



RESEARCH ARTICLE

Style transfer generative adversarial networks to harmonize multisite MRI to a single reference image to avoid overcorrection

Mengting Liu^{1,2}  | Alyssa H. Zhu² | Piyush Maiti² | Sophia I. Thomopoulos²  | Shruti Gadewar² | Yaqiong Chai² | Hosung Kim² | Neda Jahanshad² | for the Alzheimer's Disease Neuroimaging Initiative

¹School of Biomedical Engineering, Sun Yat-sen University, Shenzhen, China

²USC Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, University of Southern California, Los Angeles, California, USA

Correspondence

Mengting Liu, School Biomedical Engineering, Sun Yat-sen University, Shenzhen 518107, China.

Email: liumt55@mail.sysu.edu.cn

Hosung Kim and Neda Jahanshad, USC Mark and Mary Stevens, Neuroimaging and Informatics Institute, Keck School of Medicine of the University of Southern California, Los Angeles, CA, USA.

Email: hosung.kim@loni.usc.edu and neda.jahanshad@ini.usc.edu

Funding information

BrightFocus Foundation, Grant/Award Number: A2019052S; Foundation for the National Institutes of Health, Grant/Award Numbers: P41EB015922, R01AG059874, RF1AG057892, U01AG068057; Alzheimer's Disease Neuroimaging Initiative (ADNI); National Institutes of Health, Grant/Award Number: U01 AG024904; DOD ADNI; Department of Defense, Grant/Award Number: W81XWH-12-2-0012

Abstract

Recent work within neuroimaging consortia have aimed to identify reproducible, and often subtle, brain signatures of psychiatric or neurological conditions. To allow for high-powered brain imaging analyses, it is often necessary to pool MR images that were acquired with different protocols across multiple scanners. Current retrospective harmonization techniques have shown promise in removing site-related image variation. However, most statistical approaches may over-correct for technical, scanning-related, variation as they cannot distinguish between confounded image-acquisition based variability and site-related population variability. Such statistical methods often require that datasets contain subjects or patient groups with similar clinical or demographic information to isolate the acquisition-based variability. To overcome this limitation, we consider site-related magnetic resonance (MR) imaging harmonization as a style transfer problem rather than a domain transfer problem. Using a fully unsupervised deep-learning framework based on a generative adversarial network (GAN), we show that MR images can be harmonized by inserting the style information encoded from a single reference image, without knowing their site/scanner labels a priori. We trained our model using data from five large-scale multisite datasets with varied demographics. Results demonstrated that our style-encoding model can harmonize MR images, and match intensity profiles, without relying on traveling subjects. This model also avoids the need to control for clinical, diagnostic, or demographic information. We highlight the effectiveness of our method for clinical research by comparing extracted cortical and subcortical features, brain-age estimates, and case-control effect sizes before and after the harmonization. We showed

The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Human Brain Mapping* published by Wiley Periodicals LLC.

that our harmonization removed the site-related variances, while preserving the anatomical information and clinical meaningful patterns. We further demonstrated that with a diverse training set, our method successfully harmonized MR images collected from unseen scanners and protocols, suggesting a promising tool for ongoing collaborative studies. Source code is released in USC-IGC/style_transfer_harmonization (github.com).

KEYWORDS

harmonization, GAN, MRI, style-transfer

1 | INTRODUCTION

Neuroimaging studies often need to collect data across multiple sites to ensure the sample size and power required to obtain reliable and robust results. Combining multisite magnetic resonance imaging (MRI) data, however, is a nontrivial issue as images are subject to both acquisition- and cohort-based variability. Acquisition-based variability is often due to scanning factors such as manufacturer, magnetic field strength, coil type, positioning and immobilization procedures, number of channels, and parameters such as image resolution and those related to the pulse sequence, for example, TR, TE, and flip angle. Even when imaging parameters are prospectively planned to be as consistent as possible across scanning sites, the need for retrospective harmonization is often inevitable. Scanners in long-running studies such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), a multisite initiative that has been ongoing for nearly two decades (Jack Jr et al., 2008), may undergo scanner drift (Vos et al., 2017), or scanner or software upgrades that alter image contrast or signal-to-noise ratios (SNRs). International neuroimaging consortia have shown the need for effective retrospective neuroimaging harmonization methods. For example, the ENIGMA consortium has pooled information from hundreds of data collection sites around the world for collaborative initiatives aimed at identifying brain signatures of psychiatric or neurological conditions, charting brain trajectories, and even mapping the genetic architecture of brain structure. Collaborative efforts such as ENIGMA continue to expand, incorporating additional data from various sources. However, machine learning studies that rely on such extensive datasets face particular challenges due to the influence of site-specific factors. While initial efforts focused on harmonizing analytical plans toward coordinated meta-analyses, pooling individual-level data allows researchers to pose more targeted questions about factors that may not be similarly distributed across participating sites such as disease staging, or infrequent conditions, such as clinical diagnoses for transdiagnostic studies of history of suicide attempts (Schmaal et al., 2020). These population-based differences are another source of inter-site differences. Even with the data collected from the same scanners/sites, the images collected may also show slight variance (Tong et al., 2020). These discrepancies can result from short-term repositioning influences, long-term physiological changes, and biases in repeated acquisitions due to scanner drift.

Existing retrospective data harmonization techniques have shown promise in removing site-related variance from different studies to allow for such pooled analyses. Most harmonization methods fall into two broad categories: (1) harmonization of image-derived features using statistical properties of the distribution, for example, ComBat (Chen et al., 2022; Pomponio et al., 2020; Zhao et al., 2019); (2) harmonization of the results of specific tasks, such as a regional segmentation, disease classification, or age prediction. That is, to ensure that images from various sources produce consistent and reliable outcomes when used in the same tasks. The main drawback of the first category of harmonization techniques is that it requires many statistical assumptions that may be difficult to satisfy. An extensive review of the first category of harmonization techniques was performed by Bayer et al. (2022). The second category, which seeks to circumvent the statistical assumption pitfalls by avoiding the harmonization of datasets directly but focusing on the MR image output, is largely composed of deep learning-based approaches, namely domain adaptation techniques (Guan et al., 2021; Wang, Chaudhari, & Davatzikos, 2022). Domain transfer learning and domain adversarial learning have been applied for MRI harmonization (Dinsdale et al., 2021).

While task-specific harmonization can be powerful for a particular outcome, if a wide range of tasks were to be performed on images, harmonization would need to be performed separately for each task, resulting in inconsistent, task-dependent harmonization schemes (Wang, Bashyam, et al., 2022). There are also some applications, like cortical surface construction, that cannot be directly embedded into a deep learning framework (Dong et al., 2020). Such cases require image translation, which in the context of artificial intelligence refers to the process of converting an input image from one representation or domain to another, for MRIs. Several image translation-based harmonization methods have been proposed. Supervised methods typically require traveling subjects and must be planned prospectively (Dewey et al., 2019). Unsupervised methods, such as variational autoencoders (Moyer et al., 2020) or CycleGAN (Zhao et al., 2019)—a type of GAN that employs two competing neural networks for generating realistic synthetic data—often separate MR images into well-defined domains in terms of scanners or sites. These methods may be prone to overcorrection, by which we mean correcting for biological factors in addition to, or perhaps instead of, technical scanning-related variables. As each site gets a different label, two different sites with scanning protocols that are similar, and populations that are different, may

inadvertently adjust for biological differences rather than scanner differences.

In addition to the common challenges in unsupervised harmonization, most existing domain-based harmonization algorithms may not generalize to previously unseen sites (Zuo et al., 2021). When there is a new dataset to be harmonized that was not included in the training set, the harmonization usually cannot perform well and retraining to include the new dataset is typically required. Domain-based harmonization approaches, therefore, lack flexibility. They restrict the image harmonization to a very limited number of groups with clear borders and any images beyond the scope of these borders may not adequately harmonize. This could limit the applicability of those harmonization methods in large-scale consortia settings where there are a large number of studies with relatively small sample sizes with diverse acquisition protocols that are iteratively being added to studies.

Recently, deep learning methods have successfully completed diverse image translations by disentangling the image into “content” and “style” spaces (Bashyam et al., 2022; Huang et al., 2018; Liu et al., 2021). Contents represent high-level information in images, often carrying semantic information such as the contours and orientations of objects or structures. On the other hand, styles can be considered low-level information, representing the basic properties of an image, such as colors and textures. Images within the same group (e.g., site or scanner) share the same content space but may show different styles. In MR images, we can consider the biologically defined anatomical information as the content, and the non-biological information such as intensity variation, SNR, and contrasts as styles. Dewey et al. (2020) used this breakdown to show promising results for MRI harmonization when paired image modalities from the same subjects were available to supervise the extraction of the content information; unfortunately, the same two sets of paired images are not always available across multiple datasets. In other work, Jiang and Veeraraghavan (2020) also used a similar framework to facilitate the cross-domain image translations, where each image modality (i.e., CT, T1w, and T2w) was treated as a domain. The problem faced by Jiang and Veeraraghavan (2020) was that styles that span multiple domains must be modeled together using a variational auto-encoder with a universal prior.

Here, we propose to harmonize images using a modified version of the well-established deep learning model, StarGANv2 (Choi et al., 2020). Unlike the original StarGANv2 model, which treats images from various sites as separate “domains,” we consider that every single image belongs to a unique “domain,” and can be disentangled into its own content and style. Image harmonization is considered as a pure style transfer problem rather than a domain transfer problem, which is presumed by conventional domain-based approaches. We consider anatomical patterns (or contents) from MR images collected from different sites to share the same latent space, such that it is not necessary to separate them into different “domains”. These style-irrelevant patterns can be learned using an unsupervised cycle consistency generative adversarial network (GAN) model, and thus, paired modalities or any other paired information are not needed from the same subjects. Due to scanner shifts, and software upgrades, the styles for all the images, even those collected from

the same scanner may be different. We consider every single image as a unique “domain” with its own style, and the styles can be learned flexibly using an adversarial approach, instead of using a universal prior distribution as in Jiang and Veeraraghavan (2020). Furthermore, inspired by Choi et al. (2020), we proposed that the style information needed for harmonization can be encoded from a single reference MR image directly. In short, the entire harmonization process depends solely on a source image and a reference image, which can originate from any subjects or scanners. The source image supplies the content, while the reference image provides the desired style information.

