Medical Image Analysis 88 (2023) 102799

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

ImUnity: A generalizable VAE-GAN solution for multicenter MR image harmonization

Stenzel Cackowski, Emmanuel L. Barbier, Michel Dojat, Thomas Christen *

Univ. Grenoble Alpes, Inserm, U1216, Grenoble Institut Neurosciences, 38000 Grenoble, France

ARTICLE INFO

Keywords:
Brain
Deep Adversarial Network
Data harmonization
Self-supervised learning
Radiomic features

ABSTRACT

ImUnity is an original 2.5D deep-learning model designed for efficient and flexible MR image harmonization. A VAE-GAN network, coupled with a confusion module and an optional biological preservation module, uses multiple 2D slices taken from different anatomical locations in each subject of the training database, as well as image contrast transformations for its training. It eventually generates 'corrected' MR images that can be used for various multi-center population studies. Using 3 open source databases (ABIDE, OASIS and SRPBS), which contain MR images from multiple acquisition scanner types or vendors and a large range of subjects ages, we show that ImUnity: (1) outperforms state-of-the-art methods in terms of quality of images generated using traveling subjects; (2) removes sites or scanner biases while improving patients classification; (3) harmonizes data coming from new sites or scanners without the need for an additional fine-tuning and (4) allows the selection of multiple MR reconstructed images according to the desired applications. Tested here on T1-weighted images, ImUnity could be used to harmonize other types of medical images.

1. Introduction

Magnetic Resonance (MR) data acquired from the same patient but at different acquisition sites often lead to different MR images. This is due to the qualitative nature of MR data acquisition which produces weighted images (such as T1w or T2w) that are sensitive to technical choices (hardware, sequence or reconstruction parameters) as well as scanner artifacts. Consequently, pooling images from multi-center MR studies in order to approach a particular clinical or biological question does not guarantee an increase in statistical power because of a parallel increase in non-biological variance. These unwanted variations in image intensities also prevent large dissemination of machine learning tools that are trained on a specific site and may not generalize their model to other image providers (Liu et al., 2020).

Several solutions have been proposed in the last decade to harmonize data coming from multi-site or multi-scanner MR studies (see Table 1). Their goal is to remove confounding site, scanner or protocol effects, while preserving the biological information contained in the images. Classical post-processing steps such as standardization, global scaling (Fortin et al., 2017) or intensity histogram matching (Shinohara et al., 2014) have been shown to reduce the influence of site or scanner biases. However, they also tend to remove informative local intensity variations (Fortin et al., 2016). Statistical techniques, where image intensity and datasets bias are modeled in every voxel, have been proved to be more efficient. Ravel (Fortin et al., 2016), ComBat (Fortin et al., 2017), refined ComBat versions (Pomponio et al. (2019), Beer et al. (2020)) or dictionary learning (St-Jean et al., 2020) methods have been successfully used to analyze harmonization impacts on diffusion MRIs or longitudinal structural sequences. However, it can be noted that these techniques need to be adjusted every time new sites or scanners provide images to the database. Moreover, the same clinical individual information (such as patient age, sex, etc.) needs to be available in every center of the database. More recently, 2D deep-learning models such as CycleGAN (Zhu et al., 2018), Deep-Harmony (Dewey et al., 2019) or Calamity (Zuo et al., 2021) have shown encouraging results for structural MR image (T1w, T2w or FLAIR) harmonization but have also shown limitations. Briefly, CycleGAN, which consists in two Generative Adversarial Networks (GANs) working together, is restricted to the harmonization of 2 sites, and needs to be fine-tuned for every pair of sites. Deep-Harmony, a U-Net (Ronneberger et al., 2015) harmonization adaptation network, has the disadvantage of requiring for its training, traveling subjects (subjects who have been scanned successively at different sites or scanners) for all scanners/sites; a condition barely met in practice even in prospective studies. Similarly to CycleGAN, it is limited to two sites and needs to be fine-tuned. Calamity, an unsupervised deep-learning method, needs two different MR sequences

* Corresponding author.
E-mail addresses: sten.cackowski@gmail.com (S. Cackowski), emmanuel.barbier@univ-grenoble-alpes.fr (E.L. Barbier), michel.dojat@inserm.fr (M. Dojat), thomas.christen@univ-grenoble-alpes.fr (T. Christen).

https://doi.org/10.1016/j.media.2023.102799
Received 14 September 2021; Received in revised form 10 March 2023; Accepted 14 March 2023
Available online 24 March 2023
1361-8415/

Table 1
Versatility of deep-learning harmonization models.

	Traveling subjects	Fine-tuning for new clinical question	Fine-tuning for unseen sites	Max. number of target sites
Zhu et al. (2018) (CycleGAN)	Not required	Not required	Required	N = 2
Dewey et al. (2019) (Deep-Harmony)	Required	Not required	Required	N = 2
Zuo et al. (2021) (Calamity)	Not required	Not required	Required	N = number of training sites
Dinsdale et al. (2020), Guan et al. (2021)	Not required	Required	Not required	N > number of training sites
ImUnity (this study)	Not required	Not required	Not required	N > number of training sites

as inputs for every subject and needs fine-tuning when data from new sites are considered. Finally, Dinsdale et al. (2020) and Guan et al. (2021) have proposed to include unlearning modules or domain discriminators directly into their classification networks. As such, they learn how to remove datasets biases during their analyses without reconstructing harmonized MR images. They have shown improvements in brain tissues segmentation and brain disorder classification after harmonization. However, these techniques clearly require to be trained for every new clinical question while the former approaches harmonize data once for all.

We propose in this paper a new type of harmonization method, called ImUnity, based on 2.5D deep-learning, which extends previous techniques to offer a fast and flexible harmonization solution. ImUnity generates 'corrected' MR images that can then be utilized for various population imaging studies. To avoid the need for traveling subjects or multiple MR sequences in the database, our self-supervised Variational AutoEncoder (VAE-GAN) architecture uses for its training, multiple slices from the same individual and randomized image contrast transformations. It also unlearns center bias using a confusion module connected to its bottleneck while an optional biological module intends to preserve clinical features in the latent space. Once trained, this architecture should allow data coming from new sites or scanners to be harmonized without the need for fine-tuning. The architecture also allows estimates towards multiple target sites and then, users can choose multiple MR image reconstructions according to the chosen target domain (site or scanner).

To overcome the intrinsic problem of 2D generative models, i.e. the discontinuity in final outputs along the third axis, we introduce a 2.5D model that combines the outputs of 3 models, each one trained along a specific axis. This approach was first introduced to the MR harmonization field by Dewey et al. (2019), highlighting the great potential of such an approach.

