MISPEL: A deep learning approach for harmonizing multi-scanner matched neuroimaging data

Mahbaneh Eshaghzadeh Torbati[1], Davneet S. Minhas[2], Charles M. Laymon[2], Pauline Maillard[3], James D. Wilson[4], Chang-Le Chen[5], Ciprian M. Crainiceanu[6], Charles S. DeCarli[3], Seong Jae Hwang[7], and Dana L. Tudorascu[4,8]

[1]Intelligent System Program, University of Pittsburgh School of Computing and Information, Pittsburgh, PA 15213, USA

[2]Department of Radiology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA

[3]Department of Neurology, University of California Davis, Davis, CA 95816, USA

[4]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA

[5]Department of Bioengineering, University of Pittsburgh, Pittsburgh,

PA 15213, USA

[6]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

[7]Department of Artificial Intelligence, Yonsei University, Seoul, South Korea

[8]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15213, USA

Abstract

Large-scale data obtained from aggregation of already collected multi-site neuroimaging datasets has been brought benefits such as higher statistical power, reliability, and robustness to the studies. Despite these promises from growth in sample size, substantial technical variability stemming from differences in scanner specifications exists in the aggregated data and could inadvertently bias any downstream analyses on it. Such a challenge calls for data normalization and/or harmonization frameworks, in addition to a comprehensive criteria to estimate the scanner-related variability and evaluate the harmonization frameworks. In this study, we propose MISPEL (Multi-scanner Image harmonization via Structure Preserving Embedding Learning), a supervised multi-scanner harmonization method. Unlike existing techniques, MISPEL does not assume a perfect coregistration across the scans, and the framework is naturally extendable to more than two scanners. We also designed a set of comprehensive criteria to investigate the scanner-related technical variability and evaluate the harmonization techniques. As an essential requirement of our criteria, we introduced a multi-scanner matched dataset of four 3T MRI T1 images, which, to the best of our knowledge is the first dataset of its kind. We also investigated our evaluations using two popular segmentation frameworks: FSL and segmentation in statistical parametric mapping (SPM). Lastly, we compared MISPEL to popular methods of normalization and harmonization, namely White Stripe, RAVEL, and CALAMITI. MISPEL outperformed these methods and is promising for many other neuroimaging modalities.

Keywords: MRI, Technical variability, Scanner effects, Normalization, Harmonization

1 Introduction

There is a growing interest in the neuroimaging community to combine imaging data from a variety of diverse datasets so as to enable large-scale multi-study analyses that have high statistical power, reliability, and robustness (Madan, 2021; Mar et al., 2013; Madan, 2017; Milham et al., 2018). Despite the promise of massive data aggregation initiatives, large-scale neuroimaging analyses from such data collections often suffer from issues of technical variability due to scanner- and individual-based heterogeneity across studies, which may introduce bias in imaging-derived measures (Kruggel et al., 2010; Potvin et al., 2019; Torbati et al., 2021a) and alterations of the biological signals of clinical interest (Shinohara et al., 2014a, 2017), among other unwanted and unexpected artifacts.

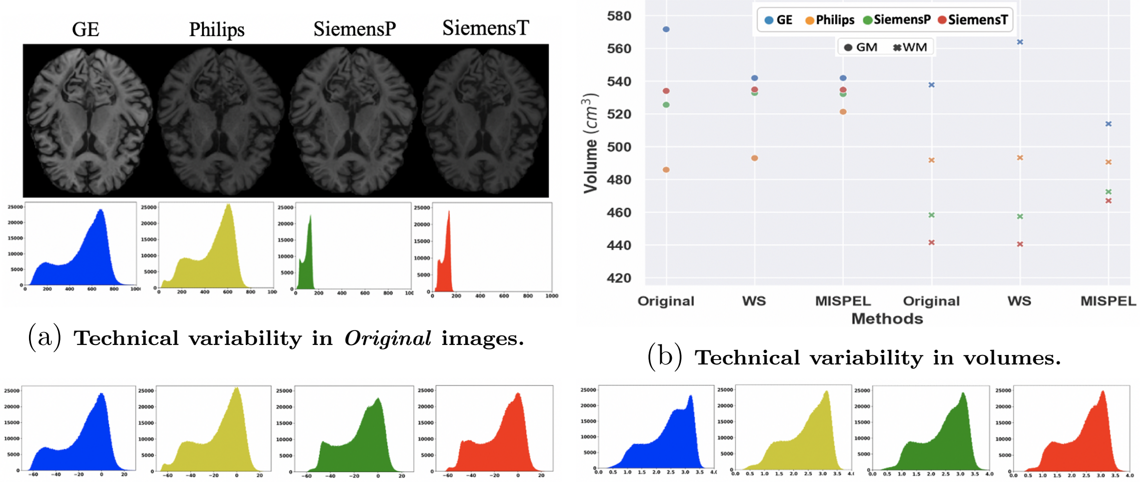
Intensity unit effects are due to the arbitrary nature of image intensity scale, which can cause variability in interpretations of intensity units and thus make the direct quantitative analysis of image intensities difficult (Wrobel et al., 2020). Intensity unit effects have been long recognized and addressed by intensity normalization methods, such as White Stripe (WS) (Shinohara et al., 2014b), a well-known normalization method in neuroimaging. A comprehensive review of the initial intensity normalization methods can be also found in (Shah et al., 2011).

Scanner effects refer to any post-normalization inter or intra-scan variation that is not biological in nature (Fortin et al., 2016) and stems from scanner and acquisition differences (Dinsdale et al., 2021). So far, these differences have been recognized across scanner manufacturer (Takao et al., 2014), scanner upgrade (Han et al., 2006), scanner drift (Takao et al., 2011), scanner strength (Han et al., 2006), and gradient non-linearities (Jovicich et al., 2006). An example of such effects can be seen in tissue type volumes extracted from White Stripe (WS)-normalized images in Figure 1b. The group of methods that aim to remove scanner effects is referred to as harmonization. Harmonization is a complex and challenging task due to (1) lack of thorough understanding of scanner effects, and (2) lack of standardized criteria for assessment of scanner effects and evaluation of harmonization.

In this specific study, our main interest lies in understanding technical variability of images, specifically the scanner effects. Scanner effects cannot be easily removed by simple intensity distribution matching (Fortin et al., 2016) or a linear transformation of images (Wrobel et al., 2020). Even though there has been a noticeable growth in the number of studies focused on scanner effects and harmonization recently (Dewey et al., 2019, 2020; Liu et al., 2021; Cackowski et al., 2021; Zuo et al., 2021), there is a lack of insight with respect to how these scanner effects appear on images. One main reason could be the lack of ground truth for these studies, which leaves them with no standard evaluation criteria and consequently makes their observations partly incoherent and hard to compare. Based on the observations confirmed by several of these studies, it is now known that scanner effects can vary across the voxels of an individual image (Chen et al., 2020a). Furthermore, it is also known that scanner effects change the tissue contrast and consequently affect the results of tissue segmentations (Meyer et al., 2019). Torbati et al. (2021a) has shown that scanner effects can affect different regions of brain differently and result in regional summary measures with varying degree of scanner effects .

The best experimental design setup to understand and quantify scanner effects is to conduct a paired study by having subjects travel to different sites/scanners, to collect the *paired* dataset (Dewey et al., 2019; Zuo et al., 2021). A paired dataset is a set of *paired* images that are the images of each individual scanned on *two* scanners with short time gap. Paired images are expected to be images of biologically similar brain with differences solely due to scanner effects. Using a paired dataset, scanner effects and harmonization can be estimated as similarities and dissimilarities within paired images, respectively. As such, a ground truth is not necessary.

Figure 1 illustrates an example from a *matched* dataset, a paired dataset with more than two scanners. An example of technical variability across MRI scanners can be observed as dissimilar contrast and voxel intensity histograms of these matched images in Figure 1a, as well as their mismatched volumes for both gray matter (GM) and white matter (WM) tissue types in Figure 1b. Moreover, Figure 1c depicts the histograms of the WS-normalized version of the matched images. The scanner effects can then be observed in the WS-normalized images as their dissimilar histograms in Figure 1c, as well as their mismatched volumes in Figure 1b.



(a) **Technical variability in *Original* images.**

(b) **Technical variability in volumes.**

(c) **Histograms of WS-normalized images.**

(d) **Histograms of MISPEL-harmonized images.**

Figure 1: **Example of technical variability, White Stripe normalization, and MISPEL harmonization in *matched* images.** For this example, *Original* images are referred to as matched T1 MRIs scanned on four different 3T scanners: GE, Philips, SiemensP, and SiemensT. Specifications of these scanners can be found in Table 1. Figure (a) depicts the technical variability of the Original images as dissimilarity in contrast of their axial slices, and discrepancy among histograms of their whole brain. Figure (b) shows the technical variability of matched images in terms of their tissue type volumetric dissimilarity. The volumes were depicted for the Original images as well as their WS-normalized and MISPEL-harmonized versions. Figures (c) and (d) depict the histograms of whole brain in WS-normalized and MISPEL-harmonized matched images, respectively. Histograms of matched images have identical axes and correspond (from left to right) to GE, Philips, SiemensP, and SiemensT scanners. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

From a methodological and more specifically, a machine learning perspective, paired and unpaired datasets are considered respectively as labeled and unlabeled data for the task of harmonization. Accordingly, the harmonization methods developed based on paired and unpaired data are called the supervised and unsupervised methods (Dewey et al., 2019; Zuo et al., 2021; Torbati et al., 2021a; Liu et al., 2021). The majority of research on harmonization is currently focused on the unsupervised methods, due to scarcity of matched or even paired datasets. However, there exist two supervised methods, namely DeepHarmony (Dewey et al., 2019) and mica (Wrobel et al., 2020). DeepHarmony is a contrast harmonization method that maps images of two scanners to a middle-ground space in which images are harmonized by having similar contrast. However, DeepHarmony is limited to harmonizing images of just two scanners. On the other hand, mica is a multi-scanner (i.e., more than two scanners) harmonization method that harmonizes images by adapting their intensity distributions to that of the *target* scanner. Even though adapting images to a target scanner seems to simplify the task of harmonization, it introduces the new challenge of determining the "best" scanner in the pooled data. Selecting such scanner is not a trivial task when, for example, motion artifacts in images could be of concern (Alexander-Bloch et al., 2016; Torbati et al., 2021a).

Depending on the type of data that harmonization can be applied to, it could also fall into two broad categories: (1) harmonization of image-derived measures, and (2) harmonization of images. The methods of the first category can be described as ComBat (Johnson et al., 2007) and its extensions (Beer et al., 2020; Chen et al., 2020b; Pomponio et al., 2020; Reynolds et al., 2022). ComBat is a location and scale adjustment method used in neuroimaging for harmonizing image-derived measures and has been applied to images of different modalities: DTI (Fortin et al., 2017), MRI (Fortin et al., 2018), and fMRI (Nielson et al., 2018). Even though ComBat is a straightforward method which showed success in harmonization of image-derived measures in many studies (Yu et al., 2018; Radua et al., 2020; Foy et al., 2020; Torbati et al., 2021a), its performance cannot be easily depicted in terms of harmonization accuracy at the image level.

RAVEL was proposed as the first normalization and harmonization framework removing inter-subject variability from MRIs at the image and voxel level (Fortin et al., 2016). Harmonization methods using deep learning techniques have subsequently been proposed. The unsupervised deep-learning-based methods treat harmonization as the task of domain or style transfer learning, in which images of scanners are mapped to the domain or style of one selected scanner, called *target* scanner (Dewey et al., 2020; Zuo et al., 2021). As well as the challenge of selecting the target scanner, these methods have other limitations based on the deep learning network they used for transfer. For example, methods using CycleGAN (Modanwal et al., 2020) or DualGAN (Zhong et al., 2020) networks are limited to harmonization of just two scanners. Another example is CALAMITI with a disentanglement network limited to harmonizing inter-modality paired dataset. This data consists of paired images that are images of two predetermined modalities taken from an individual on the *same* scanner with short time gap. In addition, methods using style transfer (Liu et al., 2021; Liu and Yap, 2021) are prone to mapping images to biological or clinical information of the target scanner, if images across scanners are confounded by this information. Another major group of unsupervised methods proposed generating scanner-invariant latent representations for synthesizing harmonized images (Moyer et al., 2020) or training the neuroimaging tasks on images (Aslani et al., 2020;

Dinsdale et al., 2021). However, these methods are prone to lose the information of images during harmonization as their generated representation has been proven to be limited to the least informative scanner (Moyer and Golland, 2021).