To illustrate the clinical effectiveness of our approach, we train our model on healthy subsets of five publicly available neuroimaging datasets, including: the UK Biobank (UKBB), Parkinson's Progression Markers Initiative (PPMI), ADNI, Adolescent Brain Cognitive Development (ABCD), and International Consortium for Brain Mapping (ICBM). We use automated software, specifically FreeSurfer, to extract several commonly used metrics of interest and compare the features extracted before and after the harmonization. We show that our harmonization method successfully removes the site-related variances, while preserving the anatomical information as demonstrated by retaining case/control effect sizes before and after harmonization. Using subjects from the site that is not involved in training phase, we further illustrate that our model successfully harmonized MR images collected from unseen scanners and protocols, suggesting a promising tool for ongoing collaborative studies.

2 | MATERIALS AND METHODS

2.1 | The architecture of style-encoding GAN

Let X be a set of single slices from full brain MR images. Given an image $x \in X$, our goal is to train a single generator G that can generate diverse images that correspond to the image x with a style code s , where s is associated with the style (non-biological) patterns from another image. The style code s is generated by a mapping network M from sampling a given latent vector z ($s = M(z)$), which is then injected into different layers of the generators to control various levels of detail and style features in the synthesized image. Karras et al. (2021) explain the rationale for using s instead of z . Then, the generator G translates an input image x into an output image $G(x, s)$ that reflects the style of s . To validate that the style code s has been successfully injected into the output image $G(x, s)$, another style encoding network E was designed to encode the style of s from images. That is, given an image x , the encoder E extracts the style code $s = E(x)$ of x . The style code s is a 1×64 vector in our experiment to ensure E can produce diverse style codes using different images. This also allows G to synthesize an output image reflecting the style, s , from different reference images of X . The goal of the network is to train E so that $E(G(x, s)) = s$, meaning that if an image was generated based on style code s , then s can also be extracted when this image was input into the style encoder E . Adaptive instance normalization (Huang & Belongie, 2017) was used to inject s into G .

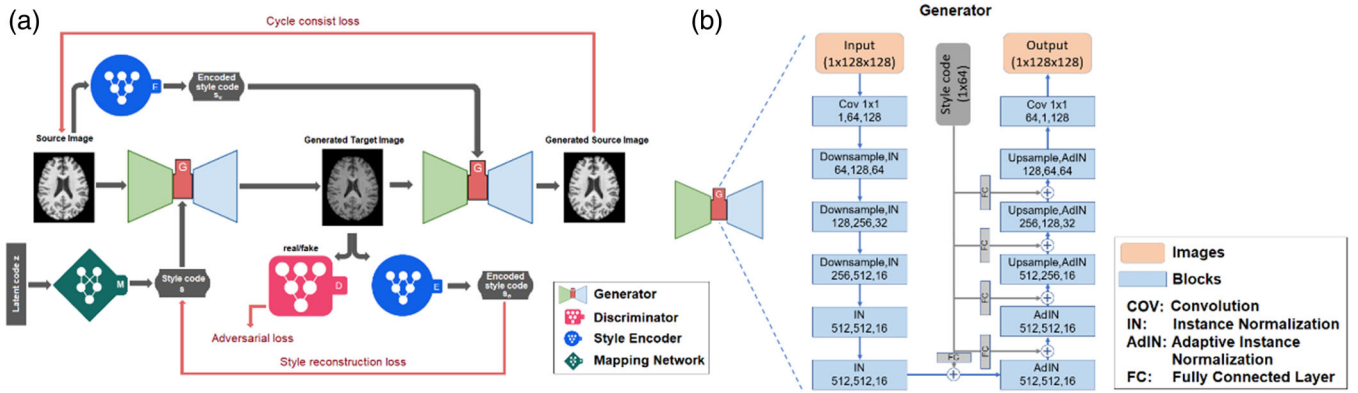


FIGURE 1 (a) The architecture of the style-encoding generative adversarial network (GAN). The generator learns to generate an image by inputting a source image and a style code. The style code s is generated by a mapping network from sampling a given latent vector, or encoded by a style encoder. The quality of the generated image is controlled by the discriminator which learns a binary classification determining whether an image is a real image or a fake image. (b) The detailed architecture of the generator in the network. In each of the blocks in the process, the three numbers represent the number of input channels, number of output channels, and the image size. The style code is injected into different layers of the generators to control various levels of detail and style features in the synthesized image. An instance normalization normalizes a mini-batch of data across each channel for each observation independently.

Finally, the discriminator D learns a binary classification determining whether an image is a real image or a fake image as produced by G , $G(x, s)$. As with Choi et al. (2018), our model includes only one generator, one discriminator, and one style encoder (Figure 1).

2.1.1 | Network training

Given an image $x \in X$, we train our framework using the following objectives: an adversarial loss; a cycle consistency loss; a style reconstruction loss; and a diversification loss, all of which we describe below.

- **Adversarial loss.** During training, we sample a latent code $z \in Z$ randomly, and the mapping network M learns to generate a target style code $s = M(z)$. The generator G takes an image x and style s as inputs and learns to generate an output image $G(x, s)$ that is indistinguishable by the discriminator D from real images via an adversarial loss:

$$L_{GAN} = E_x[\log D(x)] + E_{x,z}[\log(1 - D(G(x, s)))]$$

- **Cycle-consistency loss.** To guarantee that the generated images are consistent with the original images and properly preserving the style-irrelevant characteristics (e.g., anatomical patterns) of input x , an additional cycle consistency loss (Zhu et al., 2017) is defined as the difference between original and reconstructed images:

$$L_{cyc} = E_{x,z}[\|x - G(G(x, s), s_x)\|_1]$$

where $s_x = E(x)$ is the estimated style code of the input image x . By encouraging the generator G to reconstruct the input image x with

the estimated style code s_x , G learns to preserve the original characteristics of x while changing its style faithfully.

- **Style reconstruction loss.** In order to enforce the generator G to use the style code while generating the image $G(x, s)$, we incorporate a style reconstruction loss:

$$L_{sty} = E_{x,z}[\|s - E(G(x, s))\|_1]$$

Our learned encoder E allows G to transform an input image x , to reflect the style of a reference image.

- **Style diversification loss.** To further enable the generator G to produce diverse images, we explicitly regularize G with the diversity sensitive loss (Wang et al., 2018):

$$L_{div} = E_{x,z_1,z_2}[\|G(x, s_1) - G(x, s_2)\|_1]$$

where the target style codes s_1 and s_2 are produced by M conditioned on two random latent codes z_1 and z_2 (i.e., $s_i = M(z_i)$ for $i \in \{1, 2\}$ (Jack Jr et al., 2008)). Maximizing the regularization term forces G to explore the image space and discover meaningful style features to generate diverse images.

Put together, our full objective function can be summarized as follows:

$$L(G, M, E, D) = L_{GAN} + \lambda_{cyc} L_{cyc} + \lambda_{sty} L_{sty} - \lambda_{div} L_{div},$$

where λ_{cyc} , λ_{sty} , and λ_{div} are hyperparameters for each term. If the groups of images to be harmonized are confounded by demographic or clinical/pathological differences, such biological differences could

also be inadvertently learned during the harmonization. To avoid this, we tuned the λ_{cyc} in our model to preserve the style-irrelevant characteristics, including anatomical structure.

We set $\lambda_{cyc} = 10$ to make sure the biological content can be well preserved. We also set $\lambda_{sty} = 10$ and $\lambda_{div} = 1$. To stabilize the training, the weight λ_{div} is linearly decayed to zero over the 200K iterations. We adopted the non-saturating adversarial loss (Goodfellow et al., 2014) with R1 regularization (Mescheder et al., 2018) using $\gamma = 1$. We used the Adam (Kingma & Ba, 2014) optimizer with $\beta_1 = 0$ and $\beta_2 = .99$. The learning rates for G, D, and E are set to 10^{-4} , while that of M is set to 10^{-6} .

2.1.2 | 3D image reconstruction

Due to the GPU memory limitations, our network architecture has been designed for 2D images. This is in contrast to modern MRI scans which are 3D volumes (Dewey et al., 2019). We hence reconstructed the MRI 3D volumes by stacking the 2D slices. The model can be extended to a fully 3D deep network if GPU memory allows. Stacking 2D slices does not work well for slices at the more peripheral edges of the structure of interest, in this case, the brain; these slices contain fewer brain tissue types/contrasts to help ensure the model learns style features properly. To balance the GPU memory limitation and the edge-slice effect, we applied the harmonization on partial-3D-volumes, stacking three consecutive slices together. Using a sliding window, we generated $n_s - 2$ such partial-3D-volumes from each MRI image that contains n_s slices. This also allows for a larger training data pool since each image slab is unique. On the other hand, three natural orientations—axial, sagittal, and coronal—are available for use. To avoid the bias from the three orientations and to provide additional robustness to artifacts that may exist in the data, we pooled all partial-3D-volumes from the three orientations together during the training process. So, an $n_s \times n_s \times n_s$ MRI image can yield $3 \times (n_s - 2)$ training samples. For each of the orientations, we generated a 3D volume by stacking all the partial-3D-volumes together. In this case, if one slice belongs to multiple partial-3D-volumes the output is generated by averaging the specific slice from all partial-3D-volumes. The final MRI volume is the average of the three 3D volumes generated

using all three orientations. Furthermore, during the harmonization, we applied a slice-matched harmonization strategy which relies on the brain registration. That is, after the registration, the reference slice selected for harmonization lies in the same coordinates as the source slice to be harmonized.

2.2 | Model inputs

2.2.1 | Datasets

We obtained T1-weighted brain MR images from five publicly available datasets: UKBB, PPMI, ADNI (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment [MCI], and early Alzheimer's disease [AD]), ABCD, and ICBM. See Acknowledgments section for more information on datasets.

Scans used in this study were collected from subsets of disease-free participants (UKBB: $n = 200$, age range 45–55 years old; PPMI: $n = 76$ age range 67–70 years old; ADNI: $n = 42$, age range 67–70 years old; ABCD: $n = 200$, age range 9–11 years old; and ICBM: $n = 200$, age range 19–54 years old), among which 90% were used as training/validation sets and 10% testing sets. To ensure we can validate the disentangling of style from biological variables, we kept some cohorts overlapping in age, a demographic variable with very large effects. This way, the styles extracted from the cohorts overlapping in age can be compared directly without concern for whether they were determined by major biological factors, specifically age. As in most organic cases, the number of scans per dataset was not kept equal. All image acquisition information for these public resources can be found elsewhere, but briefly they vary in terms of scanner manufacturer, field strength, TR/TE, type of sequence, voxel size, and more, often within the same study. A list of manufacturers and field strength of the five datasets used in our study can be found in Table 1.