To evaluate the efficiency and flexibility of our harmonization tool, we tested the approach using 3 open source databases that contain images from multiple acquisition sites, scanner vendors or strength of magnetic fields, and a large range of patients ages. For most of the experiments, ImUnity was trained using data from only one of the databases and then applied to the other two to evaluate generalization of the model. Quality of the reconstructed images, capacity of removing site or scanner bias and ability to classify patients were evaluated after data harmonization.

2. Materials and methods

2.1. Data

We used three open-source databases: (1) ABIDE, a multi-center project led by Di Martino et al. (2014), which focuses on Autism Spectrum Disorder (ASD). It gathers more than 1000 autistic patients and controls. For this study, we used T1-weighted scans from 11 different sites and scanners from 3 different constructors (3T scanners at 10 different sites and one 1.5T scanner at one site). Sites presenting data from a large range of ages (from 6 to 47 years, mean age = 12 years) were selected. In total, 621 T1-weighted scans (309 patients and 312 controls) were collected. (2) OASIS (LaMontagne et al., 2019) (Open Access Series of Imaging Studies) gathers T1-weighted scans from healthy (N=605) and Alzheimer's Disease (AD) (N=493) adult subjects who underwent several MR sessions on 4 different scanners from the same site. We used these traveling subjects (N = 1098) to validate the ability of our model to perform multi-scanners harmonization.

(3) SRPBS (Tanaka et al., 2021) is a multi-site database gathering multi-disorder subjects. We used 9 healthy adult traveling subjects to validate harmonization results between the different acquisition sites of the database (6 sites, 12 scanners from 3 different constructors). Note that SRPBS contains healthy adult brain scans while ABIDE (resp. OASIS) mainly includes healthy and pathological infant (resp. AD adult) brain scans, leading to large anatomical differences between images in the databases.

For each subject in each database, the brain was extracted using Robex (Iglesias et al., 2011) and N4Bias (Tustison et al., 2010) was used to correct for intensity inhomogeneities. MR images were first co-registered, using fsl-FLIRT (Jenkinson and Smith, 2001), to the publicly available and age specific 152-MNI templates (Sanchez et al., 2012). Then, White-Stripe normalization (Shinohara et al., 2014) was run to align white matter (WM) peaks between all subjects (each WM peak was aligned to 0.7 after rescaling the whole image between [0:1]).

After visual inspection to detect images with ROBEX defects or other artifacts, we eventually included 545, 1072 and 81 T1-weighted scans from ABIDE, OASIS and SRPBS databases respectively.

2.2. ImUnity's model

The architecture of our model derives from convolutional VAE-GANs and is described in Fig. 1. We adopted adversarial settings to ensure realistic outputs using a classical CNN as discriminator. The generator (here a VAE) learns how to represent input data into a lower dimension latent space (bottleneck). Information is then decoded to generate an output image. Inspired by Dinsdale et al. (2020), an unlearning center-bias module is connected to the bottleneck to limit the impact of site or scanner information. A biological preservation module can be inserted to maintain some biological information in the latent space representation. Technical details are provided below.

2.2.1. Modified VAE generator

Inspired by Zuo et al. (2021), our generator takes two 2D-structural images in the same orientation as input, randomly taken at two different locations in the 3D-MR stack of images of each subject to consider. The first (S_1) image is used by the first CNN to encode the 'anatomical' information using only convolutional filters to ensure preservation of spatial information. The second image (S_2) differs from S_1 because randomly taken in another part of the same brain and provides the initial 'contrast' information. At least 10 slices separate S2 from S1 (i.e. in our case 10 mm) ensuring that S_1 and S_2 have different anatomy (different location in the brain) but similar contrast (same scan). S_2 contrast is modified using a gamma function (or exponential correction): $I_\gamma = range\left(\frac{I}{range}\right)^\gamma$, where I represents voxels intensities and 'range' is the difference between maximum and minimum intensity values. The 'gamma' parameter γ is sampled uniformly between 0.5 and 1.5 for each new input 2d slice. This modified S_2^γ slice is used as input to a second CNN to encode the 'contrast information' followed by a dense layer to reduce spatial information. An example of different gamma transformations applied to MR brain scans from the same subject is given in Fig. 2. Once encoded, the two independent representations of S_1 and S_2^γ are concatenated to give a latent space

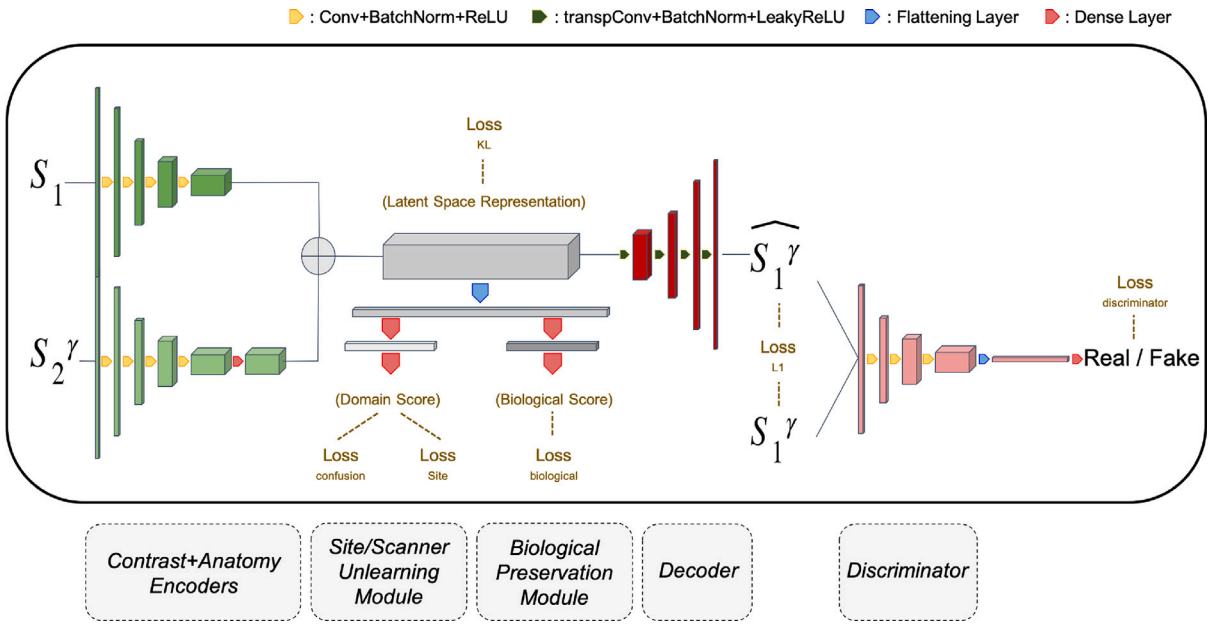

Fig. 1. ImUnity's architecture. The model involves: a modified VAE generator (2.2.1), a CNN discriminator, an additional Site Unlearning module (2.2.2) and an optional Biological module (2.2.3). Here the bottleneck corresponds to encoder's mean and variance outputs.

representation which is decoded to create the output $\hat{S_1^\gamma}$ using transposed convolutional filters. Eventually, this output is compared to the reference gamma modified slice S_1^γ. Note that this generator is trained in a self-supervised fashion as output labels are generated during the training phase and this training can be done on any MR dataset. It does not require additional information such as scanner, center or biological information.

