



# Unlearning Scanner Bias for MRI Harmonisation

Nicola K. Dinsdale<sup>1(✉)</sup>, Mark Jenkinson<sup>1,2,3</sup>, and Ana I.L. Namburete<sup>4</sup>

<sup>1</sup> Wellcome Centre for Integrative Neuroimaging, FMRIB, University of Oxford, Oxford, UK

[nicola.dinsdale@dtc.ox.ac.uk](mailto:nicola.dinsdale@dtc.ox.ac.uk)

<sup>2</sup> Australian Institute for Machine Learning (AIML), Department of Computer Science, University of Adelaide, Adelaide, Australia

<sup>3</sup> South Australian Health and Medical Research Institute (SAHMRI), North Terrace, Adelaide, Australia

<sup>4</sup> Institute of Biomedical Engineering, University of Oxford, Oxford, UK

**Abstract.** Combining datasets is vital for increased statistical power, especially for neurological conditions where limited data is available. However, variance due to differences in acquisition protocol and hardware limits our ability to combine datasets. We propose an iterative training scheme based on domain adaptation techniques, aiming to create scanner-invariant features while simultaneously maintaining overall performance on the main task. We demonstrate this on age prediction, but expect that our proposed training scheme will be applicable to any feedforward network and classification or regression task. We show that not only can we harmonise three MRI datasets from different studies, but can also successfully adapt the training to work with very biased datasets. The training scheme should, therefore, be applicable to most real-world data scenarios, enabling harmonisation for the task of interest.

**Keywords:** Harmonisation · Joint domain adaptation · MRI

## 1 Introduction

Whilst a few very large projects, such as the UK Biobank [14] now exist, the majority of dataset sizes in neuroimaging studies remain relatively small. Therefore, combining datasets from multiple sites and scanners is vital to give improved statistical power. However, this leads to greater variance in the data, largely due to differences in acquisition protocol and hardware [6]. Thus, harmonisation is required to achieve joint unbiased analysis of data from different scanners.

One popular harmonisation method is ComBat [6], which performs post-hoc normalisation using a linear model, making image-derived values comparable. This was extended to incorporate a nonlinear model in [12] and explicitly to encode bias caused by nonbiological variance in the model in [15]. The majority

of other methods for MRI harmonisation focus on making images produced on one scanner look as if they came from another, with recent studies using deep learning methods (eg. [4]). CycleGANS [18] have also been used to transform images between domains [17].

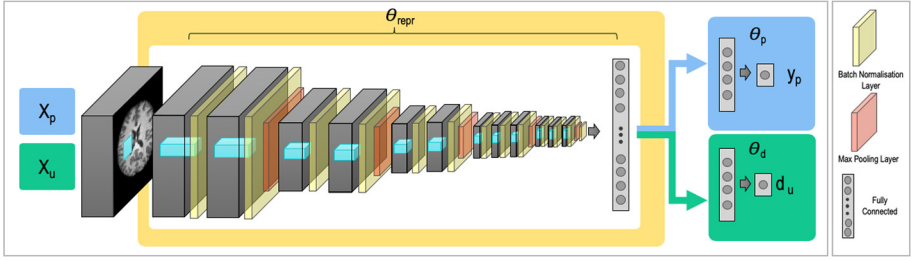
Instead of harmonising the images, we propose to harmonise the features extracted by deep learning networks using a joint domain adaptation approach. Domain adaptation assumes that we have a source domain  $D_s$  with learning task  $T_s$  and a target domain  $D_t$  with learning task  $T_t$  and either  $D_s \neq D_t$  or  $T_s \neq T_t$  [3]; the success of the domain adaptation depends on the existence of a similarity between the two domains [16]. For harmonisation, we consider the case where  $D_s \neq D_t$ ; that is, when the data was collected on distinct scanners. One of the most successful methods for domain adaptation is the DANN network [8] which uses a gradient reversal layer [7] to train a discriminator adversarially, creating a feature representation which is discriminative for the main task but indiscriminate as to the domains. There is, however, little exploration of the effect of domain adaptation on the performance on the *source domain* data, whereas for harmonization it is vital that the network performs well across all the datasets.