TABLE 1 Scanner manufacturer, field strength, and age range specifications for the five dataset subsets used for training the model. The selection of images used here might not be representative of the entire dataset.

Datasets	Scanner manufacturers				Field strength			Number of subjects	Age range (years)
	GE	PHILIPS	SIEMENS	ELSCENT	1.5 T	2 T	3 T		
ADNI3			✓				✓	42	67–70
ICBM	✓	✓		✓	✓	✓	✓	200	19–54
UKBB			✓				✓	200	45–55
PPMI	✓	✓	✓		✓		✓	76	67–70
ABCD	✓	✓	✓				✓	200	9–13

Abbreviations: ABCD, Adolescent Brain Cognitive Development; ADNI, Alzheimer's Disease Neuroimaging Initiative; ICBM, International Consortium for Brain Mapping; PPMI, Parkinson's Progression Markers Initiative; UKBB, UK Biobank.

2.2.2 | Image processing

Many image processing steps are well established, can be implemented quickly, and can help reduce unnecessary variability in site differences that might be attributable to style. Rather than start from raw MRI images, all the images were skull-stripped using HD-BET (Isensee et al., 2019), nonuniformity corrected using N3 approach, and linearly registered to the 1 mm³ MNI152 template using FSL *flirt* (nine degrees of freedom). The images were then resized to 0.8 mm³ isotropic 256 × 256 × 256 voxels to help prepare for 3D processing in all orientations.

2.3 | Experimental datasets for model evaluation

We evaluated our harmonization model on images that were not included in the training set, but were part of the datasets used in training (see details in Table 2). To make sure the images did not have major biological differences, we selected MR images from healthy subjects who were scanned between 55 age and 65 years from three datasets: UKBB ($n = 300$; age = 60.06 ± 2.96 years old); PPMI ($n = 185$; age = 59.68 ± 3.17 years old); and ADNI ($n = 290$; 60.09 ± 2.55 years old). To test whether the harmonization would over-correct the pathological alterations, we further harmonized scans from ADNI participants diagnosed with dementia within the same age range ($n = 350$; age = 59.96 ± 2.74 years old) referenced by a random healthy scan from the UKBB. Within the ADNI dataset, we tested whether dementia versus control differences remained consistent after harmonization.

To further validate our harmonization model, we applied the trained harmonization model to two traveling subjects datasets. The first dataset was a select portion of the ADNI-1 dataset: 44 subjects scanned with both a 1.5 T scanner and a 3 T scanner within 30 days of each other. We note, no scans from the 1.5 T ADNI-1 dataset were used in model training. This dataset was used to test whether our model can harmonize the images from the same subjects scanned at different field strength MRIs to extract the same metrics. The second was the data described in Tong et al. (2020), where three subjects were scanned 12 times at 10 different sites within 13 months. Among the 12 scans, 9 were at unique sites, and the remaining 3 scans were at the same site for every subject. The established acquisition protocols, including the type of scanner, imaging sequences, imaging parameters and positioning of the subject during scanning, are stringently adhered to in all scan sessions.

2.4 | Evaluation metrics

The following evaluations were conducted on the experimental dataset, to compare datasets before and after the harmonization.

2.4.1 | Histogram comparisons

To quantitatively compare the intensity histogram of the harmonized images and reference images, we select all test images from ADNI,

TABLE 2 Datasets used for each of the validation experiments. Testing images denote the images are from the 10% testing dataset listed in Table 1. Extra images denote the images are not included in Table 1.

Validation experiments							
	Histogram comparison	Intra-inter subject comparison	t-SNE	Cortical features comparison	Brain age prediction	Traveling subjects (1.5 T vs. 3 T)	Traveling subjects (12 scans)
ADNI	Testing images N = 4; age range = 67-70						
PPMI	N = 8; age range = 67-70						
UKBB		N = 10; age range = 45-55					
ADNI	Extra images		N = 290; age range = 55-65	N = 290; age range = 55-65	N = 220; age range = 55-75	N = 44; age range = 55-72	
PPMI			N = 185; age range = 55-65	N = 185; age range = 55-65	N = 190; age range = 47-76		
UKBB			N = 300; age range = 55-65	N = 300; age range = 55-65	N = 200; age range = 47-78		
Tong							N = 3; age range = 23-26

Abbreviations: ADNI, Alzheimer's Disease Neuroimaging Initiative; PPMI, Parkinson's Progression Markers Initiative; UKBB, UK Biobank.

and harmonize them with all test images (age and sex matched) from PPMI. Participants were matched for age and sex to ensure approximately similar volumes of each tissue type. We compared the histograms of all paired brain-extracted MR images (one source ADNI scan and one target PPMI scan), and the histogram of the image after the ADNI scan was translated to the PPMI domain, using Jensen-Shannon (JS) divergence, which represents a symmetric, smoothed measure of divergence between two probability distributions, quantifying their similarity or dissimilarity.

2.4.2 | Intrasubject and intersubject image similarity

Ideally, MR image harmonization methods should not only remove the site-related variances, but also rigorously maintain the anatomical information within subjects. To test whether the anatomical information in MR images were preserved after harmonization, we compared the intrasubject similarity and intersubject differences across the harmonizations. We selected 10 random subjects from the UKBB testing set as source images, and 100 randomly selected images from other datasets as reference images. That is, for each subject, we generated 100 harmonized images.

To determine how well the intrasubject similarity was preserved, we compared the similarity between images from the same subject across harmonizations, and the similarity of the images between pairs of subjects harmonized using identical reference images. The similarity was measured using intensity correlation (r), the structural similarity index measure (SSIM), and/or the peak SNR (PSNR) as metrics (Table 3).

To determine whether the intersubject differences were also preserved, we first quantified the intersubject differences by using the intensity differences to compute a Euclidean distance (Zhao et al., 2019) between any two scans, forming a distance matrix, denoted as $D_{ij}^{k \times k} = \|I_i - I_j\|_2$, where $k = 10$ is the number of scans, and I the whole-image voxel intensity vectors for scans i and j . The goal was to estimate how the distances were preserved relative to each other before and after harmonization. We computed the correlation r between the two distance-matrices (only upper triangle) before and after harmonization.

2.4.3 | T-distributed stochastic neighbor embedding of style-codes

To illustrate whether the 1×64 style code was successfully injected into the harmonized images, a t-distributed stochastic neighbor embedding (t-SNE) plot (Van der Maaten & Hinton, 2008) was used to visualize the style representations of images randomly selected from the ADNI, UKBB and PPMI datasets, respectively. Briefly, t-SNE is a nonlinear dimensionality reduction method for visualizing high dimensional data, where more similar data points are closer together, and dissimilar points further apart. The style code was extracted from the style encoder trained in the model before and after the images were harmonized.

2.5 | Task-specific evaluation analyses

2.5.1 | Comparisons of FreeSurfer derived cortical and subcortical measures

Cortical surface reconstruction and subcortical volume segmentation were performed using the freely available FreeSurfer 7.1.0 image analysis software (<http://surfer.nmr.mgh.harvard.edu/>) (Fischl, 2012). The stream encompasses various stages in reconstructing the cortical surface, which include motion correction, intensity normalization, skull stripping, tissue segmentation, surface reconstruction, surface registration, cortical parcellation, and thickness estimation. Volumes in regions of interest (ROIs), thicknesses of cortical ROIs were obtained. FreeSurfer's cortical thickness algorithm calculates the mean distance between vertices of a corrected, triangulated estimated GM/WM surface and GM/CSF (pial) surface (Fischl & Dale, 2000).

Our FreeSurfer features of interest included: lateral ventricle volumes, hippocampal volumes, cerebral gray matter (GM) and cerebral white matter (WM) volume, and cortical GM thickness. These features were used to compare case/control effect sizes before and after harmonization, and used to determine site-related similarities in extracted feature values for subjects scanned across multiple scanners.

TABLE 3 Three metrics used in measuring the similarities between images, their brief explanations and formulas.

Metric	Explanation	Formula
IIC	Measures the linear relationship between pixel intensities in two images.	$IIC = cov(X, Y) / (\sigma(X) * \sigma(Y))$, where cov is the covariance, and σ is the standard deviation.
SSIM	Evaluates the structural, luminance, and contrast differences between two images.	$SSIM(X, Y) = (2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2) / ((\mu_X^2 + \mu_Y^2 + c_1)(\sigma_X^2 + \sigma_Y^2 + c_2))$, where μ is the mean, σ^2 is the variance, σ_{XY} is the covariance, and c_1 and c_2 are constants.
PSNR	Compares the quality of a compressed or reconstructed image to the original, with a higher value indicating better quality.	$PSNR = 20 \log(\max(I)) - 10 \log(MSE)$, where $\max(I)$ is the maximum possible pixel value, and MSE is the mean squared error between the images.

Abbreviations: IIC, image intensity correlation; PSNR, peak signal-to-noise ratio; SSIM, structural similarity index measure.

2.5.2 | Brain age

Brain age is a relatively new concept aimed at providing the age that would be predicted for an individual given their brain MRI scan. In older individuals, a brain age prediction higher than the individual's true age may suggest accelerated aging (Smith et al., 2019). We used a deep-learning based brain age prediction model as in Gupta et al. (2021) to predict the brain age before and after the harmonization. The brain age prediction model applied in this study took a 3D scan as input and encodes each slice using a 2D-convolutional neural network encoder. Next, it combined the slice encodings using an aggregation module, resulting in a single embedding for the scan. Finally, this embedding was passed through the feed-forward layers to predict the brain age. In short, the model involved feeding the entire 3D MR volume into the model to generate an age prediction directly. The model was trained end-to-end using MSE loss. An overview of the network architecture of the brain age prediction model can be found in Gupta et al. (2021).

Here, the brain age prediction model was trained using the brain MR images of 1400 healthy UKBB participants between the ages of 45 and 76 years old. After the model was trained, it was applied to the T1-weighted images of a separate set of healthy subjects from the UKBB ($n = 200$; age range 47.3–77.4 years old), ADNI ($n = 220$; age range 55.6–74.8 years old), and PPMI ($n = 190$, age range: 47.2–75.5 years old) datasets. We then harmonized the ADNI and PPMI images using a reference image randomly selected from the UKBB dataset and applied the brain age prediction model for a pre- and post-harmonization comparison of predicted brain age. We hypothesized that following the harmonization process, the mean absolute error (MAE) for ADNI and PPMI test images will be comparable to UKBB images and will be reduced in comparison to their pre-harmonization values.