In this work, we present MISPEL (**M**ulti-scanner **I**mage harmonization via **S**tructure **P**reserving **E**mbedding **L**earning), which is a supervised multi-scanner harmonization method that maps images of scanners to a middle-ground harmonized space on images. Figures 1d and b depict the result of MISPEL on harmonizing our example of matched images. In this study, we introduce a multi-scanner dataset of four *matched* 3T MRI T1 images, which, to the best of our knowledge is the first study of its kind. In addition, we provide an extensive set of experiments assessing scanner effects and evaluating harmonization on our unique set of matched data using two different commonly used MR image processing and segmentation software packages, FSL (Zhang et al., 2001) and SPM (Ashburner and Friston, 2005). Lastly, we compare MISPEL with three well-known methods of normalization and harmonization, White Stripe, RAVEL, and CALAMITI.

2 Materials and Methods

2.1 Study population and image acquisition

The sample used in this study consists of 18 participants which are part of an ongoing project (UH3 NS100608 grant to J. Kramer and C. DeCarli). The median age of the participants was 72 years (range 51-78 years) and 44% (N = 8) were males. All participants were cognitively normal. Each participant was scanned for T1-weighted

Table 1: **Scanner specifications.**

Scanner Name	**GE**	**Philips**	**SiemensP**	**SiemensT**
Manufacturer	General Electrics	Philips	Siemens	Siemens
Scanner Hardware	DISCOVERY-MR750w 3T	Achieva-dStream 3T	Prisma-fit 3T	TrioTim 3T
Scanner software	27-LX-MR-Software-release: DV26.0-R03-1831.b	5.6.1-5.6.1.0	syngo-MR-E11	syngo-MR-B17
Receive Coil	32Ch-Head	MULTI-COIL	BC	32Ch-Head
T1-w Sequence Type	BRAVO	ME-MPRAGE	ME-MPRAGE	ME-MPRAGE
Resolution (mm)	$1.0 \times 1.0 \times 0.5$	$1.0 \times 1.0 \times 1.0$	$1.0 \times 1.0 \times 1.0$	$1.0 \times 1.0 \times 1.0$
TE/ΔTE (ms)	3.7	1.66/1.9	1.64/1.86	1.64/1.86
TR (ms)	9500	2530	2530	2530
TI (ms)	600	1300	1100	1200

(T1-w) MRIs on four different 3T scanners: GE, Philips, SiemensP, and SiemensT (Table 1). Scans of each subject are called *matched images* and accordingly the images of all individuals are referred to as a *matched dataset*. These matched images were taken at most four months apart, a time period over which we assume no biological changes could occur in the brain and differences observed between any pairs of scans are solely due to the scanner effects. In a matched dataset, the scanner and harmonization effects can be estimated based on the dissimilarity and similarity of matched images, respectively. The details of estimation of scanner effects and evaluation of harmonization methods are provided in Section 2.5.

2.2 Image preprocessing

We use RAVEL as one of our harmonization methods in this study. In order to prevent confounding our evaluation with inconsistent preprocessing steps, we preprocessed all images using the pipeline prescribed for RAVEL (Fortin et al., 2016). Therefore, we first used a non-linear symmetric diffeomorphic image registration algorithm (Avants et al., 2008) to register images to a high-resolution T1-w image atlas (Oishi et al., 2009), followed by applying N4 bias correction method (Tustison et al., 2010) to correct for spatial intensity inhomogeneity. As the last step of the pipeline, all images were masked using a brain mask provided in (Fortin et al., 2016) for skull stripping. We also scaled images in one additional step in which intensity values of each image were divided by their average intensity value. Throughout this manuscript, these preprocessed images are referred to as RAW and used as input to our models.

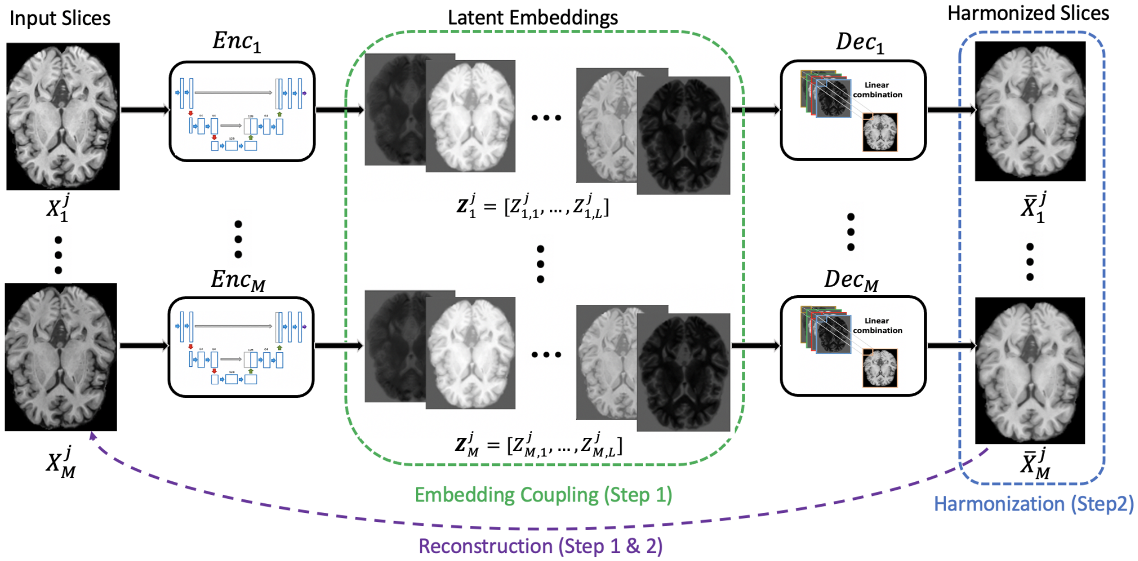


Figure 2: **Illustration of MISPEL.** For each of the $i = 1 : M$ scanners and the $j = 1 : N$ input axial slices, Enc_i (2D U-Net) decomposes the corresponding latent embeddings: $\mathbf{Z}_i^j = Enc_i(X_i^j)$. The corresponding Dec_i (linear function) then maps the embeddings to the output: $\bar{X}_i^j = Dec_i(\mathbf{Z}_i^j)$. **Step 1** Embedding Learning: Enc_i-Dec_i unit reconstructs the input images for each scanner i. In this step, $Enc_{i=1:M}$ and $Dec_{i=1:M}$ are updated using the Embedding Coupling loss and the Reconstruction loss. **Step 2** Harmonization: the Dec_i synthesizes the harmonized images for each scanner i. In this step, only $Dec_{i=1:M}$ are updated using the Harmonization loss and the Reconstruction loss.

2.3 MISPEL

Our proposed framework, MISPEL, is a convolutional deep neural network for harmonizing *matched* images from multiple scanners with potential scanner effects. We show that MISPEL (1) harmonizes matched images by making them similar across scanners, (2) preserves the structural (anatomical) information of the original brains, and (3) generalizes to multiple (more than two) scanners. In order to achieve these goals, we designed a two-step training framework for MISPEL which consists of units of encoder and decoder modules for each of the scanners (Figure 2). More detail on MISPEL were provided in (Torbati et al., 2021b) and the code is publicly available[1].

2.3.1 Implementation

We consider M scanners for RAW data, i.e., the preprocessed matched images which are registered to the same template space. The axial slices across all RAW scans are combined for a total of N slices for each scanner i, $i = 1 : M$ denoting $i \in \{1, \ldots, M\}$. The dataset thus consists of $X_{i=1:M}^{j=1:N}$, where X_i^j is the slice j from scanner i and $X_1^j, X_2^j, \ldots, X_M^j$ are the matched axial slices. Our goal is to achieve the harmonized axial slices, referred to as $\bar{X}_{i=1:M}^{j=1:N}$, by making them similar across scanners, i.e., achieving $\bar{X}_1^j \approx \ldots \approx \bar{X}_M^j$, for $j = 1 : N$.

For having a network generalizable to multiple scanners, we designed a separate unit of encoders and decoders for each of the scanners. This helps MISPEL to be easily expandable to any number of scanners by adding a unit for each. We designed Enc_i (the encoder for scanner i) as a 2D U-Net (Ronneberger et al., 2015), which

[1] https://github.com/Mahbaneh/MISPEL

decomposes slice X_i^j into its set of L latent embeddings $\mathbf{Z}_i^j = [Z_{i,1}^j, \dots, Z_{i,L}^j]$. Dec_i is then designed as a linear function combining the components of latent embeddings, $Z_{i,1}^j, \dots, Z_{i,M}^j$, to map $\mathbf{Z}_i^j$ to $\bar{X}_i^j$.

2.3.2 Network Training

Each Enc_i-Dec_i unit reconstructs $\bar{X}_i^j$ from X_i^j for each scanner i and slice j and cannot reach harmonization by itself. Thus, we employ another mechanism in order to make the synthesized images, $\bar{X}_{i=1:M}^{j=1:N}$, similar across the scanners and achieve harmonization. One way to do that would be to train all *Enc-Dec* units to directly impose similarity of matched slices by a loss function. However, this may result in modification of brain structures, as we noticed that even our matched slices which were co-registered in the preprocessing, have small structural differences. Thus, we implemented a two-step training for MISPEL which preserves the brain structure. In Step 1, we first learned the embeddings with structural information, and in Step 2, we harmonized the intensities of embeddings without modifying the structure of brains.

Step 1: Embedding Learning. For learning embeddings that could preserve the structural information of the brain, we train the *Enc-Dec* units to reconstruct their corresponding input slices. For example, for scanner i and slice j, the goal for Enc_i-Dec_i is to achieve $X_i^j \rightarrow \mathbf{Z}_i^j \rightarrow \bar{X}_i^j$, in which $X_i^j \approx \bar{X}_i^j$. To enforce all units to image reconstruction, we used Reconstruction loss ($\mathcal{L}_{recon}$). $\mathcal{L}_{recon}$ should enforce all units to reconstruct their input images. To use this specific Reconstruction loss, we first compute the pixel-wise mean absolute error (MAE) between X_i^j and $\bar{X}_i^j$ for

$i = 1 : M$ and then sum them over. In addition to this image reconstruction strategy, the Dec_i modules maintain the structural information of brain by *linearly* combining the embeddings.

Making the latent embeddings similar across scanners will improve the results of harmonization later in Step 2. By this similarity, for example for scanner i and slice j, the goal is to obtain $Z^j_{1,l} \approx \ldots \approx Z^j_{M,l}$, for $l = 1 : L$. For enforcing the similarity, we designed the Embedding Coupling loss ($\mathcal{L}_{coup}$) to "couple" the embeddings of the M scanners. We first calculated the pixel-wise variance among the lth embeddings of all M scanners. We conducted this step for all L embeddings. We then calculated $\mathcal{L}_{coup}$ as the mean of these variances over all embeddings and their pixels. The loss for Step 1 is then calculated as $\mathcal{L}_{step1} = \lambda_1 \mathcal{L}_{recon} + \lambda_2 \mathcal{L}_{coup}$, where $\lambda_1 > 0$ and $\lambda_2 > 0$ are the weights. We trained our units of *Enc-Dec* for $j = 1 : N$ slices. The units trained simultaneously for T_1 times or until the model reconstructs accurately.

