SFHarmony: Source Free Domain Adaptation for Distributed Neuroimaging Analysis

Nicola K Dinsdale[1] Mark Jenkinson[2,3,4] Ana IL Namburete[1,2]

1. Oxford Machine Learning in NeuroImaging (OMNI) Lab, Department of Computer Science, University of Oxford, UK
2. Wellcome Centre for Integrative Neuroimaging, FMRIB, University of Oxford, Oxford, UK
3. Australian Institute for Machine Learning (AIML), Department of Computer Science, University of Adelaide, Adelaide, Australia
4. South Australian Health and Medical Research Institute (SAHMRI), North Terrace, Adelaide, Australia

nicola.dinsdale@cs.ox.ac.uk

Abstract

To represent the biological variability of clinical neuroimaging populations, it is vital to be able to combine data across scanners and studies. However, different MRI scanners produce images with different characteristics, resulting in a domain shift known as the 'harmonisation problem'. Additionally, neuroimaging data is inherently personal in nature, leading to data privacy concerns when sharing the data. To overcome these barriers, we propose an Unsupervised Source-Free Domain Adaptation (SFDA) method, `SFHarmony`*. Through modelling the imaging features as a Gaussian Mixture Model and minimising an adapted Bhattacharyya distance between the source and target features, we can create a model that performs well for the target data whilst having a shared feature representation across the data domains, without needing access to the source data for adaptation or target labels. We demonstrate the performance of our method on simulated and real domain shifts, showing that the approach is applicable to classification, segmentation and regression tasks, requiring no changes to the algorithm. Our method outperforms existing SFDA approaches across a range of realistic data scenarios, demonstrating the potential utility of our approach for MRI harmonisation and general SFDA problems. Our code is available at* `https://github.com/nkdinsdale/SFHarmony`.

1. Introduction

Deep learning (DL) models have proved to be powerful tools for neuroimage analysis. However, the majority of neuroimaging datasets remain small, posing a challenge for the training of sophisticated architectures with many parameters. Thus, it is common practice to combine data from multiple sites and MRI scanners, both to increase the amount of data available for training, and to represent the breadth of biological variability that can be expected in diverse populations. However, the combination of data across MRI scanners with different acquisition protocols and hardware leads to an increase in non-biological variance [25, 26, 50], which can be large enough to mask the biological signals of interest [49], even after careful pre-processing with state-of-the-art neuroimaging pipelines [23]. The development of *harmonisation* methods is therefore vital to enable the joint unbiased analysis of neuroimaging data from different scanners and studies.

The key goal for harmonisation methods is to be discriminative for the main task of interest whilst creating shared feature representations of the data across acquisition scanners, clearly mirroring the goal of domain adaptation (DA) [15]. The majority of deep learning based harmonisation methods are based on DA methods, either using adversarial approaches to create shared feature embeddings [15, 24], or using generative approaches to create harmonised images [11, 63].

However, the vast majority of existing methods fail to be applicable in many realistic data scenarios. For example, MR images are inherently personal information so their sharing is protected by legislation, such as GDPR [9] and HIPAA [39]. Thus, the assumption of centralised data stores for model training is infeasible, particularly when working with clinical imaging data, which will be essential in order to produce representative models [14, 52]. Distributed learning offers a promising solution, but the few proposed distributed harmonisation methods [8, 16] assume the simultaneous presence of the source and target data. The source data may not be available for the adaptation phase, for instance, due to confidentiality agreements, loss of the source

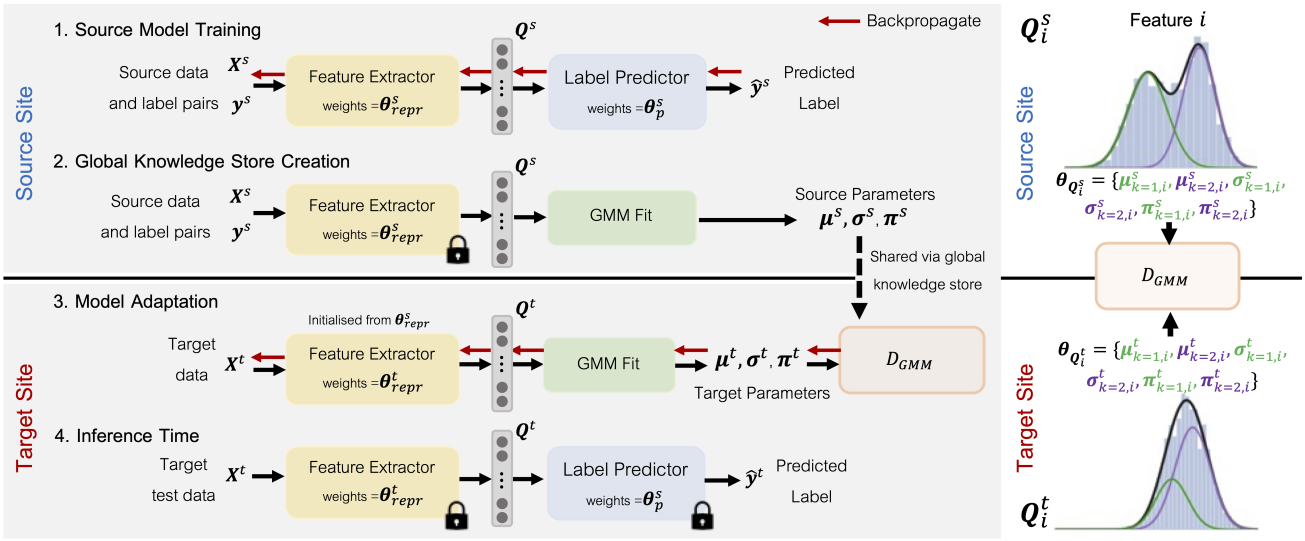

Figure 1. Schematic of the proposed SFHarmony method. The method fits a GMM to the source features, shares these via a global model store, and then completes SFDA by aligning the source and target feature distributions utilising a modified Bhattacharyya distance. $\boldsymbol{Q}^s$ is the source feature representation and $\boldsymbol{Q}^t$ is the target feature representation, and the figure shows the GMM setup, for a single feature i, when we are working with K, the number of components, being 2. This trivially generalises to more or less components.

data, or computational constraints [3]. Further, federated DA methods such as [16] would require retraining of the model to incorporate any new sites, which is infeasible and computationally expensive.

Therefore, we explore an unsupervised DA setting where only the source model, instead of the source data, is provided to the unlabelled target domain for harmonisation, known as Source Free Domain Adaptation (SFDA). This setting inherently protects individual privacy, whilst allowing the efficient incorporation of new sites without requiring target labels. We propose a simple yet effective solution, termed SFHarmony, which aims to match feature embeddings from the source and target, through characterising the embeddings as a Gaussian Mixture model (GMM) and the use of a modified Bhattacharyya distance [4]. This requires no modifications to the training of the source model, and the only additional communication is of summary statistics of the source feature embedding, allowing it to be simply applied to existing architectures. The summary statistics contain no information about individuals.

