DISARM++: Beyond scanner-free harmonization

Caldera, L.; Cavinato, L.; Cirone, A.; Cama, I.; Garbarino, S.; Lodi, R.; Tagliavini, F.; Nigri, A.; De Francesco, S.; Cappozzo, A.; Piana, M.; Ieva, F.;

DISARM++: Beyond scanner-free harmonization

Luca Caldera[a], Lara Cavinato[a], Alessio Cirone[d], Isabella Cama[b,c], Sara Garbarino[b,d], Raffaele Lodi[e,f], Fabrizio Tagliavini[g], Anna Nigri[h], Silvia De Francesco[i], Andrea Cappozzo[j], Michele Piana[b,d], Francesca Ieva[a,k], RIN-Neuroimaging Network[l], Alzheimer's Disease Neuroimaging Initiative[1]

[a] *MOX, Department of Mathematics, Politecnico di Milano, Via Bonardi 9, Milan, 20133, Italy, Italy*
[b] *Department of Mathematics, Università di Genova, Via Dodecaneso 35, Genova, 16146, Italy, Italy*
[c] *Department of Neuroscience, Rehabilitation, Ophthalmology, Genetics and Maternal-Child Sciences, Università di Genova, Largo Paolo Daneo 3, Genova, 16132, Italy, Italy*
[d] *IRCCS Ospedale Policlinico San Martino, Largo R. Benzi 10, Genova, 16132, Italy, Italy*
[e] *Department of Biomedical and Neuromotor Sciences, University of Bologna, Via Zamboni 33, Bologna, 40126, Italy, Italy*
[f] *IRCCS Institute of Neurological Science of Bologna, Bellaria Hospital, Via Altura 3, Bologna, 40139, Italy, Italy*
[g] *Unit of Neurology (V) and Neuropathology, IRCCS Istituto Neurologico Carlo Besta, Via Celoria 11, Milan, 20133, Italy, Italy*
[h] *Neuroradiology Unit, IRCCS Istituto Neurologico Carlo Besta, Via Celoria 11, Milan, 20133, Italy, Italy*
[i] *Laboratory of Neuroinformatics, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Via Pilastroni 4, Brescia, 25125, Italy, Italy*
[j] *Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1, Milan, 20123, Italy, Italy*
[k] *Health Data Science Centre, Human Technopole, V.le Rita Levi-Montalcini 1, Milan, 20157, Italy, Italy*
[l] *RIN-Neuroimaging Network, Via Clericetti 2, Milan, 20133, Italy, Italy*

Abstract

Harmonization of T1-weighted MR images across different scanners is crucial

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (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.

for ensuring consistency in neuroimaging studies. This study introduces a novel approach to direct image harmonization, moving beyond feature standardization to ensure that extracted features remain inherently reliable for downstream analysis. Our method enables image transfer in two ways: (1) mapping images to a scanner-free space for uniform appearance across all scanners, and (2) transforming images into the domain of a specific scanner used in model training, embedding its unique characteristics. Our approach presents strong generalization capability, even for unseen scanners not included in the training phase. We validated our method using MR images from diverse cohorts, including healthy controls, traveling subjects, and individuals with Alzheimer's disease (AD). The model's effectiveness is tested in multiple applications, such as brain age prediction ($R^2 \simeq 0.60 \pm 0.05$), biomarker extraction, AD classification (Test Accuracy $\simeq 0.86 \pm 0.03$), and diagnosis prediction (AUC $\simeq 0.95$). In all cases, our harmonization technique outperforms state-of-the-art methods, showing improvements in both reliability and predictive accuracy. Moreover, our approach eliminates the need for extensive preprocessing steps, such as skull-stripping, which can introduce errors by misclassifying brain and non-brain structures. This makes our method particularly suitable for applications that require full-head analysis, including research on head trauma and cranial deformities. Additionally, our harmonization model does not require retraining for new datasets, allowing smooth integration into various neuroimaging workflows. By ensuring scanner-invariant image quality, our approach provides a robust and efficient solution for improving neuroimaging studies across diverse settings. The code is available at this link.

Keywords: Image harmonization, I2I Translation, Magnetic Resonance Imaging, Noise Disentanglement, Scanner-free Imaging, Downstream Tasks

1. Introduction

As brain Magnetic Resonance Imaging (MRI) datasets from various research centers become increasingly accessible, there is a growing opportunity to gain valuable insights into brain-related diseases. These insights have the potential to enhance medical practices by providing robust statistical evidence to be translated into clinical practice. However, variations in MRI data across different centers and scanners, due to unstandardized protocols, scanner- and acquisition-specific variabilities, can lead to significant inconsistencies in the extracted biomarkers, thereby affecting their repeatability and reproducibility. Differences in hardware, software configurations, calibration procedures, maintenance practices, and operators experience can cause MRI scanners to produce images with varying contrast, brightness, and spatial resolution, i.e., voxel intensity distribution. This variability, particularly in multicenter studies, can introduce confounding effects that compromise the reliability of the results [39, 36, 50]. To address this issue, it is essential to harmonize MRI data across centers and scanners to ensure consistency and comparability of the datasets, thus strengthening the integrity of subsequent analyzes. By minimizing inter- and intra-scanner variability, harmonization enables researchers to reliably combine data from multiple sites, facilitating the development of robust machine learning and deep learning models that depend on large, high-quality datasets.

2. Related Works

Several methodologies have been proposed to harmonize images and specifically multicenter MRI data. The approaches can generally be divided into two primary categories: feature-based and image-based approaches.

2.1. Feature-Based Approaches

Feature-based harmonization aims to align extracted features, such as cortical thickness, functional connectivity, or diffusion metrics, across batches rather than directly modifying the original images. A key method in this category is the widely used ComBat and its extensions, which leverage empirical Bayes frameworks to adjust for mean and variance shifts caused by batch effects [41, 30]. These methods have been extensively validated across various datasets and data types. Typically, they rely on a linear model framework to harmonize features while preserving biologically meaningful variance. Extensions like ComBat-GAM [29] and CovBat [6] further refine these approaches by addressing nonlinear covariate effects and multivariate dependencies, respectively. Deep learning approaches have also been explored for feature-based harmonization. For example, Conditional Variational Autoencoders (CVAEs) have been applied to learn batch-invariant latent representations of imaging features, facilitating their reconstruction in a harmonized space [2, 24, 5].

2.2. Image-Based Approaches

Image-based methods operate directly on raw imaging data, using machine learning techniques to adjust for batch effects at the voxel or pixel level, creating visual consistency across datasets collected from multiple centers. In the domain of image-based harmonization, several methods have been explored. These techniques can be categorized into three main frameworks: Transformers, Image-to-Image (I2I) translation, and Style Transfer. Transformers [10] are a class of models built around a self-attention mechanism, which enables them to capture long-range dependencies within the data. However, despite their success in many applications, transformers have been noted for their limitations in retaining fine-grained, high-frequency details, which could be critical for accurately identifying biology- and pathology-related features. For instance, studies have shown that transformers may overlook important high-frequency information, thus potentially missing crucial signals relevant for medical imaging analysis [45]. Image-to-Image (I2I) translation involves mapping an input image to an output image in a way that preserves specific semantic properties while adjusting the image to match the target domain. The core of this method lies in generative models that are trained to produce images that resemble those drawn from the target distribution. I2I methods can be further classified according to the type of supervision (e.g., supervised, unsupervised) or the nature of the translation process (e.g., one-to-one, one-to-many, many-to-many) [1]. Style Transfer [20], on the other hand, treats the harmonization process as a domain adaptation problem, focusing on transferring the style of an image from one dataset to another. In this fully unsupervised framework, the primary goal is to maintain the content of the original images while adjusting their style to align with the desired target. Image-based harmonization methods have already demonstrated considerable potential. Notable examples include CALAMITI [49], MURD [21], IGUANe [32], and STGAN [7], which have advanced the field by applying deep learning techniques to tackle the challenges of multi-center data harmonization, thus improving the reproducibility and reliability of medical imaging analyses.

2.3. Contribution and Differences from the Conference Paper

Despite significant advancements, the existing approaches face several challenges. Feature-based methods rely on accurate feature extraction and often assume simplistic statistical relationships, which may fail to capture the inherent complexity of imaging datasets. Image-based methods, while promising, often require large training datasets and struggle to balance visual consistency with the preservation of biologically meaningful information. Moreover, they rely on heavy pre-processing steps and do not generalize to unseen cases.

To overcome the aforementioned issues, in this work we introduce DISARM++, a novel model for harmonizing 3D MR images by addressing inter-scanner variability. DISARM++ disentangles anatomical structure from scanner-specific information to generate *scanner-free* images. This approach preserves the original anatomical structures and biologically informative features, ensuring robust generalizability across different scanners. Our goal is to enable researchers to integrate MRI data from diverse sources without concerns about inconsistencies, while enhancing the extraction of biologically meaningful information. To achieve this, we develop a model that can harmonize images without the need for a new training phase for previously unseen data, allowing seamless integration into various neuroimaging workflows. Unlike traditional preprocessing pipelines, our method retains full-head information without the need for skull-stripping, providing a more comprehensive and less intrusive preprocessing.

The present work is an extended version of the conference paper in [4], incorporating several significant improvements: (1) we refined the network architecture and introduced a new loss function, enhancing performance as demonstrated in the ablation study; (2) we trained the model on a larger dataset of MR images; (3) we evaluated the model on a broader range of MR images, including data from healthy individuals, patients with Alzheimer's Disease (AD), and traveling subjects, comparing our approach to state-of-the-art methods; and (4) we conducted several comprehensive downstream analyses to further benchmark our proposal against existing approaches.

3. Methodology

In this section, we present the novel proposal designed to harmonize 3D T1-weighted MRI data that extends the DISARM framework introduced in [4]. The proposed model belongs to the category of I2I translation methods, with its baseline architecture inspired by [18]. The model aims to mitigate batch effects in clinical images acquired from different sources by directly working on the images at the voxel level. Specifically, our approach focuses on the *scanner-free* generation of 3D MRI data with two key objectives: (1) ensuring robust generalizability across a wide range of scanners, including those not seen during training, and (2) eliminating the need for time-consuming preprocessing steps.

3.1. Mathematical Formulation

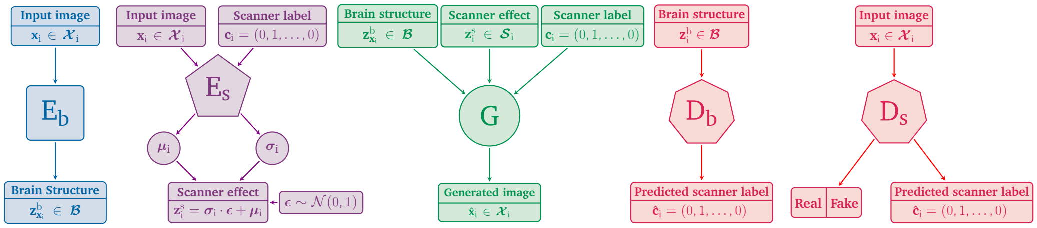
Consider a set of MR images $\mathcal{X} = \bigcup_{i=1}^{N} \mathcal{X}_i \in \mathbb{R}^{1 \text{ x H x W x D}}$, where $\mathcal{X}_i$ represents the collection of images acquired from the i-th scanner domain among N distinct domains. Since the images are grayscale 3D images, they have dimensions of 1 (channel), H (height), W (width), and D (depth). We assume that the images can be disentangled into two distinct latent spaces $(\mathcal{B}, \mathcal{S})$. Here, $\mathcal{B}$ represents the space that encodes information related to the anatomical structure of the brain, while $\mathcal{S}$ represents the scanner space, which aims to capture information about scanner effects. Thus, an image $\boldsymbol{x}$ drawn from $\mathcal{X}$ can be obtained as a combination of $\mathcal{B}$ and $\mathcal{S}$. The *scanner-free* harmonization involves eliminating scanner-specific effects by replacing $\mathcal{S}$ with random Gaussian noise $\mathcal{N}(0, 1)$, and combining it with $\mathcal{B}$ into a generator. This process aims to remove scanner-dependent effects while preserving only the anatomical noise-free features encoded in $\mathcal{B}$. We denote the space in which the images are transferred after *scanner-free* harmonization as $\mathcal{F}$.

3.2. Network Architecture

The model architecture consists of five modules, as shown in Figure 1. Let $\mathcal{C} = \{\mathbf{c} \in \{0, 1\}^N : \|\mathbf{c}\|_1 = 1\}$ represent the space of one-hot encoded vectors that describe scanner domains, and $\mathcal{L} = \{\mathbf{l} \in \{0, 1\}^2 : \|\mathbf{l}\|_1 = 1\}$ represent the space of labels distinguishing real and generated images. The brain encoder (Figure 1a)), $E_b : \mathcal{X} \to \mathcal{B}$, maps an image $\boldsymbol{x} \in \mathcal{X}$ to a lower-dimensional space $\mathcal{B}$, encoding information related to the anatomical structure into a latent vector $\boldsymbol{z}^b_{\boldsymbol{x}}$. The scanner encoder (Figure 1b)), $E_s : (\mathcal{X}, \mathcal{C}) \to \mathcal{S}$, takes as input an image $\boldsymbol{x} \in \mathcal{X}$ and its associated scanner label $\boldsymbol{c} \in \mathcal{C}$. Operating as a variational autoencoder, it aims to capture the scanner effect in the image by producing a parametric distribution that models such effect. Specifically, the scanner encoder outputs the mean and variance, characterizing its distribution. For an image drawn from $\mathcal{X}_i$, which is acquired using the i-th scanner, the corresponding latent scanner effect vector is denoted as $\boldsymbol{z}^s_i$ and is given by:

$$\boldsymbol{z}^s_i = \boldsymbol{\sigma}_i \cdot \boldsymbol{\epsilon} + \boldsymbol{\mu}_i, \qquad \boldsymbol{z}^s_i \in \mathcal{S}; \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1). \tag{1}$$

The generator $G : (\mathcal{B}, \mathcal{S}, \mathcal{C}) \to \mathcal{X}$ (Figure 1c)) produces an image $\hat{\boldsymbol{x}} \in \mathcal{X}$ that preserves a specified brain structure within the space $\mathcal{B}$ while incorporating a scanner attribute from the space $\mathcal{S}$ associated to its label within the space $\mathcal{C}$. The brain discriminator (Figure 1d)) $D_b : \mathcal{X} \rightarrow \mathcal{C}$ processes an image $\boldsymbol{x} \in \mathcal{X}$ and aims to predict the scanner label $\boldsymbol{c} \in \mathcal{C}$, indicating the scanner used to acquire the image. Finally, the scanner discriminator $D_s : \mathcal{X} \rightarrow (\mathcal{L}, \mathcal{C})$ (Figure 1e)) takes an image $\boldsymbol{x} \in \mathcal{X}$ or $\hat{\boldsymbol{x}} \in \mathcal{X}$ as input and attempts to determine whether the image is real or generated by the generator, as well as its associated scanner label $\boldsymbol{c} \in \mathcal{C}$.


Input image
$\mathbf{x}_i \in \boldsymbol{\mathcal{X}}_i$
E_b
Brain Structure
$\mathbf{z}^b_{\mathbf{x}_i} \in \boldsymbol{\mathcal{B}}$
Input image
$\mathbf{x}_i \in \boldsymbol{\mathcal{X}}_i$
Scanner label
$\mathbf{c}_i = (0, 1, \ldots, 0)$
E_s
μ_i
σ_i
Scanner effect
$\mathbf{z}^s_i = \boldsymbol{\sigma}_i \cdot \boldsymbol{\epsilon} + \boldsymbol{\mu}_i$
$\epsilon \sim \mathcal{N}(0, 1)$
Brain structure
$\mathbf{z}^b_{\mathbf{x}_i} \in \boldsymbol{\mathcal{B}}$
Scanner effect
$\mathbf{z}^s_i \in \boldsymbol{\mathcal{S}}_i$
Scanner label
$\mathbf{c}_i = (0, 1, \ldots, 0)$
G
Generated image
$\hat{\mathbf{x}}_i \in \boldsymbol{\mathcal{X}}_i$
Brain structure
$\mathbf{z}^b_i \in \boldsymbol{\mathcal{B}}$
D_b
Predicted scanner label
$\hat{\mathbf{c}}_i = (0, 1, \ldots, 0)$
Input image
$\mathbf{x}_i \in \boldsymbol{\mathcal{X}}_i$
D_s
Real
Fake
Predicted scanner label
$\hat{\mathbf{c}}_i = (0, 1, \ldots, 0)$

(a) (b) (c) (d) (e)

Figure 1: From left to right, we have a) the brain encoder E_b, b) the scanner encoder E_s, c) the generator G, d) the brain discriminator D_b and e) the scanner discriminator D_s.

Unlike the baseline DISARM model, the proposed approach processes thinner 3D volumes consisting of 26 slices-wide moving window rather than full 3D image volumes of the MR images. Volumes are then merged at inference time. This modification reduces computational complexity and memory requirements, allowing the inclusion of more sophisticated model layers that enhance feature extraction capabilities while maintaining the ability to learn meaningful representations from spatial contexts. A further key innovation is the integration of channel and spatial attention layers into the modules E_b, G, and D_b. This attention mechanism is crucial for highlighting significant features, thereby enhancing the model's ability to embed anatomical structures. This ensures that critical structural details are better preserved during the generation of harmonized images, leading to higher fidelity in the final output. In addition, a novel loss function is introduced compared to the baseline model, to improve the *scanner-free* space.

3.3. Training Process

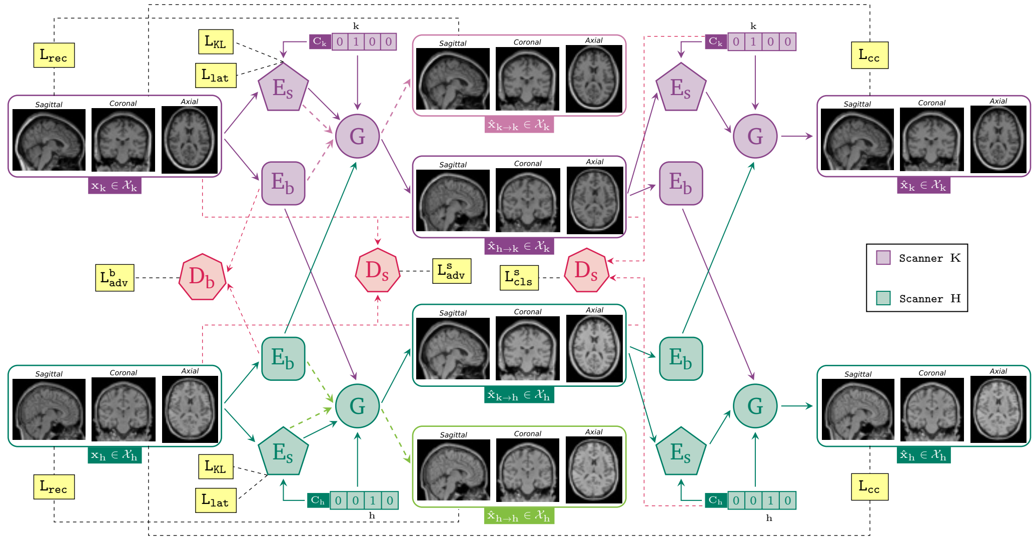


Figure 2: A high-level functional diagram of the model training procedure. The encoders and the generator are single modules, despite being depicted in different colors. The varying colors for the encoders highlight the specific acquisition scanner domain of the images they process. For the generator, the different colors emphasize the scanner domain to which it is transferring the image. The images used to illustrate the procedure are acquired with the `Prisma` (purple) and `Gyroscan Intera` (green) scanners.