In [9] a method is proposed to solve both domain transfer and task transfer simultaneously. Similarly to DANN, they complete domain adaptation using adversarial methods but, rather than using a gradient reversal layer to update the domain predictor in opposition to the task, they use an iterative training scheme: they alternate between learning the best domain classifier with a given feature representation and then minimise a confusion loss which aims to force the domain predictions to become closer to a uniform distribution so it ‘maximally confuses’ the domain classifier [9]. Compared to DANN-style networks it is also better at ensuring an equally uninformative classifier across the domains [2] because of the confusion loss, which is desirable for the harmonisation scenario. This network is applied in [2] where iterative unlearning creates classifiers that are blind to spurious variations in the data. These together form the inspiration for this work.

In this work, we apply a framework similar to that introduced in [9] for harmonisation, by posing the problem as a joint domain adaptation problem. We aim to create a feature representation that is invariant to the scanner from which the data were acquired and show that this network is still able to perform the task of interest. By taking a joint domain adaptation approach, we also require that the network is successful at all tasks and is not just driven by the larger dataset; thus, we explore the effect of training with different datasets. We show that scanner information can be successfully ‘unlearned’ in realistic scenarios, allowing us to harmonise data for the main task of interest. The code is available at: [https://github.com/nkdinsdale/Unlearning\\_for\\_MRI\\_harmonisation](https://github.com/nkdinsdale/Unlearning_for_MRI_harmonisation).

## 2 Method

We explore two different data regimes and show that, by adapting the loss functions, the same learning framework can be used in all scenarios. We consider



**Fig. 1.** Network architecture.  $\mathbf{X}_p$  and  $\mathbf{X}_u$  represent the input data for the network,  $\mathbf{X}_p$  for the main task and  $\mathbf{X}_u$  for unlearning. These can be the same data, subsets of each other, or different datasets. For  $\mathbf{X}_p$  the labels  $\mathbf{y}_p$  are the labels for the main task and for  $\mathbf{X}_u$  the labels are the domain labels  $\mathbf{d}_u$ .  $\Theta_{repr}$  are the parameters of the convolutional layers and first fully connected layer that form the encoder;  $\Theta_p$  are the parameters of the fully connected layers which predict the main task;  $\Theta_d$  are the parameters of the fully connected layers in the domain predictor.

both training with fully supervised data with similar distributions, and training when the data distributions are biased for the main task label both for the task of age prediction.

## 2.1 Standard Supervised Training

This data regime corresponds to the scenario where we have training labels for the main task available for data from all scanners and the data distributions for the main task are similar. In this case,  $\mathbf{X}_p$  and  $\mathbf{X}_u$  are a single dataset  $\mathbf{X}$  which is used to evaluate all of the training iterations.

The aim of the 3D network shown in Fig. 1 is to find a feature representation  $\Theta_{repr}$  that maximises the performance on a primary task while minimising the performance of a discriminator, which aims to predict the site of origin of the data. In this case, we use age prediction (based on T1-weighted MRI scans) as an example task, but the training procedure should generalise to any feedforward architecture and task.  $\Theta_{repr}$  represents the parameters of the encoder which are shared between the two output branches;  $\Theta_p$  represents the parameters for the primary age prediction task, and  $\Theta_d$  represents the parameters of the domain prediction branch. We consider the case of three datasets, each with input images  $\mathbf{X} \in \mathbb{R}^{W \times H \times D \times 1}$  and task labels  $\mathbf{y} \in \mathbb{R}$ , with different domains  $\mathbf{d}$ , representing scans acquired from three distinct scanners.