Our contributions are as follows: 1) We propose a new method for SFDA, SFHarmony, based on aligning feature embeddings, utilising a modified Bhattacharyya distance, requiring no changes to source training; 2) We demonstrate the method's applicability to classification, segmentation and regression tasks, and show that the approach outperforms existing SFDA methods for domain shifts experienced when working with neuroimaging data; 3) We demonstrate the robustness of the method to additional challenges likely to be faced when working with real world imaging data: differential privacy and label imbalance.

2. Related Work

Unsupervised Domain Adaptation (UDA): UDA aims to exploit the knowledge learned from a source dataset to help to create a discriminative model for a related but unlabelled target dataset [33]. DL-based UDA approaches can broadly be split into three categories [33]: discrepancy-based, reconstruction-based, and adversarial. Discrepancy based approaches aim to minimise a divergence criterion, which measures the distance between the source and target data distributions encoded in a learned feature space [10, 27, 35, 48]. Reconstruction-based approaches, instead, use reconstruction as a proxy task to enable the learning of a shared representation for both image domains [5, 22, 38]. Finally, adversarial approaches deploy a discriminator that aims to identify the source of the data; the model is trained both to do the task and to trick the discriminator, creating domain invariant features [21, 51]. These methods all assume simultaneous access to the source and target data, which poses data privacy challenges.

Federated Learning and Domain Adaptation: Federated learning (FL) has been proposed as a method to train models on distributed data [36]. The data are kept on their local servers, and users train local models with private data and communicate the weights or gradients between sites for aggregation. Many FL approaches focus on minimising the impact of distribution shifts between clients [28, 45, 55]; however, most approaches assume that the data at all sites are fully labelled. However, federated DA enables the in-

corporation of an unlabelled site into the federation without sharing data. FADA [40] is a federated DA method, where features are shared between sites in a global knowledge store. The sharing of features, however, still poses privacy concerns as images may be recoverable from the features [19]. Thus, instead FedHarmony [16] encodes the features as Gaussian distributions and thus only the mean and standard deviations of the features need to be shared. Both of these methods still assume access to the source data during training and rely on adversarial approaches that are often unstable and hard to train. Other federated DA methods produce domain-specific models or ensembles [19, 41, 59, 61], meaning that the final predictions depend on the domain of the data.

Source Free Domain Adaptation: SFDA takes the federated approach a step further and assumes that there is no access to the source data available at all: only the source model is available for model adaptation. The majority of SFDA methods have been developed for classification [1, 13, 17, 32, 33, 43, 54, 60], with a few being proposed for segmentation [30, 31, 34, 44, 57]. There are two main approaches taken for SFDA. The first set of approaches are generative, aiming to create source samples using the source model weights [32, 34]. These approaches, however, pose concerns about individual privacy, especially when working with medical images and low numbers of samples [19] and cannot be simply applied to complex target tasks, limiting their utility when working with MRI data [53]. The second set aim to minimise model entropy to improve predictions, guided by various pseudo labelling or uncertainty techniques to prevent mode collapse [1,13,17,29,33,43,60]. These methods are often effective, but largely limited to classification tasks, and may require changes to the source model training to be effective [33, 60]. AdaMI [3] was proposed directly for medical image segmentation, but requires an estimate of the proportion of each label to prevent mode collapse. This ratio is hard to estimate for labels with high variability across populations, such as tumours or lesions. We could not identify any methods proposed for regression, where the lack of softmax outputs limit the direct application of methods based on entropy minimisation.

Harmonisation: Many existing harmonisation approaches are based on COMBAT [7, 20, 42], which uses a linear model to represent the scanner effects on image-derived features. DL-based approaches for harmonisation generally utilise a DA approach, with many being generative, aiming to produce 'harmonised' images [6, 11, 37, 62, 63], while the other branch uses adversarial approaches to harmonise the learned model features for a given task [15, 24]. All of these methods assume simultaneous access to the source and target data, with some even requiring paired data [11]. The only existing methods for harmonisation which consider data privacy are Distributed COMBAT [8] and FedHarmony [16]; however, both assume constant communication with the source site.

3. Method

The aim of this work is to create a SFDA method applicable to neuroimaging tasks, and to demonstrate its suitability for MRI harmonisation. Thus, the goal is to create a model where two images with the same label would share a feature embedding, regardless of the acquisition scanner – the domain of the data. We thus follow the framework of [51] and consider the network to be formed of a feature extractor, with parameters $\mathbf{\Theta}_{repr}$, and a label predictor, $\mathbf{\Theta}_p$. This network architecture is the same across source and target sites. The general schematic for training is shown in Fig. 1.

3.1. Creation of the Source Model

The first stage is the training of the source model. This assumes the availability of a labelled training dataset $D^s = \{\mathbf{X}^s, \mathbf{y}^s\}$, where the image and label pairs depend on the task of interest. Unlike some existing methods [33, 60], our proposed approach requires no changes to the training of the source model or to the architecture. The model can thus be flexibly trained following the standard training procedure for the source data, with the goal being to create a well-trained source model. In our experiments, we consider the simplest source training, minimising a loss function (L_{task}) dependent on the task of interest with full supervision:

$$L(\mathbf{X}^s, \mathbf{y}^s; \mathbf{\Theta}^s_{repr}, \mathbf{\Theta}^s_p) = \frac{1}{N_s} \sum_{i}^{N_s} L_{task}(\mathbf{X}^s_i, \mathbf{y}^s_i) \quad (1)$$

where N_s is the total amount of labelled source data.

3.2. Global Information Store

For successful SFDA, we need to align the learned feature embedding, $\mathbf{Q}^s = f(\mathbf{X}^s, \mathbf{\Theta}^s_{repr})$ for the source and target data. To achieve this without requiring the source data, we propose to follow the precedent of existing privacy-preserving medical imaging approaches [8,16,17] and, thus, create a global knowledge store to share summary statistics of the features. In [16], it is proposed that the features can be encoded as Gaussian distributions, and thus the statistics to be shared would be a mean and standard deviation per feature. We hypothesise that, for many tasks, especially classification tasks with discrete categories, simple Gaussian distributions are unlikely to sufficiently characterise $\mathbf{Q}^s$. We thus propose to describe the features using a Gaussian mixture model (GMM), with each feature being encoded as an independent 1D GMM, such that, for feature $i \in N_{Q^s}$, where N_{Q^s} is the number of features in $\mathbf{Q}^s$:

$$\mathbf{Q}^s_i \sim \sum_{k=1}^{K} \boldsymbol{\pi}^s_{k,i} \mathcal{N}(\boldsymbol{X}^s; \boldsymbol{\mu}^s_{k,i}, \boldsymbol{\sigma}^{s^2}_{k,i}) \quad (2)$$