Step 2: Harmonization. We continue the training process with Step 2 in which for each scanner i and slice j, the goal is to achieve $X^j_i \rightarrow \mathbf{Z}^j_i \rightarrow \bar{X}^j_i$. Unlike Step 1, the $\bar{X}^j_i$ will be the harmonized slice in this step. For harmonizing slices, we froze the encoders during the training and updated just the decoders to synthesize similar matched slices, i.e., achieving $\bar{X}^j_1 \approx \ldots \approx \bar{X}^j_M$. For enforcing the similarity, we used the Harmonization loss ($\mathcal{L}_{harm}$). We first calculated the MAEs between the images of all unique scanner pairs. $\mathcal{L}_{harm}$ was then the mean of these MAEs. For example, for slice j, the $\mathcal{L}_{harm}$ is the mean of MAEs for $\{(\bar{X}^j_i, \bar{X}^j_k) \mid i, k \in \{1, \ldots, M\} \text{ and } i < k\ \}$. In the loss for Step 2, we also incorporate $\mathcal{L}_{recon}$ to ensure that harmonized images do not deviate from their originals. Thus, we have $\mathcal{L}_{step2} = \lambda_3 \mathcal{L}_{recon} + \lambda_4 \mathcal{L}_{harm}$, where

$\lambda_3 > 0$ and $\lambda_4 > 0$. With $\mathcal{L}_{step2}$, we trained the decoders of all units for $j = 1 : N$ slices and repeat it for T_2 times or until $\mathcal{L}_{step2}$ does not change anymore. By the end of this step, the synthesized images, $\bar{X}_{i=1:M}^{j=1:N}$, are the desired harmonized ones.

As explained, we started with training the model for Step 1 and then continue it with Step 2. In Step 1, we fixed $\lambda_1 = 1$ and trained for $\lambda_2 \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $L \in \{4, 6, 8\}$. We then selected $L = 6$, $\lambda_1 = 1$, and $\lambda_2 = 0.3$ based on $\mathcal{L}_{step1}$ and the quality of the reconstructed images. In Step 2, we fixed the model for $\lambda_3 = 1$ and trained it for $\lambda_4 = \{1, 2, 3, 4, 5, 6\}$. We ended up having $\lambda_3 = 1$ and $\lambda_4 = 4$, when we compared results based on $\mathcal{L}_{step2}$ and the quality of the final harmonized images. The training was conducted on NVIDIA RTX5000 for $T_1 = 100$ and $T_2 = 100$ with the batch size of 4. For both steps, we used ADAM optimizer (Kingma and Ba, 2014) with a learning rate of 0.01. The training took approximately 200 and 30 minutes for Step 1 and Step 2, respectively. We trained MISPEL on RAW images: 12 and 6 subjects for train and validation, respectively. We then reported our evaluation on all RAW images using the trained model to be comparable to RAVEL as one of our competing methods. RAVEL should be applied on all images at once and will be more discussed in the next section.

2.4 Competing methods

We compared MISPEL with one method of intensity normalization, White Stripe (WS), and two methods of harmonization, RAVEL and CALAMITI. We selected WS and RAVEL as they (1) have been widely applied to MRI neuroimaging data, (2) can be applied to multiple (more than two) scanners, and (3) do not require specification of a *target* scanner. We selected CALAMITI as it (1) can be slightly modified and applied to matched data, and (2) could be regarded as one of the state-of-the-art methods in harmonization. We should emphasize that determining the ultimate state-of-the-art harmonization method is not trivial as harmonization lack a standardized evaluation criteria currently. All methods were applied to RAW images.

White Stripe (WS) is an individual-level intensity normalization method for removing discrepancy of intensities across subjects within tissue types (Shinohara et al., 2014b). It first extracts the normal-appearing white matter voxels of the image and estimates moments of their intensity distribution. It then uses these moments in the z-score transformation for normalizing the voxels of all brain tissue types.

RAVEL is an intensity normalization and harmonization framework (Fortin et al., 2016). It initializes with a WS normalization step and then applies a voxel-wise harmonization strategy to images. In the harmonization strategy, RAVEL first estimates the components of scanner effects by applying singular value decomposition (SVD) to cerebrospinal fluid (CSF) voxels of images. These voxels are known to be unassociated with disease status and clinical covariates and are representative of scanner effects. RAVEL then uses these voxels to estimate scanner effects and harmonizes the images by removing the estimated scanner effects from the voxel intensities. For building our model, we adjusted RAVEL for status of the subjects (cognitively normal with low or high degree of SVD in (Wilcock et al., 2021)). We also set the components of scanner effects to 1, as suggested in the original work (Fortin et al., 2016).

CALAMITI is an unsupervised deep-learning method for harmonizing multi-scanner inter-modality paired dataset (Zuo et al., 2021). It is a domain adaptation approach mapping the images of scanners to the domain of a *target* scanner. Intermodality paired dataset consists of images of two predetermined modalities taken from one individual on the *same* scanner with short time gap. This dataset can have paired images of multiple scanners. For simplicity, we refer to these images as *paired* in description of this method. CALAMITI should be first trained on paired images of two scanners, one of which should be the *target* scanner. It could then be fine-tuned to map images of other scanners to the target domain. During the training, CALAMITI (I) gets the paired images as inputs and generates a disentangled representation that captures the mutual scanner-invariant anatomical information (β) of images as well as the contrast information (θ)s of their modalities and scanner; and (II) synthesizes the input paired images using their generated mutual β and θs. For harmonizing an input image, the trained model is used to generate the β of the image and θ of one random image from the target scanner. The model then synthesize the harmonized image using these two components.

We used CALAMITI as a supervised method by simply training it on our *inter-scanner* paired data. Likewise MISPEL, we trained CALAMITI on RAW images: 12 and 6 subjects for train and validation, respectively. We then reported our evaluation on all RAW images using the trained model. Following its original paper, we went through one step of normalization and trained CALAMITI using the WS-normalized RAW images. Instead of conducting fine-tuning, we went for a simpler approach and trained 3 individual models to map GE, Philips, and SiemensP to SiemensT. We used the same machines used for MISPEL and trained CALAMITI with the hyper-parameters reported in its original paper. For being comparable and fair to other methods, we trained CALAMITI on 2D axial slices and skipped its super-resolution preprocessing step and post-harmonization slice-to-slice consistency enhancement step.

2.5 Data analysis

A harmonization method is expected to remove scanner effects in order to enable unbiased neuroimaging analyses on the cross-site/scanner pooled data. In our specific matched dataset, the matched images are assumed to be biologically identical but differ entirely due to scanner differences. Thus, the scanner effects can be estimated as dissimilarity of the matched images and removing the scanner effects can be regarded as increasing the similarity of them. We investigated the similarity and dissimilarity of matched images using three evaluation criteria: (1) image similarity, (2) GM-WM contrast similarity, and (3) volumetric and segmentation similarity.

We performed our evaluation metrics for all five methods: RAW, White Stripe, RAVEL, CALAMITI, and MISPEL. The entire matched dataset was used in evaluating each method, unless otherwise mentioned. Many of our evaluation metrics require pairwise image-to-image comparison, for which we considered all possible combinations of *scanner pairs*: {(GE, Philips), (GE, SiemensP), (GE, SiemensT), (Philips, SiemensP), (Philips, SiemensT), and (SiemensP, SiemensT)}. Throughout this manuscript, the two matched images of each scanner pair are referred to as *paired* images. To determine the statistical significance of any comparisons, we used paired t-test with $p < 0.05$ denoting the significance.

We first investigated the **image similarity**. For this, we assessed the visual quality of the matched *slices* for all methods. We also quantified the similarity of *all* paired images using the structural similarity index measure (SSIM). SSIM is a pairwise metric which compares two images in terms of their luminance, contrast, and structure. A harmonization method is expected to increase the visual and structural similarity of paired images.

Second, we investigated the **GM-WM contrast similarity** of images. The GM-WM contrast can highly influence the quality of segmentation methods and increased contrast is expected to result in more accurate segmentation. The GM-WM contrast of an image can be estimated as separability of its histograms of GM and WM voxels. This separability was used as classification of GM and WM voxels of an image in (Torbati et al., 2021a) and reported as area under the receiver operating characteristic (AUROC) values, with AUROC = 1 denoting perfect classification (complete separation of histograms) and AUROC = 0.5 showing random classification (complete overlap of histograms). For calculating these AUROC values, we conducted the procedure explained in (Torbati et al., 2021a) for each of the images. We first labeled GM and WM voxels of the image using the tissue mask in the EveTemplate package (Oishi et al., 2009). We then classified these voxels using intensity thresholds selected from the range of intensity values of the GM and WM voxels. Lastly, we formed the AUC curve of the image using the result of each classification.

Third, we investigated the **volumetric and segmentation similarity** criterion for images. The most practical benefit of harmonization is to enable unbiased multi-scanner neuroimaging analyses with minimal scanner effects. The neuroimaging measures are the basis of these analyses, therefore, the volumetric and segmentation similarity of these measures across paired images is an appropriate metric for evaluating harmonization. We segmented and measured the volumes of the two brain tissue types: GM and WM. We then analyzed the similarity of *each of* these two tissue types *separately* and in four ways: (1) volume distributions, (2) volumetric bias, (3) volumetric variance, and (4) segmentation overlap. For volumetric distributions, we compared the distributions of volumes of each tissue type across their four scanners. Most of the metrics used in the three other criteria are pairwise comparisons, thus we applied them *separately* to all of the 6 *scanner pairs* we just enumerated. For volumetric bias, we calculated the absolute differences between volumes of paired images of each scanner pair and evaluated the harmonization based on the mean of these differences over all individuals of the scanner pair. We used root-mean-square deviation (RMSD) for estimating the volumetric variance of paired images of all individuals in each scanner pair. RMSD of a scanner pair denotes the deviation of volumes of one scanner, from that of the other scanner. Lastly, we used dice similarity score (DSC) to estimate the overlap of tissue segmentation of paired images of each scanner pair. The mean of these DSC values over paired images of all subjects was used as an evaluation metric for harmonization. A harmonization method is expected to result in (1) similar distribution of volumes across scanners, (2) minimal (ideally zero) bias, (3) minimal (ideally zero) variance, and (4) maximal (ideally complete) segmentation overlap; for both tissue types and all scanner pairs.

We conducted the volumetric and segmentation similarity evaluation criterion for two segmentation tools: (1) FSL FAST (version 6.0.3) (Zhang et al., 2001), and (2) segmentation in statistical parametric mapping (SPM12) (Ashburner and Friston, 2005). These frameworks are widely used for tissue segmentation in studies, however the results of these two segmentation algorithms could have moderate to large differences (Tudorascu et al., 2016). We therefore assessed volumes from each segmentation tool independently. Originally, the output of WS, RAVEL, CALAMITI and MISPEL methods were images in a template space, as all of them used RAW images as input. The RAW images were non-linearly registered to a T1-w image atlas (Oishi et al., 2009) in the preprocessing step, Section 2.2. Using their inverse transformations, processed images of all methods were transferred to their native space and then used as inputs of the two segmentation tools for tissue volume extraction and then volumetric similarity evaluation. On the other hand, for having a meaningful tissue segmentation overlap, segmentations, and accordingly their images should remain in their template space. Thus, we also ran the two segmentation frameworks on the template-space images to generate the segmentations and then evaluate the segmentation overlap similarity. For all runs of the segmentation frameworks, we set the tissue class probability thresholds to 0.8.

