS	DIR	0.616 $\pm$ 0.233	0.739 $\pm$ 0.221	7.018$\pm$3.894	27.716 $\pm$ 30.375
DS	+ $\mathcal{L}_{disc}$	0.678 $\pm$ 0.125	0.784 $\pm$ 0.104	7.419 $\pm$ 14.193	17.812 $\pm$ 15.125
DS	+ $\mathcal{L}_{seg}$	0.674 $\pm$ 0.078	0.777 $\pm$ 0.072	9.22 $\pm$ 23.128	16.734 $\pm$ 14.750
DS	+ $\mathcal{L}_{disc}$ + $\mathcal{L}_{seg}$	0.680$\pm$0.095	0.786$\pm$0.082	11.182 $\pm$ 9.219	16.062$\pm$13.607

Table 5. Content alignment (CA) and downstream segmentation (DS) performance for DRYAD $\rightarrow$ OBHC, tested on OBHC. Bold: best performance.

subsequent columns show predictions from segmentation models trained under different data augmentation strategies. Specifically, we compare models trained on: (i) source DRYAD data without augmentation (second column: OoD), (ii) DRYAD with standard affine transformations (third column: Affine), (iii) DRYAD with deformable augmentation (fourth column: Deform), and (iv) DRYAD

enhanced with target-like images generated by our proposed content alignment (CA) network for DRYAD $\rightarrow$ HarP and DRYAD $\rightarrow$ OBHC transfers (fifth column: Ours). The results demonstrate that segmentation models trained on content-aligned data achieve superior performance both quantitatively, as evidenced by higher Dice coefficients, and qualitatively, through improved preservation of hippocampal morphology and boundary delineation.

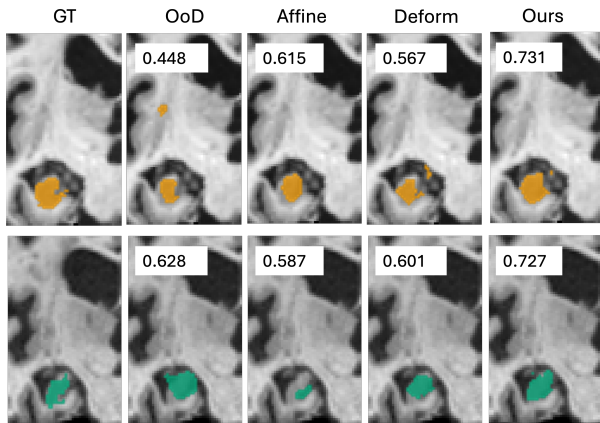


Figure 5. Predicted segmentation masks and corresponding dice scores comparing our method to image-space baselines tested on a target sample from HarP (orange) and OBHC (green). GT: Ground Truth; OoD: Out-of-Domain; Affine: Affine augmentations; Deform: Deformable augmentation.

6. Conclusion

In this paper, we introduced an image-space UDA framework that tackles both *style* and *content* shifts between source and target domains for improved hippocampus segmentation. We harmonised MRI appearance variations through intensity and spatial preprocessing to address style shift. We also developed a novel content alignment strategy utilising bidirectional deformable image registration with region-specific attention and adversarial constraints to generate target-like transformations. Extensive validation across diverse datasets showed marked performance gains, particularly in the challenging clinical scenario, where our method surpassed augmentation-based approaches by 15%. These findings confirm our framework’s efficacy in overcoming domain adaptation challenges, especially in clinical contexts featuring substantial hippocampal atrophy, thereby advancing automated hippocampus segmentation for real-world applications.

Acknowledgements

The authors are grateful for support from: the University of Oxford Department of Computer Science Scholarship (HK), the Bill and Melinda Gates Foundation (NKD, AILN) and the Presidential Postdoctoral Fellowship (Nanyang Technological University) (PHY). We are grateful to the operations team of the OBHC. The OBHC data collection and analysis is supported by the NIHR Oxford Health Biomedical Research Centre (NIHR203316) - a partnership between the University of Oxford and Oxford Health NHS Foundation Trust, the NIHR Oxford Cognitive Health Clinical Research Facility, and the Wellcome Centre for Integrative Neuroimaging (203139/Z/16/Z, 203139/A/16/Z). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