2.2.2. Site/scanner-bias unlearning module

To ensure the task of "removing site or scanner bias", a module is directly connected to the encoders' outputs (latent space representation of inputs). The module can be seen as a domain (site or scanner) discriminator and is trained independently from the encoder to predict the scan's origin based on the latent space representation. On the other side, the encoder is trained in an adversarial fashion. A confusion loss is used to unlearn domain information. This principle has been introduced in the field of domain adaptation by Ganin et al. (2016) and has been adapted to medical imaging studies by Dinsdale et al. (2020). Originally, the module was incorporated directly in the model to unlearn datasets bias and to improve predictions. Here, it is used in the bottleneck as a "datasets bias filter", forcing the encoder to learn a domain-invariant data representation. Note that the architecture of ImUnity differs from that of Dinsdale et al. (2020), and so does the position of the bottleneck. Overall, the generator learns a shared latent space that encodes all information needed to generate harmonized scans. We also chose to avoid skip-connections in our network to ensure that site/scanner related information (present in input data) does not flow directly through these connections (as it would be the case in a UNet (Ronneberger et al., 2015) for example). The loss function for the site/scanner unlearning module is:

$$l_{site}(P,Y) = -(1/N)\sum_{i=1}^{N}\sum_{s=1}^{S} \mathbb{1}(y_i = s) log(p_i^s) \quad (1)$$

While the confusion loss used in the encoders' training is :

$$l_{confusion}(P) = -(1/N)\sum_{i=1}^{N}\sum_{s=1}^{S} log(p_i^s)/S \quad (2)$$

Here $P = [p_1; ...; p_S]$ is the softmax output from the module, corresponding to the probability to belong to different sites (1, ..., S), Y is the ground truth site affiliation vector, and N is the sample size.

2.2.3. Biological preservation module

An optional module seek to "preserve" biological information. It acts as a classifier of available biological information. For instance, features such as age or the presence of diseases can be introduced. Contrary to the unlearning module, the encoder is trained to minimize its loss function. This module is not mandatory, and a fully self-supervised learning model can be adopted when it is turned off. Then, the loss function of the biological preservation module for our particular application using the ABIDE database is:

$$l_{biological}(P,Y) = -(1/N)\sum_{i=1}^{N}\sum_{f\in features}\sum_{i=1}^{N} y_i^f log(p_i^f) + (1-y_i^f) log(1-p_i^f) \quad (3)$$

Here P represents module predictions for biological features of interest, N is the sample size, and Y is the ground truth vector. Note that in this study, the binary cross entropy formulation was used for the loss function because only two features (age and patient status, i.e. ASD) were considered.

2.2.4. A 2.5D solution

The presented architecture is in 2D. As presented in Figure Appendix E-right panel, one can see that a 2D model generates high-quality images along the training axis. However, its final 3D reconstruction suffers from artifacts along the two other axis. This is why in this study, we propose to use the above model in a 2.5D way. It consists of using three 2D models along each axis and to combine their results in order to have sharper output. We fuse predictions using the median value, an approach less sensitive to outliers than the mean value used in Dewey et al. (2019). Using a 2.5D approach generates higher quality images while keeping a number of parameters reasonable ($\sim 5.10E6$ for each model) without requiring more training datasets. The introduction of a 3D architecture would highly increase the number of parameters to estimate ($\sim 30.10E6$), therefore requiring more training datasets.

2.3. Training

For each model, training involves several independent steps, due to the adversarial context and the use of the additional modules.

- Training the discriminator consists in minimizing binary cross-entropy $l_{discriminator}$ between its predictions and the labels corresponding to the nature of the inputs (real or fake). Adversarially,

the generator learns how to maximize this loss function, forcing the generation of realistic outputs.

- Training the site/scanner unlearning module consists in minimizing the categorical cross-entropy (Eq. (1)) between its predictions and the site-affiliation labels. Adversarially, the generator is trained to minimize the confusion loss (Eq. (2)). It forces a site and scanner invariant representation of the dataset in the latent space, leading to uniform outputs of the unlearning module.
- In our ABIDE experiment, training the biological preservation module consisted in minimizing binary cross-entropy losses associated with each biological feature taken into account (here sex and patient status). Unlike the previous module, the loss $l_{biological}$ was directly integrated into the generator. This was intended to preserve some biological features in the latent space.
- In addition to previous loss functions involved in training the generator, a l_1 loss function is used to ensure a good mapping between input $(S_1; S_2^\gamma)$ and the generated output $\hat{S}_1^\gamma$ ($l_1 = mean(|\hat{S}_1^\gamma - S_1^\gamma|)$). Moreover, the use of the Kullback–Leibler divergence l_{KL} (see Eq. (5)) between features distributions and a Gaussian distribution ensures a dense data representation in latent space. Therefore, the global generator's loss function to minimize is:

$$l_{generator} = -\lambda_1 l_{discriminator} + \lambda_2 l_{confusion} + \lambda_3 l_{biological} + \lambda_4 l_1 + \lambda_5 l_{KL} \quad (4)$$

$$l_{KL}(P,Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \quad (5)$$

Here, λ factors control the relative contribution of each loss. In our study, we used: $\lambda_1 = 1$; $\lambda_2 = 1$; $\lambda_3 = 1$; $\lambda_4 = 100$; $\lambda_5 = 10^{-3}$ found empirically and with P and Q, two discrete distributions. Note that because the generator searches to fool the site-classification module by forcing an uniform prediction, we integrate the confusion loss (2) and not the module's loss (1) in Eq. (4).

2.4. Inference

When inferring harmonized data, the model only requires two slices as input (the one to harmonize and the reference one). Biological information is not fed into the model while inferring, but only used during the training phase. When biological information is not available the corresponding module is not active.

