Unsupervised harmonization of brain MRI using 3D CycleGANs and its effect on brain age prediction

Dheeraj Komandur  
*Imaging Genetics Center,  
Keck School of Medicine,  
University of Southern California,  
Marina del Rey, CA, USA  
komandur@usc.edu

Umang Gupta  
*Information Sciences Institute,  
University of Southern California,  
Marina del Rey, CA, USA  
umanggup@usc.edu

Tamoghna Chattopadhyay  
*Imaging Genetics Center,  
Keck School of Medicine,  
University of Southern California,  
Marina del Rey, CA, USA  
tchattop@usc.edu

Nikhil J. Dhinagar  
*Imaging Genetics Center,  
Keck School of Medicine,  
University of Southern California,  
Marina del Rey, CA, USA  
dhinagar@usc.edu

Sophia I. Thomopoulos  
*Imaging Genetics Center,  
Keck School of Medicine,  
University of Southern California,  
Marina del Rey, CA, USA  
sthomopo@usc.edu

Jiu-Chiuan Chen  
*Depts. of Population and Public Health Sciences and Neurology,  
University of Southern California,  
Los Angeles, CA, USA  
jcchen@usc.edu

Dan Beavers  
*Dept. of Statistical Sciences,  
Wake Forest University,  
University of Southern California,  
Winston-Salem, NC, USA  
beaversd@wfu.edu

Greg ver Steeg  
*Information Sciences Institute,  
University of Southern California,  
Marina del Rey, CA, USA  
gregv@isi.edu

Paul M. Thompson  
*Imaging Genetics Center,  
Keck School of Medicine,  
University of Southern California,  
Marina del Rey, CA, USA  
pthomp@usc.edu

Abstract—Deep learning methods trained on brain MRI data from one scanner or imaging protocol can fail catastrophically when tested on data from other sites or protocols - a problem known as domain shift. To address this, here we propose a domain adaptation method that trains a 3D CycleGAN (cycle-consistent generative adversarial network) to harmonize brain MRI data from 5 diverse sources (ADNI, WHIMS, OASIS, AIBL, and the UK Biobank - a total of N=4,941 MRIs, age range: 46-96 years). The approach uses 2 generators and 2 discriminators to generate an image harmonized to a specific target dataset given an image from the source domain distribution and vice versa. We train the CycleGAN to jointly optimize an adversarial loss and cyclic consistency. We use a patch-based discriminator and impose identity loss to further regularize model training. To test the benefit of the harmonization, we show that brain age estimation - a common benchmarking task - is more accurate in GAN-harmonized versus raw data. t-SNE maps show the improved distributional overlap of the harmonized data in the latent space.

I. INTRODUCTION

Deep learning methods are now widely applied to brain MRI data for diagnostic classification, disease staging, and prognosis for a range of neurological and neurodevelopmental diseases. Even so, brain MRI protocols vary widely, and models trained on data from one scanner can fail when tested on data from a new site or protocol. To tackle this ‘domain shift' problem, domain adaptation methods have been developed to adjust multisite brain MRI data to match a reference dataset or training data to facilitate machine learning on downstream tasks.

Two broad categories of domain adaptation methods have emerged: (1) adversarial methods that map source data into a site-invariant latent space [1]–[3], where features are optimized for the main task (e.g., disease detection) but also adapted to defeat an adversary that tries to predict which site the data came from; and (2) synthetic methods that also synthesize a new image to appear as if it came from another scanner, often using neural style transfer methods [4], [5]. Such reconstruction methods can also be extended to cross-modal data synthesis (e.g., simulating PET or CT scans from MRI) or for image enhancement with super-resolution [6].

Several GAN-based approaches have shown promising results for domain harmonization. For instance, Liu et al. [4] used style transfer methods to match brain MRI scans to a reference dataset. Sinha et al. [5] used attention-guided GANs for harmonization and demonstrated improvements in Alzheimer’s disease classification with harmonized data. Dinsdale et al. [2] developed a deep learning-based model to remove dataset bias while improving performance on a downstream task of brain age prediction.