In this section, we outline the training procedure referring to Figure 2 and Figure 3. During each iteration, the training procedure involves randomly selecting two scanner domains from the pool of N domains. For clarity, we describe the procedure assuming that two images have been sampled: one from the scanner domain K (represented in purple in Figure 2), denoted by the pair $(\boldsymbol{x}_k, \boldsymbol{c}_k)$, and another from scanner domain H (represented in green in Figure 2), denoted by the pair $(\boldsymbol{x}_h, \boldsymbol{c}_h)$.

In the initial step, for each image $\boldsymbol{x}_i$ (where $i = \{h, k\}$), we extract the anatomical structure embedding $\boldsymbol{z}^b_{\boldsymbol{x}_i} = E_b(\boldsymbol{x}_i)$ using the brain encoder E_b, and the scanner effect embedding $\boldsymbol{z}^s_i = E_s(\boldsymbol{x}_i)$ using the scanner encoder E_s. Note that we denote the scanner effect embedding as $\boldsymbol{z}^s_i$ rather than $\boldsymbol{z}^s_{x_i}$ because it represents the characteristics of the scanner-specific image family as a whole, rather than the noise embedded in an individual image. At this point, the anatomical embeddings $\boldsymbol{z}^b_{\boldsymbol{x}_k}$ and $\boldsymbol{z}^b_{\boldsymbol{x}_h}$ are swapped and used as inputs to the generator. The generator then synthesizes two new images: $\hat{\boldsymbol{x}}_{h\to k} = G(\boldsymbol{z}^b_{\boldsymbol{x}_h}, \boldsymbol{z}^s_k, \boldsymbol{c}_k)$, which combines the anatomical structure of $\boldsymbol{x}_h$ with the scanner effect of $\boldsymbol{x}_k$, and $\hat{\boldsymbol{x}}_{k\to h} = G(\boldsymbol{z}^b_{\boldsymbol{x}_k}, \boldsymbol{z}^s_h, \boldsymbol{c}_h)$, which integrates the scanner effect of $\boldsymbol{x}_h$ with the anatomical structure of $\boldsymbol{x}_k$. Thereafter, the process of swapping anatomical structure embeddings is repeated—this time for the newly generated images $\hat{\boldsymbol{x}}_{h\to k}$ and $\hat{\boldsymbol{x}}_{k\to h}$—resulting in the cyclic reconstruction of the input images as $\hat{\boldsymbol{x}}_k = G(\boldsymbol{z}^b_{\hat{\boldsymbol{x}}_{k\to h}}, \boldsymbol{z}^s_{h\to k}, \boldsymbol{c}_k)$ and $\hat{\boldsymbol{x}}_h = G(\boldsymbol{z}^b_{\hat{\boldsymbol{x}}_{h\to k}}, \boldsymbol{z}^s_{k\to h}, \boldsymbol{c}_h)$. The generator is also employed to generate $\hat{\boldsymbol{x}}_{i\to i} = G(\boldsymbol{z}^b_{\boldsymbol{x}_i}, \boldsymbol{z}^s_i, \boldsymbol{c}_i)$ (where $i = \{h, k\}$) by combining the latent representations extracted from the same image, enabling the direct reconstruction of the input images.

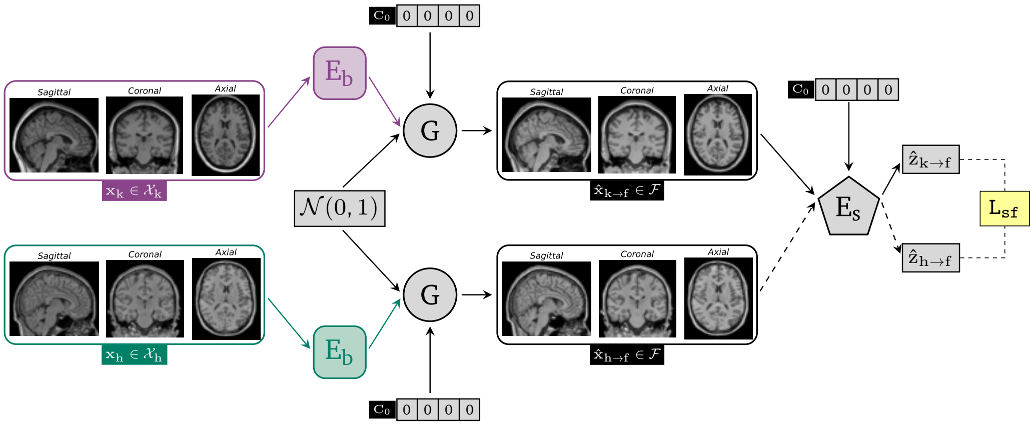

$\mathrm{C_0}$ 0 0 0 0
$\mathrm{E_b}$
G
Sagittal
Coronal
Axial
$\mathrm{x_k} \in \mathcal{X}_\mathrm{k}$
$\mathcal{N}(0,1)$
$\hat{\mathrm{x}}_{\mathrm{k}\to\mathrm{f}} \in \mathcal{F}$
$\mathrm{C_0}$ 0 0 0 0
$\mathrm{E_S}$
$\hat{\mathrm{z}}_{\mathrm{k}\to\mathrm{f}}$
$\mathrm{L_{sf}}$
$\hat{\mathrm{z}}_{\mathrm{h}\to\mathrm{f}}$
Sagittal
Coronal
Axial
$\mathrm{x_h} \in \mathcal{X}_\mathrm{h}$
$\mathrm{E_b}$
G
$\hat{\mathrm{x}}_{\mathrm{h}\to\mathrm{f}} \in \mathcal{F}$
$\mathrm{C_0}$ 0 0 0 0

Figure 3: Training procedure part related to the *scanner-free* loss. The images used to illustrate the procedure are acquired with the `Prisma` (purple) and `Gyroscan Intera` (green) scanners.

Moreover, as shown in Figure 3, during each iteration, the anatomical embeddings $\boldsymbol{z}^b_{\boldsymbol{x}_k}$ and $\boldsymbol{z}^b_{\boldsymbol{x}_h}$ for both images are fed into the generator, along with the same random noise $\mathcal{N}(0, 1)$. This enables the *scanner-free* generation of two new images, $\hat{\boldsymbol{x}}_{i\to f} = G(\boldsymbol{z}^b_{\boldsymbol{x}_i}, \boldsymbol{\epsilon}, \boldsymbol{c}_0)$ (where $i = \{h, k\}$). All these reconstructions enable the formulation of various loss functions, each addressing different aspects of the model, which will be detailed in Section 3.4.

3.4. Loss Functions

In this section, we define the objective function of the model whose minimization guides the optimization process. By minimizing this function during each iteration of the training procedure, we can iteratively adjust and refine the parameters of the network modules.

Cycle Consistency Loss (L_{cc})

As the proposed model follows a cycle-GAN architecture and training process, the first loss we introduce ensures cross-cycle consistency within the process described in Figure 2. To do this, we enforce the similarity between the original images $\boldsymbol{x}_k$ and $\boldsymbol{x}_h$ and their reconstructed versions $\hat{\boldsymbol{x}}_k$ and $\hat{\boldsymbol{x}}_h$:

$$L_{cc} = \mathbb{E}_{\boldsymbol{x}_k,\boldsymbol{x}_h}\Big[\, \big\|\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\big\| + \big\|\hat{\boldsymbol{x}}_h - \boldsymbol{x}_h\big\|\Big]. \tag{2}$$

Self Reconstruction Loss (L_{rec})

Similar to the previous loss, we seek to ensure that the input images, $\boldsymbol{x}_k$ and $\boldsymbol{x}_h$, closely resemble their directly reconstructed counterparts, $\hat{\boldsymbol{x}}_{k\to k}$ and $\hat{\boldsymbol{x}}_{h\to h}$:

$$L_{\text{rec}} = \mathbb{E}_{\boldsymbol{x}_k,\boldsymbol{x}_h}\Big[\, \big\|\hat{\boldsymbol{x}}_{k\to k} - \boldsymbol{x}_k\big\| + \big\|\hat{\boldsymbol{x}}_{h\to h} - \boldsymbol{x}_h\big\|\Big]. \tag{3}$$

Brain Structure Adversarial Loss (L^b_{adv})

In order to produce scanner-independent anatomical structure embeddings, we aim at mapping them into a shared space $\mathcal{B}$ where domain membership is indistinguishable. To achieve this, adversarial training for the brain encoder E_b is employed. Specifically, the latent representations of the brain structure $\boldsymbol{z}^b_{\boldsymbol{x}_k}$ and $\boldsymbol{z}^b_{\boldsymbol{x}_h}$ are input to the discriminator D_b (Figure 2), which learns to discriminate between their scanner memberships. Meanwhile, the encoder E_b learns to produce anatomical structure embeddings that are indistinguishable in terms of the membership of the scanner domain by D_b. The loss function is formally defined as

$$L^{\text{b}}_{\text{adv}} = \frac{1}{2}\,\mathbb{E}_{\boldsymbol{x}_k}\Big[\log\Big[D_b(\boldsymbol{z}^b_{\boldsymbol{x}_k})(1 - D_b(\boldsymbol{z}^b_{\boldsymbol{x}_k}))\Big]\Big] \;+ \frac{1}{2}\,\mathbb{E}_{\boldsymbol{x}_h}\Big[\log\Big[D_b(\boldsymbol{z}^b_{\boldsymbol{x}_h})(1 - D_b(\boldsymbol{z}^b_{\boldsymbol{x}_h}))\Big]\Big]. \tag{4}$$

Scanner Classification Loss ($L^{\text{s}}_{\text{cls}}$)

To force the generator G to adopt the desired scanner-related information, i.e., the one plugged into it, during the generation of a new image, the discriminator D_s is trained to predict the scanner label of the generated images $\hat{\boldsymbol{x}}_{h\to k}$ and $\hat{\boldsymbol{x}}_{k\to h}$ as follows

$$L^{\text{s}}_{\text{cls}} = \mathbb{E}_{\boldsymbol{x}_k}\Big[-\log\big[D_c(\boldsymbol{c}_k|\hat{\boldsymbol{x}}_{h\to k})\big]\Big] \;+\; \mathbb{E}_{\boldsymbol{x}_h}\Big[-\log\big[D_c(\boldsymbol{c}_h|\hat{\boldsymbol{x}}_{k\to h})\big]\Big]. \tag{5}$$

Scanner Adversarial Loss ($L^{\mathrm{s}}_{\mathrm{adv}}$)

Similarly, with the goal of training the generator G to produce realistic images in each specific scanner domain $\mathcal{X}_i$ through adversarial training as illustrated in Figure 2, the discriminator D_s receives both real images $\boldsymbol{x}_k$ and $\boldsymbol{x}_h$ and generated images $\hat{\boldsymbol{x}}_{h\to k}$ and $\hat{\boldsymbol{x}}_{k\to h}$. It then attempts to discriminate between real and generated images within their respective scanner domains $\mathcal{X}_k$ and $\mathcal{X}_h$, as follows

$$L^{\mathrm{s}}_{\mathrm{adv}} = \frac{1}{2}\,\mathbb{E}_{\boldsymbol{x}_k}\Big[\log\Big[D_s(\boldsymbol{x}_k)(1 - D_s(\hat{\boldsymbol{x}}_{h\to k}))\Big]\Big] + \frac{1}{2}\,\mathbb{E}_{\boldsymbol{x}_h}\Big[\log\Big[D_s(\boldsymbol{x}_h)(1 - D_s(\hat{\boldsymbol{x}}_{k\to h}))\Big]\Big] \tag{6}$$

Scanner-Free Loss (L_{sf})

We introduce a novel loss specifically designed to ensure that the generator consistently reproduces the same scanner effect when provided with identical random noise inputs. To achieve this, $\hat{\boldsymbol{x}}_{k\to f}$ and $\hat{\boldsymbol{x}}_{h\to f}$, generated as described in Section 3.3, are fed to the scanner encoder E_s along with a null vector $\boldsymbol{c}_0$. The encoder extracts their respective latent representations of the scanner effect, $\hat{\boldsymbol{z}}_{k\to f}$ and $\hat{\boldsymbol{z}}_{h\to f}$, and the loss enforces these representations to be as similar as possible (Figure 3):

$$L_{\mathrm{sf}} = \mathbb{E}_{\boldsymbol{x}_k,\boldsymbol{x}_h}\Big[\Big\|E_s\big(\hat{\boldsymbol{x}}_{k\to f}\big) - E_s\big(\hat{\boldsymbol{x}}_{h\to f}\big)\Big\|\Big]. \tag{7}$$

Total Objective Function

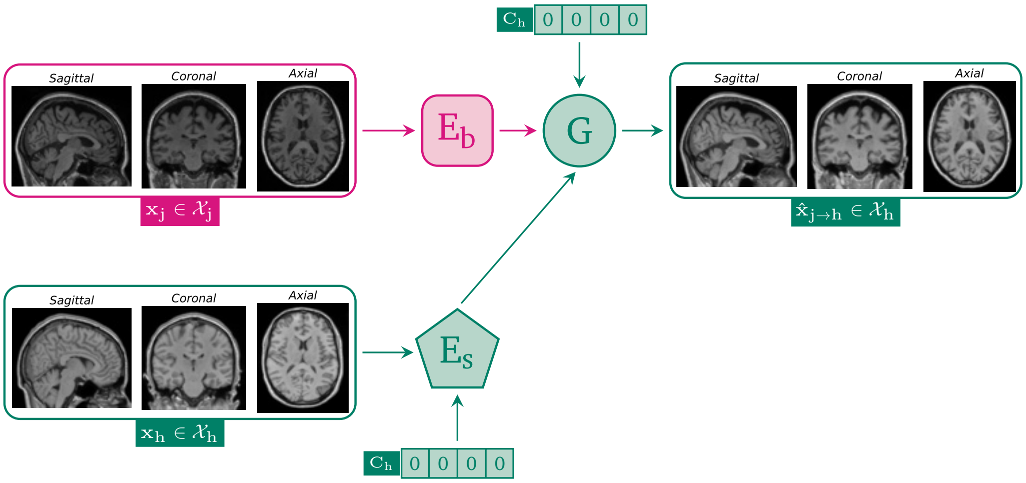
The overall model loss is defined as follows

$$L_{\mathrm{tot}} = \lambda_{\mathrm{cc}}L_{\mathrm{cc}} + \lambda_{\mathrm{rec}}L_{\mathrm{rec}} + \lambda_{\mathrm{lat}}L_{\mathrm{lat}} + \lambda_{\mathrm{KL}}L_{\mathrm{KL}} + \lambda_{\mathrm{sf}}L_{\mathrm{sf}} - \lambda^{\mathrm{b}}_{\mathrm{adv}}L^{\mathrm{b}}_{\mathrm{adv}} - \lambda^{\mathrm{s}}_{\mathrm{cls}}L^{\mathrm{s}}_{\mathrm{cls}} - \lambda^{\mathrm{s}}_{\mathrm{adv}}L^{\mathrm{s}}_{\mathrm{adv}} \tag{8}$$

where, the Kullback–Leibler divergence loss L_{KL} focuses on aligning the scanner effect embeddings with a standard Gaussian prior and L_{lat} ensures that the mean vector $\boldsymbol{\mu}_i$, produced by the scanner encoder E_s, remains close to a standard Gaussian distribution.

3.5. Inference

Consider a new image, $\boldsymbol{x}_{\mathrm{j}} \in \mathcal{X}_{\mathrm{j}}$, which may be acquired either from one of the training scanners ($j \in \{1, \ldots, N\}$) or from a previously unseen scanner not included in the pool of training scanners ($j \notin \{1, \ldots, N\}$). During inference, the image can be harmonized in two distinct ways. One approach involves transferring the new image into the space of one of the training scanners, using it as a reference. To achieve this, we employ the brain encoder E_b to extract the anatomical embedding of the new image $\boldsymbol{z}^b_{\boldsymbol{x}_j}$. This representation is then combined, using the generator G, with the scanner effect $\boldsymbol{z}^s_i$, where $i \in \{1, \dots, N\}$, corresponding to one of the training scanners (Figure 4). By harmonizing all desired images into the same scanner space, this method ensures a uniform scanner effect across the harmonized dataset. The second approach involves transferring the new image into the new *scanner-free* space. Similar to the first method, we use the brain encoder E_b to extract the anatomical embedding $\boldsymbol{z}^b_{\boldsymbol{x}_j}$. However, instead of plugging a scanner effect, we use a random Gaussian noise $\mathcal{N}(0, 1)$ to be fed into the generator G together with the brain information (Figure 5). This approach effectively harmonizes all desired images into a shared, scanner-independent space, ensuring uniformity across the dataset without introducing scanner-specific variations. This configuration behaves like a denoising step.



$\mathrm{C_h}$ 0 0 0 0
Sagittal
Coronal
Axial
$\mathrm{E_b}$
G
$\mathrm{x_j} \in \mathcal{X}_\mathrm{j}$
Sagittal
Coronal
Axial
$\hat{\mathrm{x}}_{\mathrm{j}\to\mathrm{h}} \in \mathcal{X}_\mathrm{h}$
Sagittal
Coronal
Axial
$\mathrm{E_s}$
$\mathrm{x_h} \in \mathcal{X}_\mathrm{h}$
$\mathrm{C_h}$ 0 0 0 0

Figure 4: Inference to a reference scanner for a new image $\boldsymbol{x}_\mathrm{j} \in \mathcal{X}_\mathrm{j}$. Specifically, the figure illustrates the transfer of an image acquired with the `Skyra Fit` (pink) scanner to the reference `Gyroscan Intera` (green).

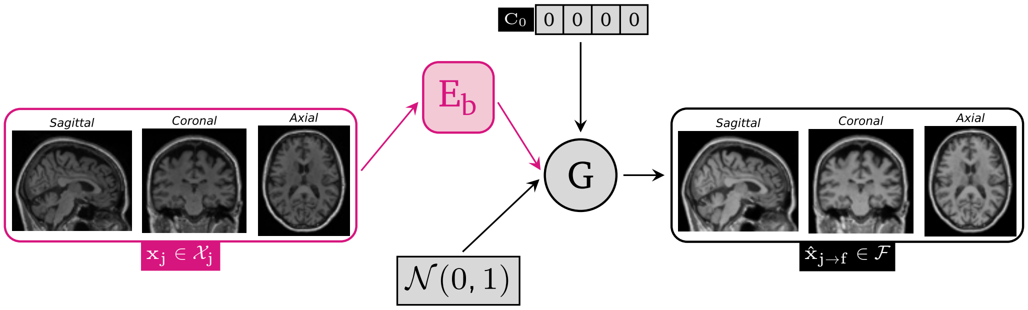


Figure 5: Inference to *scanner-free* for a new image $\boldsymbol{x}_{\mathrm{j}} \in \mathcal{X}_{\mathrm{j}}$. Specifically, the figure illustrates the transfer of an image acquired with the `Skyra Fit` (pink) scanner to the *scanner-free* space (black).

4. Datasets and Preprocessing

4.1. Datasets

We obtained T1-weighted MR images of healthy controls from five distinct datasets. Specifically, the collection includes 313 images from the Alzheimer's Disease Neuroimaging Initiative (ADNI3) [13] acquired using five different scanners, 60 images from the Parkinson's Progression Markers Initiative (PPMI) [23] obtained with three scanners, 581 images from the IXI Brain Development Dataset (IXI) [3] acquired with three scanners, 494 images from the Southwest University Adult Lifespan Dataset (SALD) [47] captured using a single scanner, and 117 images from a private dataset provided by the Italian Neuroimaging Network (RIN) [25], collected with six different scanners. We also considered patients with dementia exhibiting atrophy, including 41 MR scans from the private NeuroArtP3 [22] dataset and 61 MR scans from the public ADNI3 dataset of AD patients. In addition, we included seven subjects from the Strategic Research Program for Brain Science (SRPBS) traveling subjects dataset [40], with each subject having between 8 and 11 MR scans collected at different sites. Detailed descriptions of the datasets — including scanner types, manufacturers, number of images, field strengths, and participants' age — are provided in Appendix A of the Supplementary Materials.