Three loss functions are used in the training of the network. The first loss is for the main task and is conditioned on each domain:

$$L_p(\mathbf{X}, \mathbf{y}, \mathbf{d}; \Theta_{repr}, \Theta_p) = \sum_{n=1}^N \frac{1}{S_n} \sum_{j=1}^{S_n} L_n(\mathbf{y}_{j,n}, \hat{\mathbf{y}}_{j,n}) \quad (1)$$

where  $N$  is the number of domains and  $S_n$  is the number of subjects from domain  $n$  such that  $\mathbf{y}_{j,n}$  is the true task label for the  $j^{th}$  subject from the  $n^{th}$

domain and  $L_n$  is the loss function for the main task evaluated for the data from domain  $n$ . This loss takes the form of mean squared error (MSE) for the age prediction task. The loss is calculated for each domain separately to prevent the performance being driven by the larger dataset, especially when there is a large imbalance in sample numbers. The domain information is then unlearned using two loss functions in combination. The domain loss is simply the categorical cross entropy:

$$L_d(\mathbf{X}, \mathbf{d}, \Theta_{repr}; \Theta_d) = - \sum_{n=1}^N \mathbb{1}[d = n] \log(p_n) \quad (2)$$

which assesses how much domain information remains in  $\Theta_{repr}$ .  $p_n$  are the softmax outputs of the domain classifier and are used in the confusion loss to remove information, by penalising deviations from a uniform distribution:

$$L_{conf}(\mathbf{X}, \mathbf{d}, \Theta_d; \Theta_{repr}) = - \sum_{n=1}^N \frac{1}{N} \log(p_n) \quad (3)$$

Therefore the overall method minimises the total loss function  $L = L_p + \alpha L_d + \beta L_{conf}$  where  $\alpha$  and  $\beta$  represent weights of the relative contributions of the different loss functions. Eqs. (2) and (3) directly oppose each other and therefore cannot be optimised in a single step. Therefore, we iterate through updating the different loss functions, resulting in three forward passes per batch.

## 2.2 Biased Domains

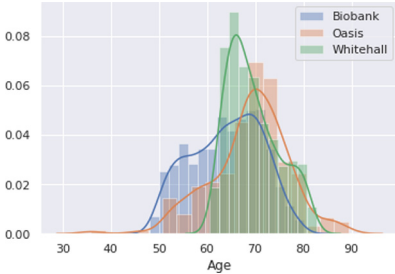
Finally, we consider the scenario where there exists a large difference between the two domains such that the main task label is highly indicative of the scanner and, thus, unlearning scanner information leads to unlearning important information for the main task. For example, consider the scenario where the age distributions for two studies are only slightly overlapping or the scenario where nearly all the subjects with a given condition were collected on one of the scanners: we show that this problem can be reduced by unlearning the domain, using different data to that used to train the main task. This is simple to train within the same learning framework, only requiring Eqs. (2) and (3) to be evaluated across a different dataset or subset of the data. For instance, in the case of slightly overlapping age distributions, the domain information would be unlearned, only using the overlapping section. For the case of a dataset being biased with more subjects with a given pathology being from one scanner, the unlearning would only be completed with the controls. In this case, the equations become:

$$L_d(\mathbf{X}_u, \mathbf{d}_u, \Theta_{repr}; \Theta_d) = - \sum_{n=1}^N \mathbb{1}[d = n] \log(p_n) \quad (4)$$

$$L_{conf}(\mathbf{X}_u, \mathbf{d}_u, \Theta_d; \Theta_{repr}) = - \sum_{n=1}^N \frac{1}{N} \log(p_n) \quad (5)$$

where  $\mathbf{X}_u$  and  $\mathbf{d}_u$  are the input images and domain labels for the intersection data, and can either be a subset of  $\mathbf{X}_p$  or a further dataset of controls only used to unlearn scanner information. It should be noted that task labels are not required for this data, making it much easier to obtain. The main task is still evaluated across the whole dataset  $\mathbf{X}_p$ .