where K is the number of components in the GMM, $\boldsymbol{\mu}^s_{k,i}$ and $\boldsymbol{\sigma}^{s^2}_{k,i}$ are the mean and variance defining the k^{th} Gaussian component of the i^{th} feature for the source site, and $\boldsymbol{\pi}^s_{k,i}$ is the weighting factor for this k^{th} Gaussian (which sum to one across components). Note that the features are considered before the activation function. The same number of components, K, are fit for all features. Thus, the GMM for feature i is defined by the parameters:

$$\boldsymbol{\Theta}^s_i = \{\boldsymbol{\pi}^s_{k,i}, \boldsymbol{\mu}^s_{k,i}, \boldsymbol{\sigma}^{s^2}_{k,i}\}, k = 1..K \tag{3}$$

and these parameters can be determined using Expectation Maximisation (EM), by finding the maximum likelihood estimate (MLE) of the unknown parameters:

$$\mathcal{L}(\Theta_i) = \sum_{n=1}^{N_{s,i}} \log(\sum_{k=1}^{K} \boldsymbol{\pi}^s_{k,i} \mathcal{N}(\boldsymbol{X}^s_n; \boldsymbol{\mu}^s_{k,i}, \boldsymbol{\sigma}^{s^2}_{k,i})) \tag{4}$$

for each feature i in $\boldsymbol{Q^s}$. This, therefore, produces three parameter arrays that fully define the GMMs of the source features, which are communicated alongside the source weights to target sites:

$$\boldsymbol{\Theta}_{Q^s} = \{\boldsymbol{\mu}^s \in \mathbb{R}^{K \times N_{Q^s}}; \boldsymbol{\sigma}^{s^2} \in \mathbb{R}^{K \times N_{Q^s}}; \boldsymbol{\pi}^s \in \mathbb{R}^{K \times N_{Q^s}}\}. \tag{5}$$

These parameters contain no individually identifying information, as they represent aggregate statistics across the whole population.

3.3. Target Model Adaptation

Given that we now have a well trained source model, with parameters $\boldsymbol{\Theta}^s_{repr}$ and $\boldsymbol{\Theta}^s_p$, and the source GMM parameters, $\boldsymbol{\Theta}_{Q^s}$, we can now adapt the model at any target site. We assume access to an unsupervised target, with only data samples $\boldsymbol{X}^t$ and no labels available.

We initialise the target model using the source trained weights. Model adaptation only involves finetuning the feature extractor to match the learned feature distribution across the two sites. In adversarial approaches, a discriminator is added to the overall architecture that aims to distinguish between source and target samples. We could utilise this approach, following [16], by drawing feature samples, using the source GMM parameters $\boldsymbol{\Theta}_{Q^s}$, but adversarial approaches are notoriously unstable and difficult to train. We therefore, instead, propose to minimise the difference between the source feature distribution and target feature distribution using the GMM parameters directly.

Therefore, the first step of model adaptation is to calculate the current target features, $\mathbf{Q}^t = f(\mathbf{X}^t, \boldsymbol{\Theta}^t_{repr})$, and then, using the same EM approach as above, we can create the parameters of the target GMM fit:

$$\boldsymbol{\Theta}_{Q^t} = \{\boldsymbol{\mu}^t \in \mathbb{R}^{K \times N_{Q^t}}; \boldsymbol{\sigma}^{t^2} \in \mathbb{R}^{K \times N_{Q^t}}; \boldsymbol{\pi}^t \in \mathbb{R}^{K \times N_{Q^t}}\}. \tag{6}$$

We propose to use a modified Bhattacharyya distance [4] as the loss function. The Bhattacharyya distance measures the similarity of two probability distributions, which for continuous probability distributions is defined as:

$$D_B(p, q) = -\ln(BC(p, q)) \tag{7}$$

where

$$BC(p, q) = \int_x \sqrt{p(x)q(x)}dx. \tag{8}$$

The Bhattacharyya distance has a simple closed form solution when the two probability distributions are both Gaussian. If $p \sim \mathcal{N}(\mu_p, \sigma^2_p)$ and $q \sim \mathcal{N}(\mu_q, \sigma^2_q)$ then:

$$D_B(p, q) = \frac{1}{4}\frac{(\mu_p - \mu_q)^2}{\sigma^2_p + \sigma^2_q} + \frac{1}{2}\ln\Big(\frac{\sigma^2_p + \sigma^2_q}{2\sigma_p\sigma_q}\Big). \tag{9}$$

There is, however, no equivalent closed form solution for a GMM. In [46] they propose an approximation for the GMM as a sum of the Bhattacharyya distances for each pair of Gaussians in the mixture model, weighted by the associated $\boldsymbol{\pi}$ values. We suggest that this is not the most appropriate reformulation: we are more interested in the corresponding pairs of Gaussians than in the cross-relationships, as we do not wish to minimise the difference between cross pairs. Rather, we wish specifically to make the target distribution match the source. Thus, if we consider our target and source GMM distributions, parameterised by $\boldsymbol{\Theta}_{Q^s}$ and $\boldsymbol{\Theta}_{Q^t}$, we propose to use the following approximation:

$$D_{GMM}(\boldsymbol{\Theta}_{Q^s}, \boldsymbol{\Theta}_{Q^t}) = \sum_{k=1}^{M} \boldsymbol{\pi}^s_k \boldsymbol{\pi}^t_k \Big(\frac{1}{4}\frac{(\boldsymbol{\mu}^s_k - \boldsymbol{\mu}^t_k)^2}{\boldsymbol{\sigma}^{s^2}_k + \boldsymbol{\sigma}^{t^2}_k} + \frac{1}{2}\ln\Big(\frac{\boldsymbol{\sigma}^{s^2}_k + \boldsymbol{\sigma}^{t^2}_k}{2\boldsymbol{\sigma}^s_k\boldsymbol{\sigma}^t_k}\Big)\Big) \tag{10}$$

such that we find the weighted sum of the Bhattacharyya distances between each corresponding pair of Gaussians, where k is the component in the GMM. It can further be seen that this approximation retains the desirable property that when $\boldsymbol{\Theta}_{Q^s} = \boldsymbol{\Theta}_{Q^t}$, then $D_{GMM}(\boldsymbol{\Theta}_{Q^s}, \boldsymbol{\Theta}_{Q^t}) = 0$. The correspondence between Gaussians can be ensured by simply ordering the parameters by the mean estimates.

Thus, the feature extractor is finetuned for the target site by minimising D_{GMM} averaged across all of the features in $\boldsymbol{Q}^s$. However, for each training iteration, only a fixed size batch is available to estimate the parameters, and for neuroimaging applications the maximum batchsize achievable is often small due to the relatively large image size [14], which affects the estimate of the GMM parameters. As EM is sensitive to initialisation, to mitigate the small batch effect, we initialise the EM algorithm only once per training epoch, using the previous batch estimate as the initialisation for the next, providing memory between batches. The EM algorithm is reinitialised for validation (needed to calculate validation loss), preventing data leakage.