4.2. Preprocessing Steps

Image preprocessing was conducted using the FSL library [38, 15]. Initially, the images were standardized to a common orientation. Next, bias-field correction was applied to address magnetic field variations in the MRI scanner, which can lead to blurring, loss of detail, and altered pixel intensities. The `FAST` algorithm within FSL was used for this correction, combining the non-parametric N3 bias field correction algorithm with a parametric model-based approach. The images were then registered to a standard space using the `FLIRT` command in FSL, which employs an affine transformation model to align the images through translations, rotations, and scaling, mapping them to the Standard MNI152-T1-1mm space. After preprocessing, the resulting images had dimensions of (1, 182, 218, 182).

5. Experiments

We validated our approach using T1-weighted MR images from healthy controls. Specifically, we trained DISARM++ on a dataset comprising 701 images from the ADNI3 [13], PPMI [23], and IXI [3] datasets (Table 1). These training images were acquired using $N = 5$ different scanner models: 167 images with Prisma Fit, 69 with Prisma, 30 with Achieva dStream, 250 with Gyroscan Intera, and 185 with Intera. To improve the robustness of the model, we used a data augmentation procedure through dense random elastic deformation [28]. Additional details about the training procedure are provided in Appendix B.

	Scanner Model	**Manufacturer**	**Dataset**	**Img. #**	**Total**
Training	Prisma Fit	SIEMENS	ADNI3	167/167	701
	Prisma	SIEMENS	ADNI3	69/69	
	Achieva dStream	Philips	ADNI3	26/26	
	Achieva dStream	Philips	PPMI	5/5	
	Gyroscan Intera	Philips	IXI	250/322	
	Intera	Philips	IXI	185/185	

Table 1: Description of the healthy controls training datasets, including details on scanner types, manufacturers, source datasets, and number of images. Values in the ”Img. #” column represent the number of images used out of the total available for that scanner and dataset combination.

In these sections, we first describe the evaluation procedures used in our experiments (Section 5.1), followed by the ablation study (Section 5.2), a comparison with state-of-the-art methods (Section 5.3), and downstream analyses (Section 5.4).

5.1. Evaluation Metrics and Statistical Tests

In this section, we detail the metrics used in our evaluations. For a more in-depth explanation of these metrics and their computation within our specific context, refer to Appendix C of the Supplementary Materials.

5.1.1. Anatomical Structure Metrics

To evaluate the preservation of anatomical structures and the quality of generated images (see Table 2), we employed four key metrics: (1) the structural component of the Structural Similarity Index Measure (SSIM) [46], known as Struct-SSIM, (2) the complete SSIM index [46], (3) the Learned Perceptual Image Patch Similarity (LPIPS) metric [48], and (4) the Fréchet Inception Distance (FID) [11]. The SSIM metrics and LPIPS are computed between image pairs, whereas FID compares two sets of images. The complete SSIM index assesses similarity based on luminance, contrast, and structure, with higher scores indicating greater overall similarity. In contrast, Struct-SSIM focuses specifically on structural similarity by evaluating local spatial patterns independently of luminance and contrast, thereby capturing fine details and spatial coherence; higher Struct-SSIM values signify better structural similarity. LPIPS measures perceptual similarity by comparing feature activation maps from a deep neural network, which makes it sensitive to semantic and textural differences that align with human visual perception; lower LPIPS scores indicate higher similarity. Finally, FID evaluates the distributional similarity between real and generated images by comparing feature embeddings extracted from a pre-trained Inception network. It effectively assesses the overall image quality and realism, with lower values reflecting closer alignment to the real data distribution. Notably, FID, LPIPS, or complete SSIM consider not only structural differences but also luminance and contrast, making them sensitive to the extent of harmonization performed by different models. As a result, using these indices would make it difficult to compare the preservation of anatomical structure across different models. We use FID and LPIPS in the ablation study (Section 6.1), where the level of harmonization across different DISARM++ configurations remains comparable, and the complete SSIM in the traveling subjects evaluation (Section 6.2.3), where images similarity can be assessed in terms of contrast, luminance, and structure.

5.1.2. Harmonization Metrics

To evaluate the transfer of scanner characteristics, we assessed the similarity of voxel intensity distributions before and after harmonization using three metrics: Jensen-Shannon Divergence (JSD) [19], Hellinger Distance (HD) [17, 31], and Wasserstein Distance (WD) [43] (see Table 2). Practically, given the MRI scans having dimension $(1, H, W, D)$ — where H represents height, W represents width, and D represents depth — we computed these metrics based on the set of $1 \times H \times W \times D$ voxel intensity distributions derived from MRI scans, considering pre- and post-harmonization data. Specifically, we compute the empirical distribution by estimating the underlying probability distribution of the voxel intensities for each unfolded image, and we average the distributions of images belonging to the same scanner. In this way, we obtain G distributions — with G being the number of the test scanners — from pre-harmonization images and G distributions from post-harmonization images. In each of the two sets of G distributions, we computed the values for the three metrics across all possible pairs to quantify the similarity between them. Therefore, when reporting these metrics, we present the mean and standard deviation of all pairwise comparison values, providing a general assessment for all scanners. We perform this step for both pre-harmonization and post-harmonization sets, enabling the comparison of the similarity between the distributions before and after harmonization. To statistically evaluate the effectiveness of harmonization, we perform a paired t-test comparing the values of the similarity metrics before and after harmonization. This test assesses whether the mean difference is significantly different from zero. If the differences do not follow a normal distribution, we employ a bootstrap paired t-test, resampling the data to estimate the distribution of differences. We report the 95% confidence interval (CI) for these differences; if the CI does not include zero, it indicates a significant effect of harmonization on voxel intensity similarity.

Besides the three aforementioned indexes, in certain analyses, we also employed the K-sample Anderson-Darling test (AD-test) [33], a non-parametric test that evaluates whether multiple samples originate from the same distribution. This test was applied to the set of G distributions, separately for pre- and post-harmonization images. If we accept the null hypothesis, it suggests no significant difference between distributions, indicating successful harmonization. Conversely, rejecting the null hypothesis implies that at least one distribution differs, suggesting incomplete harmonization.

Metric	Description	Range	Interpretation	Formula
FID	*Measures the distribution distance between real and generated image features.*	[0, ∞]	*10–20 (Good)* *50+ (Poor)*	$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}\right)$
LPIPS	*Measures perceptual similarity between images based on learned deep features.*	[0,1]	*0.1–0.3 (Good)* *0.5+ (Poor)*	$\text{LPIPS}(x, y) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l (f^l(x)_{hw} - f^l(y)_{hw})\|^2$
Struct-SSIM	*Measures the structural similarity between two images.*	[-1,1]	*Higher values, Higher similarity*	$\text{Struct-SSIM}(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3}$
SSIM	*Measures the similarity considering luminance, contrast, and structure.*	[-1,1]	*Higher values, Higher similarity*	$\text{SSIM}(x, y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$

(a)

Metric	Description	Range	Interpretation	Formula
JSD	*Measures the similarity between two probability distributions based on KL divergence.*	[0, log(2)]	*Lower values, Higher similarity*	$\text{JSD}(P, Q) = \frac{1}{2}\text{KL}(P\|M) + \frac{1}{2}\text{KL}(Q\|M)$
HD	*Measures the similarity between two probability distributions using their square root.*	[0, 1]	*Lower values, Higher similarity*	$\text{HD}(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_x \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2}$
WD	*Measures the minimum cost of transporting mass to transform one distribution into another.*	[0, ∞]	*Lower values, Better match*	$\text{WD}(P, Q) = \int_{\mathbb{R}} \left\|F_P(x) - F_Q(x)\right\| dx$

(b)

Table 2: Evaluation metrics. a) Anatomical structure metrics: For FID [11], μ_r and μ_g represent the means, and Σ_r and Σ_g are the covariances of the real and generated image features. For LPIPS [48], $f^l(x)$ and $f^l(y)$ are the deep features at layer l for images x and y, w_l are the learned weights, and H_l and W_l denote the height and width of the feature map at layer l. For SSIM (Structural) [46], σ_{xy} is the covariance between x and y, σ_x and σ_y are their standard deviations, and C_3 is a stabilization constant. For SSIM (Complete), μ_x and μ_y are the means, σ_x^2 and σ_y^2 are the variances, σ_{xy} is the covariance, and C_1 and C_2 are constants. b) Harmonization metrics: For Jensen-Shannon Divergence (JSD) [19], Hellinger Distance (HD) [26], and Wasserstein Distance (WD) [27], P and Q are the probability distributions being compared, and x denotes a specific outcome in the probability space. $M = \frac{1}{2}(P + Q)$ is the mixed distribution, and $\text{KL}(P\|M)$ represents the Kullback-Leibler divergence between P and M. $F_P(x)$ and $F_Q(x)$ denote the cumulative distribution functions of P and Q, respectively.

5.2. Ablation Study

In this section, we present the ablation study of the proposed model by systematically removing or modifying specific components. We begin by evaluating the performance of the baseline DISARM model [4]. Next, starting with the complete model, we analyze the effects of removing key components: the *scanner-free* loss (L^{sf}), the KL divergence loss (L^{KL}), the latent loss (L^{lat}), and the attention layers. The training for all configurations was conducted using 701 images from six different scanners, as described in Section 5. The evaluation was conducted by harmonizing the 250 test images from five different scanners detailed in Table 3 into the *scanner-free* space across all model configurations. Here we compare the different configurations in terms of preservation of anatomical structure and image quality using Struct-SSIM, LPIPS, and FID metrics. To assess harmonization performance, we used the JSD and the AD-test [33], considering the mean voxel intensity distributions from the five test scanners. The results of the ablation study are presented in Section 6.1.

Evaluation type	**Scanner Model**	**Manufacturer**	**Dataset**	**Images No.**
Ablation study	Prisma	SIEMENS	RIN	40/40
	Triotim	SIEMENS	PPMI	41/41
	Gyroscan Intera	Philips	IXI	72/322
	Unknown	GE	IXI	74/74
	Ingenia CX	Philips	RIN	23/23

Table 3: Description of the test images utilized in the ablation study, including details on scanner types, manufacturers, source datasets, and number of images.

5.3. Comparison with State-of-the-Art

This section outlines the evaluation of the harmonization results produced by the proposed model, comparing them to benchmark methods (Section 5.3.1). We detail the datasets used for each evaluation and describe the evaluation process. The results are then provided in Section 6.2.

5.3.1. Benchmarking Methods

To benchmark our model, we compare the performance of the proposed model with the following image-based approaches:

- **STGAN** [7]: We utilize the pre-trained STGAN model, which was trained on 718 images from five different dataset subsets: ADNI3 (42 images), ICBM (200 images), UKBB (200 images), PPMI (76 images), and ABCD (200 images). The preprocessing steps involved skull-stripping the images [34] and correcting for nonuniformity using the N3 method [37] in Freesurfer. Additionally, the images were linearly registered to the standard MNI template (ICBM 152 Nonlinear Symmetric atlas [9]) and resized to isotropic 1 mm^3 voxels. Following the STGAN implementation, harmonization was performed on a sliding window of three image slices with a stride of 1. The final T1 harmonized MRI 3D volumes were reconstructed by combining these harmonized partial volumes. The model assumes that each image belongs to a unique domain and can be decomposed into its content and style. Therefore, for comparison with other models, we harmonize the test images by using an image from the `Gyroscan Intera` scanner as a reference.

- **IGUANe** [32]: We utilize the IGUANe model, trained on a dataset of 4,347 T1-weighted brain MRI images from 11 distinct scanners across eight public studies: SALD, IXI, OASIS-3, NKI-RS, NMorphCH, AIBL, HCP, and ICBM. Preprocessing involved skull-stripping using HD-BET [12], bias field correction with N4ITK [42], linear registration to the MNI 1 $\text{mm}^{\mathbf{3}}$ space using FSL-FLIRT [14], cropping to 160×192×160 voxels, and standardizing intensities by dividing by the median brain intensity. Images were then scaled so that the median brain intensity matched a value of 1 while maintaining a background intensity of 0. The IGUANe model, based on a 3D extension of the CycleGAN framework, implements a many-to-one adversarial training strategy to harmonize MRI data from various acquisition sites into a common reference domain, represented by the SALD dataset. The authors selected the SALD dataset as a reference because it contains a large number of MR images and covers a wide age range.

5.3.2. Evaluation Setup

To evaluate the harmonization results obtained with the proposed model in comparison to benchmark methods, we use the results from the set detailed in Table 4, which included 796 healthy control MR images from the RIN, IXI, SALD and PPMI datasets acquired from 10 different scanners, along with the SRPBS traveling subject dataset.

The first analysis focused on visualizing the harmonization results by presenting slices from each direction (axial, coronal, and sagittal). For each model, we display 10 original images (one from each test scanner) alongside their harmonized counterparts. Additionally, we provide heatmaps illustrating pixel-wise differences between the harmonized images and their originals.

For the proposed model, we report the visualization of the transfer of images to the *scanner-free* space. The results are presented in Section 6.2.1.

In the second analysis, we evaluate the preservation of the anatomical structure using the Struct-SSIM for each pair of original and harmonized images. Notably, we do not employ FID, LPIPS, or complete SSIM for this evaluation, as the degree of harmonization varies across the different models, as discussed in Section 5.1.1. We analyze the transfer of scanner-specific characteristics using the JSD, HD, and WD metrics. These metrics are calculated based on mean voxel intensity distributions from the 10 test scanners and statistical significance is determined through paired bootstrap t-tests between pre- and post-harmonization, as outlined in Section 5.1. We assess the performance of the proposed procedure against competing models using two distict harmonization approaches: (1) transferring images to the *scanner-free* space and (2) transferring images to one of the training scanners, selecting the Gyroscan Intera scanner as the reference due to its largest representation in the training dataset. The results are detailed in Section 6.2.2.

For the third analysis, we compare the proposed approach against benchmark models using the traveling subject dataset. We compute the pairwise SSIM for each subject across images acquired with different scanners, both before and after harmonization. To evaluate the statistical significance of the harmonization effect, we conduct a bootstrap t-test and report the 95% confidence interval (95% CI) for the difference in SSIM values between pre- and post-harmonization images, focusing solely on the proposed model's performance when transferring images to the *scanner-free* space. The results are reported in Section 6.2.3.

	Scanner Model	Manufacturer	Dataset	Img. #	Total
Test	Trio	SIEMENS	SALD	494/494	796
	Achieva dStream	Philips	RIN	17/17	
	DISCOVERY MR750	GE	RIN	23/23	
	Ingenia CX	Philips	RIN	23/23	
	Prisma	SIEMENS	RIN	40/40	
	Skyra	SIEMENS	RIN	12/12	
	Skyra Fit	SIEMENS	RIN	3/3	
	Triotim	SIEMENS	PPMI	41/41	
	Gyroscan Intera	Philips	IXI	72/322	
	Unknown	GE	IXI	74/74	

Table 4: Description of the healthy controls test datasets, including details on scanner types, manufacturers, source datasets, and number of images.

5.4. Downstream Analysis

This section describes the downstream analysis conducted to evaluate the impact of harmonization on extracting biologically meaningful information from MRI data, comparing the proposed model with benchmark methods. We specify the datasets used for each analysis and detail the procedure. The results are presented in Section 6.3. Specifically, we perform a series of downstream analyses to assess whether harmonization enhances the reliability and predictive power of MRI-based biomarkers:

1. **Age Prediction:** We explore whether harmonization enhances the accuracy of predicting personal age based on MRI-derived features.
2. **Reduction of Inter-Scanner Variability:** We quantify how harmonization reduces variability in MRI-derived brain volumes caused by differences in scanning protocols and equipment.
3. **AD vs. Healthy Classification:** We examine whether harmonization improves the ability to distinguish between AD patients and healthy controls, focusing on predictive performance.
4. **Diagnosis Prediction:** We investigate the impact of harmonization on classifying mild cognitive impairment (MCI) due to AD and AD dementia.

These analyses are designed to assess the degree to which harmonization improves the consistency and robustness of MRI-based biomarkers, ultimately enhancing their clinical and research applications. The volumetric variables used in these analyses were extracted using a custom pipeline implemented in FreeSurfer [8].

5.4.1. Age Prediction Task

The objective of the analysis is to determine whether harmonization enhances the accuracy of personal age prediction based on MRI-derived features. Specifically, we employed a simple linear model to predict age using a set of volumetric variables extracted from MR images. The variables selected are those most strongly associated with aging [35, 16, 44], including total gray matter volume (TGV), subcortical gray volume (SGV), supratentorial volume, cortex volume (CV), cerebral white matter volume, left lateral ventricle volume, left hippocampus volume (LHV), left amygdala volume, and left putamen volume (LPV). We considered the test images from healthy controls described in Table 4. The model was applied to volumes extracted from raw MRI scans before any preprocessing or harmonization, as well as after applying each harmonization method under evaluation. To ensure robust comparisons, we perform 10-fold cross-validation and report the mean and standard deviation of the coefficient of determination ($\mathrm{R}^{\mathbf{2}}$), the root mean square error (RMSE), and the Bayesian Information Criterion (BIC). The results are presented in Section 6.3.1).

5.4.2. Inter-Scanner Variability in MRI-Derived Brain Volumes

We aimed to evaluate the extent to which harmonization reduces inter-scanner variability in MRI-derived brain volumes. Specifically, we focused on a subset of the variables used in Section 5.4.1, consisting of five volume variables extracted from MR images: TGV, SGV, CV, LHV, and LPV. To quantify the reduction in inter-scanner variability, we employed linear mixed-effects models (LMM) to predict these volumes based on age, treating scanner groups as random effects. The LMM is defined as follows:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + \epsilon_{ij}$$

where y_{ij} and x_{ij} represent the age and volume, respectively, for individual i scanned using scanner j. Here, β_0 is the fixed intercept, and β_1 is the fixed-effect coefficient for volume. The term u_j represents the random effect for scanner group j, assumed to be normally distributed with variance σ_u^2. ϵ_{ij} is the residual error term, assumed to be normally distributed with variance σ_ϵ^2. The random effect u_j captures between-group variance, whereas the residual error ϵ_{ij} captures within-group variance.

The model was applied to both pre-harmonization and post-harmonization data. In either case, we calculated three metrics: (1) the intraclass correlation coefficient (ICC) to measure the proportion of variability of the dependent variable due to scanner group differences. An ICC value close to 1 suggests that most of the variability in the outcome is due to differences between scanner groups, indicating a strong group-level influence on the dependent variable. This implies that a substantial portion of the variance in the predicted age is attributable to the scanner, reducing the reliability of volume as a biomarker of age. Conversely, an ICC value near 0 implies that variability is primarily within groups, suggesting that scanner effects have minimal impact on the outcome; (2) the marginal R^2 (R^2_m), which reflects how much of the total variance is explained by the fixed effects, excluding the contribution of the random effects; and (3) the BIC for both the LMM and the corresponding linear model (LM) that do not consider explicitly the random effects. We report the difference in BIC between the LM and the LMM (denoted as ΔBIC). A lower ΔBIC indicates that incorporating the scanner variable as a random effect yields minimal improvement in model fit. A higher ΔBIC suggests a significant contribution of the scanner variable as a random effect. For this analysis, we considered the test images from healthy controls described in Table 4, excluding scanners with fewer than 15 images. Thus, $j \in \{$Achieva dStream; DISCOVERY MR750; Ingenia CX; Prisma; Gyroscan Intera; Unknown (IOP); Triotim (PPMI); Triotim (SALD)$\}$. The results are presented in Section 6.3.2.

5.4.3. Classification of Alzheimer's Disease (AD) vs. Healthy Patients

We evaluated the effectiveness of harmonization in enhancing the biological information for a classification task distinguishing between healthy and AD patients. The variability introduced by the scanner effects can negatively affect classifier training, leading to a less accurate model. To evaluate this, we selected 102 MR images of healthy patients from the RIN, IXI, and PPMI test datasets (Table 4) and 102 images of AD patients from the NeuroArtP3 and ADNI3 datasets, both before and after harmonization. We trained a vanilla 3D CNN classifier 10 times with 10 different random splits, each composed of 132 MR images for training and 72 for testing. The splits were chosen to maintain an equal number of healthy control and AD images in both the training and test sets. For both our model and the two competing approaches, we employed the same classifier architecture and identical training configurations in each iteration to ensure a fair and consistent comparison. The results are presented in Section 6.3.3).