Note that for all the following experiments, ImUnity-2.5d was used.

The code of ImUnity is in open source access here: https://github.com/nifm-gin/dl_generic.

2.5. Experiments

The datasets extracted from the three databases were used to evaluate different aspects of our model. The impact on image quality in multi-site or multi-scanner harmonization was assessed using data from traveling subjects (ground truth) from the OASIS and SRPBS datasets. Ability to remove site information was evaluated using the ABIDE dataset. Finally, the benefits of harmonization between data provider centers were assessed using autism disorder prediction in children from the ABIDE dataset. To demonstrate the flexibility of ImUnity, all experiments were performed with the same model trained on data coming from the ABIDE database, unless specified. OASIS and SRPBS were used for the validation parts only. Each model was trained on 2D slices with at least 1% of brain tissue voxels. Training was run on a Nvidia GeForce 2080 RTX for 300 epochs using a learning rate of 10^{-4} and Adam optimizer.

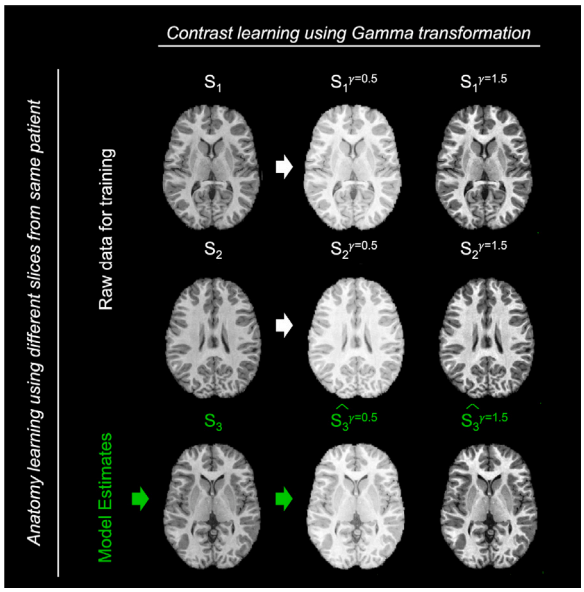

Fig. 2. Inputs and outputs of the model: Different slices from the same patient are used to encode the anatomical information. Gamma transformations are used to encode the contrast information. Rows represent different anatomical slices (S_1, S_2 and S_3) taken from the same subject. The first two rows present Gamma transformations used to train the model (S_1^γ and S_2^γ). The last raw shows model outputs ($\hat{S}_3^\gamma$) for estimated Gamma transformations for the slice S_3 not present in the training set. From left to right columns: original slices, Gamma modified slices with parameter 0.5, Gamma modified slices with parameter 1.5. Three 2D models along each direction are combined for the 2.5D approach.

2.5.1. Experiment 1 : Harmonization on traveling subjects (OASIS+SRPBS)

We first evaluated the ability of our model to transform images from one domain (site or scanner) to their equivalent in another domain. As SRPBS and OASIS databases contain traveling subjects, ground truth was available to assess ImUnity performance. In practice, one domain (acquisition site or scanner) was first selected as the reference for every subject. Individual scans were co-registered to their equivalent in the reference domain (to avoid variations between acquisitions due to movement). Then, all the images were transformed by the model into the reference domain. During this step, the slices to be harmonized (anatomy) were fed to the model along with the corresponding computed contrast slice from the reference domain. Finally, results obtained after transformation were compared to the ground truth, i.e. images acquired in the reference domain (traveling subject). Visual verification, image intensities histograms, and the Structural Similarity Index Metric (SSIM, Wang et al. (2003)) were used to assess image likeness. Paired t-test were used for statistical significance. Furthermore, the same model, trained on ABIDE data, was used for every site/scan of the other two databases to evaluate ImUnity's ability to generalize to sites never seen before. This experiment also evaluates ImUnity's versatility, either for the source domain or for the target domain (last two columns of Table 1). Additionally, we also trained and tested ImUnity on different sites combinations to evaluate the impact of the sample size and the population type.

2.5.2. Experiments 2 : Harmonization's effects on sites classification (ABIDE)

The second experiment evaluated the ability to detect the origin of data before and after harmonization. Harmonized data were obtained using a 5-fold cross-validation procedure on the ABIDE data. As no ground truth was available for this experiment, we considered harmonization impacts on classification algorithms. Standard Support Vector Machine (SVM) with a radial basis function kernel was used to classify the ABIDE data. The classifier worked on all radiomic features (N=101)

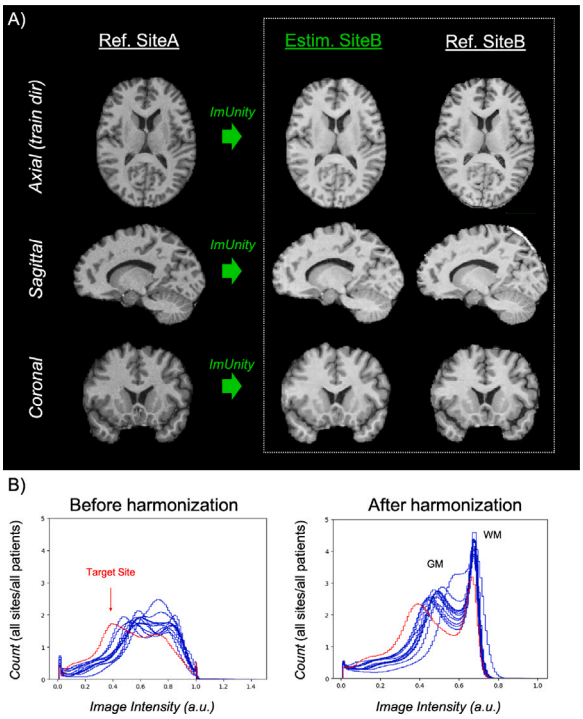

Fig. 3. Harmonization result on traveling subjects from the SRPBS database. (A) Left: 3D images from one patient (axial, sagittal, coronal views) acquired in site A before harmonization, Middle: ImUnity's harmonization to fit with acquisition at site B, Right: image acquired at site B (ground truth). (B) Images intensity distributions (all patients) before (Left) and after ImUnity's harmonization (Right). The red histogram corresponds to the site taken as reference for the harmonization process (target site).
GM = gray matter; WM = white matter.

extracted using the pyradiomics python API (van Griethuysen et al., 2017). These features aim to represent different aspects of MRI images, such as shape, contrast, or texture, and are known to be sensitive to site effects (Orlhac et al., 2019). The most 'correlated features' with sites affiliations before harmonization were selected for classification using Pearson tests (ran independently for each feature) using 10^{-3} as p-value threshold (30 features in total). Accuracy and Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve were used to evaluate the specificity and sensitivity of the site classifier.