3.4. Inference Time

Finally, inference for the test data simply involves combining the finetuned feature encoder, $\boldsymbol{\Theta}^t_{repr}$, and the frozen source label predictor, $\boldsymbol{\Theta}^s_p$, such that $\hat{\boldsymbol{y}}^t = f(\boldsymbol{X}^t, \boldsymbol{\Theta}^t_{repr}, \boldsymbol{\Theta}^s_p)$. This therefore ensures that, given data from the source or target domain with the same feature embedding, the same label prediction is achieved across sites.

4. Experimental Results

To validate the effectiveness of our SFDA framework, we conduct a range of experiments with both simulated data with known domain shifts, and real multisite MRI datasets, and we demonstrate the applicability of the method to classification, segmentation and regression tasks.

4.1. Datasets:

Further details for each dataset and model architectures are available in the Supplementary Materials. Example images from each dataset can be seen in Fig. 2.

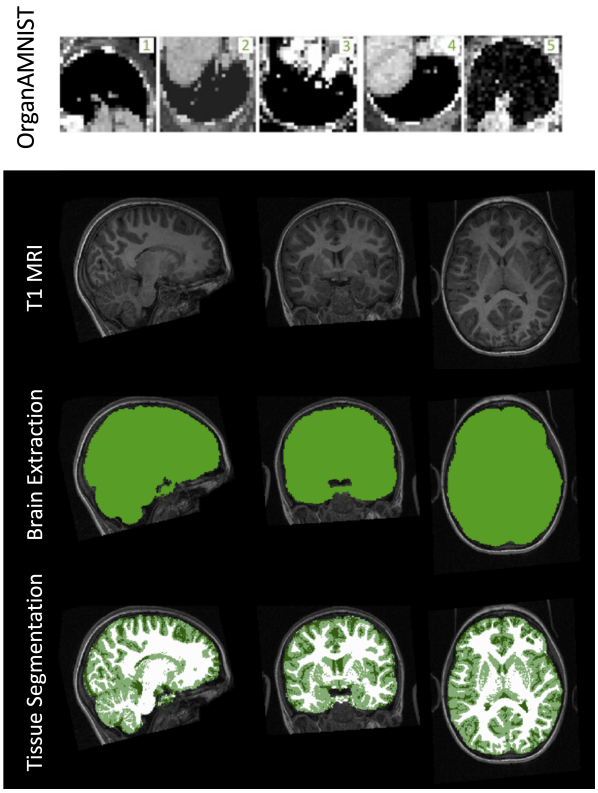

Figure 2. Example images from each dataset. OrganAMNIST: example image from the same class for each of the applied domain shifts. The T1 MRI images were used for the brain extraction, tissue segmentation and age prediction tasks.

OrganAMNIST [56] (Classification): curated as part of MedMNIST [58], we use OrganAMNIST as a test dataset. All images were pre-processed to 28×28 (2D) with the corresponding classification labels for 11 classes. We created simulated known domain shifts, to enable exploration of the method, with the strength of each shift designed to be such that a degradation in performance was seen across the sites. The dataset was split into 5 sets, each with 5000 samples for training and 2000 for testing and the following domain shifts applied: 1) no shift (source site), 2) decreased intensity range, 3) increased intensity range, 4) Gaussian blurring, 5) salt and pepper noise, to model shifts likely across imaging sites. The backbone architecture took the form of a small VGG-like classifier, with categorical crossentropy as the task loss. Code to reproduce the data is provided.

CC359 [47] (Segmentation): The dataset consists of brain images of healthy adults (29-80 years) acquired on MRI scanners from three vendors: Siemens, Philips and GE, at both 1.5 and 3T, with approximately 60 subjects per vendor and magnetic field strength. A 2D UNet was trained on slices from each site, then the performance when applied to the remaining sites was compared. As a result, the Phillips 1.5T was chosen as the source site as it had the largest performance drop. No additional preprocessing was applied to the images apart from image resizing so that each subject volume was $128 \times 240 \times 160$. The data were split at the subject level per site, such that 40 subjects were available for training and 20 for testing. The segmentation task was skull stripping, using masks from the original study, and Dice loss was used as the task loss function. Further details and example segmentation masks can be found in the Supplementary Material.

ABIDE [12] (Segmentation and Regression): Four sites (Trinity, NYU, UCLA, Yale) were used, so as to span age distributions and subject numbers. The data were split into training/test sets as 80%/20%, yielding a maximum of 127 subjects for training (NYU) and a minimum of 35 (Trinity). NYU was the largest site, spanning the age distribution of all of the other sites, and so was chosen as the source site. For segmentation, we considered tissue segmentation (grey matter (GM), white matter (WM), CSF), using labels automatically generated using FSL ANAT. We used a 2D UNet trained on slices with Dice as the main task loss function. Dice score was averaged across the three tissues. For age prediction, a separate network was trained, following the setup and architecture in [16], with MSE as the main task loss. Further details and example labels can be found in the Supplementary Material.

Implementation Details: All comparison methods used the same task-specific backbone architecture as the proposed method. Features were extracted in the second-to-last layer, before the activation function, following the result in [15]. Model architectures were chosen to give good source performance while allowing the use of large batch-sizes, but most standard architectures could be used. Train-

ing was completed on an A10 GPU, using PyTorch 1.12.0. All models were trained with five-fold cross validation and results are presented on the holdout test set. A learning rate of 1×10^{-6} was used for all datasets for adaptation with an AdamW optimiser.

4.2. Classification: OrganAMNIST

Baselines: For the classification task, we first compare our approach to supervised oracles: source model only, centralised data, and target finetuning with frozen label predictor. We then compare to DeepCORAL [48], and two federated DA approaches: FADA [40] and FedHarmony [16], both of which require the presence of the source data. Finally, we compare to SFDA methods: entropy minimisation; SHOT [33], USFAN [43], and gSFDA [60]. We do not compare to any generative SFDA methods, as the ability to create source data would not meet privacy requirements for many applications [53], especially given that GANs often replicate training images when trained with small datasets [19]. Details are provided in the Supplementary Material.

Methods Comparison: We first demonstrate the method for a range of batchsizes (5, 50 and 500) because methods that minimise entropy are expected to be more stable when using large batchsizes, which are rarely achievable when working with MR images due to the memory constraints posed by large image sizes [14]. Thus, robustness to the batchsize is vital if a SFDA method is to be used for harmonisation. We use a single source model to allow fair comparison, trained with a batchsize of 50. We wish to maximise performance across all sites: as harmonisation is normally framed as a joint domain adaptation problem [15], the average performance across all sites is reported.

The results can be seen in Table 1, alongside the baseline methods. It can be seen that `SFHarmony` outperforms the existing SFDA methods, especially when a small batchsize was used for training (86.22% for batchsize 5). SHOT [33] showed comparable performance to `SFHarmony` when trained with a batchsize of 500 (85.27%), but was highly dependent on the modified source training. Interestingly, several of the SFDA approaches outperformed the adversarial approaches despite them having access to the source data, possibly due to the instability of such approaches.