5.4.4. Diagnosis Prediction

We explored the effect of harmonization on classification performance in distinguishing between MCI and AD dementia. The dataset used for this analysis includes the 41 subjects from the NeuroArtP3 dataset. We assessed performance by comparing volumes extracted from the 41 raw MRI scans before harmonization with volumes processed through preprocessing and harmonization using each of the evaluated methods. To include as much information as we could, we considered 57 volume-related variables from the brain regions extracted from the images. To reduce dimensionality, we applied principal component analysis (PCA) and retained the number of components that explain 70% of the variance, which provides an optimal trade-off between variance explained and model complexity across all scenarios. We then use the principal components (PCs) for logistic regression modeling of the classification task. For performance evaluation, we report the area under the ROC curve (AUC), which quantifies the model's ability to differentiate between MCI and AD dementia. An AUC of 0.5 corresponds to random performance, while an AUC of 1.0 indicates perfect classification. Higher AUC values reflect better performance in distinguishing the two conditions. The results are presented in Section 6.3.4).

6. Results

6.1. Ablation Study Results

We present the results of the ablation study as provided in Table 5. Before harmonization, the mean and standard deviation of the JSD values were 0.17 ± 0.08 and thus we reject the AD-test null hypothesis (p-value $\ll 0.05$). In Table 5, for each model setup, considering data after harmonization, we indicate whether the null hypothesis of the AD-test is accepted or rejected, with acceptance suggesting the effectiveness of harmonization. The results of the ablation study show that all tested components are crucial for enhancing the model's performance. After harmonization, the similarity between distributions across all models increases significantly in terms of JSD, resulting in comparable means and standard deviations. For all evaluated models, except the one without attention layers, we accept the null hypothesis of the AD test, indicating successful harmonization. Regarding the quality of generated images and structural preservation, the model incorporating all components performs the best, as evidenced by higher SSIM scores and lower LPIPS and FID scores. In addition, the ablation study reveals substantial improvements over the baseline DISARM in both image quality and structural preservation. The findings also confirm that the evaluation metrics effectively capture different aspects of the generated images. For instance, while the model without latent loss has a lower SSIM than the one without attention layers, it achieves a better FID score.

Setup	L^{sf}	L^{KL}	L^{lat}	**Att. Lay.**	**Part. Vol.**	**SSIM**	**FID**	**LPIPS**	**JS-div (Post)**	**AD-test**
DISARM++	✓	✓	✓	✓	✓	0.983 ± 0.006	18.6	0.10 ± 0.02	0.008 ± 0.002	✓
w/o L^{sf}	✗	✓	✓	✓	✓	0.964 ± 0.008	30.5	0.15 ± 0.03	0.004 ± 0.001	✓
w/o L^{KL}	✓	✗	✓	✓	✓	0.960 ± 0.007	35.5	0.14 ± 0.03	0.003 ± 0.002	✓
w/o L^{lat}	✓	✓	✗	✓	✓	0.975 ± 0.005	25.4	0.14 ± 0.03	0.006 ± 0.002	✓
w/o Att.Lay.	✓	✓	✓	✗	✓	0.979 ± 0.007	35.3	0.13 ± 0.02	0.012 ± 0.005	✗
DISARM	✗	✓	✓	✗	✗	0.958 ± 0.009	35.0	0.20 ± 0.03	0.007 ± 0.002	✓

Table 5: Summary of the ablation study results, with each row representing a different model configuration and its corresponding evaluation results.

6.2. Harmonization Results

This section presents the results of the harmonization assessments described in Section 5.3.2.

6.2.1. Visual Assessment of Harmonization

Figures 6, 7, and 8 present the visualizations described in Section 5.3.2. DISARM++ demonstrates a significantly stronger harmonization effect than both IGUANe and STGAN, leading to a more uniform visual appearance across all directions in the images acquired with the ten different scanners. This is further emphasized by the heatmaps, which reveal a stronger effect for DISARM++, followed by STGAN, while IGUANe's heatmaps are the least pronounced.

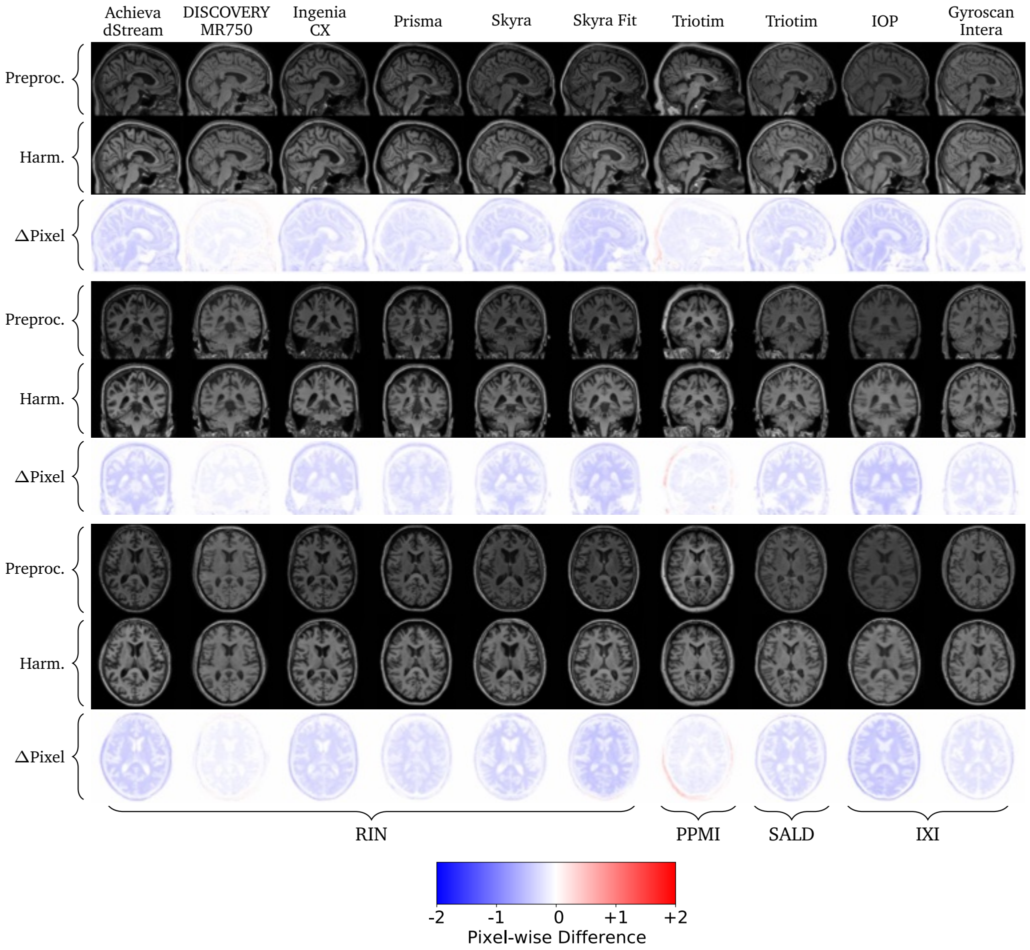


Figure 6: DISARM++ harmonization visual assessment. The figure displays slices from the axial, coronal, and sagittal dimensions for 10 original images — one per test scanner — alongside their corresponding harmonized slices. Heatmaps illustrate the pixel-wise differences between the harmonized images and their original counterparts.

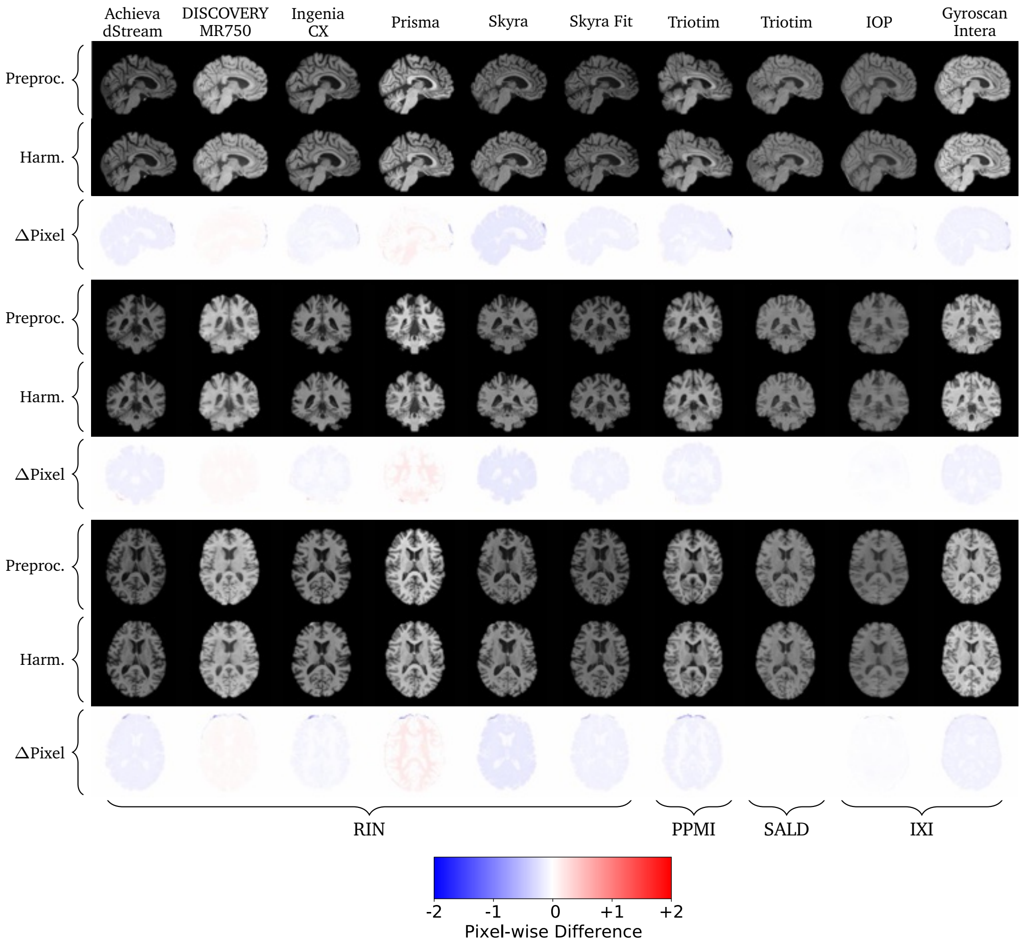


Figure 7: IGUANe harmonization visual assessment. The figure displays slices from the axial, coronal, and sagittal dimensions for 10 original images — one per test scanner — alongside their corresponding harmonized slices. Heatmaps illustrate the pixel-wise differences between the harmonized images and their original counterparts.

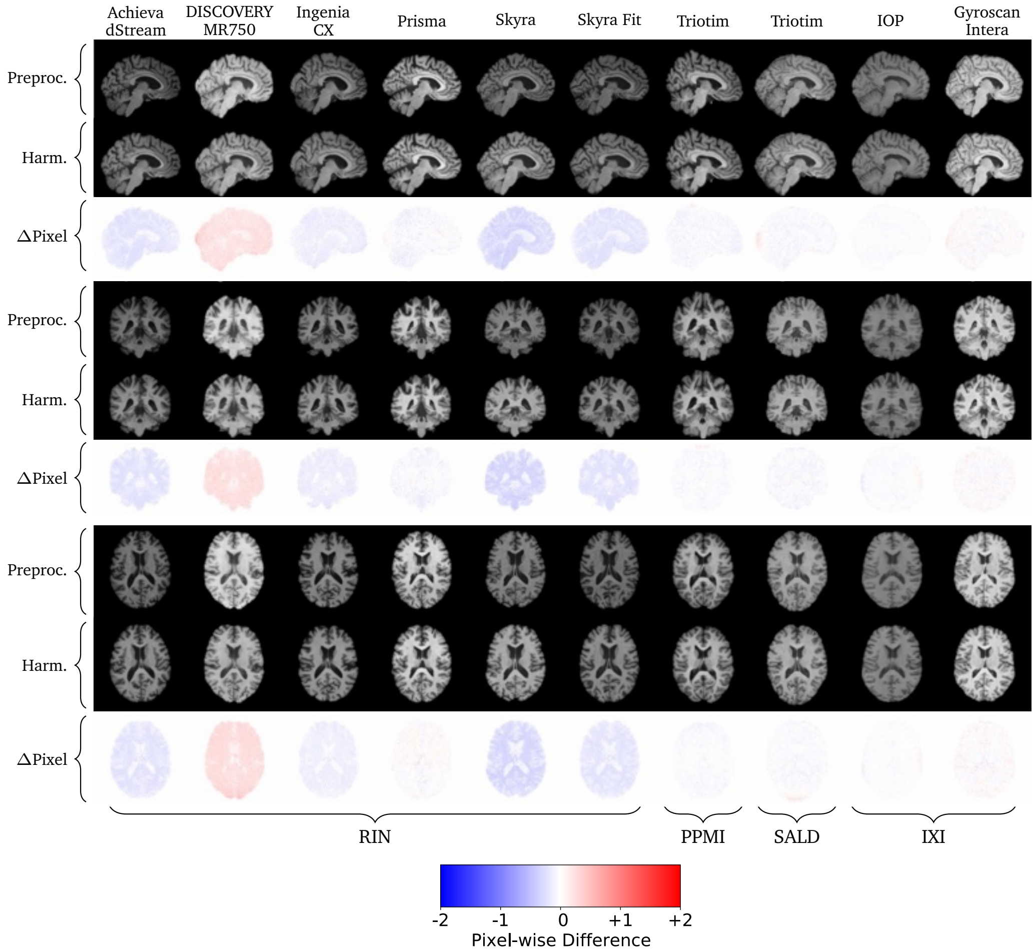


Figure 8: STGAN harmonization visual assessment. The figure displays slices from the axial, coronal, and sagittal dimensions for 10 original images — one per test scanner — alongside their corresponding harmonized slices. Heatmaps illustrate the pixel-wise differences between the harmonized images and their original counterparts.

6.2.2. Assessment of Anatomical Structure Preservation and Scanner Characteristics Transfer

In this section, we evaluate the proposed model's ability to preserve anatomical structures and transfer scanner-specific characteristics as outlined in Section 5.3.2, comparing it with the benchmark models.

Regarding the preservation of anatomical structure, Table 6 presents the means and standard deviations of Struct-SSIM across all test images. Both IGUANe and STGAN achieve slightly higher scores than DISARM++. However, when assessing anatomical preservation, the presence of large black borders in skull-stripped images—compared to non-skull-stripped images—presents a notable challenge. These black regions, resulting from the removal of non-brain tissues during skull-stripping, are uniform and low-intensity, which can inflate similarity scores. Consequently, the Struct-SSIM may yield high values as it considers the black borders structurally similar between the original and harmonized images. This likely explains why IGUANe and STGAN exhibit slightly higher Struct-SSIM scores compared to DISARM++, even though the results remain visually similar, as shown in the previous section. Moreover, we deliberately avoid skull-stripping in our method for several reasons. First, since skull-stripping would need to be performed both before and after harmonization—unlike in other models—BET might inconsistently remove some non-brain areas at different stages, leading to discrepancies that negatively impact the Struct-SSIM scores. Second, skull-stripping is computationally expensive; for instance, HD-BET [12] takes approximately ten minutes per image on a CPU, significantly increasing processing time. Finally, we choose to retain the skull to maintain the integrity of the original anatomical structures, which is particularly beneficial for applications beyond brain tissue analysis, such as studies involving head trauma or cranial deformities.

In terms of transferring scanner-specific characteristics, Table 6 summarizes the means and standard deviations of all pairwise distribution comparisons after harmonization in terms of JSD, HD, and WD. Comprehensive heatmaps displaying all pairwise values for each model before and after harmonization across all three metrics can be found in Appendix D of the Supplementary Materials. DISARM++ — including both the *scanner-free* and reference scanner harmonization approaches — demonstrates the best overall performance, yielding the lowest metric values. For both harmonization strategies, the bootstrap t-test results in a p-value $\ll 0.05$, confirming a significant increase in distribution similarity across all three metrics. The negative confidence intervals (Table 6) highlight the substantial increase in similarity. IGUANe shows higher means and standard deviations for JSD, HD, and WD compared to DISARM++, with bootstrap t-tests indicating a statistically significant but minor improvement in distribution similarity. Similarly, STGAN exhibits higher means and standard deviations for JSD, HD, and WD than DISARM++ but demonstrates statistically significant improvements across all three metrics — greater than IGUANe, yet still lower than DISARM++.

Figure 9 visually compares the mean voxel intensity distributions for each test scanner before and after harmonization for the three models. This visual comparison supports the previously described metric results and statistical test outcomes, demonstrating that DISARM++ has a more substantial effect between pre- and post-harmonization and leads to better alignment of the distributions after harmonization.

The metric results and the visualizations of the voxel intensity distributions presented in this section further support the visual assessment discussed in the previous section.

Metric	**DISARM++** (*Scanner-free*)	**DISARM++** (Gyroscan Intera)	**IGUANE** (SALD)	**STGAN** (Gyroscan Intera)
Struct-SSIM	0.986 ± 0.005	0.986 ± 0.005	0.996 ± 0.001	0.997 ± 0.001
JSD (Post)	0.009 ± 0.004	0.010 ± 0.004	0.1281 ± 0.1004	0.0612 ± 0.0373
JSD (95% CI)	$[-0.23, -0.16]^*$	$[-0.22, -0.16]^*$	$[-0.04, -0.01]^*$	$[-0.06, -0.02]^*$
HD (Post)	0.154 ± 0.038	0.162 ± 0.037	0.6082 ± 0.2643	0.4154 ± 0.1496
HD (95% CI)	$[-0.68, -0.52]^*$	$[-0.65, -0.51]^*$	$[-0.09, -0.02]^*$	$[-0.16, -0.05]^*$
WD (Post)	2.173 ± 0.962	2.059 ± 0.880	14.329 ± 7.891	13.548 ± 5.814
WD (95% CI)	$[-6.71, -4.55]^*$	$[-6.66, -4.61]^*$	$[-4.05, -1.26]^*$	$[-5.38, -1.71]^*$

Table 6: Comparison of harmonization methods on healthy controls MR scans concerning the preservation of anatomical structures and the transfer of scanner characteristics. The asterisks denote significant values.