2.5.3. Experiments 3 : Harmonization's effects on autism syndrome disorder prediction (ABIDE)

Similarly to Experiment 2, Experiment 3 evaluated the ability of our classifier to detect patients with ASD from the ABIDE database, before and after harmonization. Here, results were obtained following a 10-fold cross-validation procedure. The same trained model was used for different numbers of sites (and different combinations of sites) included in the ABIDE database.

3. Results

Experiment 1: Fig. 3A shows the results obtained for one traveling subject from the SRPBS database (2.5.1). Images are shown for one acquisition at site (A) before harmonization (3A, left), corrected by ImUnity to fit with acquisition at site B (3A middle) and the corresponding ground truth acquired at site B (3A right). One can notice the difference in image contrast between the 2 sites, highlighting the need for image harmonization, as well as the visual similarity between the harmonized image and the ground truth. It is interesting to observe that the anatomical structures of the input contrast reference are not propagated through the model, which explains small anatomical differences (e.g. superior sagittal sinus) between the model estimates and the ground truth. It is also worth noting that although each model was trained on 2D slices, combining outputs by taking the median for each voxel gives a final 3D reconstruction of the estimates of high quality in each orientation. In SM, Figure Appendix E highlights the positive impact of the 2D models fusion. For some subjects, 2D models present artifacts along the third axis that disappear after the median 2.5D combination. Fig. 3B shows the effects of ImUnity's harmonization on image intensity distributions for all selected subjects from the SRPBS database. The model was used to harmonize every image to a target site (indicated in red). An alignment of histograms is clearly observed after harmonization, with both gray and white matter peaks shifted. Changes in intensity distribution of the site of reference are due to preprocessing (see details in SM, Figure A.6, top row). The images obtained after ImUnity's harmonization of the OASIS datasets are also provided in SM, Figure B.7.

Quantitative results obtained with the SSIM metric in all traveling subjects are summarized in Table 2. Both multi-site (SRPBS) and multi-scanner (OASIS) experiments are shown. For the latter, results from the literature are also given for reference. It can be seen that when trained on the ABIDE database, ImUnity significantly increases the structural similarity in all cases and provides better performances compared to other deep learning approaches. Moreover, results from multi-scanner harmonization show that ImUnity performs well independently of the chosen reference domain. The last 2 rows in Table 2 present results obtained after training ImUnity on OASIS (to better match literature protocols) and SRPBS (to highlight sample size impact on the model's generalization) data. These models were used to harmonize OASIS as well as SRPBS data. Note here that multi-scanner harmonization results were obtained on healthy subjects only to match literature results. Similar results were obtained when including AD subjects (not shown).

Experiment 2: Fig. 4 shows ImUnity's harmonization effects on site classification on the ABIDE datasets (2.5.2) using tSNE (Maaten and Hinton, 2008), a dimension reduction algorithm, on radiomic features. Before harmonization, the presence of site clusters is clear. Once the data are harmonized using ImUnity, the points are shuffled and the accuracy of the SVM site prediction decreases from 0.74 to 0.37 (before and after harmonization respectively). This confirms the removal of site bias by ImUnity as the classifier is no longer able to correctly separate the sites. Note that Fig. 4 also shows that small clusters remain after harmonization, which could be explained by remaining site or scanner bias or by difference in demographic (age or sex)or biological (pathology) features between the respective groups of subjects. Additional results on the influence of the preprocessing step on sites classification are provided in SM Fig. A.6.

Experiment 3: Fig. 5 shows the capacity of our model to improve ASD prediction from the ABIDE datasets. Here, we used the same trained models to test the influence of different numbers of sites included in the database (from 2 to 11) as well as different combinations of those sites (for example 55 combinations of 2 sites taken among the 11 sites available). In every case, we observed a clear improvement of classification of autistic patients after harmonization as shown by increases in AUC provided by the SVM classifier. We show the results obtained with the best combination of sites as well as average and standard deviation of AUC with all combinations of sites. The preprocessing also has a positive impact on the prediction as shown in SM (Fig. A.6, bottom row).

4. Discussion

We have presented ImUnity, an original 2.5D harmonization tool for multi-center MRI databases. ImUnity shows high performances in term of quality of the generated harmonized images, as well as clear removal of the idiosyncratic bias attached to site-dependent image acquisition conditions. Moreover, the performed experiments clearly demonstrate ImUnity's versatility. By training ImUnity's model on datasets extracted

Table 2
SSIM in traveling subjects for multi-scanner (healthy subjects from the OASIS database) and multi-site (SRPBS database) harmonization. Results are compared to the literature when available.

Task	Multi-scanner harmonization.		Multi-site harmonization
Dataset	OASIS scanner F → scanner E	OASIS all scanners → scanner E	SRPBS all sites → UTO site
Raw data	0.871 ± 0.045	0.845 ± 0.059	0.853 ± 0.021
Zhu et al. (2018) CycleGAN[a]	0.873 ± 0.046	–	–
Zuo et al. (2021) Calamity[a]	0.884 ± 0.046	–	–
ImUnity[b]	0.918 ± 0.071*	0.920 ± 0.067*	0.882 ±0.060*
ImUnity[a]	**0.951 ± 0.013***	**0.942 ± 0.011***	**0.901 ± 0.035***
ImUnity[c]	0.832 ± 0.067	0.835 ± 0.063	0.883 ± 0.059

[a]Model trained on OASIS database (n=1072).
[b]Model trained on ABIDE database (n=545)
[c]Model trained on SRPBS database (n=81).
*Significant improvement ($p << 10^{-5}$).