### 3 Experimental Setup

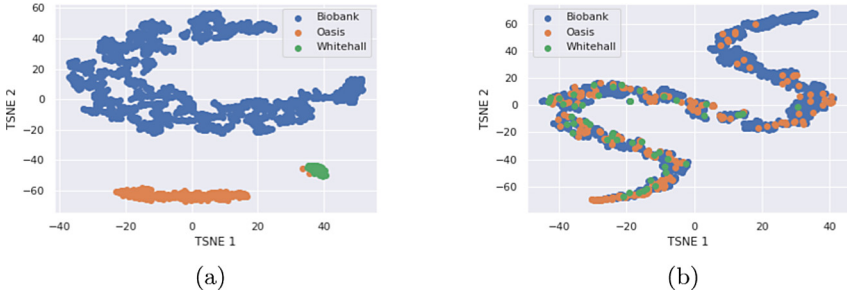


**Fig. 2.** Data distributions for the three datasets, normalised so that the distributions of the smaller datasets can be seen.

retained, leaving 32 slices in the  $z$  direction, chosen so as to maximise coverage across the whole brain whilst minimising redundancy, and allowing a larger batch size and number of filters to be used. The inputs were also normalised to have zero mean and unit standard deviation. We chose to investigate the age prediction task as this is a task for which accurate task labels are easy to obtain. The data distributions can be seen in Fig. 2.

The network was implemented in Python 3.6 using PyTorch (1.0.1) and is based on the VGG-16 architecture [13]; however, our proposed training procedure is applicable to any feedforward network. A batch size of 16 was used throughout, with each batch constrained to contain at least one example from each dataset, increasing the stability during training. To achieve this, the smaller datasets were oversampled.  $\alpha$  and  $\beta$  were empirically set for the different experiments, taking values between 1 and 20.

For the experiments in this work, T1 weighted MRI scans from three datasets were used: UK Biobank [14] (Siemens Skyra 3T) which had been processed using the UK Biobank Pipeline [1] (5508 training, 1377 testing); healthy subjects from the OASIS dataset [11] (Siemens Tesla Vision 1.5T) at multiple time points, split into training and test sets at the subject level (813 training, 217 testing), and subjects from the Whitehall II study [5] (Siemens Magnetom Verio 3T) (452 training, 51 testing). The input images for all datasets were resized to  $128 \times 128 \times 128$  voxels and then every fourth slice was



**Fig. 3.** a) T-SNE [10] plot of the fully connected layer in  $\Theta_{repr}$  from before unlearning. It can be seen that the domains can be almost entirely separated, except for two datapoints grouped incorrectly, showing that data from each scanner has its own distinct distribution. b) T-SNE plot of the fully connected layer in  $\Theta_{repr}$  after unlearning. It can be seen that, through the unlearning, the distributions become entirely jointly embedded.

**Table 1.** Results comparing unlearning to training the network in different combinations on the datasets using Mean Absolute Error (MAE) as the metric for the task of age regression. Scanner accuracy is the accuracy achieved by a domain predictor given the fixed feature representation at convergence, evaluating only for the datasets the network was trained on. The number in brackets indicates random chance. B = Biobank, O = OASIS, W = Whitehall

Training data	Biobank MAE			OASIS MAE			Whitehall MAE			Scanner classification accuracy (%)
	B	O	W							
Normal training										
1.	✓	×	×	$3.25 \pm 2.36$	$16.50 \pm 6.77$	$13.81 \pm 5.42$				–
2.	×	✓	×	$5.61 \pm 3.52$	$4.27 \pm 3.79$	$6.73 \pm 4.82$				–
3.	×	×	✓	$5.61 \pm 3.65$	$5.22 \pm 4.83$	$3.15 \pm 2.81$				–
4.	✓	✓	×	$3.30 \pm 2.50$	$4.00 \pm 2.78$	$4.71 \pm 3.42$				98 (50)
5.	✓	×	✓	$3.31 \pm 2.49$	$4.45 \pm 3.53$	$3.05 \pm 2.84$				100 (50)
6.	×	✓	✓	$5.71 \pm 3.59$	$4.05 \pm 3.71$	$3.21 \pm 2.94$				100 (50)
7.	✓	✓	✓	$3.24 \pm 2.47$	$4.19 \pm 3.50$	$2.89 \pm 2.70$				96 (33)
Unlearning										
8.	✓	✓	×	$3.41 \pm 2.04$	$3.79 \pm 2.99$	$4.60 \pm 3.47$				48 (50)
9.	✓	×	✓	$3.41 \pm 2.58$	$4.07 \pm 4.12$	$2.81 \pm 2.57$				52 (50)
10.	×	✓	✓	$3.38 \pm 2.64$	$3.91 \pm 3.53$	$2.82 \pm 2.65$				50 (50)
11.	✓	✓	✓	$3.38 \pm 2.64$	$3.90 \pm 3.53$	$2.56 \pm 2.47$				34 (33)