2.6 | Traveling subjects evaluation analyses

2.6.1 | Traveling subjects who were scanned on 1.5 T and 3 T scanners

To provide a ground truth, we applied our harmonization model on two traveling subjects cohorts. One is from the ADNI dataset, who were scanned on 1.5 T and 3 T scanners. FreeSurfer metrics were extracted from all the images and compared between 1.5 T images and 3 T images before and after the harmonization. The percent volume differences (delta volume) were calculated by dividing the volume differences values by the average of the two volumes (average of 1.5 T and 3 T), and then compared. To prove the effectiveness of our method, we further harmonized the structural volume using a classic harmonization method for image derived features, namely ComBat (Fortin et al., 2018), and then compared the volume differences between ComBat and our method. Specifically, for each cortical structure, the structural volumes were divided into two groups (1.5 T

and 3 T) referring to the field strength. The ComBat then harmonized the structural volumes by removing the group difference according to their distributions. Sex, age, and whether the participants were healthy/MCI/Dementia were considered as the covariates in ComBat analysis.

2.6.2 | Traveling subjects who were scanned at 10 sites

For the evaluation on the second traveling subjects cohort from Tong et al. (2020), we highlighted how the reference can be to an image from a dataset not in the initial model training. One randomly selected image from the PPMI dataset was chosen as the reference image to harmonize all other MR images for all subjects of the Tong cohort. The PSNR and SSIM were measured between all pairs of images for the same subjects for comparisons between images before and after harmonization.

2.7 | Comparisons with other harmonization networks

We compared our method to two other state-of-the-art unsupervised deep learning harmonization methods referred to as cycleGAN and starGAN. The images of 10 ADNI subjects were harmonized to a single UKBB subject. Input and output images were subtracted to compare preservation of anatomical structure. Furthermore, to quantitatively compare the results, we trained another two models of cycleGAN and starGAN using 40/44 of the ADNI 1.5 T–3 T traveling subjects' cohort and use the left 4/44 for test. The PSNR and SSIM between images before and after harmonization were compared across methods.

3 | RESULTS

3.1 | Hyperparameter tuning

Figure 2 shows an example of an image harmonized using the same model but with different λ_{cyc} values. In this example, the source image is from ADNI and the reference image is from ABCD. There is a 58-year age difference between these subjects, and very evident differences in anatomical structure, including larger lateral ventricles in the older adults. If $\lambda_{cyc} = 0$, meaning none of the style-irrelevant characteristics are needed, the model learns everything from the reference image, generating an image completely identical to the reference. If $\lambda_{cyc} = 1$, then the model learns the style from the reference image but also some biological patterns, such as smaller lateral ventricles and thicker GM cortices. If $\lambda_{cyc} = 10$, then the model learns only the style information from the reference and rigorously maintains the style-irrelevant characteristics (i.e., ventricle and other regional volumes) from the source images.

3.2 | Image-wide evaluations

3.2.1 | Comparisons across cohorts before and after harmonization

Figure 3 illustrates the harmonized images among the five datasets according to nine randomly selected reference images from across

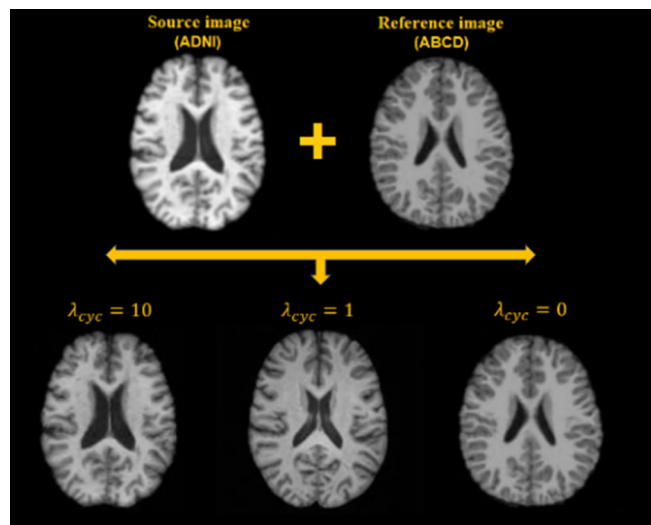


FIGURE 2 Small cycle consistency loss coefficients bias the generated images toward the reference images, while larger cycle consistency coefficients rigorously maintain the structure of the source images, only altering the style.

the datasets. Qualitatively, harmonized images are noticeably more similar in contrast and intensity to those of the reference images, and their anatomical structures were well-maintained. In a quantitative comparison of the age- and sex-matched participants of the ADNI and PPMI datasets, the JS divergence between the histograms of the ADNI and the translated ADNI \rightarrow PPMI image (0.047 ± 0.005) was significantly higher than that of the PPMI and the ADNI \rightarrow PPMI translated image (0.023 ± 0.006 ; $p < .0001$), suggesting the histograms of the ADNI image harmonized to PPMI has an average intensity profile more similar to PPMI than ADNI, from which it came.

Our t-SNE results of the style representations reveal that before harmonization, the style features produced are separable by datasets, especially the PPMI dataset. After the harmonization the style features become jointly embedded and the style feature embedding is not informative of datasets (Figure 4), suggesting the effectiveness of the proposed method.

3.2.2 | Intrasubject similarity

When testing the intrasubject similarities between UKBB source subjects and reference subjects from other datasets, the average correlation between the intensity of the images from the same subjects across harmonizations was 0.991 ± 0.013 and the average SSIM was 0.801 ± 0.068 . The average correlation between the intensity of the images from two different subjects in identical harmonization were $r = .889 \pm .048$ and $SSIM = 0.517 \pm 0.087$. For both sets of

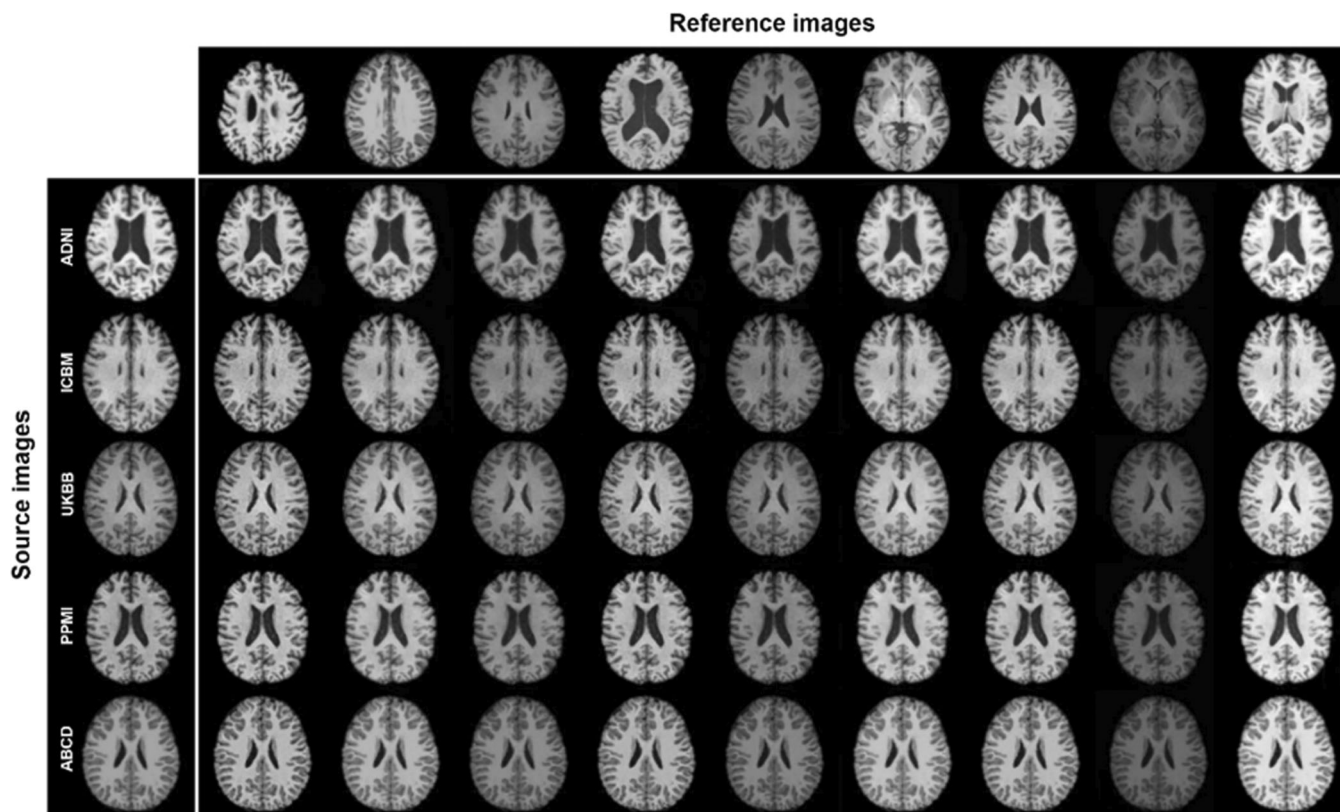


FIGURE 3 The style-encoding generative adversarial network (GAN) can harmonize images based on a single reference image.

values, the intrasubjects similarities after different harmonizations were significantly higher than the intersubject similarities after identical harmonizations, indicating the intrasubject anatomical information was preserved after the harmonization compared to intersubject variances.

3.2.3 | Intersubject differences

When computing the correlation r between the two image-to-image wide Euclidean distance-matrices as described, before and after harmonization, our model achieved an average correlation of $r = .979$ (range: [.954, .994]) between the distance-matrices before harmonization and the 100 distance-matrices after harmonization, indicating the intersubject difference was reliably preserved after harmonization.