3 Results

3.1 Image similarity

The similarity of images across normalization and harmonization methods is depicted in Figures 3 and 4. Visual assessment of processed images in Figure 3 revealed that (1) scanner effects are present in the matched RAW images and appear most significantly as differences in image contrast, (2) White Stripe made matched images more similar,

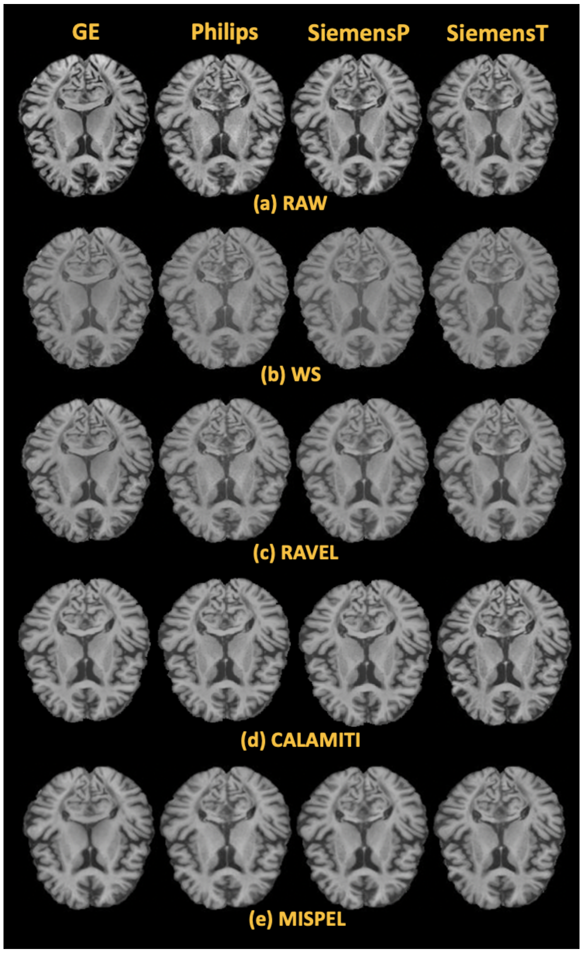


Figure 3: **Visual assessment of matched images of a slice.** Rows and columns correspond to methods and scanners respectively. All four methods: WS, RAVEL, CALAMITI, and MISPEL made the matched slices of RAW more similar, with CALAMITI and MISPEL preserving their contrast the most.

but at the expense of decreased contrast, (3) RAVEL improved upon WS by increasing contrast relative to WS-normalized images, (4) CALAMITI improved similarity of the matched images by adapting contrast across all scanners to that of the RAW

SiemensT, and (5) MISPEL improved similarity of images likewise CALAMITI, but made images smooth to some extent.

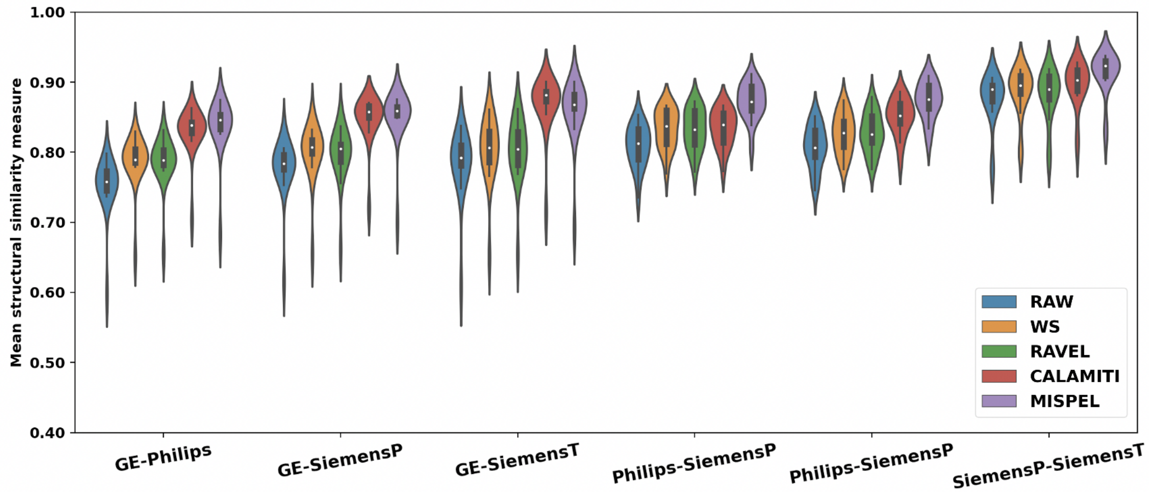


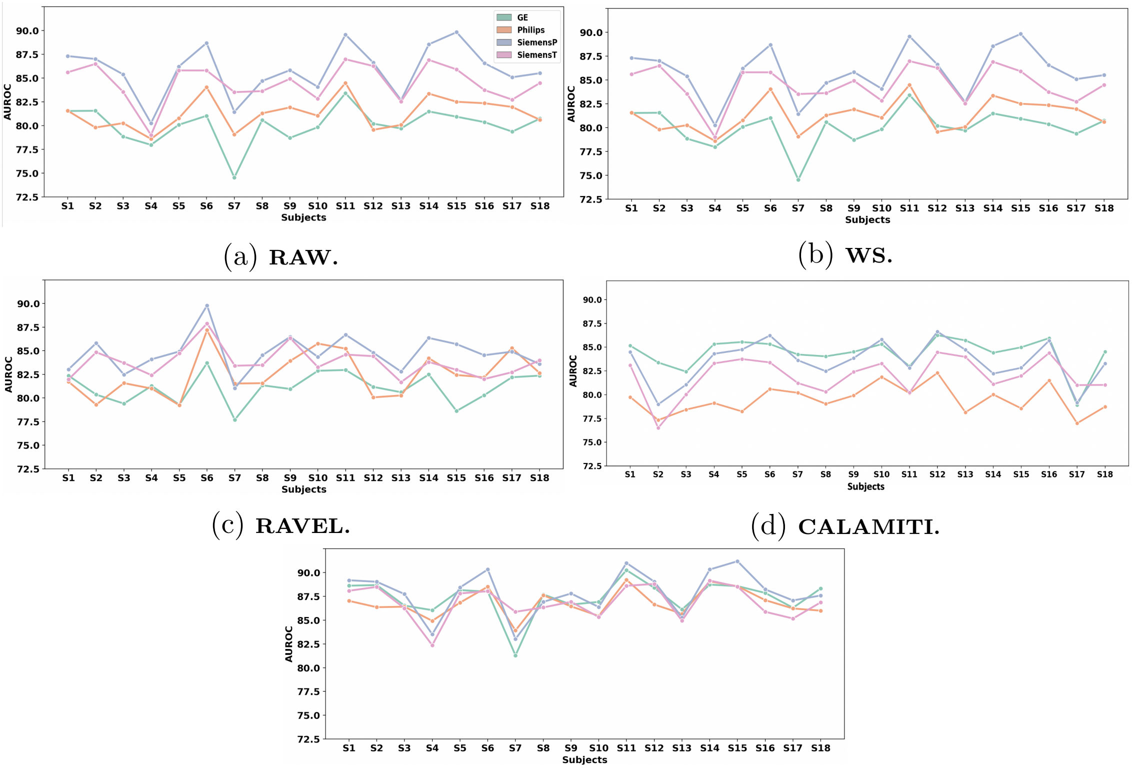
Figure 4: **Structural similarity index measures (SSIM) for matched dataset.** The SSIM distribution of images of scanner pairs were depicted as violin plots for each of the methods. A harmonization method is expected to have the highest mean of SSIM. All four methods of WS, RAVEL, CALAMITI, and MISPEL improved the mean SSIM of RAW, with MISPEL outperforming the other three. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

For a quantitative understanding of similarity of images, we explored the SSIM distribution of the matched images of all subjects, for the 6 *scanner pairs* enumerated in Section 2.5. These distributions are depicted as violin plots for the five methods: RAW, WS, RAVEL, CALAMITI and MISPEL, in Figure 4. The violin plots with the smallest SSIM mean belong to RAW, indicating scanner effects exist in our matched dataset. Scanner pairs including GE have long-tailed distributions, which indicates that GE images are most dissimilar to others. Moreover, the SiemensP-SiemensT scanner pair had the largest SSIM mean, indicating that these two are the most similar scanners. Lastly, we observed that WS, RAVEL, CALAMITI, and MISPEL improved SSIM of RAW for all of its scanner pairs, with MISPEL outperforming the other three methods. All these comparisons were statistically significant, as assessed using paired t-tests.

3.2 GM-WM contrast similarity

We estimated the GM-WM contrast of an image, using the AUROC values denoting the separation of histograms of its GM and WM voxels. High AUROC indicates higher contrast, with 100% the highest. In Figure 5, we depicted the spaghetti plots of AUROC values of images of all subjects across the four scanners. A harmonization method is expected to (1) make the AUROC of matched images similar, i.e., results in overlapped lines, and (2) not deteriorate the AUROC of images.

We show in Figure 5a that scanner effects exist in RAW data and appeared as dissimilarity of GM-WM contrast in matched dataset, i.e., distant lines in this plot. WS in Figure 5b does not seem to change AUROCs of RAW and CALAMITI seems to just increase AUROCs of GE and makes its line much closer to that of SiemensP and SiemensT in Figure 5d. On the other hand, RAVEL and MISPEL resulted in more overlapped lines in respectively Figures 5c and 5e, with MISPEL having the highest overlap. Figure 6 shows the bar plots indicating the mean AUROC of images of each scanner. MISPEL is the only method that increased the mean AUROC of RAW images for all scanners. We also observed that: (1) WS did not change the mean AUROC value of RAW, (2) RAVEL improved the contrast for GE and Philips, but made it worse for SiemensP and SiemensT, and (3) CALAMITI improved the mean



(a) **RAW.**

(b) **WS.**

(c) **RAVEL.**

(d) **CALAMITI.**

(e) **MISPEL.**

Figure 5: **GM-WM contrast spaghetti plots.** The GM-WM contrast was estimated as AUROC values and was depicted for images of all subjects as spaghetti plots. In these plots, each line corresponds to one scanner. A harmonization method that performs well should show overlap of the lines. Plots showed that MISPEL outperformed WS, RAVEL, and CALAMITI with the highest overlapped of the lines. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

AUROC of GE and did not affect that of other scanners. In addition to these results, MISPEL seems to be the most successful method in bringing the mean AUROC of the scanners closer to each other. In summary, we show that MISPEL is the only method that satisfied both harmonization criteria determined for GM-WM contrast similarity.

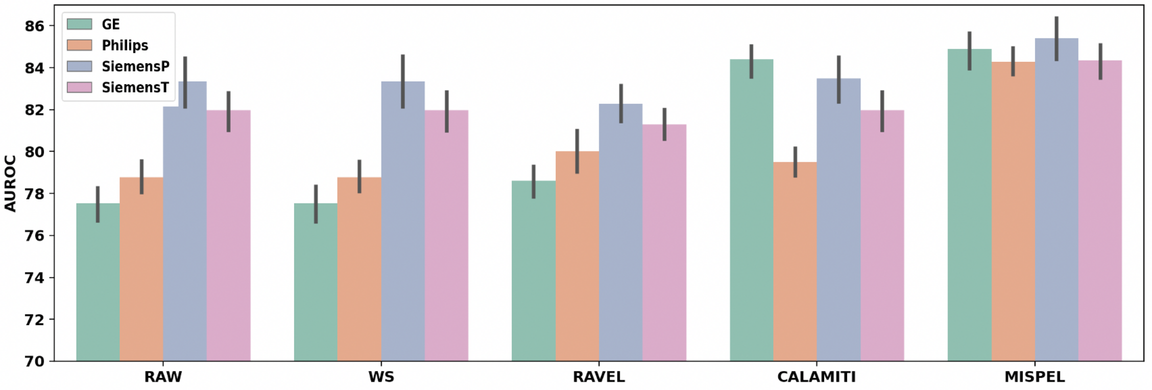


Figure 6: **GM-WM contrast bar plots.** Each bar indicates the mean AUROC of images of each scanner, with error bars denoting the standard deviation for each method. A harmonization method is expected to not deteriorate the GM-WM contrast of images. Plots show that MISPEL outperformed WS, RAVEL, and CALAMITI reflected in the similarity of the boxplots. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