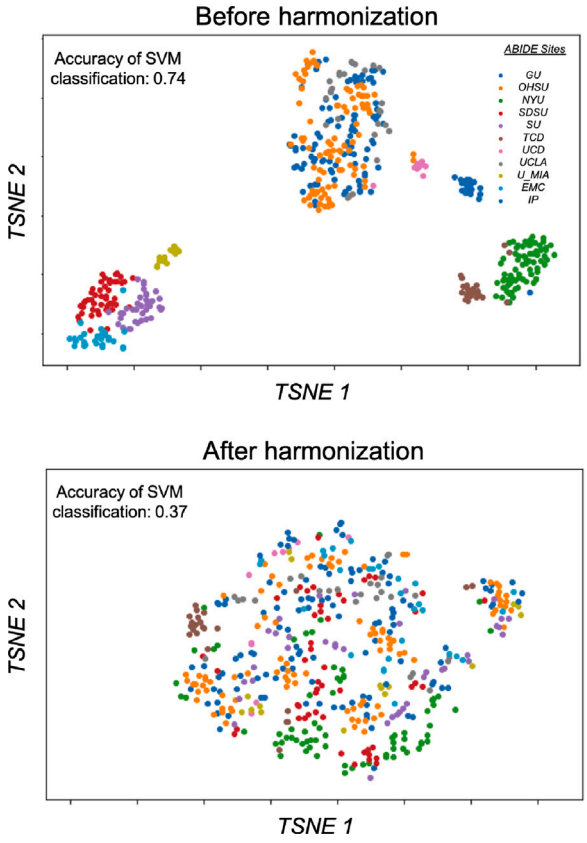

Fig. 4. Harmonization effects on ABIDE sites classification. 2D scans representation of ABIDE database using tSNE reduction algorithm are presented before and after harmonization. Colors correspond to different sites.

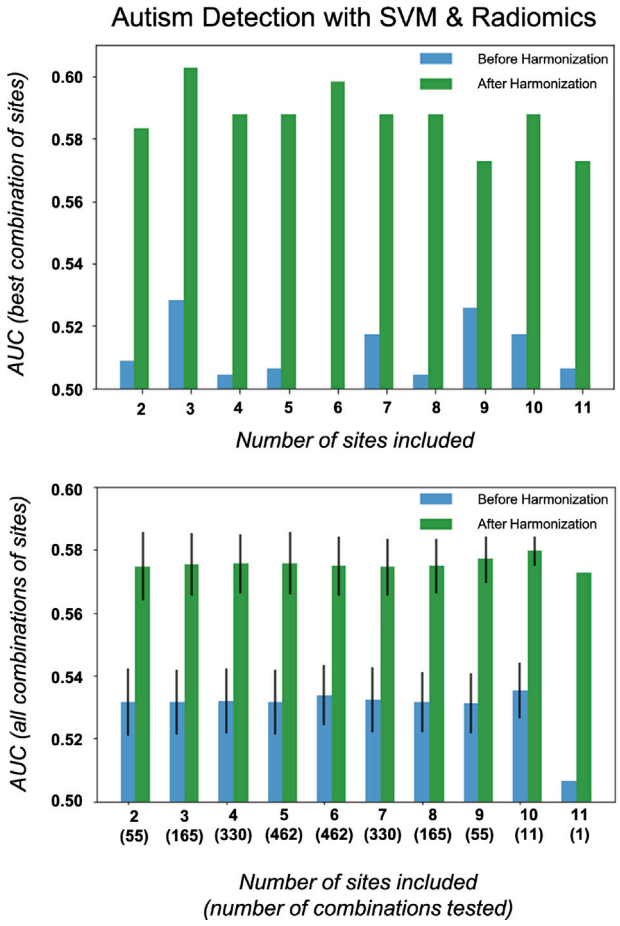

Fig. 5. Harmonization effects on ABIDE patients classification. AUC metrics for classification of patients with autism spectrum disorder (ASD) using SVM and extracted radiomic features (Experience 3, 2.5.3), are shown for different numbers of sites included in the database (from 2 to 11 sites). Top row: Results obtained from the best (largest change in AUC before and after harmonization) combination of sites. Bottom row : Average and standard deviation of AUC estimates for all combination of sites. Same trained model and harmonized data were used for different site combinations.

from one database (here ABIDE) and looking at images harmonized from traveling subjects provided by two different databases (here OASIS and SRPBS), we show that ImUnity does not require new training phase to generalize to unseen sites or scanners (see Fig. 3). The performances were maintained independently of the site selected as reference (see Table 2). While the model was trained on ABIDE data only, it provided better results than the state-of-the-art methods in terms of image quality (+4%, see Table 2).

The last two rows of Table 2 present SSIM metrics obtained when training ImUnity on the other 2 datasets (OASIS and SRPBS). As no biological features were available in these databases, the biological module was disabled and the model was trained in a self-supervised way. First, we noted additional improvements for scanner harmonization when the model was trained and applied on the same database (here OASIS, +2.5%). Second, the score obtained for multi-site harmonization (SRPBS) was the highest when trained on OASIS(N = 1098) data (with a slightly better score than with the other databases). It is interesting to observe the impact on these scores of the dataset

size, the number of site/scanner involved in the training, the use of the biological module and the anatomical differences between datasets (ABIDE mainly contains children data while OASIS and SRPBS focuses on adults). While OASIS results suggest a better generalization on unseen data because more training data were available, ABIDE results suggest that anatomical differences could be compensated by a large training dataset presenting more site/scanner variability than OASIS (11 sites for ABIDE vs. 4 sites for OASIS). On the other hand, results from multi-scanners harmonization depict the difficulty of the model trained on SRPBS data to generalize its training to the OASIS data. This indicates an over-fitting effect in this situation, as there was not enough training data (here N=81 distributed over 9 sites). This suggests that ImUnity may not be adapted to small sample size scenarios, which provides useful information for understanding why the model is less effective in some contexts. Note that data augmentation could have been performed to improve the results in this experiment.

To better estimate the impact of the biological module, we have also run Experiment 3 (2.5.3) with this option disabled. We found (see SM Fig Appendix D) that the biological module has a positive impact on the results with a contribution representing about 20 percent of the total harmonization effect and suggesting that additional input features along the image could lead to better harmonized outputs. Additionally, in Experience 1, we tested different combinations of training and testing data. In the situation: ' training = OASIS' and 'testing = SRPBS', no biological features were available, thus the biological module was not used. Our solution still showed good generalization on the testing dataset. Based on this set of results, we may assume that the addition of some biological information to the model would also lead to better results on traveling subjects. They are still inherent limitation to the biological module including (a) redundancy in signal by the use of two 2D slices from the same subject; (2) information present across slices that is not considered and (3) possible remaining confounders. A throughout study, with rigorous sampling of the initial biological data, could be done to see how much of the information is actually maintained at the end of the harmonization process.

Because ImUnity is designed to reconstruct images and to create a new harmonized database, it does not need new training for new clinical or biological questions. Beyond classification, new clinical data investigation should be conducted with ABIDE (or other multi-center clinical databases) to have better understanding on the impact of our method on clinical research studies.

To estimate the impact of the slice of reference on the harmonization performances, we performed several experiments under different conditions. The corresponding results are in appendix F. For these results, we conclude that the impact of the slice of reference is small. However, it still seems that there is an optimal contrast for the reference related to the task of interest. We recommend the Imunity's users to try several slices of reference for their application. The choice of the reference frames at test time will allow the customization of Imunity depending on the user's objective. As this step does not require retraining the model, it should not be time consuming.