3.3 | Task-specific evaluation of downstream analyses on 3D reconstructions

3.3.1 | FreeSurfer cortical thickness and regional volumes

We illustrated the downstream applications of our harmonization method using cortical features extracted from automated processing software, in this case, FreeSurfer. Cortical thickness, surface area for all cortical regions, and volumes of key subcortical brain structures were generated from healthy subjects aged 55 to 65 years across three datasets (UKBB, ADNI, and PPMI). Figure 5 displays a comparison of cortical measurements between any pair of datasets. The first row in paired-cohort comparison indicates Cohen's d scores in cortical regions with significant differences, without adjusting for multiple testing, while the second row represents Cohen's d scores in

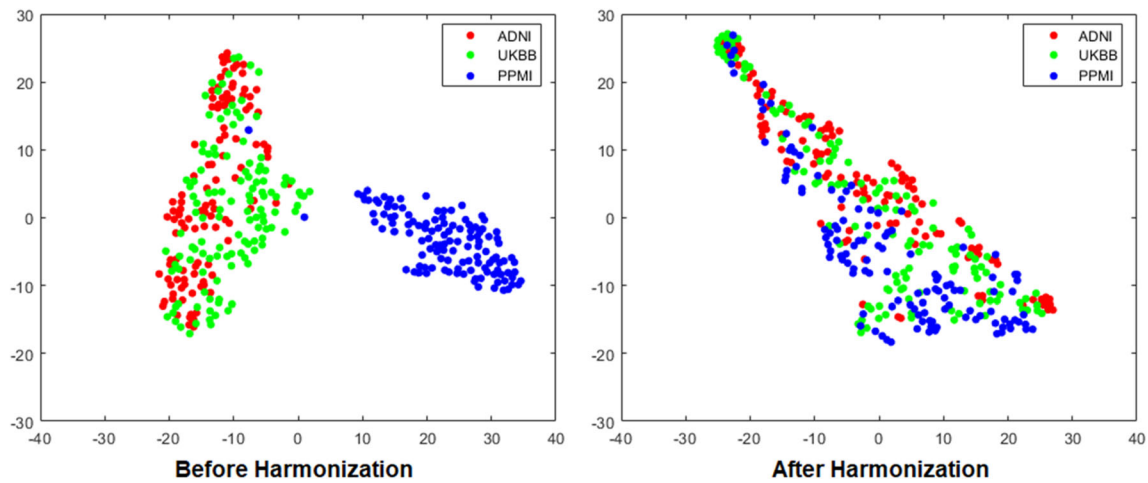


FIGURE 4 T-distributed stochastic neighbor embedding (t-SNE) representation of the style code extracted in images from three datasets (ADNI, UKBB, and PPMI) before and after the harmonization.

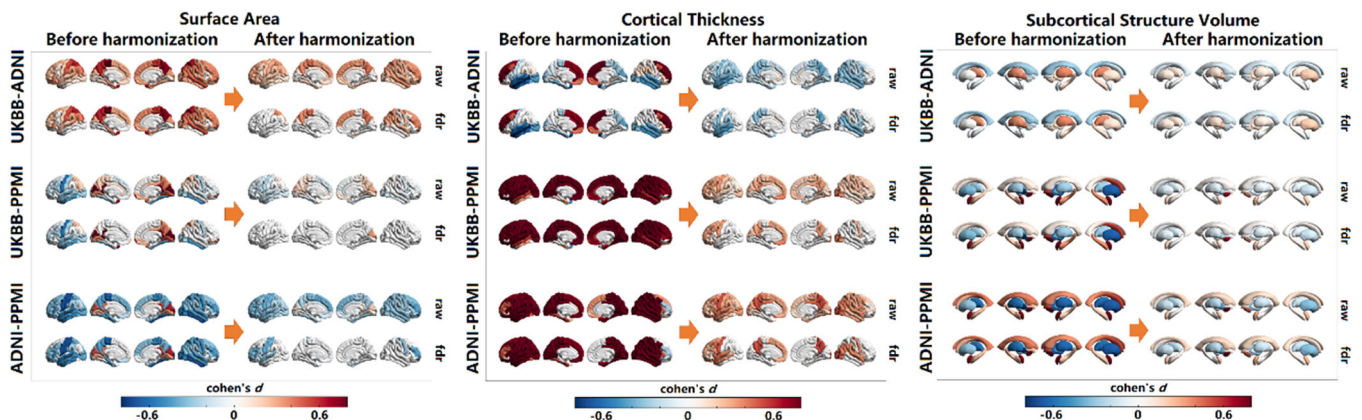


FIGURE 5 Surface are, cortical thickness, and brain structural volumes comparisons among the three datasets before and after the harmonization.

cortical regions after applying multiple testing correction using the false discovery rate approach. It is evident that following harmonization, the differences in cortical measurements between pairs of cohorts are substantially reduced overall, although not entirely eliminated.

3.3.2 | Maintaining case/control effect size differences

To illustrate that the pathological features can be well-preserved after harmonization using our method without overcorrection, we compared the hippocampal volumes between AD patients and healthy subjects in the ADNI dataset, before and after the harmonization to a UKBB reference. Age, sex, and intracranial volume were incorporated as covariates. Hippocampal volume was extracted for left and right hemispheres respectively. Before harmonization, the left and right hippocampal volumes in AD patients (left: $2127.9 \pm 985.8 \text{ mm}^3$, right: $2676.29 \pm 1026.6 \text{ mm}^3$) were significantly smaller than that of healthy subjects (left: $3130.76 \pm 1009.1 \text{ mm}^3$; right: $3835.6 \pm 1023.8 \text{ mm}^3$; left: $p < .0001$, Cohen's $d = -0.98$; right: $p < .0001$, Cohen's $d = -1.14$). As depicted in Figure 9, these differences remained robust after harmonization as well; the hippocampal volumes in AD patients (left: $2298.6 \pm 1010.1 \text{ mm}^3$; right: $2858.6 \pm 923.5 \text{ mm}^3$) compared to controls (left: $3129.9 \pm 1008.3 \text{ mm}^3$, right: $3930.2 \pm 953.0 \text{ mm}^3$) remained significant with effect size differences nearly identical to those before harmonization (left: $p < .0001$, Cohen's $d = -0.83$, right: $p < .0001$, Cohen's $d = -1.13$) (Figure 6).

3.3.3 | Brain age prediction

In the UKBB healthy brain age test set, an MAE of 3.47 years was achieved between the true chronological age and the predicted brain age, and the Pearson correlation coefficient was 0.82. We observed poor generalization ability to other datasets before harmonization (for ADNI: MAE = 4.9 years, Pearson correlation coefficient = 0.45; for PPMI: MAE = 4.8 years, Pearson correlation = 0.79). After harmonization of images from ADNI and PPMI to a reference image from UKBB, we found an improvement in the generalization performance of our predictor (for ADNI: MAE = 3.8 years, Pearson correlation coefficient = 0.58; and for PPMI: MAE = 3.9 years, Pearson correlation = 0.84; Figure 7).

3.4 | Traveling subjects evaluations

3.4.1 | Harmonization of traveling subjects scanned on 1.5 T and 3 T scanners

After harmonization, volume differences between 1.5 T image and 3 T images were smaller than before harmonization for all the brain structures we evaluated. Paired t-tests indicate that after harmonization, volume difference between 1.5 T and 3 T images is significantly smaller than before harmonization for hippocampal volume ($p = .002$). comparisons of the volume differences between ComBat and our method showed that our method outperformed the ComBat for significantly smaller volume difference between 1.5 T and 3 T for hippocampal volume ($p = .004$). Our method also exhibited comparable

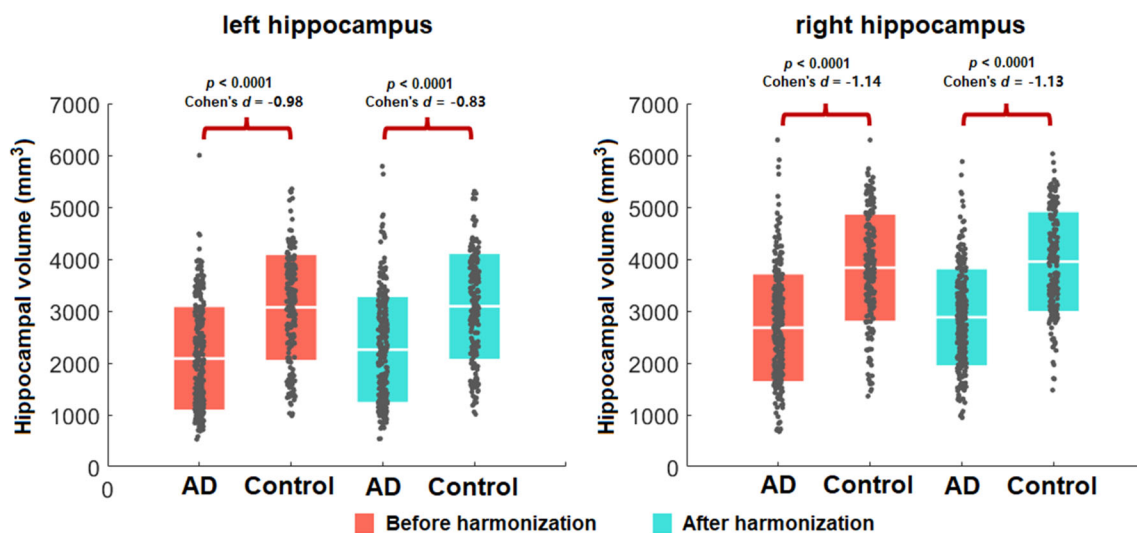


FIGURE 6 Hippocampal volume comparisons among participants with an Alzheimer's disease (AD) diagnosis to cognitively healthy controls within the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, before and after the harmonization. Harmonization does not affect the within-cohort statistical case/control differences.

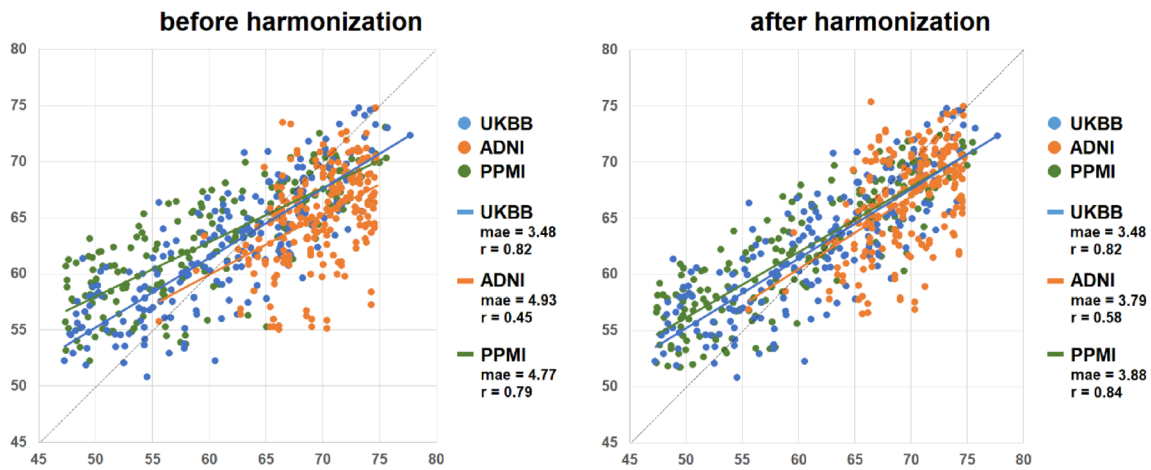


FIGURE 7 Brain age prediction comparisons among the three datasets before and after the harmonization. The brain age prediction model was trained using 1400 healthy scans from the UK Biobank (UKBB) dataset. Before harmonization, the mean absolute errors (MAEs) were much lower when the test set was from the same data collection site as the training data than other datasets, but after harmonization, the errors from the other cohorts were minimized.