Dewey et al. [7] presented DeepHarmony, a UNET-based architecture that requires a paired dataset for training. Many...
works limit the harmonization training to 2D slices rather than the whole 3D volume [2], [5], [7]. Zuo et al. [3] developed CALAMITI, which uses information bottleneck theory to learn a disentangled latent space that contains both anatomical and contrast information, enabling controllable harmonization in a parametrized protocol space. [8]–[11] have employed CycleGAN models for harmonization. [8] showed that harmonization improved lesion segmentation, and [9] improve the harmonization by incorporating knowledge about downstream tasks. [11] applied CycleGAN for PET scan harmonization.

Here we study an unsupervised CycleGAN method for domain adaptation which does not need any paired data across domains. Our model uses 3D convolutions in the generators and discriminators to harmonize full 3D MRI scans. The main goal of harmonization is to improve predictive performance by pooling MRIs from other sources (different scanners, hospitals, etc.) To this end, we evaluate the harmonized scans by training machine learning models on a common benchmark downstream task (brain age estimation). We train downstream models for brain age estimation with harmonized and non-harmonized scans from five datasets. Our results illustrate improved brain age prediction after harmonization, suggesting that harmonization can improve the predictive performance of deep learning models in multi-site predictive modeling.

II. DATASETS

Table 1 summarizes the datasets analyzed in our study. We used 5 datasets: UK Biobank [12], ADNI [13], AIBL [14], OASIS-1 [15], and WHIMS [16]. Apart from WHIMS, we used 1,000 samples from each dataset. ADNI, AIBL, OASIS-1, and WHIMS datasets include both controls and patient scans. The patients are diagnosed with Alzheimer’s disease in ADNI, AIBL, and OASIS-1 datasets, and with Parkinson’s disease in the WHIMS dataset. UK Biobank has only controls. These datasets were used to train 4 CycleGAN models, mapping each of the four datasets (target datasets) ADNI, AIBL, OASIS-1, and WHIMS to (source dataset) UK Biobank and vice-versa.

All 3D T1-weighted MRI brain scans were pre-processed using a sequence of steps detailed in [17]. These included non-parametric intensity normalization (N4 bias field correction), skull stripping, 6 degrees-of-freedom registration to a template, and isotropic voxel resampling to 2 mm. Pre-processed images of size \(91 \times 109 \times 91\) were resized to \(64 \times 64 \times 64\), and intensities were linearly mapped to \([0, 1]\) using min-max normalization.

III. METHODS

Fig. 1 summarizes our CycleGAN architecture. Similar to [18], it consists of two generators \((G_X : X \rightarrow Y\) and \(G_Y : Y \rightarrow X\)) and two discriminators \((D_X\) and \(D_Y\)) for source domain \(X\) and target domain \(Y\). With \(G_X\), we want to generate an image from the target distribution given an image from the source domain distribution and vice versa with \(G_Y\). We train the CycleGAN with an adversarial GAN loss and cyclic consistency loss. We use a patch-based discriminator and impose identity loss to regularize model training.

A. Model Architecture

The generator first transforms the input using a 3D convolution layer with 32 output channels and then passes it to the downsampling block. The downsampling block consists of two 3D convolution layers with 64 and 128 output channels that project the \(64 \times 64 \times 64\) input to \(16 \times 16 \times 16\). This is followed by 5 residual blocks, each having two 3D convolution layers with 128 output channels, and residual connections are made between each block [19]. Next, this output is passed to the upsampling block, which has two 3D transposed convolution layers with 64 and 32-channel outputs. Like UNET, we concatenate the downsampling block’s output with the upsampling blocks’ input. Finally, we concatenate the output of the upsampling block with the input image and pass it through a convolution layer to compute the output. We use instance normalization and ReLU non-linearity for all layers. For all convolution layers, we use padding and stride of 1, and kernel size 3 except for upsampling and downsampling blocks, for which stride was 2. The network’s output has the same dimensions as the input. The discriminator uses a PatchGAN architecture [20]–[22] and has five 3D Convolution layers with 32, 64, 128, 256, and 1 output channel with kernel size 4. All layers except the first are followed by instance normalization and have a stride of 2, and all the layers use ReLU non-linearity except the last, which uses sigmoid activation. The output shape of the discriminator is \(6 \times 6 \times 6\) when the input dimension of the scan is \(64 \times 64 \times 64\).
### B. Loss Function