Like the majority of deep networks used for medical image analysis, the MR images used as inputs of our network were first pre-processed for intensity normalization, co-registration or brain extraction. Usually, the impact of these transformations is not examined in harmonization studies. In SM, Fig. A.6 highlights the fact that these steps are already able to remove some of the sites and scanners biases with positive impacts on intensity distributions across sites or patients classification. Intrinsically, the use of White-Stripe normalization (Shinohara et al., 2014) forces the alignment of intensity distributions. Yet, a perfect alignment is not the ultimate goal of harmonization as we also seek to preserve informative biological variations which should persist independently across sites. Eventually, we observe that the best results are obtained for all experiments after the whole ImUnity process, with better intensity distributions alignments, removal of persistent datasets noises, and most importantly improvement in patients classification results. On the opposite, other experiments (not reported) also showed that the VAE-GAN network alone performed poorer when the pre-processing steps were omitted, suggesting that these steps are needed to simplify the training process and improve generalization of the results.

As harmonization is a direct application of domain adaption, we could investigate different architectures like Huang and Belongie (2017), Choi et al. (2018) and include Adaptive Instance Normalization instead of classical batch normalization layers. Then, very promising results were reported with the use of StarGAN model for MR harmonization (Bashyam et al., 2021). Finally the inclusion of a cycle consistency loss as presented in Liu et al. (2021) could enhance our contrast encoder and force a good contrast-style representation in the latent space.

In our study, we only report results with anatomical T1-weighted images. We show that a single type of sequence, combined with computed image transformations with the Gamma function, are sufficient to learn contrast mapping. This greatly facilitates the use of our model because of few data requirements (the origin of each scan is the only pre-required information) and the possibility of self-supervised training. Yet, we believe that this approach is not only dedicated to T1 contrast harmonization and can easily be generalized to any MRI sequences. Presently, the model needs to be fine-tuned in order to harmonize a new medical imaging type. It could however be interesting to investigate its capacity to learn how to harmonize multiple sequences at once. This could be done by mixing sequence types in our training dataset and ensuring the conservation of this information by adding a new conservation module in the bottleneck. It could also be interesting to add other types of artificial contrast transformations (Schettini et al., 2010) for our training in order to account for other types of sites or sequences biases. Similarly we could also include global geometrical distortions to account for biases impacting images geometry. In order to show general harmonization usage of our method, an additional experiment could be conducted on brain lesion segmentation with the hypothesis that the use of Imunity improves the segmentation results on multisite datasets. In that case, the use of traveling patients would help in the definition of ground truth of lesion borders. Finally, although using a 2.5D approach is time efficient, it inherently limits the quality of generated images compared to a fully 3D approach. However, this would induce more parameters ($\sim 32.10E6$) to estimate and therefore would require more computational power and the availability of larger training datasets.

5. Conclusion

We presented ImUnity, an original and effective tool dedicated to MRI harmonization. Our proposed 2.5D model derives from the VAE-GAN architecture. It ensures realistic outputs and allows removal of idiosyncratic datasets bias while intending to preserve biological information. Our results show that the method reaches state-of-the-art performances in term of image quality on traveling patients of the OASIS and SRPBS databases and improves autistic patients classification from the ABIDE database. The proposed 2.5D model is versatile, requiring only one type of MR sequence without the need of matching subjects, can be generalized to sites unseen during the training phase and can be used to harmonize MR images to different reference domains without a new training phase.

Perspectives of enhancement still remain. The introduction of more complex contrast variations during training phase and the use of domain adaption techniques could benefit our proposed solution, especially for multiple MR sequences harmonization.

6. Compliance with ethical standards

This research study was conducted retrospectively using human subject data made available by the following open sources: ABIDE, OASIS, SRPBS. Ethical approval was not required as confirmed by the license attached with the data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgment

Stenzel Cackowski is supported by MIAI@Grenoble Alpes (ANR 19-P3IA-003).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2023.102799.

References

Bashyam, V.M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Singh, A., Habes, M., Fan, Y., Masters, C.L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S.C., Fripp, J., Koutsouleris, N., Satterthwaite, T.D., Wolf, D.H., Gur, R.E., Gur, R.C., Morris, J.C., Albert, M.S., Grabe, H.J., Resnick, S.M., Bryan, N.R., Wittfeld, K., Bülow, R., Wolk, D.A., Shou, H., Nasrallah, I.M., Davatzikos, C., The iSTAGING and PHENOM consortia, 2021. Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors. J. Magn. Res. Imag. n/a (n/a), http://dx.doi.org/10.1002/jmri.27908, URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.27908. Number: n/a _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.27908.

Beer, J.C., Tustison, N.J., Cook, P.A., Davatzikos, C., Sheline, Y.I., Shinohara, R.T., Linn, K.A., 2020. Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. NeuroImage 220, 117129. http://dx.doi.org/10.1016/j.neuroimage.2020.117129, URL: http://www.sciencedirect.com/science/article/pii/S1053811920306157.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., Choo, J., 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. Technical Report, http://dx.doi.org/10.48550/arXiv.1711.09020, arXiv. arXiv:1711.09020. [cs] type: article. URL: http://arxiv.org/abs/1711.09020.

Dewey, B.E., Zhao, C., Reinhold, J.C., Carass, A., Fitzgerald, K.C., Sotirchos, E.S., Saidha, S., Oh, J., Pham, D.L., Calabresi, P.A., van Zijl, P.C.M., Prince, J.L., 2019. DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. Magn. Reson. Imaging http://dx.doi.org/10.1016/j.mri.2019.05.041.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D.A., Gallagher, L., Kennedy, D.P., Keown, C.L., Keysers, C., Lainhart, J.E., Lord, C., Luna, B., Menon, V., Minshew, N.J., Monk, C.S., Mueller, S., Müller, R.-A., Nebel, M.B., Nigg, J.T., O'Hearn, K., Pelphrey, K.A., Peltier, S.J., Rudie, J.D., Sunaert, S., Thioux, M., Tyszka, J.M., Uddin, L.Q., Verhoeven, J.S., Wenderoth, N., Wiggins, J.L., Mostofsky, S.H., Milham, M.P., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry 19 (6), 659–667. http://dx.doi.org/10.1038/mp.2013.78, Number: 6.