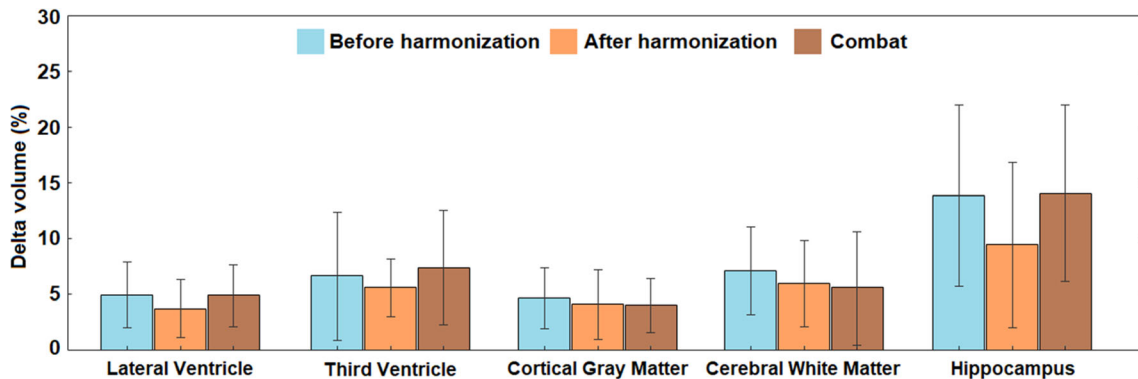


FIGURE 8 Cortical structural volume differences for traveling subjects scanned by 1.5 T scanner and 3 T scanner within 30 days. The comparison was made among the volume differences before the harmonization, after the harmonization using our method, and after harmonization using combat.

effects as ComBat for other brain structures (p 's > .21) in this application (Figure 8).

3.4.2 | Harmonization of traveling subjects from 10 sites

In the traveling subjects' cohort from Tong et al. (2020), we quantitatively highlighted how the model can also harmonize images unseen to the initial model training. One randomly selected image from the PPMI dataset was chosen as the reference image to harmonize all other MR images for all subjects in Tong's cohort. The PSNR and SSIM were measured between all pairs of images for the same subjects. That is, for each subject, we have 45 pairs of images acquired from different sites (using 9 scanners) and 6 pairs of images from the same site (with 3 scans from the same scanner). The average and standard deviation values for each subject are shown in Figure 9, which shows

that the harmonized images, either for site-related scans and same-site scans, are more similar in appearance. Quantitative results for the traveling subjects show a dramatic improvement in similarity using both SSIM (0.954 for original images vs. 0.969 for harmonized images) and PSNR ($M = 26.1$ for original images and $M = 28.2$ for harmonized images), paired t tests, p 's < .01.

To evaluate whether the harmonization was affected by the reference images, we harmonized the Tong's cohort using another five images not included in the training datasets; results showed that after harmonization, the SSIM and PSNR both increased regardless of the reference images (Table 4).

We further evaluated our model by harmonizing images from the unseen dataset (one randomly selected image from the traveling subjects in Tong et al., 2020) to the images in our training datasets. Figure 10 shows, qualitatively, that our model successfully captures the styles of the unseen/traveling subject and renders these styles correctly to the source images.

FIGURE 9 Image similarity comparisons using the structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR) for three traveling subjects, each scanned 12 times at 10 different sites, all within 13 months.

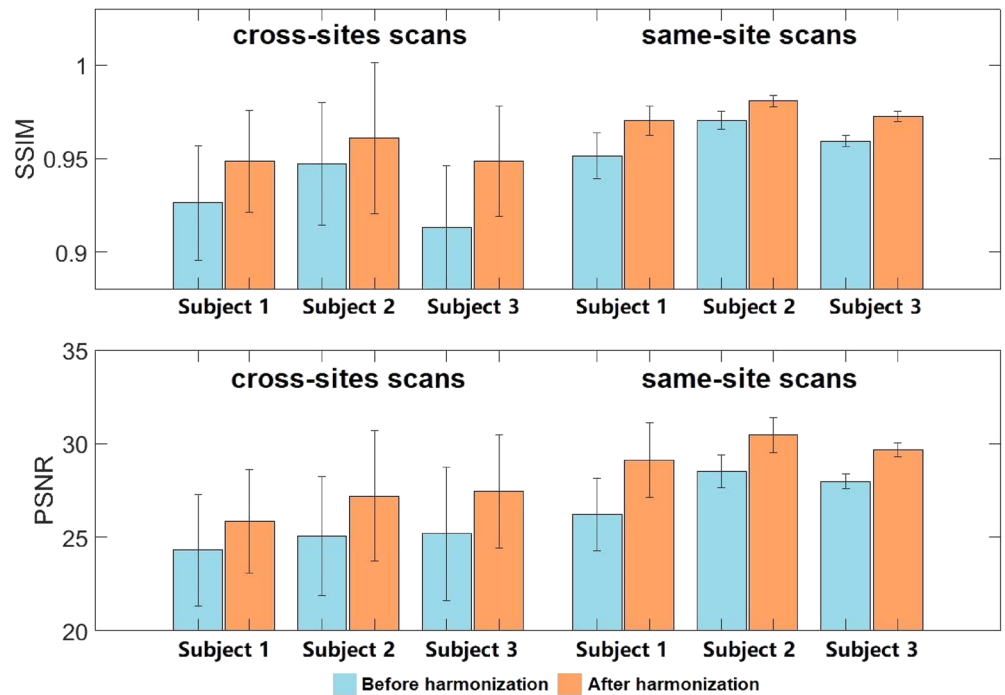


TABLE 4 Quantitative comparison of harmonization using different reference images.

	Site-related scans		Same-site scans	
	SSIM	PSNR	SSIM	PSNR
Before harm	0.928 ± 0.033	24.573 ± 2.713	0.952 ± 0.008	26.409 ± 1.078
Ref 1	0.945 ± 0.025	26.366 ± 2.098	0.969 ± 0.009	28.492 ± 1.211
Ref 2	0.942 ± 0.021	26.781 ± 2.278	0.971 ± 0.012	28.189 ± 1.009
Ref 3	0.951 ± 0.026	26.173 ± 2.109	0.965 ± 0.009	27.709 ± 1.013
Ref 4	0.948 ± 0.027	25.891 ± 1.889	0.968 ± 0.007	28.610 ± 1.468
Ref 5	0.946 ± 0.031	26.257 ± 2.179	0.973 ± 0.008	27.902 ± 1.177

Abbreviations: PSNR, peak signal-to-noise ratio; SSIM, structural similarity index.

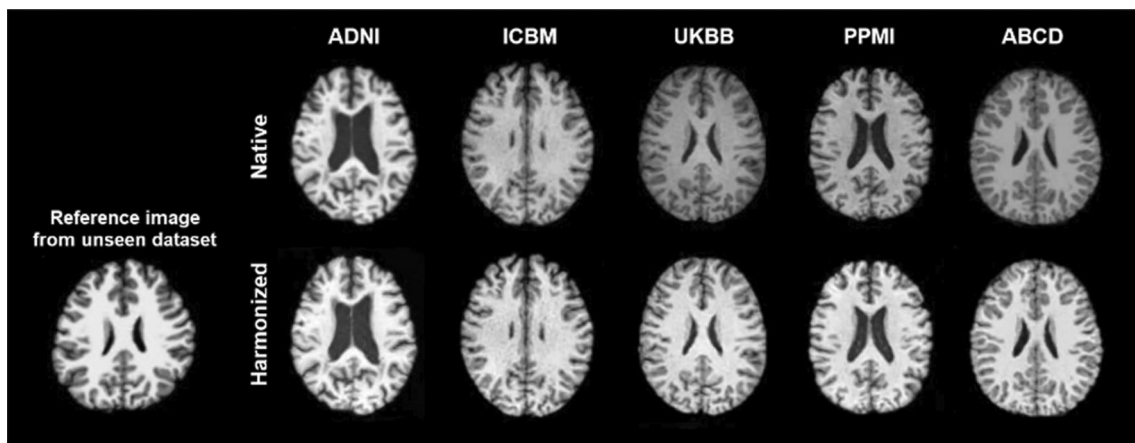


FIGURE 10 The trained style-encoding generative adversarial network (GAN) successfully captures styles of reference images from novel acquisition protocols and renders these styles correctly to the source images.