The proposed D_{GMM} loss is clearly able to align the features across sites using only the GMM summary statistics. This is demonstrated by Fig. 3, which shows the source and target features for each site before and after DA. Clearly the features overlap much more after DA, which both leads to the clear improvement in performance, and shows that the approach is achieving the harmonisation goals of the model having a shared feature embedding across sites. The change shift in the features is more visible for the intensity based shifts (1-2 and 1-3) suggesting that the two largest PCA components largely encode intensity. Thus, as the noise

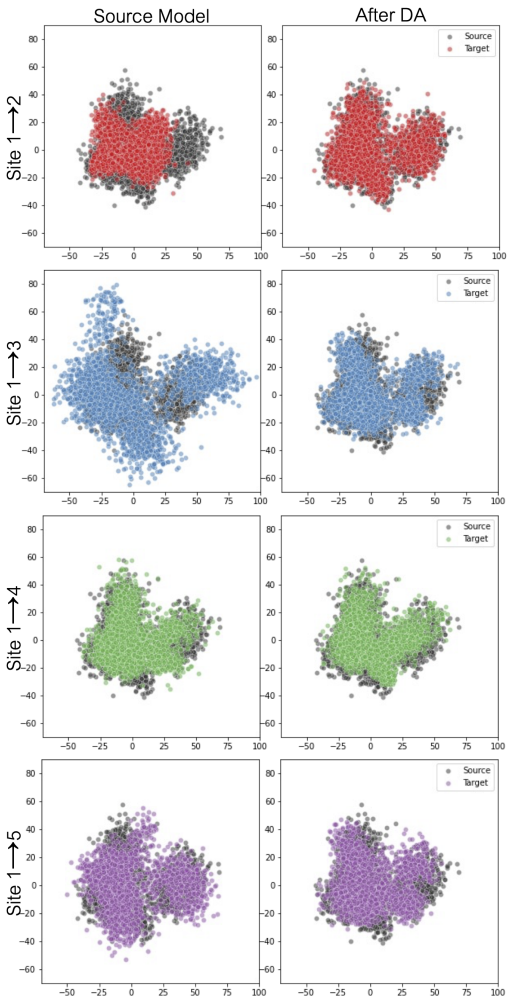

Figure 3. PCA of $\boldsymbol{Q}_s$ and $\boldsymbol{Q}_t$ for each target site, before and after domain adaptation for the OrganAMNIST data, with simulated domain shifts. Black dots are the source features which are fixed and the colour represents the features for the relevant site. (Best viewed in colour.)

augmentations do not create values outside of the existing intensity range, the noise shifts do not lead to large changes to the features in the PCA space.

We tried modelling the features with $K \in \{1, 2, 3\}$ GMM components: visual inspection of the features suggested that at least 2 components would be beneficial. This was confirmed by the results, with the best performance being achieved when modelling the features with 2 components, as shown in Table 1. However, the approach still performed well for 1 and 3 components, showing limited sensitivity to the number of components chosen. The number of components chosen is the only additional hyperparameter to be tuned with our approach, with only a single loss function to minimise. The results clearly show the robustness to the choice of batchsize, and the results were also robust to the

Method	S	T	C	Information Communicated	Average Accuracy		
					Batchsize 5	Batchsize 50	Batchsize 500
Source Model	✓	x	x	-		80.71	
Centralised Data	✓	✓	✓	All Data	88.34	91.65	91.27
Target Finetune	x	✓	✓	Model Weights	88.15	88.96	83.26
DeepCORAL [48]	✓	x	✓	All Data	82.43	83.85	83.65
FADA [40]	✓	x	x	Model Weights + Features	81.69	76.77	76.53
FedHarmony [16]	✓	x	x	Model Weights + Statistics	81.48	76.12	76.20
Minimise Entropy	x	x	x	Model Weights	42.59	83.54	83.96
SHOT [33] (no smoothing)	x	x	x	Model Weights	66.86	83.60	85.40
SHOT [33] (Source batchsize 5)	x	x	x	Model Weights	72.06	74.44	74.61
SHOT [33] (Source batchsize 500)	x	x	x	Model Weights	83.10	84.68	85.27
gSFDA [60]	x	x	x	Model Weights	60.57	85.87	84.67
USFAN [43]	x	x	x	Model Weights	26.94	79.83	83.79
SFHarmony 1 GMM Component	x	x	x	Model Weights + Statistics	85.47	85.71	86.16
w/o EM (Direct Fit)	x	x	x	Model Weights + Statistics	77.26	76.99	86.03
w/o Batch Memory	x	x	x	Model Weights + Statistics	80.25	82.13	84.60
SFHarmony 2 GMM Components	x	x	x	Model Weights + Statistics	**86.22**	**86.25**	**86.21**
SFHarmony 3 GMM Components	x	x	x	Model Weights + Statistics	86.21	85.70	85.96

Table 1. Results on the OrganAMNIST classification task. S = Source data required, T = Target labels required, C = Centralised data. The average accuracy is across all 5 sites, weighted equally, and is reported for training batchsizes of 5, 50 and 500. Best SFDA method for each batchsize is in bold, other methods are included for reference. The w/o (without) components form an ablation study.

choice of learning rate, with the accuracy staying within 1% of the best result across learning rates from 10^{-7} to 10^{-4}. Therefore, deployment of the proposed approach requires no changes for a new site, only the choice of the number of components for a new source model. This is in contrast to many existing SFDA approaches that require the balancing of several loss functions (e.g. [3, 33]).

Ablation Study: We considered the GMM with $K = 1$, allowing us to explore the w/o EM case, where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are calculated directly. We also considered removing the batch memory across the training loop, reinitialising the EM algorithm before each batch. From Table 1 it is clear that both aspects are contributing to the performance, especially for small batchsizes.

Class Imbalance: In the above experiments, the distribution of class labels was approximately equal across sites. We now consider the extreme scenario where the source site contains samples from across all classes but target sites are missing classes. This is a conceivable scenario when considering MR images, where a given clinical site specialises in a certain condition and we are trying to harmonise the data to a carefully curated research dataset. Figure 4 shows the average accuracy across sites, when the target sites had samples from an increasing number of classes removed. Each comparison method was trained using the best setting from Table 1. The proposed `SFHarmony` approach was more robust to the increased class imbalance than existing SFDA methods.

Differential Privacy (DP): Finally, we considered simulating the approach when DP is being used to further protect privacy. We simply simulated a Laplace mechanism of

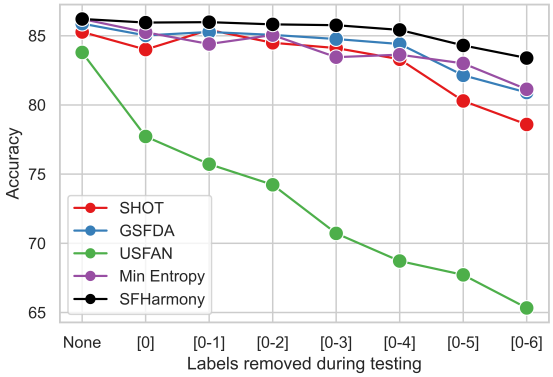

Figure 4. Average accuracy across the sites with increasing numbers of classes removed from the target site training, creating increasingly imbalanced data distributions. The x axis shows the classes that were removed.