3.3 Volumetric and segmentation similarity

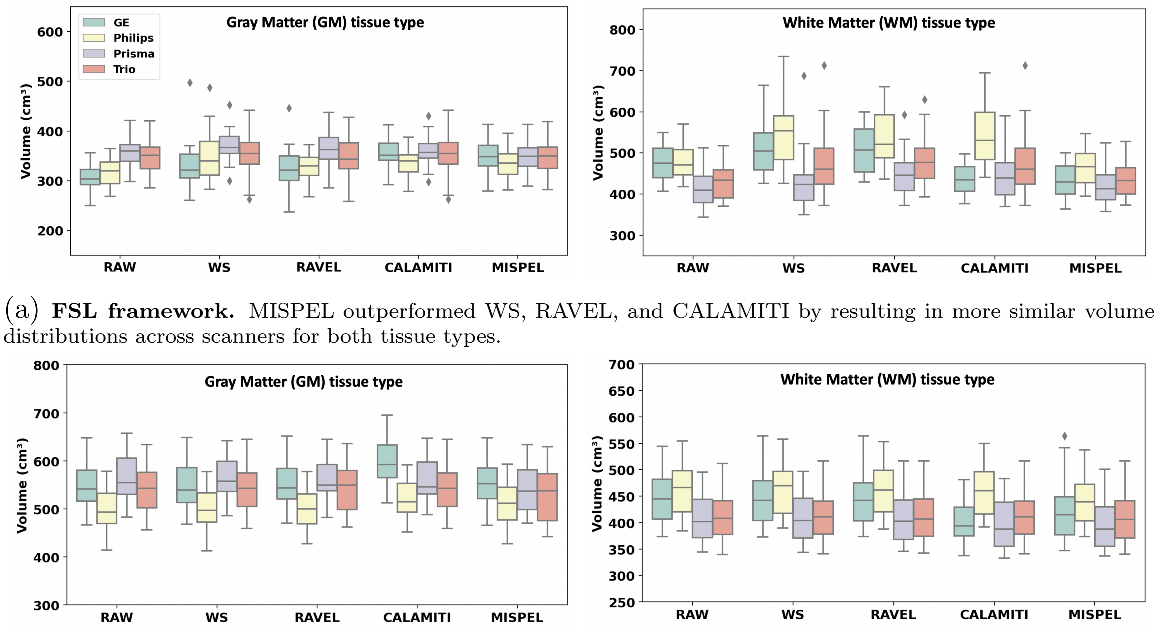
We estimated the volumetric and segmentation similarity of GM and WM tissue types based on four criteria: (1) volume distributions, (2) volumetric bias, (3) volumetric variance, and (4) segmentation overlap. We performed our evaluation for FSL and SPM segmentation frameworks and expected the harmonization methods to result in: (1) similar volume distributions across scanners, (2) minimal bias, (3) minimal variance, and (4) maximal segmentation overlap; for both tissue types and both segmentation frameworks.

3.3.1 Volume distributions

Figure 7 shows boxplots of volumes of the two tissue types, GM and WM, across the four scanners, with Figures 7a and 7b depicting these boxplots for volumes extracted by FSL and SPM frameworks, respectively. We depicted these boxplots for all five methods. Plots in Figure 7a showed that scanner effects exist in the matched volumes derived through FSL and appeared as dissimilar boxplots for RAW across scanners. When compared to RAW, WS and RAVEL resulted in more dissimilar boxplots for FSL-derived volumes of both GM and WM. On the other hand, we noticed that the use of CALAMITI and MISPEL helped towards harmonization of data. CALAMITI made GE more similar to SiemensP and SiemensT for both GM and WM, but made Philips more dissimilar to the three other scanners for WM. Among the four methods, MISPEL did the best for the FSL-derived volumes by resulting in the most similar cross-scanner distributions for both GM and WM.

Figure 7b showed that scanner effects exist in RAW volumes extracted by SPM too. Our normalization and harmonization methods though resulted in relatively minor changes in SPM-derived GM and WM volumes, with CALAMITI and MISPEL showed the most noticeable changes. For WM volumes, CALAMITI made the distribution of GE more similar to that of SiemensP and SiemensT, while the reverse was observed for the GM volumes of this scanner. Moreover, MISPEL made the distributions of GM volumes more similar across Philips, SiemensP, and SiemensT.

In summary, MISPEL outperformed WS, RAVEL, and CALAMITI in harmonizing FSL-derived volumes and none of the methods resulted in *visually significant* assessed harmonization for the SPM-derived volumes, when volumetric distribution similarity of *both* GM and WM volumes were used as the evaluation metric. Results for the statistical assessment of harmonization of FSL- and SPM-derived GM and WM volumes are presented in the next section.



(a) **FSL framework.** MISPEL outperformed WS, RAVEL, and CALAMITI by resulting in more similar volume distributions across scanners for both tissue types.

(b) **SPM framework.** No *visually significant* noticeable harmonization was observed for any of the methods.

Figure 7: **Volume distribution boxplots.** Boxplots denote the volume distribution of GM and WM tissue types for images of each scanner. These boxplots were depicted for all five methods and explored for two segmentation frameworks: (a) FSL and (b) SPM. A harmonization method is expected to result in similar distributions of volumes across scanners. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

3.3.2 Volumetric bias

Table 2 shows mean and standard deviation (SD) of cross-scanner absolute differences of all paired volumes in each scanner pair. We calculated these statistics for volumes of GM and WM tissue types extracted using FSL and SPM segmentation frameworks, for all five methods. We also presented the distributions of these differences as violin plots in Figure 8. Using paired t-test, we compared each of these distributions to their equivalent distributions in RAW.

A harmonization method is expected to result in minimal (ideally zero) mean of absolute differences (bias), with no major change in SD of the differences. The SD values indicate the consistency of harmonization across subjects. A harmonization method should harmonize images of all subjects to a comparable degree, and thus should not change the SDs drastically. Likewise visually, the violin plots in Figure 8 for harmonized images are expected to be centered as close as possible to zero.

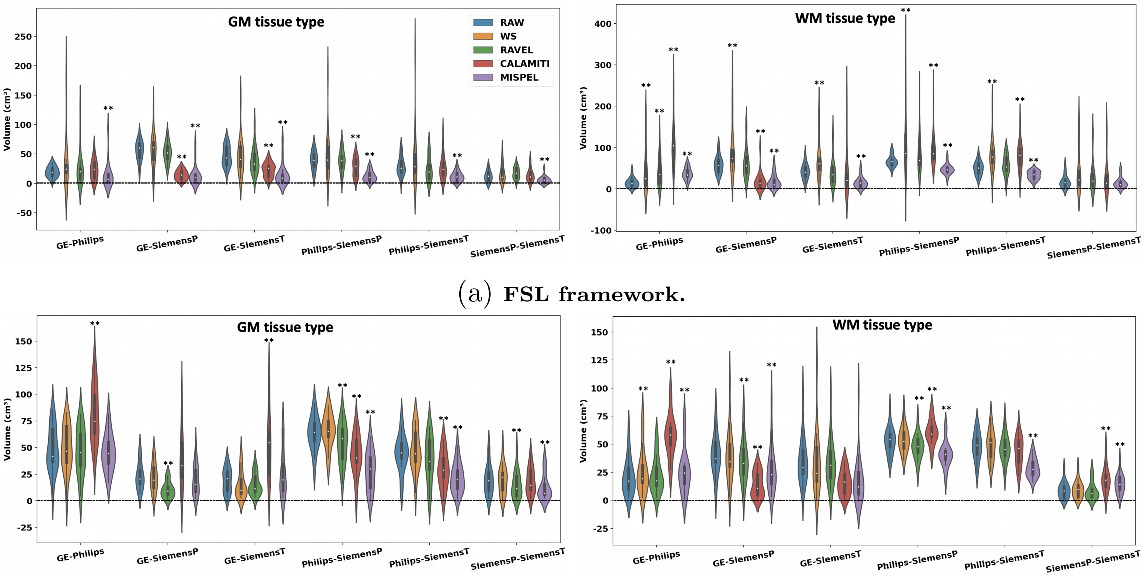
Table 2: **Mean absolute differences for matched dataset.** Mean (SD) of cross-scanner absolute differences of volumes for all scanner pairs and all methods. The volumes are for GM and WM tissue types and were extracted through two segmentation frameworks: FSL and SPM. A harmonization method that works is expected to have low values of mean and SD for all paired scanners. MISPEL outperformed WS, RAVEL, and CALAMITI by having the largest number of smallest means and not significantly increasing the values of SD, for both FSL and SPM. The distributions that showed statistically significant t-statistics when compared to RAW were marked by *.

Framework	**Tissue**	**Method**	**Mean (SD) of Volumetric Absolute Differences for Paired Dataset**					
			GE-Philips	GE-SiemensP	GE-SiemensT	Philips-SiemensP	Philips-SiemensT	SiemensP-SiemensT
FSL	GM	RAW	19.82 (9.10)	55.84 (16.54)	46.53 (16.94)	39.70 (15.28)	29.00 (15.75)	12.14 (9.17)
		WS	43.53 (56.27)	56.66 (29.56)	46.34 (31.84)	49.31 (40.21)	43.09 (49.92)	18.00 (15.37)
		RAVEL	27.53 (32.28)	52.88 (16.60)	39.22 (22.41)	38.87 (17.79)	24.65 (20.50)	17.53 (9.35)
		CALAMITI	23.20 (17.20)	**14.27** (8.19)*	22.94 (11.76)*	27.85 (15.45)*	28.14 (21.21)	14.13 (10.91)
		MISPEL	**13.90** (22.26)	14.38 (17.20)*	**15.09** (18.74)*	**12.47** (8.75)*	**11.62** (8.41)*	**5.87** (6.25)*
	WM	RAW	**15.39** (11.29)	59.59 (20.92)	42.45 (18.37)	67.30 (13.41)	50.16 (16.43)	17.89 (15.48)
		WS	46.99 (54.09)*	100.63 (64.18)*	71.35 (47.72)*	119.73 (79.23)*	81.41 (41.29)*	41.03 (50.11)
		RAVEL	43.95 (36.46)*	65.02 (37.96)	42.42 (34.98)	89.59 (49.34)	57.60 (23.77)	32.18 (39.53)
		CALAMITI	111.31 (58.26)*	19.38 (23.22)*	45.73 (64.41)	101.52 (46.16)*	79.94 (34.52)*	34.71 (48.12)
		MISPEL	37.46 (14.29)*	**18.43** (17.00)*	**15.91** (13.55)*	**47.24** (12.08)*	**33.35** (10.59)*	**15.04** (12.80)
SPM	GM	RAW	48.22 (20.82)	23.45 (12.67)	19.37 (11.23)	63.57 (15.90)	44.65 (16.94)	19.86 (13.77)
		WS	48.60 (21.35)	21.75 (13.15)	**14.94** (12.70)	65.72 (15.54)	46.84 (18.99)	19.46 (13.53)
		RAVEL	46.12 (22.48)	**10.44** (7.57)*	15.22 (9.77)	53.82 (18.48)*	39.14 (20.99)	15.26 (12.85)*
		CALAMITI	81.45 (26.94)*	39.15 (27.21)	54.84 (32.34)*	43.36 (18.64)*	29.14 (16.62)*	18.98 (12.55)
		MISPEL	**45.59** (17.39)	20.37 (14.30)	25.86 (21.99)	**28.82** (17.92)*	**21.89** (13.79)*	**12.78** (11.84)*
	WM	RAW	**21.06** (15.98)	40.40 (18.08)	35.45 (20.87)	53.16 (11.74)	48.22 (12.32)	9.06 (7.31)
		WS	25.97 (20.29)*	40.18 (23.46)	34.18 (27.71)	54.43 (11.43)	48.80 (13.16)	9.69 (7.53)
		RAVEL	22.49 (15.69)	35.60 (19.36)*	34.02 (21.03)	47.64 (10.95)*	46.48 (12.14)	**8.41** (8.00)
		CALAMITI	57.33 (18.74)*	**14.49** (10.96)*	**15.80** (10.95)	60.74 (11.07)*	43.70 (12.90)	18.87 (11.34)*
		MISPEL	25.39 (19.34)	28.26 (21.39)*	18.75 (22.44)*	**40.92** (11.40)*	**28.91** (8.33)*	14.95 (8.62)*

We observed that scanner effects exist in the RAW volumes extracted through

FSL framework and appeared for all scanner pairs as non-zero bias values. We also observed that MISPEL resulted in the largest number of smallest biases for FSL-derived volumes, when compared to the other three methods. This number was 10 out of total of 12 cases, which are the 6 scanner pairs of the 2 tissue types. 8 out of these 10 biases were significantly different than their equivalents in RAW. Moreover, we noticed that MISPEL did not significantly increase the SD of distributions, just 4 increases out of 12, in which only the SD of GM for the GE-Philips pair had a major increase. On the other hand, WS, RAVEL, and CALAMITI showed increase in SD of differences, with WS and RAVEL showed increase for all 12 distributions, CALAMITI increased SD for 10 of the cases, and WS showed the most drastic increases. In general, RAVEL and CALAMITI did harmonization for FSL-derived volumes to some extent. Comparing to RAW, RAVEL totally decreased 5 biases and CALAMITI had 1 smallest bias as well as 4 decreases. However, CALAMITI resulted in drastic biases for the WM volumes of scanner pairs that include Philips. This was expected as we already observed noticeable dissimilar distribution of Philips to that of other scanners for WM volumes in Figure 7a.