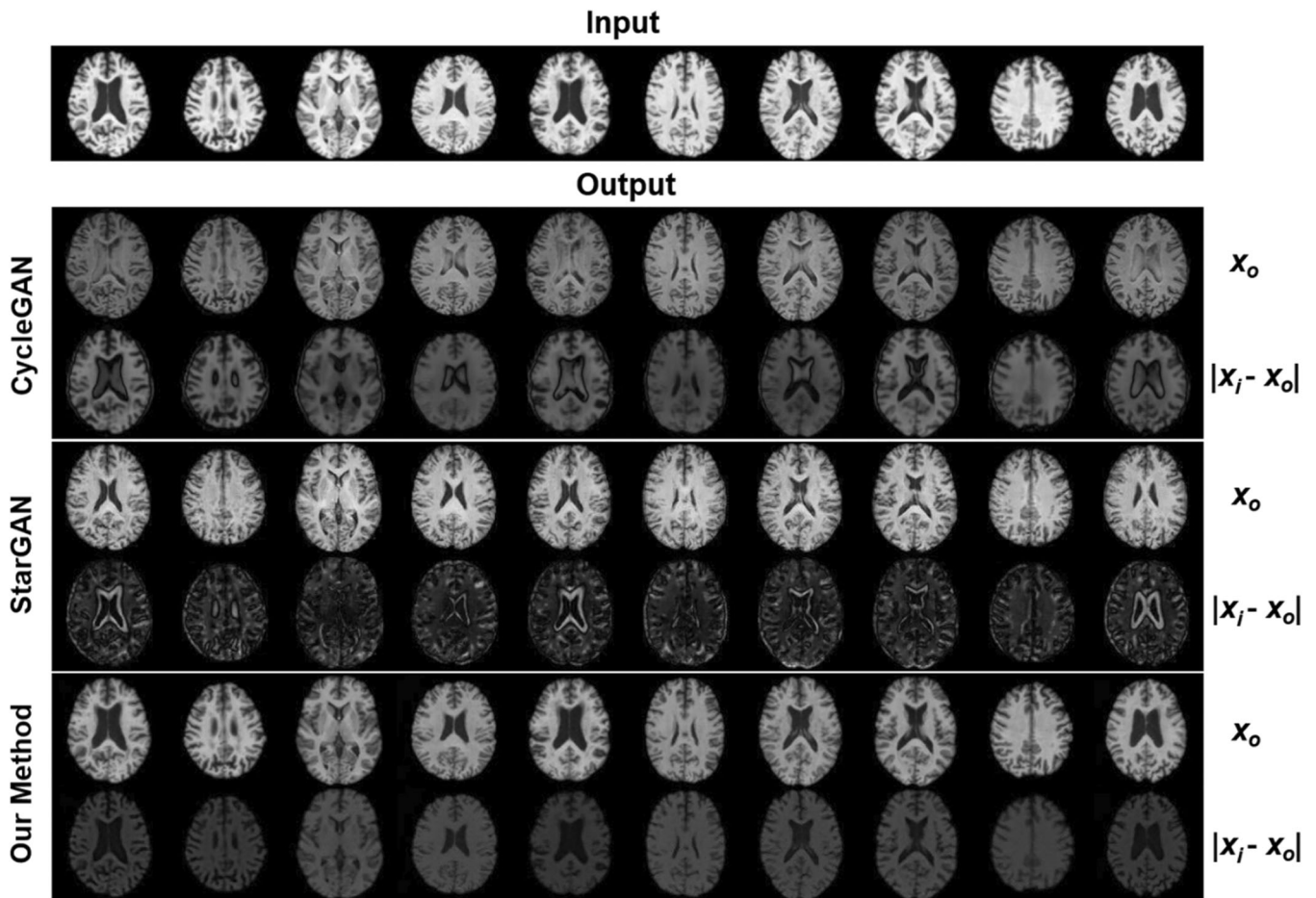


FIGURE 11 Visual comparison across three methods (CycleGAN, StarGAN, our style-encoding generative adversarial network [GAN]) of image harmonization using 10 images from subjects within the ADNI dataset, harmonized to a single subject within the UK Biobank (UKBB) dataset. For each method, the first row is the harmonized images translated from the input, and the second row is the absolute difference between the input and the harmonized images. Anatomical structure is not preserved in CycleGAN or StarGAN as specifically evident by the alterations within the ventricles and other contrast differences around tissue boundaries.

3.5 | Comparisons with other harmonization network structures

A visual inspection of harmonization from ADNI dataset to UKBB dataset can be found in Figure 11. The results revealed that both cycleGAN and starGAN have some deficit in brain ventricles, which is not found in our method. Quantitative comparisons are shown in Table 5, and we performed an inter-site harmonization using traveling subjects not included in training. We showed that our method improved the similarity after harmonization compared with the other two unsupervised harmonization methods.

4 | DISCUSSION

We have developed a harmonization approach for T1-weighted MRIs collected across multiple vendors and acquisition protocols using a style-encoding GAN. In this work, a style-harmonized and anatomically preserved image is synthesized from a single input image from an

arbitrary site, taking the style code encoded from an arbitrary reference image (usually from another site) directly as extra information, without knowing details of the acquisition protocols a priori. Our model does not rely on traveling subjects or any paired modalities of images or any other paired information from the same subjects. Furthermore, because we consider the site-related harmonization as a style transfer problem rather than a domain transfer problem, the MR images from multiple sites do not need to be categorized into specific domains (i.e., acquisition protocols, specific scanners, studies, clinical conditions, age bins, etc.). Thus, the demographic and pathological conditions do not need to be matched for the harmonization.

Unlike other harmonization approaches that work on harmonizing image derived features, and prepare harmonized features for specific tasks, our method works directly to harmonize the full brain MRI, from which already harmonized features can then be extracted. We have tested our method on several tasks including automatic image segmentation software, brain age predictions, and case/control effect size calculations. All of these applications highlighted the effectiveness of our model.

TABLE 5 Quantitative comparison between our method and two other unsupervised harmonization methods in ADNI traveling subjects cohort.

	1.5 T → 3 T			3 T → 1.5 T		
	IIC	SSIM	PSNR	IIC	SSIM	PSNR
Original	0.931 ± 0.027	0.855 ± 0.023	23.172 ± 0.983	0.931 ± 0.027	0.855 ± 0.023	23.172 ± 0.983
cycleGAN	0.954 ± 0.027	0.901 ± 0.031	25.566 ± 1.076	0.957 ± 0.025	0.905 ± 0.019	25.109 ± 1.248
starGAN	0.948 ± 0.021	0.908 ± 0.021	25.331 ± 1.145	0.955 ± 0.023	0.909 ± 0.028	25.482 ± 1.379
Ours	0.966 ± 0.014	0.922 ± 0.018	26.189 ± 0.909	0.971 ± 0.018	0.935 ± 0.024	26.015 ± 1.182

Abbreviations: ADNI, Alzheimer's Disease Neuroimaging Initiative; GAN, generative adversarial network; IIC, image intensity correlation; PSNR, peak signal-to-noise ratio; SSIM, structural similarity index.

Most of the current image-to-image harmonization methods translate images between different domains, such as specific acquisition protocols or studies (Bashyam, Doshi, et al., 2020; Moyer et al., 2020). However, “domain” itself is a complex concept. While it may seem straightforward to group images into different domains, such as the dataset or study where they come from, a study may oftentimes collect images across many sites, scanners, and acquisition protocols, suggesting more complex, nested domains. Even if the images were grouped as domains according to identical collection sites, scanners, or acquisition protocols, they may still exhibit within-domain variabilities, due to scanner drift over time as seen in the traveling subjects' dataset from Tong et al. (2020). In short, the scenarios where domain-based approaches can optimally fit include (a) if the image translation is conducted among limited groups with clear definitions, which is not realistic in many datasets, or (b) when the domain is actually limited to a single specific image, as in our case. In other words, here, we do not separate images into domains based on datasets but consider every single image as a unique “domain” with its own style. While some methods make assumptions about the distribution of styles, for example, that they match a universal prior, such as a Gaussian distribution, that spans all images and can be learned using a variational auto-encoder as in Jiang and Veeraraghavan (2020), our method does not make any such assumptions about the style distributions. We learn style codes adversarially using a GAN-like approach, which does not rely on any hypothesized prior distribution and allows us to learn style codes from every single image individually with greater flexibility and accuracy. An important caveat in biological data harmonization across data collection sites, including brain MRI harmonization, is that the biological information (i.e., brain anatomy) and non-biological information are convoluted, where demographic and clinical characteristics are often also dependent on the cohort, specifically the study inclusion and exclusion criteria. In these common instances, harmonization methods can easily “harmonize” both sets of information, leading to inadequate image harmonization and overcorrection (Dinsdale et al., 2021). Disentangling images into content and style spaces can overcome this issue. Disentangled latent spaces have been used in several past image translation studies (Dewey et al., 2020; Jiang & Veeraraghavan, 2020). These studies both extracted the brain structures as content explicitly, which requires an extra step to supervise the content learning using either

different modality of images from the same subjects (Dewey et al., 2020), or an extra content decoder (Jiang & Veeraraghavan, 2020). To preserve the content, in this case being the anatomical information in the brain MRI, we propose not to generate such an explicit content code. We preserve the content information using a cycle-consistency loss by directly matching the source image and the image translated from the target image based on source style code. In this way, no extra paired information is needed, and we can still avoid the overcorrection of image acquisition confounded by biologically relevant information.

In our study, we proved that our method is not at danger if two sites are harmonized to each other that have no overlap of their age distributions. We incorporated the ABCD dataset, which consists of participants aged 9–13 years old and has no age overlap with any of the other datasets. The overall performance on ABCD images is quite satisfactory. As per the reviewer's request, we conducted an additional analysis to examine the correlation between styles and age, which revealed no significant correlation. This suggests that the styles inserted into the images are not age-related.

An important aspect of harmonization is, therefore, to keep the relevant biological and clinical patterns in the images without this overcorrection. We provide evidence that our model can preserve these patterns using the brain age prediction and hippocampal volume comparison in AD patients compared to age and sex matched controls. Brain age estimation has become an established biomarker of overall brain health in the neuroimaging community, exhibiting overlapping neuroanatomical patterns with a variety of other pathologic processes (Bashyam, Erus, et al., 2020; Jónsson et al., 2019). Accurate brain age estimation depends on fine neuroanatomical patterns that can be obfuscated by site-related imaging variations. Therefore, brain age is an excellent candidate experiment to assess harmonization performance. When training our model on one dataset, we demonstrated improved age prediction estimates in three separate sites, following their mapping to a reference image within the dataset used to train the brain age model. We further demonstrate how case control differences in hippocampal volumes between ADNI participants with dementia compared to cognitive healthy controls were not affected by the harmonization procedure. As the harmonization was to a reference image of a healthy individual, a harmonization procedure that would overcorrect, and confound imaging and biological sources of

variability, would likely remove some of the anatomical variability due to the dementia and the case/control effect sizes would be smaller after harmonization than with the original data. That was not the case with our harmonization approach. In other words, we demonstrated that the degree of impairment captured by hippocampal volumes in the patient population was preserved after the harmonization, indicating the clinical patterns in different datasets were not over-corrected by our harmonization method.

We also show that the similarity between within-site MRI or MRI from traveling subjects may be further increased. This suggests that MRI scans from the same individual, even when taken in the same sequence, can exhibit differences between separate scans. Several factors contribute to these slight variations, including the patient's movement, physiological changes such as alterations in blood flow or tissue composition, and the natural fluctuations within the body. Additionally, MRI scanners might have minor variations in their magnetic field strengths and gradients, which can result in small differences between scans. Furthermore, it is important to consider that the subjects underwent 13 scans over a period of 13 months. As a result, age-related changes could play a significant role and should not be overlooked.

Finally, we showed our model generalizes well even to unseen samples, effectively being able to harmonize to a reference image not included in any of the datasets used for model training. This is a particularly important advantage for studies with very small sample sizes, or those using less common acquisition protocols. For example, if investigators in charge of a relatively small, or unique, cohort wanted to compare their subjects with data from a large, open resource, they may use our method to avoid overcorrection and obtain similar styled-images.