**a) Adversarial Loss:** We used the standard adversarial training to train the generator to synthesize real images. Instead of using the negative log-likelihood objective, we used least-squares loss, which provided stable training and better results. We use separate discriminators for each domain to compute the adversarial loss. The job of \( D_X \) is to distinguish between samples from the source data distribution, \( P(X) \), and the output of \( G_Y \), and similarly, \( D_Y \) distinguishes between \( P(Y) \) and the output of \( G_X \). The overall adversarial loss is

\[
L_{GAN}(G_X, G_Y, D_X, D_Y) = \mathbb{E}_{y \sim P(Y)}[\|D_Y(y) - 1\|^2 + \|D_X G_Y(y)\|^2] + \mathbb{E}_{x \sim P(X)}[\|D_X(x)\|^2 + \|D_Y G_X(x)\|^2]
\]

**b) Cycle-Consistency Loss:** To reduce the space of possible mapping functions, we adopt the Cycle Consistency loss [18] which ensures that the two generators are cycle-consistent. That is, the scan translated to the target domain \( G_X(x) \) and then translated back to the source domain - i.e., \( G_Y(G_X(x)) \) - must be mapped to the input from domain \( X \). Effectively, we want \( G_Y G_X \) to be identity mapping, i.e., \( x \rightarrow G_X(x) \rightarrow G_Y G_X(x) \approx x \) and similarly for \( G_X G_Y \). We enforce this with pixel-wise \( L_1 \) loss in both forward and backward directions, ensuring cycle consistency:

\[
L_{cycle}(G_X, G_Y) = \mathbb{E}_{x \sim P(X)}[\|G_Y G_X(x) - x\|_1] + \mathbb{E}_{y \sim P(Y)}[\|G_X G_Y(y) - y\|_1]
\]

**c) Identity Loss:** We further regularize the CycleGAN models by enforcing the identity loss similar to Taigman et al. [23], which showed that it helps to restrict the intensity range of the image for the image-to-image translation task. In the case of the grayscale scan, we noticed a slight improvement in the quality of the harmonized scans. The identity loss is computed as the pixel-wise \( L_1 \) loss between the output of the generator (e.g., \( G_X : X \rightarrow Y \)), and the corresponding input (\( y \sim P(Y) \)):

\[
L_{identity}(G_X, G_Y) = \mathbb{E}_{y \sim P(Y)}[\|G_X(y) - y\|_1] + \mathbb{E}_{y \sim P(X)}[\|G_X(y) - x\|_1]
\]

**d) Full Objective:** We introduce additional hyperparameters, \( \lambda_1 \) and \( \lambda_2 \) to balance different loss terms. These are set to 10 and 0.1 during training. The overall objective loss for CycleGAN training is

\[
L(G_X, G_Y, D_X, D_Y) = L_{GAN}(G_X, G_Y, D_X, D_Y) + \lambda_1 L_{cycle}(G_X, G_Y) + \lambda_2 L_{identity}(G_X, G_Y)
\]

The patch-based adversarial loss ensures that the generator synthesizes images from distributions similar to the original distribution. The cyclic consistency and identity loss provide regularization on the generators.

### C. Training

We trained 4 CycleGAN models, with \( X \) representing the UK Biobank source dataset and \( Y \) representing one of the ADNI, AIBL, OASIS-1, and WHIMS target datasets. Each model was trained on pre-processed scans from the source and target datasets. Both generators and discriminators are trained with the ADAM optimizer [24], with a learning rate of \( 10^{-4} \) and batch size of 4. The model was trained for 100 epochs with a multi-step learning rate scheduler with a gamma of 0.1 and steps on 35 and 75 epochs. Overall, our model has 16 million (M) parameters. Each generator and discriminator has 5 M and 3 M parameters, respectively.