Dinsdale, N.K., Jenkinson, M., Namburete, A.I.L., 2020. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. BioRxiv. 2020.10.09.332973. http://dx.doi.org/10.1101/2020.10.09.332973. URL: https://www.biorxiv.org/content/10.1101/2020.10.09.332973v1. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M.A., Ruparel, K., Roalf, D.R., Satterthwaite, T.D., Gur, R.C., Gur, R.E., Schultz, R.T., Verma, R., Shinohara, R.T., 2017. Harmonization of multi-site diffusion tensor imaging data. NeuroImage 161, 149–170. http://dx.doi.org/10.1016/j.neuroimage.2017.08.047, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5736019/.

Fortin, J.-P., Sweeney, E.M., Muschelli, J., Crainiceanu, C.M., Shinohara, R.T., 2016. Removing inter-subject technical variability in magnetic resonance imaging studies. NeuroImage 132, 198–212. http://dx.doi.org/10.1016/j.neuroimage.2016.02.036, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5540379/.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. URL: http://arxiv.org/abs/1505.07818. [Cs, Stat] arXiv:1505.07818.

Guan, H., Liu, Y., Yang, E., Yap, P.-T., Shen, D., Liu, M., 2021. Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. Med. Image Anal. 71, 102076. http://dx.doi.org/10.1016/j.media.2021.102076, URL: https://www.sciencedirect.com/science/article/pii/S1361841521001225.

Huang, X., Belongie, S., 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. Technical Report, http://dx.doi.org/10.48550/arXiv.1703.06868, arXiv. URL: http://arxiv.org/abs/1703.06868. [cs] type: article arXiv:1703.06868.

Iglesias, J.E., Liu, C.-Y., Thompson, P.M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging 30 (9), 1617–1634. http://dx.doi.org/10.1109/TMI.2011.2138152, Number: 9.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156. http://dx.doi.org/10.1016/s1361-8415(01)00036-6, Number: 2.

LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., Raichle, M.E., Cruchaga, C., Marcus, D., 2019. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. Technical Report, http://dx.doi.org/10.1101/2019.12.13.19014902, 2019.12.13.19014902. URL: https://www.medrxiv.org/content/10.1101/2019.12.13.19014902v1. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press ISSN: 1901-4902 Label: Cold Spring Harbor Laboratory Press Type: article.

Liu, Q., Dou, Q., Yu, L., Heng, P.A., 2020. MS-Net: Multi-site network for improving prostate segmentation with heterogeneous MRI data. arXiv:2002.03366. URL: http://arxiv.org/abs/2002.03366.

Liu, M., Maiti, P., Thomopoulos, S., Zhu, A., Chai, Y., Kim, H., Jahanshad, N., 2021. Style transfer using generative adversarial networks for multi-site MRI harmonization. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 313–322. http://dx.doi.org/10.1007/978-3-030-87199-4_30.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (Nov), 2579–2605, Number: Nov. URL: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Orlhac, F., Frouin, F., Nioche, C., Ayache, N., Buvat, I., 2019. Validation of A method to compensate multicenter effects affecting CT radiomics. Radiology 291 (1), 53–59. http://dx.doi.org/10.1148/radiol.2019182023, Number: 1.

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I.M., Satterthwaite, T.D., Fan, Y., Launer, L.J., Masters, C.L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S.C., Fripp, J., Koutsouleris, N., Wolf, D.H., Gur, R., Gur, R., Morris, J., Albert, M.S., Grabe, H.J., Resnick, S.M., Bryan, R.N., Wolk, D.A., Shinohara, R.T., Shou, H., Davatzikos, C., 2019. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage 116450. http://dx.doi.org/10.1016/j.neuroimage.2019.116450.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. URL: http://arxiv.org/abs/1505.04597. [Cs] arXiv:1505.04597.

Sanchez, C.E., Richards, J.E., Almli, C.R., 2012. Age-specific MRI templates for pediatric neuroimaging. Dev. Neuropsychol. 37 (5), 379–399. http://dx.doi.org/10.1080/87565641.2012.688900, Number: 5. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3399736/.

Schettini, R., Gasparini, F., Corchs, S., Marini, F., Capra, A., Castorina, A., 2010. Contrast image correction method. J. Electron. Imaging 19 (2), 023005. http://dx.doi.org/10.1117/1.3386681, Number: 2 Publisher: SPIE. URL: https://www.spiedigitallibrary.org/journals/journal-of-electronic-imaging/volume-19/issue-2/023005/Contrast-image-correction-method/10.1117/1.3386681.full.

Shinohara, R.T., Sweeney, E.M., Goldsmith, J., Shiee, N., Mateen, F.J., Calabresi, P.A., Jarso, S., Pham, D.L., Reich, D.S., Crainiceanu, C.M., 2014. Statistical normalization techniques for magnetic resonance imaging. NeuroImage Clin. 6, 9–19. http://dx.doi.org/10.1016/j.nicl.2014.08.008, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4215426/.

St-Jean, S., Viergever, M.A., Leemans, A., 2020. Harmonization of diffusion MRI data sets with adaptive dictionary learning. Hum. Brain Mapp. http://dx.doi.org/10.1002/hbm.25117, URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.25117.

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., Imamizu, H., 2021. A multi-site, multi-disorder resting-state magnetic resonance image database. Sci. Data 8 (1), 227. http://dx.doi.org/10.1038/s41597-021-01004-8, URL: http://www.nature.com/articles/s41597-021-01004-8. Bandiera_abtest: a Cc_license_type: cc_publicdomain Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Diagnostic markers;Neural circuits;Neurological disorders;Psychiatric disorders Subject_term_id: diagnostic-markers;neural-circuit;neurological-disorders;psychiatric-disorders.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320. http://dx.doi.org/10.1109/TMI.2010.2046908, Number: 6.

van Griethuysen, J.J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G.H., Fillion-Robin, J.-C., Pieper, S., Aerts, H.J.W.L., 2017. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 77 (21), e104–e107. http://dx.doi.org/10.1158/0008-5472.CAN-17-0339, Number: 21.

Wang, Z., Simoncelli, E., Bovik, A., 2003. Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003. Vol. 2. pp. 1398–1402. http://dx.doi.org/10.1109/ACSSC.2003.1292216.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2018. Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv:1703.10593. URL: http://arxiv.org/abs/1703.10593.

Zuo, L., Dewey, B.E., Carass, A., Liu, Y., He, Y., Calabresi, P.A., Prince, J.L., 2021. Information-based disentangled representation learning for unsupervised MR harmonization. In: Feragen, A., Sommer, S., Schnabel, J., Nielsen, M. (Eds.), Information Processing in Medical Imaging. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 346–359. http://dx.doi.org/10.1007/978-3-030-78191-0_27.