DP [18], by injecting noise onto the weights before communication, modelled as: $\boldsymbol{w} = \boldsymbol{w} + Lap(|\boldsymbol{w}|f)$ where f was varied to create increasing levels of noise. As the GMM is fit at the local site, $\boldsymbol{\Theta}_{Q^s}$ can be calculated before the noise is applied. The comparison methods were again all trained using the best setting from Table 1. Although this is a very simple model of DP, with many more sophisticated approaches existing, Fig. 5 demonstrates that many existing methods for SFDA are very sensitive to the applied noise. SHOT [33] is the most dramatically affected, with the pseudo-labelling approach suffering a significant degradation in performance. Our proposed approach maintained

Method	S	T	C	Information Communicated	CC359 Average Dice			ABIDE Average Dice		
					Bs 5	Bs 50	Bs 500	Bs 5	Bs 50	Bs 500
Source Model	✓	x	x	-		0.832			0.775	
Centralised Training	✓	✓	✓	All Data	0.983	0.985	0.983	0.884	0.885	0.875
Target Finetune	x	✓	x	Model Weights	0.981	0.982	0.982	0.883	0.884	0.885
DeepCORAL [48]	✓	x	✓	All Data	0.768	-	-	0.523	-	-
FADA [40]	✓	x	x	Model Weights + Features	0.967	0.964	0.959	0.830	0.827	0.825
FedHarmony [16]	✓	x	x	Model Weights + Statistics	0.965	0.962	0.950	0.825	0.810	0.822
Minimise Entropy	x	x	x	Model Weights	0.767	0.849	0.951	0.570	0.542	0.659
AdaEnt [2]	x	x	x	Model Weights	0.827	0.817	0.962	0.625	0.656	0.682
AdaMI [3]	x	x	x	Model Weights	0.820	0.835	0.965	0.606	0.657	0.660
Direct Fit	x	x	x	Model Weights + Statistics	0.648	0.696	0.873	0.615	0.803	0.830
SFHarmony 1 GMM Component	x	x	x	Model Weights + Statistics	0.950	0.949	0.959	0.831	**0.832**	0.831
SFHarmony 2 GMM Components	x	x	x	Model Weights + Statistics	0.970	**0.970**	**0.970**	0.832	**0.832**	**0.832**
SFHarmony 3 GMM Components	x	x	x	Model Weights + Statistics	**0.972**	0.968	**0.970**	**0.833**	**0.832**	**0.832**

Table 2. Results on the CC359 dataset for brain extraction, and the ABIDE dataset for the tissue segmentation. S = Source data required, T = Target labels required, C = Centralised data, Bs = batchsize. The average Dice score is the performance across all 5 (CC359) /4 (ABIDE) sites, weighted equally, and is reported for training batchsizes of 5, 50 and 500. The best performing SFDA method for each batchsize for each segmentation task is in bold.

Features	65536 (Full)	10000	1000	100	10
Average Dice	0.972	0.970	0.968	0.968	0.890

Table 3. Results on the CC359 dataset using only a subset of features to complete the domain adaptation for 3 GMM components and a batchsize of 5.

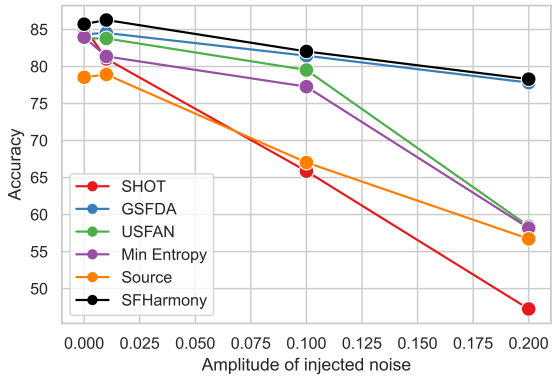

Figure 5. Average accuracy across the sites with increasing magnitudes of noise injected into the source weights before communication. Amplitude is as a proportion of the source weights magnitude.

performance well across the applied noise levels, despite the frozen label predictor imposing a ceiling on performance.

4.3. Segmentation: CC359 and ABIDE datasets

We now demonstrate our approach on two multisite MRI datasets for segmentation tasks: brain extraction (CC359) with two labels (brain/background) and tissue segmentation (ABIDE) with four labels (WM, GM, CSF, background).

Baselines: There are far fewer existing methods for SFDA, and we again did not compare to generative approaches. Thus, we compared to supervised oracles: source model only, centralised data and target finetuning with frozen label predictor; semisupervised approaches: DeepCORAL [48], FADA [40] and FedHarmony [16]. Then, for SFDA approaches, we compared to minimising entropy, AdaEnt [2] and AdaMI [3], and Direct Fit (w/o EM in ablation study). We were unable to train DeepCORAL with a batchsize of more than 5 due to memory constraints.

Methods Comparison: Table 2 shows the results for both tasks. In the classification task there were only 32 features in the fully connected layer; however, now there are many more, for instance for the CC359 data there are 65536 features across all of the convolutional filters. Despite this increase in features, `SFHarmony` was able to complete the DA for both segmentation tasks, leading to an improved Dice score over the existing methods, across the batchsizes considered. Again, the existing SFDA methods were very sensitive to batchsize, and AdaMI [3] was also sensitive to the choice of tissue ratio prior: as we were completing the segmentation tasks on 2D slices, different slices had varying amounts of the target label present and we had to create a prior that was dependent on slice depth to achieve reasonable performance. The ABIDE tissue segmentation task was more challenging, as can be seen by the comparatively lower Dice Scores, especially due to the large imbalance in tissues, which affected the performance of AdaMI.

No changes needed to be made to the approach compared to the classification task, including the learning rate, showing the generalisability of the method across tasks.