Results of RAW volumes extracted by SPM showed that scanner effects exist in volumes of this segmentation framework too. Comparing to FSL, almost no major increase of SD was observed for SPM-derived volumes of any of the methods, except for WM volumes of GE-SiemensP and GE-SiemensT pairs of CALAMITI. This can be observed in Table 2 as well as Figure 8b, in which violin plots of methods for each scanner pair have more comparable shapes than that of FSL. MISPEL resulted in the largest number of smallest biases for SPM too, 6 out of 12 cases. 5 of these smallest



(a) **FSL framework.**

(b) **SPM framework.**

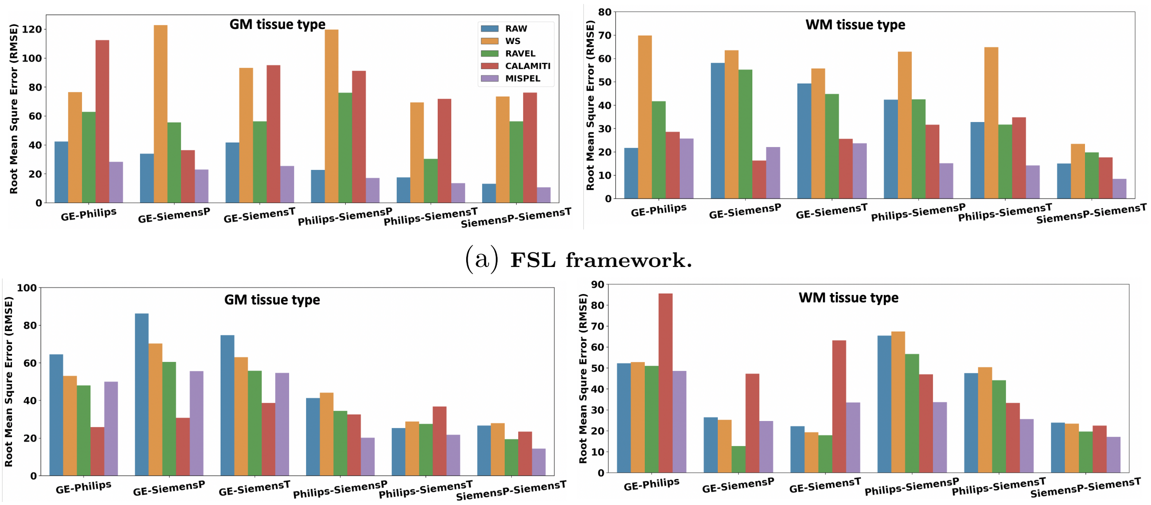
Figure 8: **Absolute difference violin plots.** The distributions of absolute differences of paired volumes as violin plots for all scanner pairs. The volumes are for GM and WM tissue types and extracted using two segmentation frameworks: (a) FSL and (b) SPM. A harmonization method is expected to result in short and fat (wide) violins, with mean values centered at zero. MISPEL outperformed WS, RAVEL, and CALAMITI by having largest number of these violin plots for both FSL and SPM. The distributions that showed statistically significant t-statistics when compared to RAW were marked by **. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

biases are statistically significantly different than their equivalent distributions in RAW. WS, RAVEL, and CALAMITI harmonized the SPM-derived volumes to some extent by decreasing the biases of 5, 11, and 6 cases, respectively. They also resulted in few smallest biases which are 1, 2, and 2 cases for WS, RAVEL, and CALAMITI, respectively.

Summarizing Table 2 and Figure 8, we observed that MISPEL outperformed WS, RAVEL, and CALAMITI when FSL and SPM were used for extracting volumes and volumetric bias was used as the harmonization evaluation metric.

3.3.3 Volumetric variance

Figure 9 shows bar plots that indicate the RMSD of paired volumes in each of the scanner pairs. We calculated these values for volumes of GM and WM tissue types and depicted them for all five methods. Figure 9 contains these sets of bar plots for volumes extracted through FSL and SPM frameworks in Figures 9a and 9b, respectively. Ideal harmonization would result in a zero RMSD for each scanner pair.



(a) **FSL framework.**

(b) **SPM framework.**

Figure 9: **Root-mean-square deviation (RMSD) bar plots.** Bar plots indicate the RMSD of paired volumes in scanner pairs. These values were calculated for volumes of GM and WM tissue types and depicted for all five methods. These set of bar plots were depicted for volumes extracted through two segmentation frameworks: (a) FSL and (b) SPM. A harmonization method is expected to lower values of RMSDs. MISPEL outperformed WS, RAVEL, and CALAMITI by having the largest number of smallest RMSD values for volumes of both FSL and SPM. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

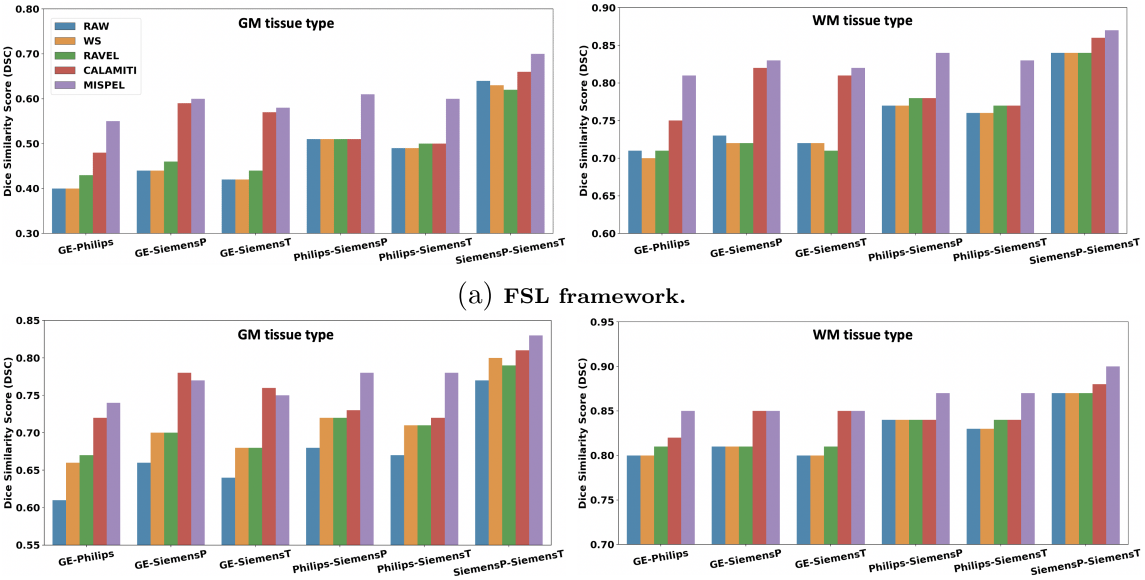
We observed that scanner effects exist in RAW volumes for both segmentation frameworks and appeared as non-zero RMSD values. Also, MISPEL outperformed WS, RAVEL, and CALAMITI, showing the smallest RMSD values: 11 and 7 out of 12 cases for FSL and SPM, respectively. These statistics are 1 and 3 for CALAMITI as well as 0 and 2 for RAVEL. We also observed that WS did not improve the RMSD values of any 12 scanner pairs for FSL, when compared to RAW. However, it performed better for SPM by decreasing these number of worse cases to 6. RAVEL, CALAMITI, and MISPEL deteriorated some of the RMSDs too. Among these methods, MISPEL deteriorated the least number of cases, just one for each of FSL- and SPM-derived volumes.

In summary, we observed that MISPEL outperformed WS, RAVEL, and CALAMITI when FSL and SPM were used for deriving volumes and volumetric variance was used as the harmonization evaluation metric.

3.3.4 Segmentation overlap

Figure 10 shows bar plots that indicate the mean DSC of all paired segmentations in each scanner pair. We calculated the means of DSCs for segmentations of GM and WM tissue types and depicted them for all five methods. Figure 10 contains these sets of bar plots for segmentations extracted through FSL and SPM frameworks in Figures 10a and 10b, respectively. DSC shows the overlap of two paired segmentations. A good harmonization method would result in an increased mean of DCSs for all scanner pairs, with 1 indicating the highest.

We observed in Figure 10 that scanner effects exist in RAW segmentations of



(a) **FSL framework.**

(b) **SPM framework.**

Figure 10: **Dice similarity score (DSC) bar plots.** Bar plots indicate the means of DSCs of all paired segmentations in scanner pairs. These values were calculated for segmentations of GM and WM tissue types and depicted for all four methods. These set of bar plots were depicted for volumes extracted through two segmentation frameworks: (a) FSL and (b) SPM. A harmonization method is expected to result in high mean of DSCs. MISPEL outperformed WS, RAVEL, and CALAMITI by having the largest DSC means for all scanner pairs in both FSL and SPM. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

both FSL and SPM and appeared as relatively low means of DSC values. MISPEL outperformed WS, RAVEL, and CALAMITI in harmonization, displaying the largest means of DSC: 12 and 10 cases for FSL and SPM, respectively. We compared the DSC distributions of MISPEL with their equivalents in RAW using paired t-test and all improvements of MISPEL over RAW were statistically significant. Results also showed that while WS did not change any of the means for FSL, it did better for SPM by increasing the means for 6 of the cases. RAVEL performed slightly better than WS by increasing 9 and decreasing 3 of the DSC means for FSL and improved 9 cases for SPM. On the other hand, CALAMITI did not deteriorate any of DSCs and resulted in 4 largest means for SPM, 2 of which are shared with MISPEL. Using paired t-test, we observed that these DSCs were statistically significantly larger than that of their RAW equivalents.

In summary, MISPEL outperformed WS and RAVEL, when FSL and SPM were used as segmentation frameworks and segmentation overlap was used as the harmonization evaluation metric.

4 Discussion

In this study, we proposed MISPEL, a supervised deep harmonization technique for removing scanner effects from images of multiple scanners, while preserving their biological and anatomical information. Unlike other supervised or unsupervised methods, MISPEL is a multi-scanner method mapping images to a scanner *middle-ground* space in which images are harmonized. We evaluated MISPEL against commonly used intensity normalization and harmonization methods (White Stripe, RAVEL, and CALAMITI) using a comprehensive set of evaluation criteria including image similarity, GM-WM tissue contrast, and tissue segmentation similarity in a dataset of matched T1 MR images acquired from 4 different 3T scanners. We found that (1) scanner effects appear in our dataset as dissimilarity in image appearance/contrast, GM-WM contrast, and tissue type volumetric and segmentation distributions; (2) White Stripe normalized images, but did not achieve harmonization; (3) RAVEL and CALAMITI achieved harmonization to some extent but was outperformed by MISPEL; (4) MISPEL outperformed other two methods in harmonization.