5 | LIMITATIONS

Although the model was designed to separate brain structures (contents) and anatomy-irrelevant information (styles) completely, the model sometimes may not automatically and accurately recognize which are contents and which are styles. In other words, the styles recognized by the model may contain some of the content. This is controlled by the selection of hyperparameters, more experiments are needed to test how the hyperparameters may influence the harmonization. The outcome of the harmonization is dependent on the reference image, and here, we only tested reference images from individuals without gross brain structural abnormalities. It is possible content and style may be conflated in cases where the reference image may have severe artifacts, or anatomical abnormalities such as a large lesion. Our method works on T1-weighted images only, yet a similar framework may also be applied to harmonize images for other modalities, or the multimodal image conversions.

In some cases, T1 sequence variations may cause a shift in the GM and WM boundary. While the harmonization process may help to improve the consistency of the GM/WM boundary by adjusting the intensities of the compartments, it is not intended to correct shifts in

the actual GM/WM boundary. The shifts in the boundary may be corrected by other factors, such as reduction of motion artifacts, physiological changes, or scanner variability. Another potential limitation is the influence of preprocessing steps prior to harmonization on the harmonization outcome. In some situations, harmonization may be necessary for native MRIs. Future studies may need to further validate the effects of harmonization on native MRIs and investigate how the quality of harmonization may be dependent on the preprocessing steps employed. The final limitation of the methodology is that sharing MRI data can raise ethical concerns in many retrospective studies, which may necessitate designing the proposed harmonization method with individual privacy protection in mind.

In conclusion, here, we proposed a harmonization approach for T1-weighted MRIs using a style-encoding GAN that can be used to harmonize entire images for a variety of international, multi-cohort, neuroimaging collaborations.

ACKNOWLEDGMENTS

This work was supported in part by: R01AG059874, RF1AG057892, U01AG068057, and P41EB015922. BrightFocus Research Grant award (A2019052S). This research has been conducted using the UK Biobank Resource under Application Number "11559." Data used in preparation of this article were also obtained from the ADNI database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. Data used in the preparation of this article were also obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data), the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), the Adolescent Brain Cognitive Development (ABCD) Study (<https://abcdstudy.org>), held in the NIMH Data Archive (NDA), and the traveling subjects cohort from Tong et al (<https://www.nature.com/articles/s41597-020-0493-8#Sec7>). For up-to-date information on the PPMI study, visit ppmi-info.org. PPMI—a public-private partnership—is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including (list the full names of all of the PPMI funding partners found at www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson

Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

DATA AVAILABILITY STATEMENT

The data used in these experiments are available on application to the relevant studies. The code used is available when the paper is published and weights from training are available in USC-IGC/style_transfer_harmonization (github.com).

ORCID

Mengting Liu  <https://orcid.org/0000-0003-4972-9006>

Sophia I. Thomopoulos  <https://orcid.org/0000-0002-0046-4070>

REFERENCES

- Bashyam, V. M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Habes, M., Fan, Y., Masters, C. L., Maruff, P., & Zhuo, C. (2020). Medical image harmonization using deep learning based canonical mapping: Toward robust and generalizable learning in imaging. arXiv preprint arXiv:2010.05355.
- Bashyam, V. M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Singh, A., Habes, M., Fan, Y., Masters, C. L., & Maruff, P. (2022). Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors. *Journal of Magnetic Resonance Imaging*, *55*, 908–916.
- Bashyam, V. M., Erus, G., Doshi, J., Habes, M., Nasrallah, I. M., Truelove-Hill, M., Srinivasan, D., Mamourian, L., Pomponio, R., & Fan, Y. (2020). MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain*, *143*, 2312–2324.
- Bayer, J. M. M., Thompson, P., Ching, C. R., Liu, M., Chen, A., Panzenhagen, A. C., Jahanshad, N., Marquand, A., Schmaal, L., & Saemann, P. G. (2022). Site effects how-to & when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses.
- Chen, A. A., Srinivasan, D., Pomponio, R., Fan, Y., Nasrallah, I. M., Resnick, S. M., Beason-Held, L. L., Davatzikos, C., Satterthwaite, T. D., & Bassett, D. S. (2022). Harmonizing functional connectivity reduces scanner effects in community detection. *NeuroImage*, *256*, 119198.
- Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8789–8797.
- Choi, Y., Uh, Y., Yoo, J., & Ha, J.-W. (2020). Stargan v2: Diverse image synthesis for multiple domains. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8188–8197.
- Dewey, B. E., Zhao, C., Reinhold, J. C., Carass, A., Fitzgerald, K. C., Sotirchos, E. S., Saidha, S., Oh, J., Pham, D. L., Calabresi, P. A., van Zijl, P. C. M., & Prince, J. L. (2019). DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic Resonance Imaging*, *64*, 160–170.
- Dewey, B. E., Zuo, L., Carass, A., He, Y., Liu, Y., Mowry, E. M., Newsome, S., Oh, J., Calabresi, P. A., & Prince, J. L. (2020). A disentangled latent space for site-related MRI harmonization. In *International conference on medical image computing and computer-assisted intervention* (pp. 720–729). Springer.
- Dinsdale, N. K., Jenkinson, M., & Namburete, A. I. L. (2021). Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage*, *228*, 117689.
- Dong, J., Cong, Y., Sun, G., Zhong, B., & Xu, X. (2020). What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4023–4032.
- Fischl, B. (2012). FreeSurfer. *Neuroimage*, *62*, 774–781.
- Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 11050–11055.
- Fortin, J. P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., McInnis, M., Phillips, M. L., Trivedi, M. H., Weissman, M. M., & Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, *167*, 104–120.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *63*(11), 139–144.
- Guan, H., Liu, Y., Yang, E., Yap, P. T., Shen, D., & Liu, M. (2021). Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. *Medical Image Analysis*, *71*, 102076.
- Gupta, U., Lam, P. K., Ver Steeg, G., & Thompson, P. M. (2021). Improved brain age estimation with slice-based set networks. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 840–844.
- Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510.
- Huang, X., Liu, M.-Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172–189.
- Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H. P., Heiland, S., & Wick, W. (2019). Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, *40*, 4952–4964.
- Jack, C. R., Jr., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., Whitwell, J. L., & Ward, C. (2008). The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*, 685–691.
- Jiang, J., & Veeraraghavan, H. (2020). Unified cross-modality feature disentangler for unsupervised multi-domain MRI abdomen organs segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 347–358). Springer.
- Jónsson, B. A., Bjornsdottir, G., Thorgeirsson, T., Ellingsen, L. M., Walters, G. B., Gudbjartsson, D., Stefansson, H., Stefansson, K., & Ulfarsson, M. (2019). Brain age prediction using deep learning uncovers associated sequence variants. *Nature Communications*, *10*, 1–10.
- Karras, T., Laine, S., & Aila, T. (2021). A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*, 4217–4228.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- Liu, M., Maiti, P., Thomopoulos, S., Zhu, A., Chai, Y., Kim, H., & Jahanshad, N. (2021). Style transfer using generative adversarial

- networks for multi-site mri harmonization. In *International conference on medical image computing and computer-assisted intervention* (pp. 313–322). Springer.
- Mescheder, L., Geiger, A., & Nowozin, S. (2018). Which training methods for GANs do actually converge? *International Conference on Machine Learning*. PMLR, pp. 3481–3490.
- Moyer, D., Ver Steeg, G., Tax, C. M. W., & Thompson, P. M. (2020). Scanner invariant representations for diffusion MRI harmonization. *Magnetic Resonance in Medicine*, *84*, 2174–2189.
- Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I. M., Satterthwaite, T. D., & Fan, Y. (2020). Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, *208*, 116450.
- Schmaal, L., van Harmelen, A. L., Chatzi, V., Lippard, E. T. C., Toenders, Y. J., Averill, L. A., Mazure, C. M., & Blumberg, H. P. (2020). Imaging suicidal thoughts and behaviors: A comprehensive review of 2 decades of neuroimaging studies. *Molecular Psychiatry*, *25*, 408–427.
- Smith, S. M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T. E., & Miller, K. L. (2019). Estimation of brain age delta from brain imaging. *NeuroImage*, *200*, 528–539.
- Tong, Q., He, H., Gong, T., Li, C., Liang, P., Qian, T., Sun, Y., Ding, Q., Li, K., & Zhong, J. (2020). Multicenter dataset of multi-shell diffusion MRI in healthy traveling adults with identical settings. *Scientific Data*, *7*, 157.
- Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.
- Vos, S. B., Tax, C. M., Luijten, P. R., Ourselin, S., Leemans, A., & Froeling, M. (2017). The importance of correcting for signal drift in diffusion MRI. *Magnetic Resonance in Medicine*, *77*, 285–299.
- Wang, R., Bashyam, V., Yang, Z., Yu, F., Tassopoulou, V., Sreepada, L. P., Chintapalli, S. S., Sahoo, D., Skampardon, I., & Nikita, K. (2022). Applications of generative adversarial networks in neuroimaging and clinical neuroscience. arXiv preprint arXiv:2206.07081.
- Wang, R., Chaudhari, P., & Davatzikos, C. (2022). Embracing the disharmony in medical imaging: A simple and effective framework for domain adaptation. *Medical Image Analysis*, *76*, 102309.
- Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., & Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. *Proceedings of the European Conference on Computer Vision (ECCV) workshops*.
- Zhao, F., Wu, Z., Wang, L., Lin, W., Xia, S., Shen, D., Li, G., & UNC/UMN Baby Connectome Project Consortium (2019). Harmonization of infant cortical thickness using surface-to-surface cycle-consistent adversarial networks. In *International conference on medical image computing and computer-assisted intervention* (pp. 475–483). Springer.
- Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232.
- Zuo, L., Dewey, B. E., Liu, Y., He, Y., Newsome, S. D., Mowry, E. M., Resnick, S. M., Prince, J. L., & Carass, A. (2021). Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. *NeuroImage*, *243*, 118569.

How to cite this article: Liu, M., Zhu, A. H., Maiti, P., Thomopoulos, S. I., Gadewar, S., Chai, Y., Kim, H., Jahanshad, N., & for the Alzheimer's Disease Neuroimaging Initiative (2023). Style transfer generative adversarial networks to harmonize multisite MRI to a single reference image to avoid overcorrection. *Human Brain Mapping*, *44*(14), 4875–4892. <https://doi.org/10.1002/hbm.26422>