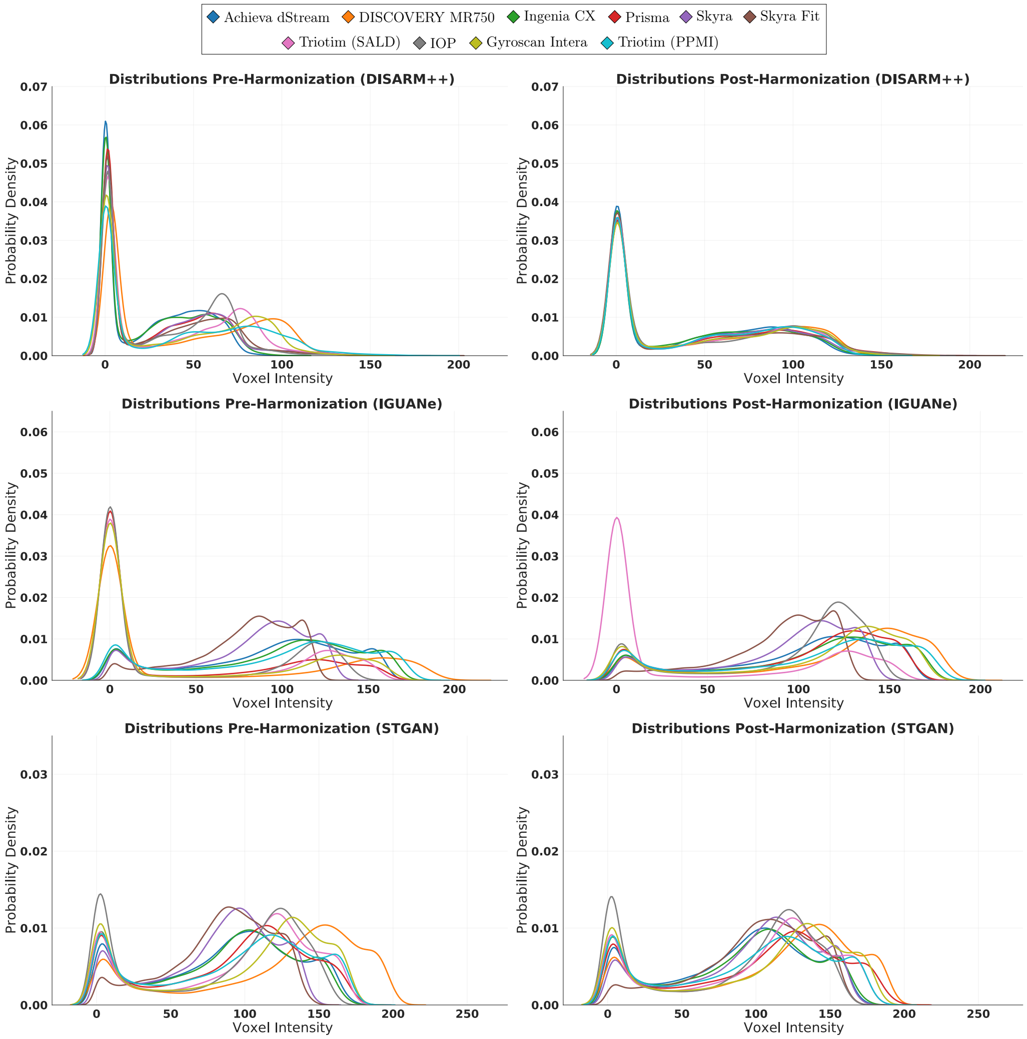


Figure 9: Comparison of mean voxel intensity distributions from the 10 test scanners before and after harmonization using DISARM++, IGUANe, and STGAN.

6.2.3. Traveling Subjects

With reference to Table 7, we evaluate the proposed model and compare it to the benchmark models using the traveling subject dataset as described in Section 5.3.2. The results show that DISARM++ leads to a significant improvement in SSIM after harmonization for all subjects in the dataset, with the 95% CI consistently ranging from 0.13 to 0.24, indicating a robust and significant improvement. In contrast, IGUANe shows a significant increase in

SSIM in only three of the seven subjects and the observed gains are considerably smaller than those achieved by DISARM++. For two subjects, there is a slight but significant decrease in SSIM post-harmonization, while for the remaining two subjects, no significant increase is observed. For STGAN, a significant improvement in SSIM is observed in all seven subjects, with 95% CI consistently ranging from 0.02 to 0.04. These confidence intervals contain substantially smaller values compared to those of DISARM++, indicating a much less pronounced improvement. Nonetheless, STGAN consistently outperforms IGUANe, where the SSIM increase is limited to just three subjects, and the 95% CI consistently ranges from 0.002 to 0.011.

Moreover, we provide a visual comparison of the harmonization results in Figure 10. For subject 1, we display a sagittal slice from three MR images acquired using different scanners, showing both pre- and post-harmonization slices, along with heatmaps of pixel-wise differences. For subject 6, we present coronal slices, and for subject 9, we show axial slices. DISARM++ demonstrates a significant improvement between pre- and post-harmonization slices, resulting in harmonized slices with a much more similar visual appearance with respect to IGUANe and STGAN. This visualization further confirm the SSIM metric results.

DISARM++

Subj.	SSIM (pre-harm)	SSIM (post-harm)	SSIM (95% CI)
1	0.633 ± 0.119	0.821 ± 0.05	$[+0.160, +0.218]^*$
2	0.659 ± 0.123	0.817 ± 0.06	$[+0.130, +0.186]^*$
3	0.642 ± 0.114	0.806 ± 0.05	$[+0.139, +0.191]^*$
6	0.602 ± 0.122	0.785 ± 0.06	$[+0.150, +0.216]^*$
7	0.605 ± 0.138	0.804 ± 0.05	$[+0.155, +0.242]^*$
8	0.609 ± 0.128	0.800 ± 0.06	$[+0.159, +0.219]^*$
9	0.617 ± 0.114	0.801 ± 0.05	$[+0.155, +0.213]^*$

IGUANe

Subj.	SSIM (pre-harm)	SSIM (post-harm)	SSIM (95% CI)
1	0.910 ± 0.03	0.911 ± 0.02	$[-0.002, +0.005]$
2	0.916 ± 0.03	0.922 ± 0.02	$[+0.003, +0.0009]^*$
3	0.917 ± 0.03	0.917 ± 0.03	$[-0.001, +0.002]$
6	0.922 ± 0.02	0.919 ± 0.02	$[-0.005, -0.002]^*$
7	0.916 ± 0.03	0.911 ± 0.03	$[-0.008, -0.002]^*$
8	0.916 ± 0.03	0.922 ± 0.02	$[+0.002, +0.008]^*$
9	0.913 ± 0.03	0.921 ± 0.02	$[+0.005, +0.011]^*$

STGAN

Subj.	SSIM (pre-harm)	SSIM (post-harm)	SSIM (95% CI)
1	0.881 ± 0.03	0.916 ± 0.03	$[+0.030, +0.040]^*$
2	0.898 ± 0.03	0.932 ± 0.02	$[+0.028, +0.040]^*$
3	0.896 ± 0.03	0.919 ± 0.03	$[+0.021, +0.026]^*$
6	0.886 ± 0.04	0.919 ± 0.02	$[+0.028, +0.039]^*$
7	0.886 ± 0.03	0.916 ± 0.03	$[+0.024, +0.035]^*$
8	0.900 ± 0.03	0.925 ± 0.02	$[+0.021, +0.029]^*$
9	0.887 ± 0.04	0.917 ± 0.03	$[+0.026, +0.036]^*$

Table 7: Mean SSIM values and standard deviations across all image pairs for each subject evaluated, both before and after harmonization, comparing DISARM++, IGUANe, and STGAN. The 95% confidence intervals indicate the differences in SSIM between pre- and post-harmonization.

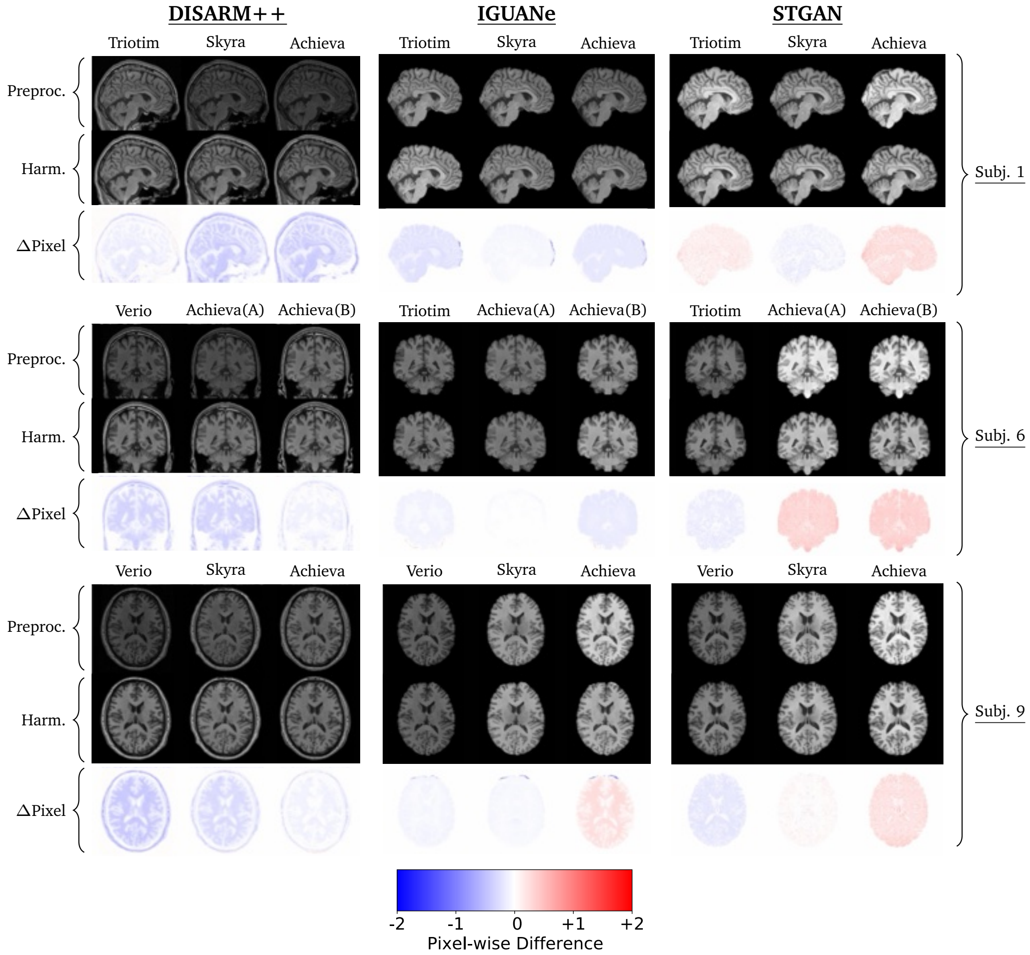


Figure 10: Visual harmonization comparison for traveling subjects. For Subject 6, we present sagittal slices of MR images acquired with three different scanners, both before and after harmonization, along with heatmaps showing pixel-wise differences. Similarly, for Subject 1, we provide coronal slices, and for Subject 9, we display axial slices.

6.3. Downstream Analysis Results

This section presents the results of the downstream analysis described in Section 5.4.

6.3.1. Age Prediction Task

We present the results of the downstream analysis described in Section 5.4.1, i.e., the age prediction task. With reference to Table 8, the DISARM++ model outperforms all others, achieving the highest R^2 score

($\simeq 0.60 \pm 0.05$), the lowest RMSE ($\simeq 0.168 \pm 0.008$), and the best BIC score ($\simeq -455.4 \pm 7.6$). These results demonstrate DISARM++'s superior predictive accuracy and model efficiency. In contrast, the baseline raw images exhibit significantly lower performance, while STGAN provides only marginal improvements over the baseline and remains inferior to DISARM++. Finally, IGUANe performs slightly worse than the baseline.

Metric	Raw Images	DISARM++ (*Scanner-free*)	IGUANE (SALD)	STGAN (Gyroscan Intera)
R^2	0.5209 ± 0.1126	$\mathbf{0.6042 \pm 0.0495}$	0.4879 ± 0.0665	0.5290 ± 0.1177
RMSE	0.1873 ± 0.0259	$\mathbf{0.1684 \pm 0.008}$	0.1941 ± 0.018	0.1871 ± 0.034
BIC	-353.6 ± 12.6	$\mathbf{-455.4 \pm 7.6}$	-266.6 ± 12.9	-374.3 ± 10.0

Table 8: Comparison of models on age prediction task. Bold font denotes the best results.

6.3.2. Inter-Scanner Variability in MRI-Derived Brain Volumes

Var.	Metrics	Raw Images	DISARM++ (*Scanner-free*)	IGUANe (SALD)	STGAN (Gyroscan Intera)
TGV	ICC (%)	60.19	**11.43**	55.91	27.59
	R_m (%)	22.37	**50.36**	32.65	41.81
	ΔBIC	+ 287	**+ 18**	+ 417	+ 95
SGV	ICC (%)	54.28	**7.21**	58.60	15.39
	R_m (%)	19.67	29.17	**32.62**	26.65
	ΔBIC	+ 246	**− 5**	+ 383	+ 25
CV	ICC (%)	57.84	**16.46**	49.21	26.88
	R_m (%)	22.87	**51.12**	27.81	40.28
	ΔBIC	+ 271	**+ 34**	+ 382	+ 93
LHV	ICC (%)	42.61	**12.80**	40.52	24.89
	R_m (%)	11.64	**13.44**	12.91	3.71
	ΔBIC	+ 204	**+ 22**	+ 223	+ 113
LPV	ICC (%)	47.28	**13.08**	39.13	23.39
	R_m (%)	16.30	**25.99**	12.85	15.79
	ΔBIC	+ 203	**+ 17**	+ 262	+ 66.39

Table 9: Comparison of inter-scanner variability in volumes extracted from images before preprocessing and harmonization, and after applying all harmonization models. The ΔBIC represents the difference between the metric computed without and with random effects. Bold font denotes the best results.

In this section, we present the results of the downstream analysis described in Section 5.4.2, concerning the inter-scanner variability assessment in MRI-derived brain volumes. The three metrics for all five variables are presented in Table 9. DISARM++ harmonization, across all five variables, significantly reduces inter-scanner variability, as evidenced by a substantial decrease in the ICC (e.g., ICC for TGV drops from $\simeq 60\%$ to $\simeq 11\%$). Furthermore, harmonization increases the proportion of variance explained by fixed effects alone (e.g., R_m for TGV rises from $\simeq 22\%$ to $\simeq 50\%$). Before preprocessing, the BIC for the models with random effects are significantly lower than that for the models without, indicating that random effects are essential for a better model fit (e.g., $\Delta\text{BIC} \simeq +287$ for TGV). Following DISARM++ harmonization, the BIC for the models without random effects become comparable to, or even lower than, that for the models with random effects, suggesting that random effects are no longer crucial for predicting the outcome (e.g., $\Delta\text{BIC} \simeq -5$ for SGV). In contrast, both STGAN and IGUANe harmonization yield a smaller reduction in inter-scanner variability, explain a lower proportion of variance via fixed effects (except for SGV in the case of IGUANe), and consistently result in worse ΔBIC values across all five variables compared to DISARM++. For instance, the ICC for TGV is approximately 28% under STGAN harmonization and 56% under IGUANe, compared to only 11% with DISARM++; the R_m is around 42% with STGAN and around 33% with IGUANe, versus 50% with DISARM++; and the ΔBIC is roughly +95 for STGAN and +419 for IGUANe, compared to +18 for DISARM++. Moreover, Figure 11 illustrates the median volumes for each scanner across the analyzed volume variables, both before harmonization and after applying the three competing models. Additionally, the figures depict the median age for each test scanner, as brain region volumes generally decline with aging. Among the scanners, IOP and Triotim (SALD) include the youngest individuals (IQR = [30, 50] and IQR = [25, 59], respectively), while Triotim (PPMI) has the oldest (IQR = [64, 71]). In the TGV plot, raw image volumes (blue) reveal that the IOP and Triotim (SALD) scanners have lower median volumes (IQR = [547, 654] and IQR = [588, 678], respectively) compared to Achieva dStream (IQR = [668, 763]), DISCOVERY MR750 (IQR = [691, 750]), Ingenia CX (IQR = [662, 743]), and Prisma (IQR = [674, 758]), which correspond to scanners with higher median ages. This observation is counterintuitive, as total brain volume (TGV) typically decreases with age. After applying DISARM++ (green), the IOP and Triotim (SALD) scanners exhibit the highest median volumes (IQR = [654, 724] and IQR = [695, 766], respectively), aligning with the expected trend of age-related volume reduction. The median values for the other scanners remain comparable across groups, reflecting their similar age distributions. Conversely, STGAN (red) produces uniform median values across all scanners, indicating that the model homogenizes volumes regardless of age distribution differences. For instance, IOP (IQR = [719, 773]) and Triotim (PPMI) (IQR = [733, 774]) display nearly identical median volumes despite their age disparities. Similarly, IGUANe (purple) results in comparable median volumes for IOP (IQR = [692, 737]), Ingenia CX (IQR = [651, 683]), and DISCOVERY MR750 (IQR = [679, 721]), despite differences in median age. Moreover, IGUANe yields the lowest median volume values for Triotim (SALD), even though this cohort has the second youngest median age among all scanners, following IOP.

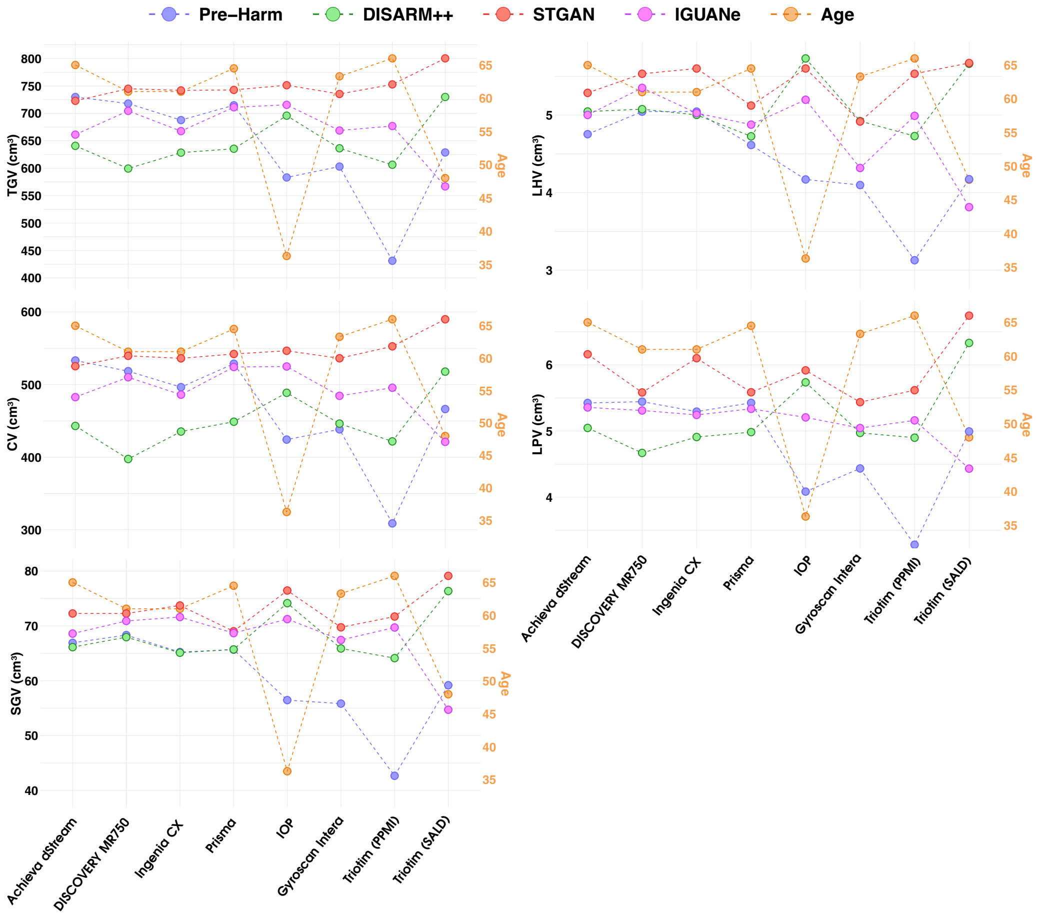


Figure 11: For each volume variable considered, we present the median volumes for each scanner before preprocessing or harmonization (orange), along with those obtained from the three competing models: DISARM++ (green), STGAN (red), and IGUANe (purple). In addition, we report the median age for each test scanner (blue).

6.3.3. *Classification of Alzheimer's Disease (AD) vs. Healthy Patients*

In this section, we present the results of the downstream analysis described in Section 5.4.3 investigating the discrimination power between pathological and healthy subjects. Table 10 summarizes the classifier's performance in terms of accuracy, precision, recall, and F1-score. DISARM++ harmonization outperforms the other methods across all four metrics. Additionally, Figure 12 displays boxplots showing the distributions of these metrics after harmonization for each evaluated model. Paired bootstrap t-tests show that DISARM++ significantly outperforms IGUANe across all four metrics. Compared to STGAN, DISARM++ shows significantly better precision, F1 score, and recall, although no significant difference in precision is observed.