Based on the evaluated harmonization metrics, we observed that images of GE were more similar to those of Philips and images of SiemensP showed more similarity to SiemensT's. We also observed that scanner effects appeared mainly as dissimilarity between pairs of GE or Philips and SiemensP or SiemensT. We observed that removing intensity unit effects using White Stripe successfully normalized images (Supplementary Figure 1) and resulted in improved image similarity, but did not majorly enhance other metrics we used for evaluating harmonization. The relative failure to harmonize may be due to the fact that White Stripe is an individual-level method. Scaling and centering the intensity distributions does not necessarily remove scanner effects; on the contrary, over-matching distributions could result in removal of other source of variability that could be of interest (Fortin et al., 2016).

Our results also show that RAVEL achieved harmonization to some extent relative to White Stripe, but was outperformed by MISPEL. RAVEL increased the similarity of images in their appearance/contrast, GM-WM contrast, and tissue type volumes and segmentation overlap when SPM framework was used. Larger variability and inconsistent behavior of RAVEL in harmonization of images across subjects was reported in (Torbati et al., 2021a), when RAVEL was used for harmonizing paired images of GE 1.5T and Siemens 3T scanners and FreeSurfer was used. Such inconsistency for WS-normalized and RAVEL-harmonized images was also observed as large SD values, when comparing volumetric differences based on FSL segmentation (Table 2 and Figure 8a).

For having fair comparison with CALAMITI as an unsupervised harmonization method, we used it in a supervised manner by applying it to our inter-scanner paired dataset instead of the inter-modality paired data. This should also help CALAMITI to perform better harmonization, confirmed by Zuo et al. (2021) that CALAMITI requires to be trained on inter-scanner paired images too. Results showed that CALAMITI achieved harmonization to some extent relative to White Stripe, but was outperformed by MISPEL. CALAMITI improved similarity of images in their appearance/contrast, GM-WM contrast, and tissue type volumes and segmentation overlap. These improvements were greater for image similarity and segmentation overlap, for which CALAMITI is the second-best model. However, CALAMITI resulted in less improvements for tissue type volumetric similarity by improving just few cases for each of FSL- and SPM-derived volumes and did worse than RAVEL (Table 2 and Figure 8). When SSIM and DSC were used for evaluating image similarity and segmentation overlap in paired dataset respectively, Zuo et al. (2021) observed that CALAMITI outperformed two unsupervised harmonization methods (Modanwal et al., 2020; Dewey et al., 2020). However, CALAMITI did not show any statistically significant improvement over non-harmonized data, when percentage volume difference was studied as volumetric bias in paired data and SLANT segmentation tool (Huo et al., 2019) used for extracting brain summary measures, including ventricles, cerebellum GM, cerebrum GM, caudate, thalamus, putamen, brainstem, cerebellum WM, and cerebrum WM.

MISPEL outperformed White Stripe, RAVEL, and CALAMITI based on all harmonization evaluation criteria. Figures 5, 6, and 7 show that MISPEL mapped images to an average harmonized space, in which GE and Philips images were more similar to those of SiemensP and SiemensT, in terms of GM-WM contrast and tissue type volumetric distributions. It should be noted that no directed mapping or a *target* scanner was selected for MISPEL harmonization, and MISPEL does not require a selected *target*. In fact, MISPEL naturally finds this average space. GE and Philips images were made more similar to SiemensP and SiemensT, with relatively minimal change made to SiemensP and SiemensT by MISPEL, likely due to SiemensP and SiemensT images being most similar and therefore biasing the average image space found by MISPEL. Results from volumetric and segmentation evaluations also show that MISPEL can perform harmonization regardless of the segmentation framework. It showed improvement for both segmentation platforms tested, FSL and SPM, which have been shown to largely differ in their segmentation results (Tudorascu et al., 2016) even in healthy volunteers.

Our study adds to the growing harmonization literature by (1) presenting MISPEL, a supervised multi-scanner harmonization method; (2) introducing a multi-scanner matched dataset of four 3T scanners, (3) providing a comprehensive set of experiments assessing scanner effects and evaluating harmonization, and (4) comparing MISPEL with White Stripe, RAVEL, and CALAMITI and analyzing the behavior of these methods in terms of harmonization. Limitations of our study include the use of a single matched-scan cohort. The generalizability of MISPEL to unmatched multi-scanner data, relative to existing and commonly used normalization and harmonization methods, was not assessed. As future work, we will study the generalizability of MISPEL to other matched datasets with different degrees of scanner effects, such as paired GE 1.5T and Siemens 3T data (Torbati et al., 2021a), as well as unmatched multi-scanner datasets. We will also study MISPEL across other modalities, such as Fluid-attenuated inversion recovery (FLAIR).

Declaration of Competing Interest

The authors declare that they have no competing interests.

Credit authorship contribution statement

Mahbaneh Eshaghzadeh Torbati: Conceptualization, Methodology, Writing – original draft. Davneet S. Minhas: Methodology, Writing – review & editing. Charles M. Laymon: Writing – review & editing. Pauline Maillard: Writing – review & editing. James D. Wilson: Writing – review & editing. Chang-Le Chen: Writing – review & editing. Ciprian M. Crainiceanu: Methodology, Writing – review & editing. Charles S. DeCarli: Funding acquisition, Writing – review & editing. Seong Jae Hwang: Funding acquisition, Methodology, Conceptualization, Supervision, Writing – review & editing. Dana L. Tudorascu: Funding acquisition, Methodology, Conceptualization, Supervision, Writing – review & editing.

Acknowledgments

This work was supported by the following NIH/NIA grants: R01 AG063752 (D.L. Tudorascu), P30 AG10129 and UH3 NS100608 (C. DeCarli), and the University of Pittsburgh Alzheimer's Disease Research Center Grant P30 AG066468 (S. Hwang). S. Hwang was also supported by Institute of Information & communications Technology

Planning & Evaluation (IITP) grant funded by the Korea government (MSIT), Artificial Intelligence Graduate Program, Yonsei University, under Grant 2020-0-01361-003, and the Yonsei University Research Fund of 2022 (2022-22-0131).

References

Christopher R Madan. Scan once, analyse many: using large open-access neuroimaging datasets to understand the brain. Neuroinformatics, pages 1–29, 2021.

Raymond A Mar, R Nathan Spreng, and Colin G DeYoung. How to produce personality neuroscience research with high statistical power and low additional cost. Cognitive, Affective, & Behavioral Neuroscience, 13(3):674–685, 2013.

Christopher R Madan. Advances in studying brain morphology: The benefits of open-access data. Frontiers in human neuroscience, 11:405, 2017.

Michael P Milham, R Cameron Craddock, Jake J Son, Michael Fleischmann, Jon Clucas, Helen Xu, Bonhwang Koo, Anirudh Krishnakumar, Bharat B Biswal, F Xavier Castellanos, et al. Assessment of the impact of shared brain imaging data on the scientific literature. Nature Communications, 9(1):1–7, 2018.

Frithjof Kruggel, Jessica Turner, L Tugan Muftuler, Alzheimer's Disease Neuroimaging Initiative, et al. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the adni cohort. Neuroimage, 49(3): 2123–2133, 2010.

Olivier Potvin, April Khademi, Isabelle Chouinard, Farnaz Farokhian, Louis Dieumegarde, Ilana Leppert, Rick Hoge, Maria Natasha Rajah, Pierre Bellec, Simon Duchesne, et al. Measurement variability following mri system upgrade. Frontiers in neurology, 10:726, 2019.

Mahbaneh Eshaghzadeh Torbati, Davneet S Minhas, Ghasan Ahmad, Erin E O'Connor, John Muschelli, Charles M Laymon, Zixi Yang, Ann D Cohen, Howard J Aizenstein, William E Klunk, et al. A multi-scanner neuroimaging data harmonization using ravel and combat. NeuroImage, 245:118703, 2021a.

Russell T Shinohara, Elizabeth M Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J Mateen, Peter A Calabresi, Samson Jarso, Dzung L Pham, Daniel S Reich, and Ciprian M Crainiceanu. Australian imaging biomarkers lifestyle flagship study of ageing, and alzheimer's disease neuroimaging initiative. statistical normalization techniques for magnetic resonance imaging. Neuroimage Clin, 6(9), 2014a.

Russell T Shinohara, Jiwon Oh, Govind Nair, Peter A Calabresi, Christos Davatzikos, Jimit Doshi, Roland G Henry, Gloria Kim, Kristin A Linn, Nico Papinutto, et al. Volumetric analysis from a harmonized multisite brain mri study of a single subject with multiple sclerosis. American Journal of Neuroradiology, 38(8):1501–1509, 2017.

J Wrobel, ML Martin, R Bakshi, PA Calabresi, M Elliot, D Roalf, RC Gur, RE Gur, RG Henry, G Nair, et al. Intensity warping for multisite mri harmonization. NeuroImage, 223:117242, 2020.

Russell T Shinohara, Elizabeth M Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J Mateen, Peter A Calabresi, Samson Jarso, Dzung L Pham, Daniel S Reich, Ciprian M Crainiceanu, et al. Statistical normalization techniques for magnetic resonance imaging. NeuroImage: Clinical, 6:9–19, 2014b.

Mohak Shah, Yiming Xiao, Nagesh Subbanna, Simon Francis, Douglas L Arnold, D Louis Collins, and Tal Arbel. Evaluating intensity normalization on mris of human brain with multiple sclerosis. Medical image analysis, 15(2):267–282, 2011.

Jean-Philippe Fortin, Elizabeth M Sweeney, John Muschelli, Ciprian M Crainiceanu, Russell T Shinohara, Alzheimer's Disease Neuroimaging Initiative, et al. Removing inter-subject technical variability in magnetic resonance imaging studies. NeuroImage, 132:198–212, 2016.

Nicola K Dinsdale, Mark Jenkinson, and Ana IL Namburete. Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal. NeuroImage, 228:117689, 2021.

Hidemasa Takao, Naoto Hayashi, and Kuni Ohtomo. Effects of study design in multi-scanner voxel-based morphometry studies. Neuroimage, 84:133–140, 2014.

Xiao Han, Jorge Jovicich, David Salat, Andre van der Kouwe, Brian Quinn, Silvester Czanner, Evelina Busa, Jenni Pacheco, Marilyn Albert, Ronald Killiany, et al. Reliability of mri-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. Neuroimage, 32(1): 180–194, 2006.

Hidemasa Takao, Naoto Hayashi, and Kuni Ohtomo. Effect of scanner in longitudinal studies of brain volume changes. Journal of Magnetic Resonance Imaging, 34(2): 438–444, 2011.

Jorge Jovicich, Silvester Czanner, Douglas Greve, Elizabeth Haley, Andre van Der Kouwe, Randy Gollub, David Kennedy, Franz Schmitt, Gregory Brown, James MacFall, et al. Reliability in multi-site structural mri studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage, 30(2):436–443, 2006.

Blake E Dewey, Can Zhao, Jacob C Reinhold, Aaron Carass, Kathryn C Fitzgerald, Elias S Sotirchos, Shiv Saidha, Jiwon Oh, Dzung L Pham, Peter A Calabresi, et al. Deepharmony: A deep learning approach to contrast harmonization across scanner changes. Magnetic resonance imaging, 64:160–170, 2019.

Blake E Dewey, Lianrui Zuo, Aaron Carass, Yufan He, Yihao Liu, Ellen M Mowry, Scott Newsome, Jiwon Oh, Peter A Calabresi, and Jerry L Prince. A disentangled latent space for cross-site mri harmonization. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 720–729. Springer, 2020.

Mengting Liu, Piyush Maiti, Sophia Thomopoulos, Alyssa Zhu, Yaqiong Chai, Hosung Kim, and Neda Jahanshad. Style transfer using generative adversarial networks for multi-site mri harmonization. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 313–322. Springer, 2021.

Stenzel Cackowski, Emmanuel L Barbier, Michel Dojat, and Thomas Christen. Imunity: a generalizable vae-gan solution for multicenter mr image harmonization. arXiv preprint arXiv:2109.06756, 2021.