### IV. Results

We evaluated our domain adaptation model (1) via t-SNE for visual analysis of the data distributions and (2) by testing if data harmonization reduced error on the downstream task of brain age estimation.

#### A. t-SNE and Clustering

Fig. 2 visualizes the results of the harmonization of the source and target domains. Scans from the target domain were harmonized with the source domain. Due to computational reasons, all the scans were resized to \( 8 \times 8 \times 8 \) before passing to the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm. The t-SNE algorithm yields a two-dimensional embedding for each scan. The first row in Fig. 2 shows the t-SNE embeddings of the unharmonized source and target domain. Without harmonization, source and target data are readily distinguished: in ADNI, all source domain points are clustered near the center, while target domain points are scattered around the source cluster. Scans from the WHIMS dataset are similarly clustered almost entirely within the UK Biobank reference data. The second row shows t-SNE results after harmonizing the target to the source domain; there are now no distinguishable clusters, and source and target distributions overlap. This result is consistent with the goal that the source of the scans is hard to distinguish once they are harmonized.

#### B. Brain Age Prediction

We tested how harmonization affects ‘Brain Age’ estimation, where a 3D CNN is trained on healthy controls’ MRI data to predict a person’s age from their scan. We used a model with 5 3D-Convolutional layers (output sizes: 32, 64, 128, 256, and 256), with a max pool in the first 3 layers. These layers were followed by LeakyReLU non-linearity and

---

**TABLE II**  
STATISTICS OF THE TRAINING DATASET USED FOR CYCLEGAN TRAINING

| Dataset | AIBL | ADNI | OASIS-1 | WHIMS  |
|---------|------|------|---------|--------|
| Baseline (Target only) | 4.557 | 3.437 | 5.109  | 3.856  |
| Harmonizing to source (UKB) e.g. UKB+AIBL -> UKB) | 4.250 | 3.772 | 5.487  | 3.645  |
| Harmonizing to target (e.g. UKB + AIBL -> AIBL) | 4.551 | 3.149 | 4.770  | 3.552  |

In the AIBL column, for example, we take 743 AIBL controls and 743 UKB controls, mix them, and split them into 1040 training and 446 test samples. After training, we report the MAE on the test samples with no harmonization (Baseline). Then we harmonize all scans to ‘look like’ UKB and report the test MAE (to Source) and harmonize all scans to look ‘like AIBL’ (to Target) and report the test MAE.
Despite population differences (Table I), unsupervised harmonization with the CycleGAN improved downstream task performance, as shown in Table II. MAE was improved when MRIs were translated to the target dataset, except for AIBL, where MAE was improved when translating to the source dataset. However, in the reverse case, e.g., translating to target for AIBL, the performance degrades or remains approximately the same. We see similar trends with mean square error (MSE), and it is omitted due to space. Visual insights after training at different numbers of epochs may be found on our GitHub repository (https://github.com/dheeraj-komandur/3D-CycleGAN-based-Harmonization).

Table III reports brain age results for ADNI and OASIS-1 datasets on a held-out set not used for training the CycleGAN. Prediction error changes little between Tables II and III, suggesting that training on harmonized data can generalize to unseen samples.

V. CONCLUSION

We trained four CycleGAN models for deep learning-based harmonization of multicohort MRI, using the UK Biobank as the source domain and ADNI, AIBL, OASIS-1, and WHIMS as targets. Each CycleGAN model was trained with a joint adversarial loss, cycle consistency, and identity loss. We visualized the positive effects of harmonization in overlaying distributions by performing t-SNE on data before and after harmonization; clusters in the data distributions, found initially, were indistinguishable after harmonization. Further, brain age estimation improved in controls across all target domains after harmonization. In the future, we will assess if GAN-based harmonization improves multisite predictive modeling in AD and MRI-based amyloid level prediction. CycleGAN is limited in that it can only be used to harmonize two datasets. In the future, we aim to analyze models such as StarGAN [25] for multisite domain harmonization instead of 1-to-1 source-to-target mapping.