- [1] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, R. C. Mohs, J. C. Morris, M. N. Rossor, P. Scheltens, M. C. Carrillo, B. Thies, S. Weintraub, and C. H. Phelps, "The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," vol. 7, no. 3, pp. 263–269. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3312024/>
- [2] O. Colliot, G. Chételat, M. Chupin, B. Desgranges, B. Magnin, H. Benali, B. Dubois, L. Garnero, F. Eustache, and S. Lehéricy, "Discrimination between alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus," vol. 248, no. 1, pp. 194–201.
- [3] R. A. Morey, C. M. Petty, Y. Xu, J. Pannu Hayes, H. R. Wagner, D. V. Lewis, K. S. LaBar, M. Styner, and G. McCarthy, "A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes," vol. 45, no. 3, pp. 855–866. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1053811908012792>
- [4] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of u-net," vol. 46, no. 12, pp. 10076–10095, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. [Online]. Available: <https://ieeexplore.ieee.org/document/10643318>
- [5] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. MIT Press, google-Books-ID: qJ0jEAAAQBAJ.
- [6] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," vol. 69, no. 3, pp. 1173–1185. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9011180/>
- [7] L. Minkova, A. Habich, J. Peter, C. P. Kaller, S. B. Eickhoff, and S. Kloppel, "Gray matter asymmetries in aging and neurodegeneration: A review and meta-analysis," vol. 38, no. 12, pp. 5890–5904. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23772>
- [8] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Information Processing in Medical Imaging*, ser. Lecture Notes in Computer Science, M. Niethammer, M. Styner, S. Aylward, H. Zhu, I. Oguz, P.-T. Yap, and D. Shen, Eds. Springer International Publishing, pp. 597–609.
- [9] J. Palladino, D. Slezak, and E. Ferrante, "Unsupervised domain adaptation via cyclegan for white matter hyperintensity segmentation in multicenter MR images," vol. 11583.
- [10] S. Kumari and P. Singh, "Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives," vol. 170, p. 107912. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S001048252301377X>
- [11] D. C. Castro, J. Tan, B. Kainz, E. Konukoglu, and B. Glocker, "Morpho-MNIST: Quantitative assessment and diagnostics for representation learning," vol. 20, no. 178, pp. 1–29. [Online]. Available: <http://jmlr.org/papers/v20/19-033.html>
- [12] J. Kulaga-Yoskovitz, B. C. Bernhardt, S.-J. Hong, T. Mansi, K. E. Liang, A. J. W. van der Kouwe, J. Smallwood, A. Bernasconi, and N. Bernasconi, "Multi-contrast submillimetric 3 tesla hippocampal subfield segmentation protocol and dataset," vol. 2, no. 1, p. 150059, publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/sdata201559>
- [13] M. Boccardi, M. Bocchetta, L. G. Apostolova, J. Barnes, G. Bartzokis, G. Corbetta, C. DeCarli, L. deToledo-Morrell, M. Firbank, R. Ganzola, L. Gerritsen, W. Henneman, R. J. Killiany, N. Malychin, P. Pasqualetti, J. C. Pruessner, A. Redolfi, N. Robitaille, H. Soininen, D. Tolomeo, L. Wang, C. Watson, H. Wolf, H. Duvernoy, S. Duchesne,