	DISARM++ (*Scanner-free*)		**IGUANe** (SALD)		**STGAN** (Gyroscan Intera)	
	Pre	**Post**	**Pre**	**Post**	**Pre**	**Post**
Accuracy	0.807 ± 0.05	0.858 ± 0.03	0.731 ± 0.05	0.738 ± 0.05	0.751 ± 0.06	0.757 ± 0.07
Precision	0.784 ± 0.07	0.859 ± 0.03	0.740 ± 0.08	0.744 ± 0.07	0.742 ± 0.06	0.804 ± 0.09
Recall	0.861 ± 0.09	0.861 ± 0.07	0.733±0.101	0.742 ± 0.09	0.775±0.114	0.700 ± 0.19
F1-score	0.816 ± 0.05	0.858 ± 0.03	0.730 ± 0.06	0.737 ± 0.05	0.754 ± 0.07	0.731 ± 0.11

Table 10: Comparison of the classifier quality from all models before and after harmonization.

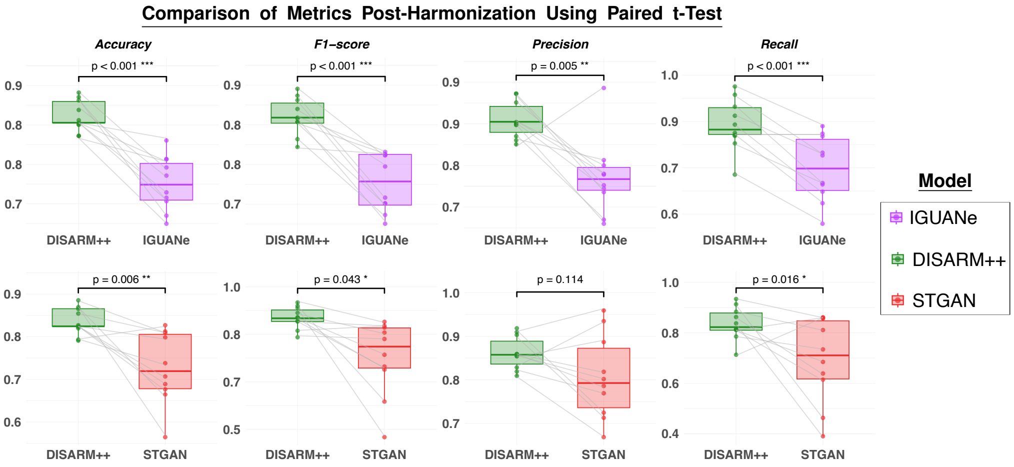


Figure 12: Comparison of metrics between DISARM++ and IGUANe and between DISARM++ and STGAN, using a paired bootstrap t-test, post-harmonization harmonization across the 10 random splits.

6.3.4. Diagnosis Prediction

This section presents the results of the downstream analysis described in Section 5.4.4 aiming at the prediction of diagnosis, i.e., AD or MCI. The LR model trained on raw MRI volumes (without preprocessing or harmonization) achieves an AUC of 0.6284. After applying DISARM++ harmonization, the AUC significantly improves to 0.9459, highlighting a substantial enhancement in classification performance. In comparison, STGAN harmonization increases the AUC to 0.7353, reflecting a more moderate improvement. IGUANe achieves an AUC of 0.8176, outperforming STGAN but not reaching the performance of DISARM++.

7. Discussion

In this study, we introduced a novel approach to harmonizing T1-weighted MR images acquired from different scanners. Unlike existing harmonization techniques that focus on standardizing image-derived features, our method directly harmonizes the brain MRI. This ensures that downstream features extracted from the harmonized images are inherently consistent, enhancing their reliability for various analysis. Our image-based approach allows for image transfer from different scanners in two distinct ways: (1) transferring images to a *scanner-free* space, ensuring consistent appearances regardless of the original scanner source; (2) mapping images to the space of one of the scanners used in the model's training, embedding the unique characteristics of the selected scanner into the transferred image. A key strength of our model lies in its ability to generalize effectively to unseen scanners not included in the training set. We evaluated our method using MR images from healthy controls across different scanners, traveling subjects, and patients with AD. Additionally, we tested the model's performance across multiple applications, including brain age prediction, biomarkers extraction, AD classification, and diagnosis prediction. In all cases, our method demonstrated significant improvements in reliability and predictive accuracy compared to existing state-of-the-art image-based approaches, such as STGAN and IGUANe. Furthermore, our harmonization model eliminates the need for time-intensive preprocessing steps, such as skull-stripping, which can introduce errors by either removing brain tissue or retaining non-brain structures. This feature makes our approach particularly advantageous for applications requiring analysis of the entire head, such as research on head trauma or cranial deformities. Notably, our method provides a robust tool for harmonizing images without the necessity of a new training phase, allowing seamless integration into various neuroimaging workflows.

8. Conclusions

Our results demonstrate that the proposed harmonization model offers superior performance over existing methods, ensuring reproducible and consistent MRI-based analyses across diverse scanning environments. Importantly, our model does not require retraining when applied to new data, making it a practical and scalable solution for large-scale neuroimaging studies. By improving the harmonization of MR images at the source level rather than at the feature level, our method enhances the accuracy and reliability of downstream analyses while reducing preprocessing complexity. The ability to harmonize images without the need for a new training phase further underscores its efficiency and adaptability across different applications. Future work will focus on further optimizing the preprocessing pipeline by eliminating the bias field correction step, which would further streamline image processing. Additionally, we aim to extend our approach to other imaging modalities and anatomical regions. To enhance the model's robustness, future studies should include experiments with a broader range of pathologies, ensuring that the harmonization method remains effective across diverse clinical and research settings.

Acknowledgments

This work has been partially supported by the Health Big Data Project (CCR-2018-23669122), funded by the Italian Ministry of Economy and Finance and coordinated by the Italian Ministry of Health and the network Alliance Against Cancer. The authors also acknowledge the support from the Italian Ministry of Health, with the project NeuroArtP3 (NET-2018-12366666). All authors acknowledge the following grants by the Italian Ministry of Health (RRC-2016-2361095; RRC-2017-2364915; RRC-2018-2365796; RCR-2019-23669119_001; RCR 2020-23670067; RCR 2022-23682285) and by the Ministry of Economy and Finance (CCR-2017-23669078) for the data acquisition. L. Caldera is funded by Health Big Data project sponsored by the Italian Ministry of Health (CCR-2018-23669122). L. Cavinato is funded by the National Plan for NRRP Complementary Investments "Advanced Technologies for Human-centred Medicine" (PNC0000003). L. Caldera, L.

Cavinato and F. Ieva acknowledge the MIUR Excellence Department Project 2023-2027 awarded to Dipartimento di Matematica, Politecnico di Milano.

RIN IRCCS List

The RIN Neuroimaging Network is constituted by the following centers: IRCCS Istituto Auxologico Italiano (Milan); IRCCS Ospedale pediatrico Bambino Gesù (Rome); Fondazione IRCCS Istituto neurologico "Carlo Besta" (Milan); IRCCS Centro Neurolesi "Bonino Pulejo" (Messina); Centro IRCCS "Santa Maria nascente" - Don Gnocchi (Milan); IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli (Brescia); IRCCS Ospedale pediatrico "Giannina Gaslini" (Genoa); IRCCS Istituto Clinico Humanitas (Milan); Istituto di Ricerche Farmacologiche Mario Negri IRCCS (Milan); Istituti Clinici Scientifici Maugeri, IRCCS (Pavia); IRCCS Eugenio Medea (Bosisio Parini); Fondazione IRCCS Istituto Neurologico "Casimiro Mondino" (Pavia); IRCCS NEUROMED – Istituto Neurologico Mediterraneo (Pozzilli); IRCCS Associazione Oasi Maria SS Onlus – Troina (Enna); Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico (Milan); IRCCS Fondazione Ospedale San Camillo (Venice); IRCCS Ospedale San Raffaele (Milan); IRCCS Fondazione Santa Lucia (Rome); IRCCS Istituto di Scienze Neurologiche (Bologna); IRCCS SDN Istituto di ricerca diagnostica e nucleare (Naples); IRCCS Fondazione Stella Maris (Pisa); IRCCS San Martino (Genova); IRCCS Gemelli (Roma).

Code and Data Availability

The implementation of the proposed approach is accessible at this link. ADNI3, PPMI, SALD, IXI, SRPBS are publicy available datasets. The private datasets that were used for testing in this study are available from the Italian Neuroimaging Network (RIN) and Ospedale Policlinico San Martino (NeurArtP3) but restrictions apply to the availability of these data, which were used under license for the current study, and so they are not publicly available. Data are however available from the authors upon reasonable request and with permission of both Italian Neuroimaging Network (RIN) and Ospedale Policlinico San Martino.

Appendix

Appendix A: Detailed Description of the MRI Datasets

This section presents a comprehensive overview of the MR T1-weighted images used in the main manuscript. Tables .11 and .12 provide detailed descriptions of the datasets for healthy controls and Alzheimer's disease (AD) patients, respectively. These tables include essential information such as scanner types, manufacturers, the number of images, field strengths, and participant age ranges. We obtained T1-weighted MR images of healthy controls from five distinct datasets. Specifically, the collection comprises 313 images from the Alzheimer's Disease Neuroimaging Initiative (ADNI3), acquired using five different scanners; 60 images from the Parkinson's Progression Markers Initiative (PPMI), obtained with three scanners; 581 images from the IXI Brain Development Dataset (IXI), acquired with three scanners; 494 images from the Southwest University Adult Lifespan Dataset (SALD), captured using a single scanner; and 117 images from a private dataset provided by the Italian Neuroimaging Network (RIN), collected with six different scanners. For AD patients, the dataset includes 41 MR scans from the private NeuroArtP3 dataset and 61 MR scans from the public ADNI3 dataset. Additionally, we include data from seven subjects in the SRPBS traveling subjects dataset [40]. Table .13 provides a detailed breakdown of this dataset, including the number of sites, scanner models, field strengths, and demographic information (age and sex) of the subjects. The "Number of Sites" column represents the locations where each subject underwent MRI scans, corresponding to the total number of images considered per subject. In the "Scanner Models" column, the numbers in parentheses indicate the count of images obtained using each scanner model across different sites.

Dataset	Scanner Model	Manufacturer	Img. #	Total	Field Strength	Age Range
ADNI3 [13]	Achieva dStream	Philips	26	313	3.0T	50-95
	Prisma Fit	SIEMENS	167			
	Prisma	SIEMENS	69			
	Skyra	SIEMENS	38			
	Achieva	Philips	13			
PPMI [23]	Achieva dStream	Philips	5	60	3.0T	60-80
	Achieva	Philips	14			
	Triotim	SIEMENS	41			
IXI [3]	Gyroscan Intera	Philips	322	581	1.5T	20-86
	Intera	Philips	185		3.0T	
	Unknown	GE	74		1.5T	
SALD [47]	Trio	SIEMENS	494	494	3.0T	19-80
RIN [25]	Achieva dStream	Philips	17	115	3.0T	49-80
	DISCOVERY MR750	GE	20			
	Ingenia CX	Philips	23			
	Prisma	SIEMENS	40			
	Skyra	SIEMENS	12			
	Skyra Fit	SIEMENS	3			

Table .11: Healthy controls datasets description, including scanner types, manufacturers, number of images, field strengths, and participant age range.

Dataset	Scanner Model	Manufacturer	Img. #	Total	Field Strength	Age Range
NeuroArtP3 [22]	Achieva	Philips	17	41	1.5T	63-81
	Genesis Signa	GE	6			
	Signa HDxt	GE	14			
	Ingenia	Philips	2			
	Intera	SIEMENS	1			
	Magnetom Espree	SIEMENS	1			
ADNI3 [13]	Prisma	SIEMENS	9	61	3.0T	55-89
	Ingenia	Philips	5			
	Prisma Fit	SIEMENS	20			
	Biograph mMR	SIEMENS	2			
	Skyra	SIEMENS	11			
	Triotim	SIEMENS	3			
	Ingenia Elition X	Philips	1			
	Verio	SIEMENS	5			
	Achieva	Philips	1			
	Achieva dStream	Philips	4			

Table .12: Alzheimer's Disease (AD) patients datasets description, including scanner types, manufacturers, number of images, field strengths, and participant age range.

Dataset	Subject	No. Sites	Scanner Models	Field Strength	Age	Sex
SRPBS [40]	1	10	Triotim, Verio (2), Spectra, Signa HDxt, Achieva (3), DISCOVERY MR750w, Skyra	3T	25	M
	2	10	Triotim (2), Verio (2), Spectra, Signa HDxt, Skyra, DISCOVERY MR750w, Achieva (2)	3T	27	M
	3	9	Triotim, Verio (2), Spectra, Signa HDxt, Achieva (3), Skyra	3T	26	M
	6	10	Triotim (2), Verio (2), Spectra, Signa HDxt, Achieva (3), Skyra	3T	24	M
	7	8	Triotim, Verio, Spectra, Signa HDxt, Achieva (3), Skyra	3T	25	M
	8	11	Triotim (2), Verio (2), Spectra, Signa HDxt, Achieva (3), DISCOVERY MR750w, Skyra	3T	28	M
	9	10	Triotim, Verio (2), Spectra, Signa HDxt, Achieva (3), DISCOVERY MR750w, Skyra	3T	30	M

Table .13: Description of the subjects from the SRPBS [40] traveling subject dataset, including the number of sites, scanner models, field strength, age, and sex. The number of sites column ("No. Sites") indicates the locations where each subject underwent MRI scans, corresponding to the total number of images considered for each subject. In the "Scanner Models" column, numbers in parentheses represent the count of images acquired with the specified scanner model across different sites.

Appendix B: Additional Details on Training Procedure

Table .14 provides details on the loss weights used during model training, the dimensions of the spaces $\mathcal{B}$ and $\mathcal{S}$, and the number of training iterations.

$\lambda_{\mathbf{cc}}$	$\lambda_{\mathbf{rec}}$	$\lambda_{\mathbf{lat}}$	$\lambda_{\mathbf{KL}}$	$\lambda_{\mathbf{sf}}$	$\lambda^{\mathbf{b}}_{\mathbf{adv}}$	$\lambda^{\mathbf{s}}_{\mathbf{cls}}$	$\lambda^{\mathbf{s}}_{\mathbf{adv}}$	**dim**($\mathcal{B}$)	**dim**($\mathcal{S}$)	**Number of Iterations**
10	10	8	0.01	7	1	3	10	$(91, 109, 91)$	16	69k

Table .14: Implementation details.

Appendix C: Additional Details on the Experiments

In this section, we provide additional details regarding the experiments discussed in Section 5 of the main manuscript.

Detailed Description of the Evaluation Metrics and Statistical Tests

We outline the metrics and statistical tests used to evaluate the performance of harmonization methods and describe how they are applied.

Anatomical Structure Metrics

For the assessment of anatomical structure preservation and the quality of generated images, we employ four metrics: (1) the structural component of the SSIM index [46], referred to as Struct-SSIM, the complete SSIM index [46], the Learned Perceptual Image Patch Similarity (LPIPS) metric [48], and the Fréchet Inception Distance (FID) [11].

Complete SSIM and Structural SSIM

The complete SSIM index [46] evaluates similarity considering luminance, contrast, and structure. A higher SSIM score indicates greater overall similarity. In our context, it is used to measure the similarity between harmonized images in the traveling subject dataset. The structural component of the SSIM index [46], referred to as Struct-SSIM, focuses on assessing the similarity of structural patterns between the reference and generated images by comparing local spatial structures independently of luminance and contrast, evaluating the preservation of fine details and spatial coherence. A higher Struct-SSIM score indicates greater structural similarity. In our context, it is used to measure the structural similarity between the original images and their harmonized versions. Since these metrics are defined for 2D images, to assess the similarity between two 3D MR scans, we calculate the metric for each slice in all three directions (sagittal, coronal, axial) and then average the results to obtain the overall similarity of the two 3D MR scans. The mathematical definitions of SSIM and Struct-SSIM are given as follows:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}; \tag{.1}$$

$$\text{Struct-SSIM}(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{.2}$$

where μ_x and μ_y are the means, σ_x^2 and σ_y^2 are the variances, and σ_{xy} is the covariance between images x and y. The terms C_1, C_2, and C_3 are small constants for numerical stability.

Learned Perceptual Image Patch Similarity (LPIPS)

The Learned Perceptual Image Patch Similarity (LPIPS) metric [48], which measures perceptual similarity between two images by comparing feature activations from a deep neural network, making it sensitive to semantic and textural differences and aligned with human visual perception. A lower LPIPS score indicates higher similarity. As the first index, we use the LPIPS to evaluate the perceptual similarity between pairs of original and harmonized images. Specifically, since LPIPS is defined for RGB 2D images, the comparison is performed by extracting 2D slices from the MRI volumes and converting them to RGB by replicating the single channel across the three color channels. Finally, given the computational intensity of LPIPS, we limit comparisons to a subset B of central slices for each of the T images considered. This process is applied across axial, sagittal, and coronal orientations, and the final LPIPS value is the average across all $3 \times B \times T$ slice comparisons. The LPIPS formula is given by:

$$\mathrm{LPIPS}(x, y) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \| w_l \big(f^l(x)_{hw} - f^l(y)_{hw} \big) \|^2 \qquad (.3)$$

where $f^l(x)$ and $f^l(y)$ are the deep features at layer l for images x and y, w_l are the learned weights, and H_l and W_l denote the height and width of the feature map at layer l.

Fréchet Inception Distance (FID)

The Fréchet Inception Distance (FID) [11], which quantifies the distributional similarity between real and generated images by comparing their feature embeddings extracted from a pre-trained Inception network, effectively assessing the overall quality and realism of the generated images, with lower values indicating closer alignment to the real data distribution. Similar to the LPIPS, the FID is defined for RGB 2D images, so we apply the same conversion process. However, since FID compares sets of real and generated images and outputs a single scalar value, we compute it by comparing the set of original images with the set of harmonized images. Specifically, we extract B central slices from each of the T images considered for both the original and harmonized sets. This results in two sets, each containing $B \times T$ slices. The FID is computed by comparing these sets separately for the axial, sagittal, and coronal orientations. Finally, the overall FID score is obtained by averaging the FID values from all three orientations. The FID formula is given by:

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}\right) \tag{.4}$$

where μ_r and μ_g represent the means, and Σ_r and Σ_g are the covariances of the real and generated image features.

Harmonization Metrics

To evaluate the transfer of scanner characteristics, we assessed the similarity of voxel intensity distributions before and after harmonization using three metrics: Jensen-Shannon Divergence (JSD) [19], Hellinger Distance (HD) [17, 31], and Wasserstein Distance (WD) [43]. Practically, given the MRI scans having dimension $(1, H, W, D)$ — where H represents height, W represents width, and D represents depth — we computed these metrics based on the set of $1 \times H \times W \times D$ voxel intensity distributions derived from MRI scans, considering pre- and post-harmonization data. Specifically, we compute the empirical distribution by estimating the underlying probability distribution of the voxel intensities for each unfolded image, and we average the distributions of images belonging to the same scanner. In this way, we obtain G distributions — with G being the number of the test scanners — from pre-harmonization images and G distributions from post-harmonization images. In each of the two sets of G distributions, we computed the values for the three metrics across all possible pairs to quantify the similarity between them. Therefore, when reporting these metrics, we present the mean and standard deviation of all pairwise comparison values, providing a general assessment for all scanners. We perform this step for both pre-harmonization and post-harmonization sets, enabling the comparison of the similarity between the distributions before and after harmonization. To statistically evaluate the effectiveness of harmonization, we perform a paired t-test comparing the values of the similarity metrics before and after harmonization. This test assesses whether the mean difference is significantly different from zero. If the differences do not follow a normal distribution, we employ a bootstrap paired t-test, resampling the data to estimate the distribution of differences. We report the 95% confidence interval (CI) for these differences; if the CI does not include zero, it indicates a significant effect of harmonization on voxel intensity similarity.