Subsampling Features: The increase in the number of features for segmentation is a potential limiting factor for the approach, especially if the approach is to be expanded

Method	S	T	C	Information Communicated	Average MAE		
					Bs 4	Bs 8	Bs 16
Source Model	✓	x	x	-	4.38		
Centralised Training	✓	✓	✓	All Data	3.52	3.38	3.36
Target Finetune	x	✓	x	Model Weights	3.57	3.60	3.58
DeepCORAL [48]	✓	x	✓	All Data	4.58	4.41	4.12
FADA [40]	✓	x	x	Model Weights + Features	3.55	3.42	3.78
FedHarmony [16]	✓	x	x	Model Weights + Statistics	3.61	3.50	3.79
Direct Fit	x	x	x	Model Weights + Statistics	4.70	4.31	4.05
SFHarmony 1 GMM Component	x	x	x	Model Weights + Statistics	4.21	4.13	3.71
SFHarmony 2 GMM Components	x	x	x	Model Weights + Statistics	3.87	**3.72**	**3.69**
SFHarmony 3 GMM Components	x	x	x	Model Weights + Statistics	**3.64**	**3.72**	3.73

Table 4. Results on the ABIDE dataset for the age prediction task. S = Source data required, T = Target labels required, C = Centralised data, Bs = Batchsize. The average MAE is the performance across all 4 sites, weighted equally, and is reported for training batchsizes of 4, 8 and 16: 16 was the largest batch achievable. The best SFDA method for each batchsize is in bold.

to 3D networks. As the features learned by convolutional layers are highly correlated, large amounts of redundancy exist between the features; thus we considered selecting only a random subset of the features. The procedure would be identical on this subset and the only additional information to be shared would be the random subsampling indices which would be the same across sites. We explored this approach for the CC359 data (3 components, batchsize 5), and the result can be seen in Table 3. It can be seen that using a smaller subset of the features leads to only a small drop in performance for the segmentation task, such that SFHarmony still outperforms the existing methods when less than 1% of the features are shared. Therefore, when expanding the approach to 3D datasets, subsampling the features is likely to be a sensible compromise between performance and the amount of information to be shared between sites.

4.4. Regression: ABIDE dataset

Baselines: We could not identify any appropriate SFDA baselines. Therefore, the only comparison methods were source model only, centralised data and target finetuning with frozen label predictor, DeepCORAL [48], FADA [40], FedHarmony [16] and Direct Fit. The maximum batchsize possible was 16, and so we tried batchsizes of 4, 8 and 16.

Methods Comparison: It can be seen from Table 4 that for the age prediction task FADA [40] outperformed our proposed approach for two of the three reported batchsizes, unlike in the other tasks. This may well be because the task was completed in 3D, and thus, a small number of samples were available, meaning that the presence of the source data supported the model training. `SFHarmony` did, however, show comparable performance, especially when modelling the features with more components. We could not identify any SFDA methods in the literature that could be directly applied to regression tasks. Our method is flexible and can be directly applied to the regression task without any change to the model architecture or DA procedure.

5. Conclusion

We have presented `SFHarmony`, a method for SFDA, motivated by the need to harmonise MRI data across imaging sites while relaxing assumptions about the availability of source data. We have demonstrated the applicability of the method to classification, regression, and segmentation tasks, and have shown that it outperforms existing SFDA approaches when applied to MR imaging data. The approach is general, allowing it to be applied across architectures and tasks. Possible limitations may arise due the increase in features when applying the approach to 3D volumes, but the initial results on subsampling the features for the segmentation task suggest that the approach will still be applicable.

6. Acknowledgements

ND is supported by a Academy of Medical Sciences Springboard Award. MJ is supported by the National Institute for Health Research, Oxford Biomedical Research Centre, and this research was funded by the Wellcome Trust [215573/Z/19/Z]. WIN is supported by core funding from the Wellcome Trust [203139/Z/16/Z]. AN is grateful for support from the Academy of Medical Sciences under the Springboard Awards scheme (SBF005/1136), and the Bill and Melinda Gates Foundation.

References

[1] Sk Ahmed, Dripta Raychaudhuri, Sujoy Paul, Samet Oymak, and Amit Roy-Chowdhury. Unsupervised multi-source domain adaptation without access to source data. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[2] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Source-relaxed domain adaptation for image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 490–499, 2020.

[3] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Source-free domain adaptation for image segmentation. *Medical Image Analysis*, 82:102617, 2022.

[4] A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 7(4):401–406, 1946.

[5] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 343–351, 2016.

[6] Stenzel Cackowski, Emmanuel Barbier, Michel Dojat, and Thomas Christen. Imunity: a generalizable vae-gan solution for multicenter mr image harmonization. *ArXiv*, 09 2021.

[7] Andrew A. Chen, Joanne C. Beer, Nicholas J. Tustison, Philip A. Cook, Russell T. Shinohara, Haochang Shou, and The Alzheimer's Disease Neuroimaging Initiative. Mitigating site effects in covariance for machine learning in neuroimaging data. *Human Brain Mapping*, 43(4):1179–1195, 2022.

[8] Andrew A. Chen, Chongliang Luo, Yong Chen, Russell T. Shinohara, and Haochang Shou. Privacy-preserving harmonization via distributed combat. *NeuroImage*, 248:118822, 2022.

[9] Intersoft Consulting. General Data Protection Regulation (GDPR), Sep 2019.

[10] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

[11] Blake E. Dewey, Can Zhao, Jacob C. Reinhold, Aaron Carass, Kathryn C. Fitzgerald, Elias S. Sotirchos, Shiv Saidha, Jiwon Oh, Dzung L. Pham, Peter A. Calabresi, Peter C.M. van Zijl, and Jerry L. Prince. Deepharmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic Resonance Imaging*, 64:160–170, 2019.

[12] Adriana di Martino, Chaogan Yan, Qingyang Li, Erin B. Denio, Francisco Xavier Castellanos, Kaat Alaerts, Jeffrey S. Anderson, Michal Assaf, Susan Y. Bookheimer, Mirella Dapretto, Ben Deen, Sonja Delmonte, Ilan Dinstein, Birgit B. Ertl-Wagner, Damien A. Fair, Louise Gallagher, Daniel P. Kennedy, Christopher Lee Keown, Christian Keysers, Janet E. Lainhart, Catherine Lord, Beatriz Luna, V. Menon, Nancy J. Minshew, Christopher S. Monk, Sophia Mueller, Ralph-Axel Müller, Mary Beth Nebel, Joel T. Nigg, Kirsten O'Hearn, Kevin A. Pelphrey, Scott J. Peltier, Jeffrey D. Rudie, Stefan Sunaert, Marc Thioux, Julian Michael Tyszka, Lucina Q. Uddin, Judith S. Verhoeven, Nicole Wenderoth, Jillian Lee Wiggins, Stewart H. Mostofsky, and Michael Peter Milham. The autism brain imaging data exchange: Towards large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19:659 – 667, 2013.

[13] Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7202–7212, 2022.

[14] Nicola K. Dinsdale, Emma Bluemke, Vaanathi Sundaresan, Mark Jenkinson, Stephen M. Smith, and Ana I.L. Namburete. Challenges for machine learning in clinical translation of big data imaging studies. *Neuron*, 110(23):3866–3881, 2022.

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

[16] Nicola K. Dinsdale, Mark Jenkinson, and Ana I. L. Namburete. Fedharmony: Unlearning scanner bias with distributed data. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 695–704, 2022.

[17] Nanqing Dong and Irina Voiculescu. Federated contrastive learning for decentralized unlabeled medical images. 09 2021.

[18] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9, 01 2013.

[19] Qianli Feng, Chenqi Guo, Fabian Benitez-Quiroz, and Aleix M. Martinez. When do gans replicate? on the choice of dataset size. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6701–6710, October 2021.