- C. R. Jack Jr., G. B. Frisoni, and E.-A. W. G. o. t. H. P. f. M. H. Segmentation, "Delphi definition of the EADC-ADNI harmonized protocol for hippocampal segmentation on magnetic resonance," vol. 11, no. 2, pp. 126–138. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2014.02.009>
- [14] M. C. O'Donoghue, J. Blane, G. Gillis, R. Mitchell, K. Lindsay, J. Semple, P. M. Pretorius, L. Griffanti, J. Fossey, V. Raymont, L. Martos, and C. E. Mackay, "Oxford brain health clinic: protocol and research database," vol. 13, no. 8, p. e067808, publisher: British Medical Journal Publishing Group Section: Neurology. [Online]. Available: <https://bmjopen.bmj.com/content/13/8/e067808>
- [15] N. K. Dinsdale, M. Jenkinson, and A. I. L. Namburete, "Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal," vol. 228, p. 117689. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1053811920311745>
- [16] M. Liu, A. H. Zhu, P. Maiti, S. I. Thomopoulos, S. Gadewar, Y. Chai, H. Kim, N. Jahanshad, and f. t. A. D. N. Initiative, "Style transfer generative adversarial networks to harmonize multisite MRI to a single reference image to avoid overcorrection," vol. 44, no. 14, pp. 4875–4892. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.26422>
- [17] H. Kalabizadeh, L. Griffanti, P.-H. Yeung, N. Voets, G. Gillis, C. Mackay, A. I. Namburete, N. K. Dinsdale, and K. Kamnitsas, "Is your style transfer doing anything useful? an investigation into hippocampus segmentation and the role of preprocessing," in *Machine Learning in Clinical Neuroimaging*, D. R. Bathula, A. Benet Nirmala, N. C. Dvornek, S. T. Govindarajan, M. Habes, V. Kumar, A. Nebli, T. Wolfers, and Y. Xiao, Eds. Springer Nature Switzerland, pp. 155–165.
- [18] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation." [Online]. Available: <http://arxiv.org/abs/1804.04732>
- [19] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, and M.-H. Yang, "DRIT++: Diverse image-to-image translation via disentangled representations," vol. 128, no. 10, pp. 2402–2417. [Online]. Available: <https://doi.org/10.1007/s11263-019-01284-z>
- [20] B. E. Dewey, L. Zuo, A. Carass, Y. He, Y. Liu, E. M. Mowry, S. Newsome, J. Oh, P. A. Calabresi, and J. L. Prince, "A disentangled latent space for cross-site MRI harmonization," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, ser. Lecture Notes in Computer Science, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racocanu, and L. Joskowicz, Eds. Springer International Publishing, pp. 720–729.
- [21] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," vol. 12, no. 1, pp. 26–41. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1361841507000606>
- [22] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "An unsupervised learning model for deformable medical image registration," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9252–9260. [Online]. Available: <http://arxiv.org/abs/1802.02604>
- [23] G. Balakrishnan, A. Zhao, M. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," vol. 38, no. 8, pp. 1788–1800. [Online]. Available: <http://arxiv.org/abs/1809.05231>
- [24] B. Kim, D. H. Kim, S. H. Park, J. Kim, J.-G. Lee, and J. C. Ye, "CycleMorph: Cycle consistent unsupervised deformable image registration," vol. 71, p. 102036. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1361841521000827>
- [25] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," vol. 11070, pp. 729–738. [Online]. Available: <http://arxiv.org/abs/1805.04605>
- [26] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8535–8545, ISSN: 2575-7075. [Online]. Available: <https://ieeexplore.ieee.org/document/8953991>
- [27] Z. Shen, Z. Xu, S. Olut, and M. Niethammer, "Anatomical data augmentation via fluid-based image registration," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, ser. Lecture Notes in Computer Science, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racocanu, and L. Joskowicz, Eds. Springer International Publishing, pp. 318–328.
- [28] Y. He, T. Li, G. Yang, Y. Kong, Y. Chen, H. Shu, J.-L. Coatrieux, J.-L. Dillenseger, and S. Li, "Deep complementary joint model for complex scene registration and few-shot segmentation on medical images," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Springer International Publishing, pp. 770–786.

- [29] Z. Xu and M. Niethammer, “DeepAtlas: Joint semi-supervised learning of image registration and segmentation.” [Online]. Available: <http://arxiv.org/abs/1904.08465>
- [30] S. Wang, S. Cao, D. Wei, R. Wang, K. Ma, L. Wang, D. Meng, and Y. Zheng, “LT-net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation.” [Online]. Available: <http://arxiv.org/abs/2003.07072>
- [31] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, “Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces,” vol. 57, pp. 226–236. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1361841519300635>
- [32] A. Krizhevsky, “Learning multiple layers of features from tiny images.” [Online]. Available: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>
- [33] M. Boccardi, M. Bocchetta, F. C. Morency, D. L. Collins, M. Nishikawa, R. Ganzola, M. J. Grothe, D. Wolf, A. Redolfi, M. Pievani, L. Antelmi, A. Fellgiebel, H. Matsuda, S. Teipel, S. Duchesne, C. R. Jack Jr., G. B. Frisoni, and EADC-ADNI Working Group on The Harmonized Protocol for Manual Hippocampal Segmentation and for the Alzheimer’s Disease Neuroimaging Initiative, “Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol,” vol. 11, no. 2, pp. 175–183. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2014.12.002>