Besides the three aforementioned indexes, in certain analyses, we also employed the K-sample Anderson-Darling test (AD-test) [33], a non-parametric test that evaluates whether multiple samples originate from the same distribution. This test was applied to the set of G distributions, separately for pre- and post-harmonization images. If we accept the null hypothesis, it suggests no significant difference between distributions, indicating successful harmonization. Conversely, rejecting the null hypothesis implies that at least one distribution differs, suggesting incomplete harmonization.

Jensen-Shannon Divergence (JSD)
The Jensen–Shannon Divergence (JSD) [19] is a symmetric and bounded metric that quantifies the difference between two probability distributions. Unlike the Kullback–Leibler (KL) divergence, which is asymmetric, the JSD improves upon it by symmetrizing the measure. This is achieved by computing the KL divergence of each distribution relative to their average distribution $M = \frac{1}{2}(P+Q)$, and then taking the mean of these two values. Formally, the JSD is defined as the average of the KL divergences from P and Q to M as:

$$\mathrm{JSD}(P, Q) = \frac{1}{2}\mathrm{KL}(P\|M) + \frac{1}{2}\mathrm{KL}(Q\|M) \tag{.5}$$

where $\mathrm{KL}(P\|M)$ is the Kullback-Leibler divergence between P and M. The JSD is bounded between 0 and log(2), where:

- $\mathrm{JSD}(P, Q) = 0$ indicates that the two distributions are identical.
- $\mathrm{JSD}(P, Q) = \log(2)$ indicates that the distributions are maximally different.

Because of its symmetry and bounded nature, the JSD is often preferred for comparing distributions when assessing their similarity or dissimilarity, especially when dealing with probability distributions that may have different supports or shapes. It has strong interpretability and guarantees a finite value even when distributions exhibit significant differences.

Hellinger Distance (HD)
The Hellinger Distance (HD) is a metric derived from the Bhattacharyya coefficient, quantifying the similarity between two probability distributions by focusing on the overlap of their supports. It is particularly useful in geometric comparisons, measuring the distance between two distributions in terms of their density overlap. The Hellinger distance is defined as:

$$\mathrm{HD}(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{x}\left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2} \tag{.6}$$

where $\sqrt{P}$ and $\sqrt{Q}$ represent the square roots of the probability densities P and Q. The Hellinger distance ranges from 0 to 1:

- $\mathrm{HD}(P,Q) = 0$ indicates that the two distributions are identical.
- $\mathrm{HD}(P,Q) = 1$ indicates that the distributions are completely disjoint.

One of the strengths of the Hellinger distance is its sensitivity to differences in the tails of distributions. This sensitivity can reveal subtle distinctions that might be overlooked by other metrics. As such, the Hellinger distance is often preferred in cases where small differences in tail behavior are of importance.

Wasserstein Distance (WD)
The Wasserstein Distance (WD) measures the minimal cost of transforming one distribution into another. This cost is computed based on how much mass needs to be moved and how far it must be transported to convert one distribution into another. The Wasserstein distance considers both the magnitude of differences between distributions and the spatial arrangement of these differences. Mathematically, the Wasserstein distance is defined as the infimum of the total transportation cost over all possible couplings between the two distributions P and Q:

$$\mathrm{WD}(P,Q) = \int_{\mathbb{R}} |F_P(x) - F_Q(x)|\, dx \tag{.7}$$

where $F_P(x)$ and $F_Q(x)$ are the cumulative distribution functions of P and Q, respectively. The Wasserstein distance is sensitive to the geometry of distributions, making it particularly useful when the spatial arrangement of data is crucial. The Wasserstein distance is unbounded and can take arbitrarily large values, especially when the supports of the two distributions are far apart.

Using all three metrics — Jensen-Shannon Divergence (JSD), Hellinger Distance (HD), and Wasserstein Distance (WD) — enables a comprehensive and nuanced assessment of probability distribution similarity. Each metric highlights different aspects of distributional differences:

- **JSD**: Offers a symmetric, bounded measure of divergence, effective for general similarity comparisons but may overlook subtle differences in distribution tails.
- **HD**: Provides sensitivity to small discrepancies, especially in tails, making it valuable for geometric structure comparisons.
- **WD**: Captures minimal transformation cost between distributions, considering both magnitude and spatial arrangement, ensuring deeper interpretability.

By leveraging all three metrics, a more holistic perspective on the similarities and dissimilarities of probability distributions can be obtained. This multimetric approach ensures a robust comparison, especially in complex analyses where a single measure may not fully capture all relevant information.

Additional details on the Downstream Analysis

This section offers further details about the downstream analysis procedures.

Inter-Scanner Variability in MRI-Derived Brain Volumes

In the analysis, we employed Linear Mixed-effects Models (LMM) to predict these volumes based on age, treating scanner groups as random effects. The LMM is defined as follows:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + \epsilon_{ij}$$

where y_{ij} and x_{ij} represent the age and volume, respectively, for individual i scanned using scanner j. Here, β_0 is the fixed intercept, and β_1 is the fixed-effect coefficient for volume. The term u_j represents the random effect for scanner group j, assumed to be normally distributed with variance σ_u^2. ϵ_{ij} is the residual error term, assumed to be normally distributed with variance σ_ϵ^2.

The random effect u_j captures between-group variance, whereas the residual error ϵ_{ij} captures within-group variance. The ICC and R_m^2 are computed as:

$$\text{ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}; \qquad R_m^2 = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_\epsilon^2 + \sigma_u^2}$$

where σ_β^2 is the variance explained by the fixed effects.

Appendix D: Additional Details on the Results

In this section, we present supplementary information and visualizations related to the results discussed in Section 6 of the main manuscript.

Visual Assessment of Harmonization

In this section, we provide further details related to Section 6.2.1 of the main manuscript. Figure .13 displays 10 original images (one from each test scanner), showing slices from axial, coronal, and sagittal dimensions. These slices are presented alongside their harmonized counterparts with DISARM++ to the Gyroscan Intera reference scanner space.

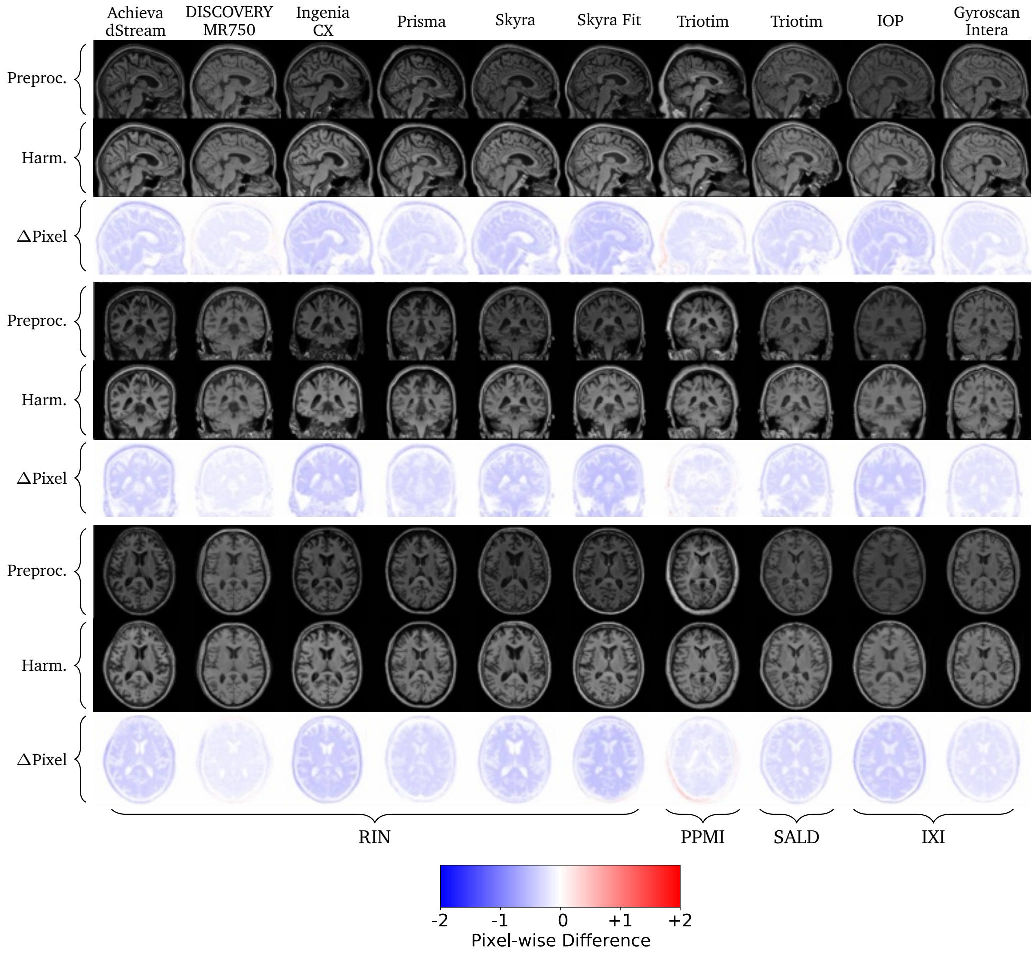


Figure .13: DISARM++ visual assessment of the harmonization to Gyroscan Intera reference scanner space. The figure displays slices from the axial, coronal, and sagittal dimensions for 10 original images—one per test scanner—alongside their corresponding harmonized slices. Heatmaps illustrate the pixel-wise differences between the harmonized images and their original counterparts.

Assessment of Anatomical Structure Preservation and Scanner Characteristics Transfer

In this section, we provide additional details related to Section 6.2.2 of the main manuscript. Figures .14, .15, and .16 present heatmaps illustrating all pairwise values for DISARM++, IGUANE, and STGAN before and after harmonization across the three metrics. Each cell in the heatmaps represents the similarity between the mean voxel intensity distributions of a pair of scanners.

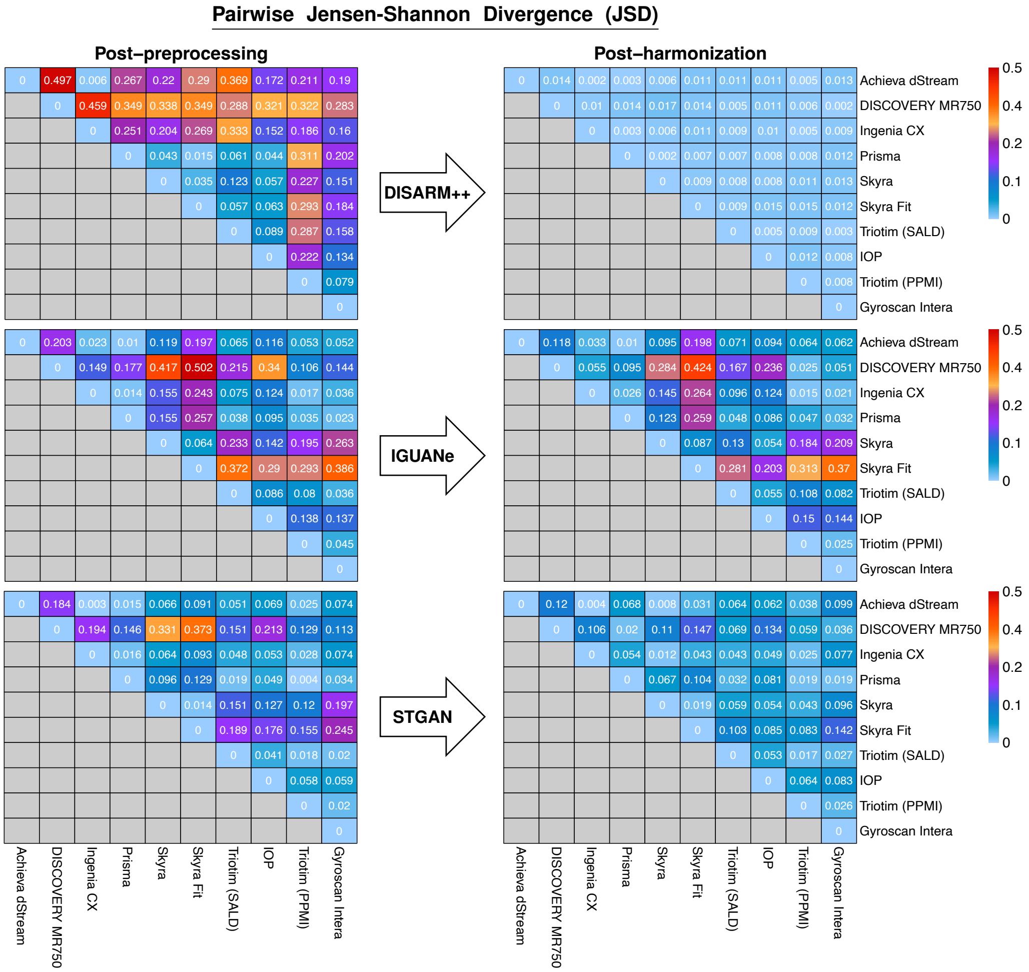


Figure .14: Heatmaps illustrating the Jensen-Shannon Divergence (JSD) between the mean distributions of each test scanner pair, both after preprocessing and post-harmonization using DISARM++, IGUANe, and STGAN, where lower values indicate greater similarity.

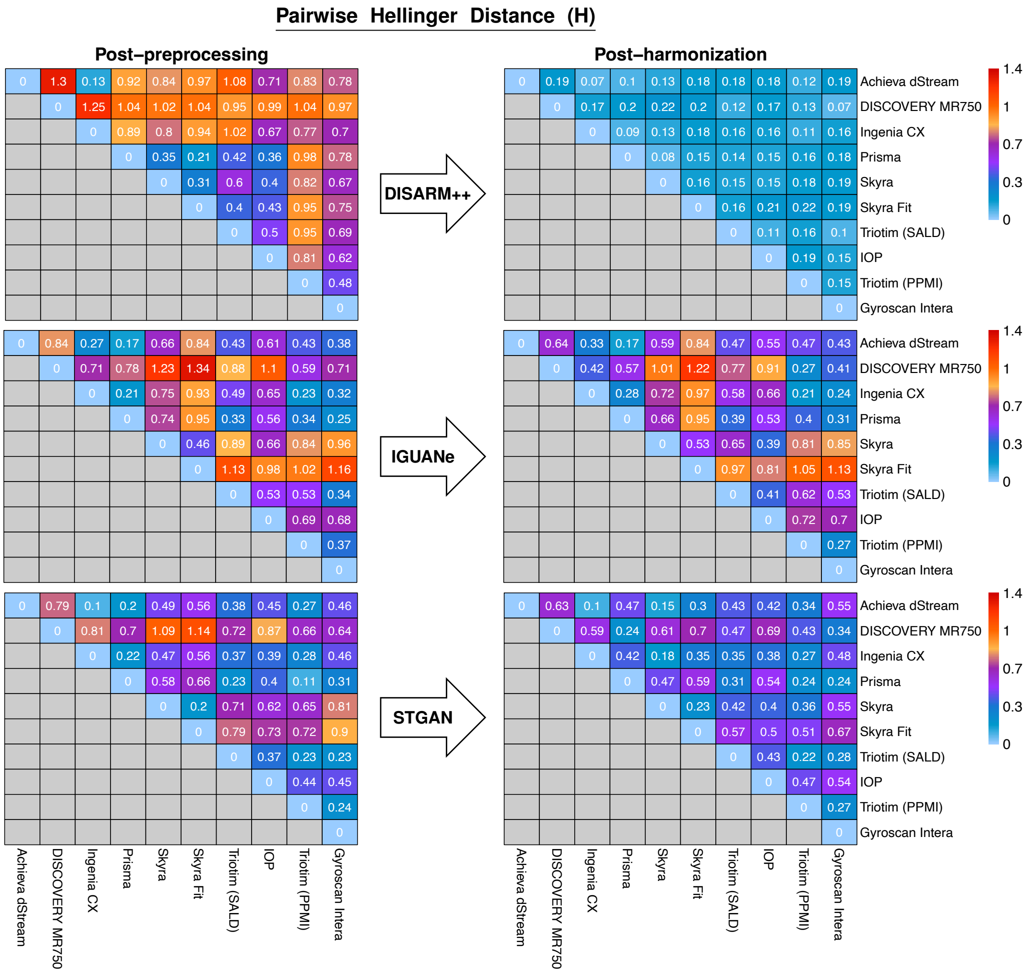
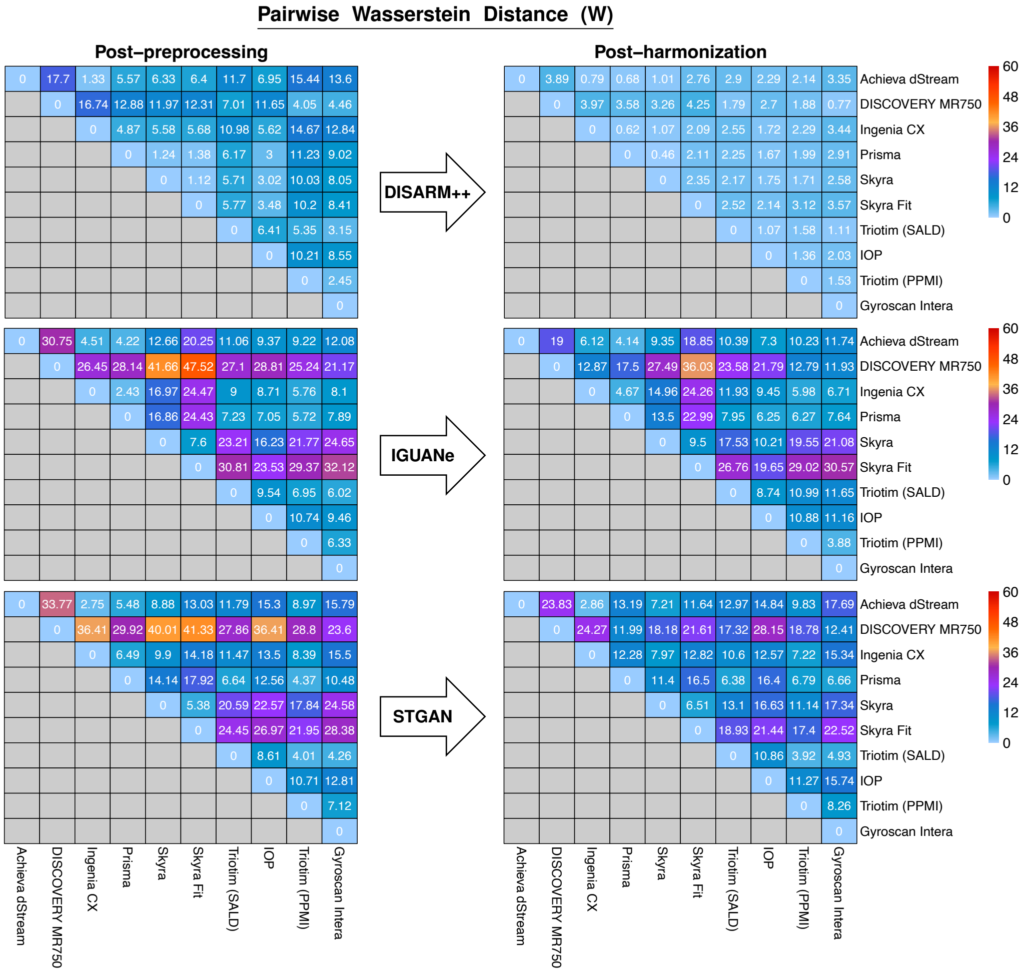


Figure .15: Heatmaps illustrating the Hellinger Distance (H) between the mean distributions of each test scanner pair, both after preprocessing and post-harmonization using DISARM++, IGUANe, and STGAN, where lower values indicate greater similarity.