[20] Jean-Philippe Fortin, Drew Parker, Birkan Tunç, Takanori Watanabe, Mark A. Elliott, Kosha Ruparel, David R. Roalf, Theodore D. Satterthwaite, Ruben C. Gur, Raquel E. Gur, Robert T. Schultz, Ragini Verma, and Russell T. Shinohara. Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, 161:149–170, 2017.

[21] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1180–1189, 2015.

[22] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *Computer Vision – ECCV 2016*, pages 597–613, 2016.

[23] Ben Glocker, Robert Robinson, Daniel Coelho de Castro, Qi Dou, and Ender Konukoglu. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *ArXiv*, abs/1910.04597, 2019.

[24] Hao Guan, Yunbi Liu, Erkun Yang, Pew-Thian Yap, Dinggang Shen, and Mingxia Liu. Multi-site mri harmonization via attention-guided deep domain adaptation for brain disorder identification. *Medical Image Analysis*, 71:102076, 2021.

[25] Xiao Han, Jorge Jovicich, David Salat, Andre van der Kouwe, Brian Quinn, Silvester Czanner, Evelina Busa, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Paul Maguire, Diana Rosas, Nikos Makris, Anders Dale, Bradford Dickerson, and Bruce Fischl. Reliability of mri-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, 32(1):180–194, 2006.

[26] Jorge Jovicich, Silvester Czanner, Douglas Greve, Elizabeth Haley, Andre van der Kouwe, Randy Gollub, David Kennedy, Franz Schmitt, Gregory Brown, James MacFall, Bruce Fischl, and Anders Dale. Reliability in multi-site structural mri studies: Effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, 30(2):436–443, 2006.

[27] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4888–4897, jun 2019.

[28] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143, 13–18 Jul 2020.

[29] Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence*, 2(6):508–518, 2021.

[30] Jogendra Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R. Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. pages 7026–7036, 10 2021.

[31] Jogendra Nath Kundu, Akshay R Kulkarni, Suvaansh Bhambri, Deepesh Mehta, Shreyas Anand Kulkarni, Varun Jampani, and Venkatesh Babu Radhakrishnan. Balancing discriminability and transferability for source-free domain adaptation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11710–11728, 17–23 Jul 2022.

[32] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9638–9647, 2020.

[33] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6028–6039. PMLR, 13–18 Jul 2020.

[34] Y. Liu, W. Zhang, and J. Wang. Source-free domain adaptation for semantic segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1215–1224, 2021.

[35] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 97–105, 2015.

[36] H. B. McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2016.

[37] Daniel Moyer, Greg Ver Steeg, Chantal M. W. Tax, and Paul M. Thompson. Scanner invariant representations for diffusion mri harmonization. *Magnetic Resonance in Medicine*, 84(4):2174–2189, 2020.

[38] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[39] Office for Civil Rights US. The hipaa privacy rule, Jul 2021.

[40] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. In *2020 International Conference on Learning Representations*, 2020.

[41] Daniel Peterson, Pallika H. Kanani, and Virendra J. Marathe. Private federated learning with domain adaptation. *ArXiv*, abs/1912.06733, 2019.

[42] Raymond Pomponio, Guray Erus, Mohamad Habes, Jimit Doshi, Dhivya Srinivasan, Elizabeth Mamourian, Vishnu Bashyam, Ilya M. Nasrallah, Theodore D. Satterthwaite, Yong Fan, Lenore J. Launer, Colin L. Masters, Paul Maruff, Chuanjun Zhuo, Henry Völzke, Sterling C. Johnson, Jurgen Fripp, Nikolaos Koutsouleris, Daniel H. Wolf, Raquel Gur, Ruben Gur, John Morris, Marilyn S. Albert, Hans J. Grabe, Susan M. Resnick, R. Nick Bryan, David A. Wolk, Russell T. Shinohara, Haochang Shou, and Christos Davatzikos. Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208:116450, 2020.

[43] Subhankar Roy, Martin Trapp, Andrea Pilzer, Juho Kannala, Nicu Sebe, Elisa Ricci, and Arno Solin. Uncertainty-guided source-free domain adaptation. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, page 537–555, 2022.

[44] Prabhu Teja S and François Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9608–9618, 2021.

[45] Anit Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks. *ArXiv*, 12 2018.

[46] Giorgos Sfikas, Constantinos Constantinopoulos, Aristidis C. Likas, and Nikolas P. Galatsanos. An analytic distance metric for gaussian mixture models with application in image retrieval. In *International Conference on Artificial Neural Networks*, 2005.

[47] Roberto Souza, Oeslle Lucena, Julia Garrafa, David Gobbi, Marina Saluzzi, Simone Appenzeller, Letícia Rittner, Richard Frayne, and Roberto Lotufo. An open, multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage*, 170:482–494, 2018. Segmenting the Brain.

[48] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision – ECCV 2016 Workshops*, pages 443–450, 2016.

[49] Hidemasa Takao, Naoto Hayashi, and Kuni Ohtomo. Effect of scanner in longitudinal studies of brain volume changes. *Journal of magnetic resonance imaging : JMRI*, 34:438–44, 08 2011.

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

[51] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. pages 4068–4076, 12 2015.

[52] Gael Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine*, 5:48, 04 2022.

[53] Fan Wang, Zhongyi Han, Yongshun Gong, and Yilong Yin. Exploring domain-invariant parameters for source free domain adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7141–7150, 2022.

[54] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8990–8999, 2021.

[55] Ming Xie, Guodong Long, Tao Shen, Tianyi Zhou, Xianzhi Wang, Jing Jiang, and Chengqi Zhang. Multi-center federated learning. *ArXiv*, abs/2108.08647, 2021.

[56] Xuanang Xu, Fugen Zhou, Bo Liu, Dongshan Fu, and Xiangzhi Bai. Efficient multiple organ localization in ct image using 3d region proposal network. *IEEE Transactions on Medical Imaging*, 38:1885–1898, 01 2019.

[57] Chen Yang, Xiaoqing Guo, Zhen Chen, and Yixuan Yuan. Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Analysis*, 79:102457, 2022.

[58] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10, 01 2023.

[59] Lixuan Yang, Cedric Beliard, and Dario Rossi. Heterogeneous data-aware federated learning. *ArXiv*, 11 2020.

[60] S. Yang, Y. Wang, J. van de Weijer, L. Herranz, and S. Jui. Generalized source-free domain adaptation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8958–8967, oct 2021.

[61] Chun-Han Yao, Boqing Gong, Yin Cui, Hang Qi, Yukun Zhu, and Ming-Hsuan Yang. Federated multi-target domain adaptation. *ArXiv*, 2021.

[62] Fenqiang Zhao, Zhengwang Wu, Li Wang, Weili Lin, Shunren Xia, and Gang Li. *Harmonization of Infant Cortical Thickness Using Surface-to-Surface Cycle-Consistent Adversarial Networks*, volume 11767, pages 475–483. 10.

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