Lianrui Zuo, Blake E Dewey, Yihao Liu, Yufan He, Scott D Newsome, Ellen M Mowry, Susan M Resnick, Jerry L Prince, and Aaron Carass. Unsupervised mr harmonization by learning disentangled representations using information bottleneck theory. NeuroImage, 243:118569, 2021.

Andrew A Chen, Joanne C Beer, Nicholas J Tustison, Philip A Cook, Russell T Shinohara, Haochang Shou, Alzheimer's Disease Neuroimaging Initiative, et al. Removal of scanner effects in covariance improves multivariate pattern analysis in neuroimaging data. bioRxiv, page 858415, 2020a.

Maria Ines Meyer, Ezequiel de la Rosa, Koen Van Leemput, and Diana M Sima. Relevance vector machines for harmonization of mri brain volumes using image descriptors. In OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging, pages 77–85. Springer, 2019.

Aaron Alexander-Bloch, Liv Clasen, Michael Stockman, Lisa Ronan, Francois Lalonde, Jay Giedd, and Armin Raznahan. Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo mri. Human brain mapping, 37 (7):2385–2397, 2016.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics, 8(1):118–127, 2007.

Joanne C Beer, Nicholas J Tustison, Philip A Cook, Christos Davatzikos, Yvette I Sheline, Russell T Shinohara, Kristin A Linn, Alzheimer's Disease Neuroimaging Initiative, et al. Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data. Neuroimage, 220:117129, 2020.

Andrew A Chen, Joanne C Beer, Nicholas J Tustison, Philip A Cook, Russell T Shinohara, Haochang Shou, Alzheimer's Disease Neuroimaging Initiative, et al. Removal of scanner effects in covariance improves multivariate pattern analysis in neuroimaging data. bioRxiv, page 858415, 2020b.

Raymond Pomponio, Guray Erus, Mohamad Habes, Jimit Doshi, Dhivya Srinivasan, Elizabeth Mamourian, Vishnu Bashyam, Ilya M Nasrallah, Theodore D Satterthwaite, Yong Fan, et al. Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage, 208:116450, 2020.

Maxwell Reynolds, Tigmanshu Chaudhary, Mahbaneh Eshaghzadeh Torbati, Dana L Tudorascu, and Kayhan Batmanghelich. Combat harmonization: Empirical bayes versus fully bayes approaches. bioRxiv, 2022.

Jean-Philippe Fortin, Drew Parker, Birkan Tunç, Takanori Watanabe, Mark A Elliott, Kosha Ruparel, David R Roalf, Theodore D Satterthwaite, Ruben C Gur, Raquel E Gur, et al. Harmonization of multi-site diffusion tensor imaging data. Neuroimage, 161:149–170, 2017.

Jean-Philippe Fortin, Nicholas Cullen, Yvette I Sheline, Warren D Taylor, Irem Aselcioglu, Philip A Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J McGrath, et al. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage, 167:104–120, 2018.

Dylan M Nielson, Francisco Pereira, Charles Y Zheng, Nino Migineishvili, John A Lee, Adam G Thomas, and Peter A Bandettini. Detecting and harmonizing scanner differences in the abcd study-annual release 1.0. BioRxiv, page 309260, 2018.

Meichen Yu, Kristin A Linn, Philip A Cook, Mary L Phillips, Melvin McInnis, Maurizio Fava, Madhukar H Trivedi, Myrna M Weissman, Russell T Shinohara, and Yvette I Sheline. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fmri data. Human brain mapping, 39(11): 4213–4227, 2018.

Joaquim Radua, Eduard Vieta, Russell Shinohara, Peter Kochunov, Yann Quidé, Melissa J Green, Cynthia S Weickert, Thomas Weickert, Jason Bruggemann, Tilo Kircher, et al. Increased power by harmonizing structural mri site differences with the combat batch adjustment method in enigma. NeuroImage, 218:116956, 2020.

Joseph J Foy, Hania A Al-Hallaq, Vincent Grekoski, Tri Tran, Kharina Guruvadoo, Samuel G Armato III, and William F Sensakovic. Harmonization of radiomic feature variability resulting from differences in ct image acquisition and reconstruction: assessment in a cadaveric liver. Physics in Medicine & Biology, 65(20):205008, 2020.

Gourav Modanwal, Adithya Vellal, Mateusz Buda, and Maciej A Mazurowski. Mri image harmonization using cycle-consistent generative adversarial network. In Medical Imaging 2020: Computer-Aided Diagnosis, volume 11314, page 1131413. International Society for Optics and Photonics, 2020.

Jie Zhong, Ying Wang, Jie Li, Xuetong Xue, Simin Liu, Miaomiao Wang, Xinbo Gao, Quan Wang, Jian Yang, and Xianjun Li. Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: application to neonatal white matter development. Biomedical engineering online, 19(1):1–18, 2020.

Siyuan Liu and Pew-Thian Yap. Learning multi-site harmonization of magnetic resonance images without traveling human phantoms. arXiv preprint arXiv:2110.00041, 2021.

Daniel Moyer, Greg Ver Steeg, Chantal MW Tax, and Paul M Thompson. Scanner invariant representations for diffusion mri harmonization. Magnetic resonance in medicine, 84(4):2174–2189, 2020.

Shahab Aslani, Vittorio Murino, Michael Dayan, Roger Tam, Diego Sona, and Ghassan Hamarneh. Scanner invariant multiple sclerosis lesion segmentation from mri. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pages 781–785. IEEE, 2020.

Daniel Moyer and Polina Golland. Harmonization and the worst scanner syndrome. arXiv preprint arXiv:2101.06255, 2021.

Y Zhang, M Brady, and S Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging, 20(1):45–57, 2001.

John Ashburner and Karl J Friston. Unified segmentation. Neuroimage, 26(3):839–851, 2005.

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

Kenichi Oishi, Andreia Faria, Hangyi Jiang, Xin Li, Kazi Akhter, Jiangyang Zhang, John T Hsu, Michael I Miller, Peter CM van Zijl, Marilyn Albert, et al. Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and alzheimer's disease participants. Neuroimage, 46(2):486–499, 2009.

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

Mahbaneh Eshaghzadeh Torbati, Dana L Tudorascu, Davneet S Minhas, Pauline Maillard, Charles S DeCarli, and Seong Jae Hwang. Multi-scanner harmonization of paired neuroimaging data via structure preserving embedding learning. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3284–3293, 2021b.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

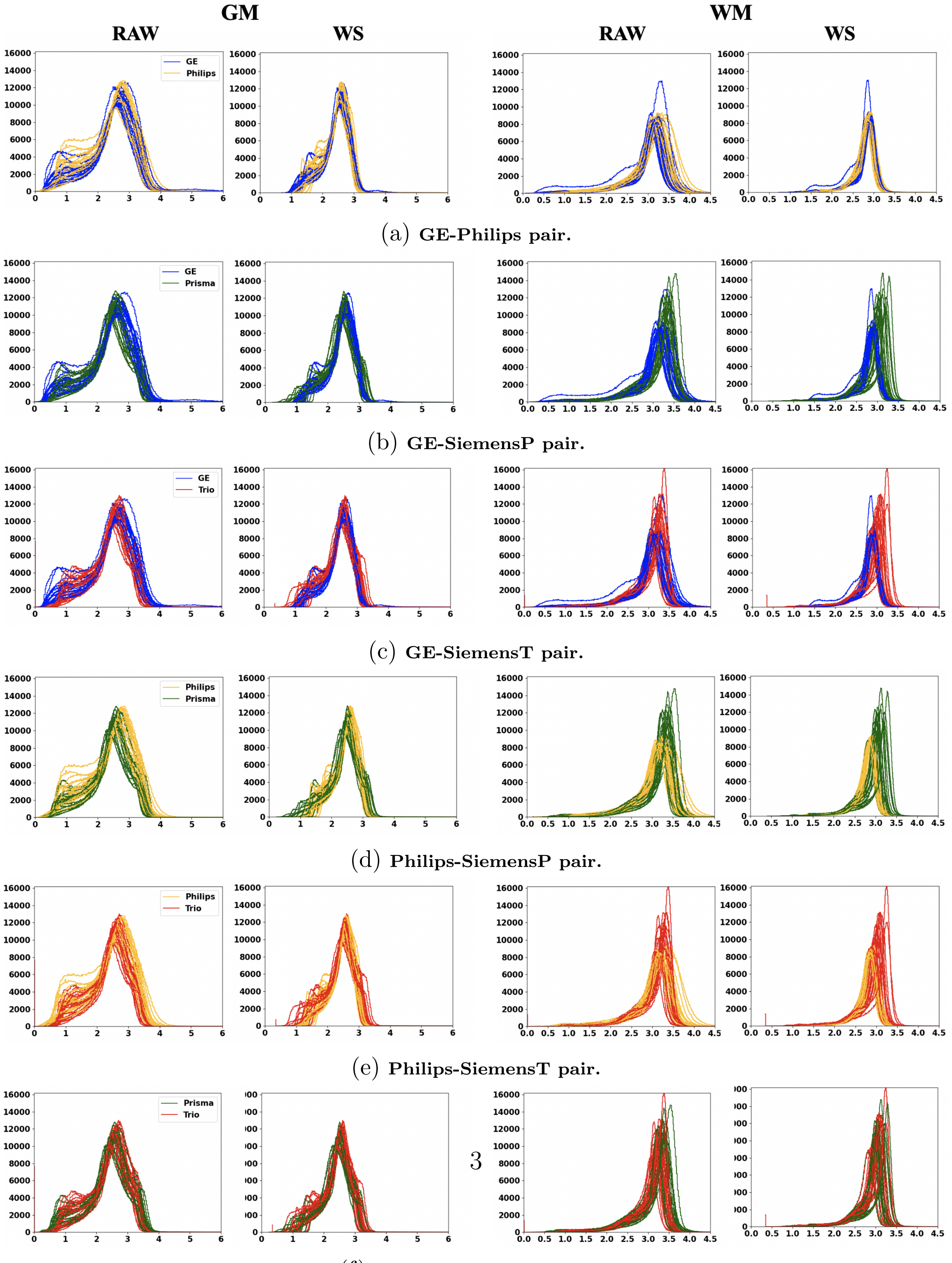
Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Donna Wilcock, Gregory Jicha, Deborah Blacker, Marilyn S Albert, Lina M D'Orazio, Fanny M Elahi, Myriam Fornage, Jason D Hinman, Janice Knoefel, Joel Kramer, et al. Markvcid cerebral small vessel consortium: I. enrollment, clinical, fluid protocols. Alzheimer's & Dementia, 17(4):704–715, 2021.

Dana L Tudorascu, Helmet T Karim, Jacob M Maronge, Lea Alhilali, Saeed Fakhran, Howard J Aizenstein, John Muschelli, and Ciprian M Crainiceanu. Reproducibility and bias in healthy brain segmentation: comparison of two popular neuroimaging platforms. Frontiers in neuroscience, 10:503, 2016.

Yuankai Huo, Zhoubing Xu, Yunxi Xiong, Katherine Aboud, Prasanna Parvathaneni, Shunxing Bao, Camilo Bermudez, Susan M Resnick, Laurie E Cutting, and Bennett A Landman. 3d whole brain segmentation using spatially localized atlas network tiles. NeuroImage, 194:105–119, 2019.

5 Appendix



(a) **GE-Philips pair.**

(b) **GE-SiemensP pair.**

(c) **GE-SiemensT pair.**

(d) **Philips-SiemensP pair.**

(e) **Philips-SiemensT pair.**

(f) **SiemensP-SiemensT pair.**

Supplementary Figure 1: Histograms of gray matter (GM) and white matter (WM) voxels for RAW and White Stripe (WS)-normalized images of all subjects. These histograms were plotted for all 6 scanner pairs. WS makes the plots more centered, overlapped and therefore comparable across subjects. WS usually outputs images with negative intensity values. for plotting the histograms, we shifted the WS-normalized images to have positive intensity values. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.