DISARM++ — Post–preprocessing

	Achieva dStream	DISCOVERY MR750	Ingenia CX	Prisma	Skyra	Skyra Fit	Triotim (SALD)	IOP	Triotim (PPMI)	Gyroscan Intera
Achieva dStream	0	17.7	1.33	5.57	6.33	6.4	11.7	6.95	15.44	13.6
DISCOVERY MR750		0	16.74	12.88	11.97	12.31	7.01	11.65	4.05	4.46
Ingenia CX			0	4.87	5.58	5.68	10.98	5.62	14.67	12.84
Prisma				0	1.24	1.38	6.17	3	11.23	9.02
Skyra					0	1.12	5.71	3.02	10.03	8.05
Skyra Fit						0	5.77	3.48	10.2	8.41
Triotim (SALD)							0	6.41	5.35	3.15
IOP								0	10.21	8.55
Triotim (PPMI)									0	2.45
Gyroscan Intera										0

DISARM++ — Post–harmonization

	Achieva dStream	DISCOVERY MR750	Ingenia CX	Prisma	Skyra	Skyra Fit	Triotim (SALD)	IOP	Triotim (PPMI)	Gyroscan Intera
Achieva dStream	0	3.89	0.79	0.68	1.01	2.76	2.9	2.29	2.14	3.35
DISCOVERY MR750		0	3.97	3.58	3.26	4.25	1.79	2.7	1.88	0.77
Ingenia CX			0	0.62	1.07	2.09	2.55	1.72	2.29	3.44
Prisma				0	0.46	2.11	2.25	1.67	1.99	2.91
Skyra					0	2.35	2.17	1.75	1.71	2.58
Skyra Fit						0	2.52	2.14	3.12	3.57
Triotim (SALD)							0	1.07	1.58	1.11
IOP								0	1.36	2.03
Triotim (PPMI)									0	1.53
Gyroscan Intera										0

IGUANe — Post–preprocessing

	Achieva dStream	DISCOVERY MR750	Ingenia CX	Prisma	Skyra	Skyra Fit	Triotim (SALD)	IOP	Triotim (PPMI)	Gyroscan Intera
Achieva dStream	0	30.75	4.51	4.22	12.66	20.25	11.06	9.37	9.22	12.08
DISCOVERY MR750		0	26.45	28.14	41.66	47.52	27.1	28.81	25.24	21.17
Ingenia CX			0	2.43	16.97	24.47	9	8.71	5.76	8.1
Prisma				0	16.86	24.43	7.23	7.05	5.72	7.89
Skyra					0	7.6	23.21	16.23	21.77	24.65
Skyra Fit						0	30.81	23.53	29.37	32.12
Triotim (SALD)							0	9.54	6.95	6.02
IOP								0	10.74	9.46
Triotim (PPMI)									0	6.33
Gyroscan Intera										0

IGUANe — Post–harmonization

	Achieva dStream	DISCOVERY MR750	Ingenia CX	Prisma	Skyra	Skyra Fit	Triotim (SALD)	IOP	Triotim (PPMI)	Gyroscan Intera
Achieva dStream	0	19	6.12	4.14	9.35	18.85	10.39	7.3	10.23	11.74
DISCOVERY MR750		0	12.87	17.5	27.49	36.03	23.58	21.79	12.79	11.93
Ingenia CX			0	4.67	14.96	24.26	11.93	9.45	5.98	6.71
Prisma				0	13.5	22.99	7.95	6.25	6.27	7.64
Skyra					0	9.5	17.53	10.21	19.55	21.08
Skyra Fit						0	26.76	19.65	29.02	30.57
Triotim (SALD)							0	8.74	10.99	11.65
IOP								0	10.88	11.16
Triotim (PPMI)									0	3.88
Gyroscan Intera										0

STGAN — Post–preprocessing

	Achieva dStream	DISCOVERY MR750	Ingenia CX	Prisma	Skyra	Skyra Fit	Triotim (SALD)	IOP	Triotim (PPMI)	Gyroscan Intera
Achieva dStream	0	33.77	2.75	5.48	8.88	13.03	11.79	15.3	8.97	15.79
DISCOVERY MR750		0	36.41	29.92	40.01	41.33	27.86	36.41	28.8	23.6
Ingenia CX			0	6.49	9.9	14.18	11.47	13.5	8.39	15.5
Prisma				0	14.14	17.92	6.64	12.56	4.37	10.48
Skyra					0	5.38	20.59	22.57	17.84	24.58
Skyra Fit						0	24.45	26.97	21.95	28.38
Triotim (SALD)							0	8.61	4.01	4.26
IOP								0	10.71	12.81
Triotim (PPMI)									0	7.12
Gyroscan Intera										0

STGAN — Post–harmonization

	Achieva dStream	DISCOVERY MR750	Ingenia CX	Prisma	Skyra	Skyra Fit	Triotim (SALD)	IOP	Triotim (PPMI)	Gyroscan Intera
Achieva dStream	0	23.83	2.86	13.19	7.21	11.64	12.97	14.84	9.83	17.69
DISCOVERY MR750		0	24.27	11.99	18.18	21.61	17.32	28.15	18.78	12.41
Ingenia CX			0	12.28	7.97	12.82	10.6	12.57	7.22	15.34
Prisma				0	11.4	16.5	6.38	16.4	6.79	6.66
Skyra					0	6.51	13.1	16.63	11.14	17.34
Skyra Fit						0	18.93	21.44	17.4	22.52
Triotim (SALD)							0	10.86	3.92	4.93
IOP								0	11.27	15.74
Triotim (PPMI)									0	8.26
Gyroscan Intera										0

Figure .16: Heatmaps illustrating the Wasserstein Distance (W) between the mean distributions of each test scanner pair, both after preprocessing and post-harmonization using DISARM++, IGUANe, and STGAN, where lower values indicate greater similarity.

References

[1] Alotaibi, A. (2020). Deep generative adversarial networks for image-to-image translation: A review. *Symmetry*, 12(10):1705.

[2] An, L., Chen, J., Chen, P., Zhang, C., He, T., Chen, C., Zhou, J. H., Yeo, B. T., of Aging, L. S., Initiative, A. D. N., et al. (2022). Goal-specific brain mri harmonization. *Neuroimage*, 263:119570.

[3] Brain-Developement (2019). Brain-developement ixi. `https://brain-development.org/ixi-dataset/`.

[4] Caldera, L., Cavinato, L., Cappozzo, A., Cama, I., Garbarino, S., Cirone, A., Lodi, R., Tagliavini, F., Nigri, A., De Francesco, S., et al. (2025). Disarm: Disentangled scanner-free image generation via unsupervised image2image translation. In *International Workshop on Machine Learning in Clinical Neuroimaging*, pages 102–112. Springer.

[5] Cavinato, L., Massi, M. C., Sollini, M., Kirienko, M., and Ieva, F. (2023). Dual adversarial deconfounding autoencoder for joint batch-effects removal from multi-center and multi-scanner radiomics data. *Scientific Reports*, 13(1):18857.

[6] Chen, A. A., Beer, J. C., Tustison, N. J., Cook, P. A., Shinohara, R. T., Shou, H., and Initiative, A. D. N. (2022). Mitigating site effects in covariance for machine learning in neuroimaging data. *Human brain mapping*, 43(4):1179–1195.

[7] Choi, Y., Uh, Y., Yoo, J., and Ha, J.-W. (2020). Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197.

[8] Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., and Dale, A. M. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33:341–355.

[9] Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., and Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102.

[10] Guo, Z., Gu, Z., Zheng, B., Dong, J., and Zheng, H. (2022). Transformer for image harmonization and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[11] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

[12] Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.-P., Heiland, S., Wick, W., Bendszus, M., H Maier-Hein, K., and Kickingereder, P. (2019). Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17):4952–4964.

[13] Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L. G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C. S., Krueger, G., Ward, H. A., Metzger, G. J., Scott, K. T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J. P., Fleisher, A. S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., and Weiner, M. W. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691.

[14] Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841.

[15] Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *Neuroimage*, 62(2):782–790.

[16] Jernigan, T. L., Archibald, S. L., Fennema-Notestine, C., Gamst, A. C., Stout, J. C., Bonner, J., and Hesselink, J. R. (2001). Effects of age on tissues and regions of the cerebrum and cerebellum. *Neurobiology of aging*, 22(4):581–594.

[17] Kailath, T. (2003). The divergence and bhattacharyya distance measures in signal selection. *IEEE transactions on communication technology*, 15(1):52–60.

[18] Lee, H.-Y., Tseng, H.-Y., Mao, Q., Huang, J.-B., Lu, Y.-D., Singh, M., and Yang, M.-H. (2020). Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128:2402–2417.

[19] Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

[20] Liu, M., Zhu, A. H., Maiti, P., Thomopoulos, S. I., Gadewar, S., Chai, Y., Kim, H., Jahanshad, N., and Initiative, A. D. N. (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.

[21] Liu, S. and Yap, P.-T. (2024). Learning multi-site harmonization of magnetic resonance images without traveling human phantoms. *Communications Engineering*, 3(1):6.

[22] Malaguti, M. C., Gios, L., Giometto, B., Longo, C., Riello, M., Ottaviani, D., Pellegrini, M., Di Giacopo, R., Donner, D., Rozzanigo, U., Chierici, M., Moroni, M., Jurman, G., Bincoletto, G., Pardini, M., Bacchin, R., Nobili, F., Di Biasio, F., Avanzino, L., Marchese, R., Mandich, P., Garbarino, S., Pagano, M., Campi, C., Piana, M., Marenco, M., Uccelli, A., and Osmani, V. (2024). Artificial intelligence of imaging and clinical neurological data for predictive, preventive and personalized (p3) medicine for parkinson disease: The neuroartp3 protocol for a multicenter research study. *Plos one*, 19(3):e0300127.

[23] Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., Poewe, W., Mollenhauer, B., Sherer, T., Frasier, M., Meunier, C., Rudolph, A., Casaceli, C., Seibyl, J., Mendick, S., Schuff, N., Zhang, Y., Toga, A., Crawford, K., Ansbach, A., De Blasio, P., Piovella, M., Trojanowski, J., Shaw, L., Singleton, A., Hawkins, K., Eberling, J., Brooks, D., Russell, D., Leary, L., Factor, S., Sommerfeld, B., Hogarth, P., Pighetti, E., Williams, K., Standaert, D., Guthrie, S., Hauser, R., Delgado, H., Jankovic, J., Hunter, C., Stern, M., Tran, B., Leverenz, J., Baca, M., Frank, S., Thomas, C.-A., Richard, I., Deeley, C., Rees, L., Sprenger, F., Lang, E., Shill, H., Obradov, S., Fernandez, H., Winters, A., Berg, D., Gauss, K., Galasko, D., Fontaine, D.,

Mari, Z., Gerstenhaber, M., Brooks, D., Malloy, S., Barone, P., Longo, K., Comery, T., Ravina, B., Grachev, I., Gallagher, K., Collins, M., Widnell, K. L., Ostrowizki, S., Fontoura, P., La-Roche, F. H., Ho, T., Luthman, J., van der Brug, M., Reith, A. D., and Taylor, P. (2011). The parkinson progression marker initiative (ppmi). *Progress in Neurobiology*, 95(4):629–635.

[24] Moyer, D., Ver Steeg, G., Tax, C. M., and Thompson, P. M. (2020). Scanner invariant representations for diffusion mri harmonization. *Magnetic resonance in medicine*, 84(4):2174–2189.

[25] Nigri, A., Ferraro, S., Wheeler-Kingshott, C. A. G., Tosetti, M., Redolfi, A., Forloni, G., D'Angelo, E., Aquino, D., Biagi, L., Bosco, P., Carne, I., De Francesco, S., Demichelis, G., Gianeri, R., Lagana, M. M., Micotti, E., Napolitano, A., Palesi, F., Pirastru, A., Savini, G., Alberici, E., Amato, C., Arrigoni, F., Baglio, F., Bozzali, M., Castellano, A., Cavaliere, C., Contarino, V. E., Ferrazzi, G., Gaudino, S., Marino, S., Manzo, V., Pavone, L., Politi, L. S., Roccatagliata, L., Rognone, E., Rossi, A., Tonon, C., Lodi, R., Tagliavini, F., and Bruzzone, M. G. (2022). Quantitative mri harmonization to maximize clinical impact: the rin–neuroimaging network. *Frontiers in Neurology*, 13:855125.

[26] Nikulin, M. S. (2001). Hellinger distance. In Hazewinkel, M., editor, *Encyclopedia of Mathematics*. Springer. https://encyclopediaofmath.org/wiki/Hellinger_distance.

[27] Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6(1):405–431.

[28] Pérez-García, F., Sparks, R., and Ourselin, S. (2021). Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208:106236.

[29] Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I. M., Satterthwaite, T. D., Fan, Y., et al. (2020). Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208:116450.

[30] Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quidé, Y., Green, M. J., Weickert, C. S., Weickert, T., Bruggemann, J., Kircher, T., Nenadić, I., Cairns, M. J., Seal, M., Schall, U., Henskens, F., Fullerton, J. M., Mowry, B., Pantelis, C., Lenroot, R., Cropley, V., Loughland, C., Scott, R., Wolf, D., Satterthwaite, T. D., Tan, Y., Sim, K., Piras, F., Spalletta, G., Banaj, N., Pomarol-Clotet, E., Solanes, A., Albajes-Eizagirre, A., Canales-Rodríguez, E. J., Sarro, S., Di Giorgio, A., Bertolino, A., Stäblein, M., Oertel, V., Knöchel, C., Borgwardt, S., du Plessis, S., Yun, J.-Y., Kwon, J. S., Dannlowski, U., Hahn, T., Grotegerd, D., Alloza, C., Arango, C., Janssen, J., Díaz-Caneja, C., Jiang, W., Calhoun, V., Ehrlich, S., Yang, K., Cascella, N. G., Takayanagi, Y., Sawa, A., Tomyshev, A., Lebedeva, I., Kaleda, V., Kirschner, M., Hoschl, C., Tomecek, D., Skoch, A., van Amelsvoort, T., Bakker, G., James, A., Preda, A., Weideman, A., Stein, D. J., Howells, F., Uhlmann, A., Temmingh, H., López-Jaramillo, C., Díaz-Zuluaga, A., Fortea, L., Martinez-Heras, E., Solana, E., Llufriu, S., Jahanshad, N., Thompson, P., Turner, J., and van Erp, T. (2020). Increased power by harmonizing structural mri site differences with the combat batch adjustment method in enigma. *NeuroImage*, 218:116956.

[31] Rao, C. R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. *Qüestiió: quaderns d'estadística i investigació operativa*.

[32] Roca, V., Kuchcinski, G., Pruvo, J.-P., Manouvriez, D., and Lopes, R. (2024). Iguane: a 3d generalizable cyclegan for multicenter harmonization of brain mr images. *arXiv preprint arXiv:2402.03227*.

[33] Scholz, F. W. and Stephens, M. A. (1987). K-sample anderson–darling tests. *Journal of the American Statistical Association*, 82(399):918–924.

[34] Ségonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., and Fischl, B. (2004). A hybrid approach to the skull stripping problem in mri. *Neuroimage*, 22(3):1060–1075.

[35] Sele, S., Liem, F., Mérillat, S., and Jäncke, L. (2020). Decline variability of cortical and subcortical regions in aging: A longitudinal study. *Frontiers in human neuroscience*, 14:363.

[36] Shinohara, R. T., Oh, J., Nair, G., Calabresi, P. A., Davatzikos, C., Doshi, J., Henry, R. G., Kim, G., Linn, K. A., Papinutto, N., Pelletier, D.,

Pham, D. L., Reich, D. S., Rooney, W., Roy, S., Stern, W., Tummala, S., Yousuf, F., Zhu, A., Sicotte, N. L., Bakshi, R., and Cooperative, N. (2017). Volumetric analysis from a harmonized multisite brain mri study of a single subject with multiple sclerosis. *American Journal of Neuroradiology*, 38(8):1501–1509.

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

[38] Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., and Matthews, P. M. (2004). Advances in functional and structural mr image analysis and implementation as fsl. *Neuroimage*, 23:S208–S219.

[39] Takao, H., Hayashi, N., and Ohtomo, K. (2011). Effect of scanner in longitudinal studies of brain volume changes. *Journal of Magnetic Resonance Imaging*, 34(2):438–444.

[40] Tanaka, S. C., Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Mano, H., Yoshida, W., Seymour, B., Shimizu, T., Hosomi, K., Saitoh, Y., Kasai, K., Kato, N., Takahashi, H., Okamoto, Y., Yamashita, O., Kawato, M., and Imamizu, H. (2021). A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific data*, 8(1):227.

[41] Torbati, M. E., Minhas, D. S., Ahmad, G., O'Connor, E. E., Muschelli, J., Laymon, C. M., Yang, Z., Cohen, A. D., Aizenstein, H. J., Klunk, W. E., Christian, B. T., Hwang, S. J., Crainiceanu, C. M., and Tudorascu, D. L. (2021). A multi-scanner neuroimaging data harmonization using ravel and combat. *Neuroimage*, 245:118703.

[42] Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320.

[43] Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer.

[44] Walhovd, K. B., Fjell, A. M., Reinvang, I., Lundervold, A., Dale, A. M., Eilertsen, D. E., Quinn, B. T., Salat, D., Makris, N., and Fischl, B. (2005). Effects of age on volumes of cortex, white matter and subcortical structures. *Neurobiology of aging*, 26(9):1261–1270.

[45] Wang, P., Zheng, W., Chen, T., and Wang, Z. (2022). Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*.

[46] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

[47] Wei, D., Zhuang, K., Chen, Q., Yang, W., Liu, W., Wang, K., Sun, J., and Qiu, J. (2017). Structural and functional mri from a cross-sectional southwest university adult lifespan dataset (sald). *bioRxiv*, page 177279.

[48] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

[49] Zuo, L., Dewey, B. E., Liu, Y., He, Y., Newsome, S. D., Mowry, E. M., Resnick, S. M., Prince, J. L., and Carass, A. (2021). Unsupervised mr harmonization by learning disentangled representations using information bottleneck theory. *NeuroImage*, 243:118569.

[50] Zuo, X.-N., Xu, T., and Milham, M. P. (2019). Harnessing reliability for neuroscience research. *Nature human behaviour*, 3(8):768–771.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

17/2026 Caldera, L.; Bottacini, G.; Cavinato, L.
MAGIC-Flow: multiscale adaptive conditional flows for generation and interpretable classification

19/2026 Caldera, L.; Cavinato, L.; Ieva, F.
Scanner-agnostic MRI harmonization via SSIM-guided disentaglement

15/2026 Zecchi, A. A.; Sanavio, C.; Cappelli, L.; Perotto, S.; Roggero, A.; Succi, S.
Block encoding of sparse matrices with a periodic diagonal structure

14/2026 Agasisti, T.; Cannistrà, M.; Paganoni, A.M.
Nudging communication for students at risk: experimental evidence from an Italian university

13/2026 Dimola, N.; Coclite, A.; Zunino, P.
Neural Preconditioning via Krylov Subspace Geometry

12/2026 Corbetta A; Logan K.M.; Ferro M; Zuccolo L; Perola M.; Ganna A.; Di Angelantionio E.;Ieva F.
Longitudinal patterns of statin adherence and factors associated with decline in over one million individuals in Finland and Italy

11/2026 Cicalese, G.; Ciaramella, G.; Mazzieri, I.; Gander, M. J.
Optimized Schwarz Waveform Relaxation for the Damped Wave Equation

10/2026 Dimola, N.; Franco, N. R.; Zunino, P.
Numerical Solution of Mixed-Dimensional PDEs Using a Neural Preconditioner

09/2026 Manzoni, V.; Ieva, F.;Larranaga, A.C.; Vetrano, D.L.; Gregorio, C.
Hidden multistate models to study multimorbidity trajectories

08/2026 Micheletti, S.
Newmark time marching as a preconditioned iteration for large SPD linear systems