## 4 Results

### 4.1 Supervised Unlearning

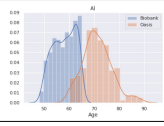
We compared our method to standard training on all three datasets individually and on combinations of datasets and compare the mean absolute errors (MAEs) between methods. The results can be seen in Table 1. It can first be seen that training on all three datasets using normal training gives the best overall performance of the different regimes for standard training as would be expected (row 7), giving the lowest MAE overall across the datasets. This, however, produces a feature representation  $\Theta_{repr}$  in which the three datasets can be separated as is shown in Fig. 3a, so information relating to the scanner is being used to inform the age prediction. This would be particularly problematic if there was a large correlation between task label and scanner as is shown in Sect. 4.2. On the other hand, it can be seen from the results of unlearning on all three datasets (row 11) that we are able to remove scanner information successfully, from the fact that the scanner accuracy is approximately random chance. Simultaneously, there is little decrease in performance across the datasets, showing that the unlearning is not detrimental to performance. In fact, a lower MAE is achieved for the two small datasets (OASIS and Whitehall) and only the performance on Biobank decreases, probably because the network is no longer driven by its much larger size. For reference, standard training using the loss function conditioned on each dataset led to MAEs of  $3.55 \pm 2.68$ ,  $3.90 \pm 3.53$  and  $2.62 \pm 2.65$  years, respectively, and so we are probably seeing improvement due to the unlearning process. This is because, in essence, this method is a domain adaptation approach and so we are harnessing information from each dataset to boost the network’s overall performance by removing scanner differences. Figure 3b) confirms the success of the harmonisation as the scanner domains can no longer be separated.

The comparison for training on two datasets and testing the trained network on the third unseen dataset (e.g. comparing row 6 and 10) also shows that the unlearning procedure helps the network to generalise better: the MAE values for the unseen dataset in both cases are improved by unlearning. This shows that by removing scanner information, preventing it from influencing the prediction, the network learns features that are more applicable to other datasets.

### 4.2 Biased Datasets

We also assessed network performance when training with biased subsets of the OASIS and Biobank datasets using: (i) a standard training regime and, (ii) naïve unlearning (unlearning on the whole of both datasets) and (ii) unlearning on just the overlap set. We considered three degrees of overlap: 5, 10 and 15 years. The resulting networks were then tested on the full testing dataset, spanning across the whole age range. Figure 4 shows the resulting errors. As expected, it can be seen that the normal training regime produces large errors, especially outside of the range of the Biobank training data, and is entirely driven by the larger Biobank training data.

**Table 2.** MAE results for Biobank and OASIS data from training with datasets with varying degrees of overlap as shown in the first column. Scanner accuracy is calculated by training a domain classifier on the fixed feature representation. Random chance is given in brackets.

	Method	Biobank MAE	OASIS MAE	Scanner Classification Accuracy (%)
<b>5 Years</b>	1. Standard	$16.5 \pm 5.94$	$15.5 \pm 6.95$	100 (50)
	2. Naïve Unlearning	$6.11 \pm 3.99$	$4.44 \pm 4.20$	58 (50)
	3. Unlearning on Overlap	$5.49 \pm 3.67$	$4.37 \pm 4.05$	53 (50)
<b>10 Years</b>	4. Standard	$9.66 \pm 5.83$	$13.6 \pm 6.58$	100 (50)
	5. Naïve Unlearning	$4.20 \pm 2.90$	$4.29 \pm 4.01$	56 (50)
	6. Unlearning on Overlap	$3.93 \pm 2.81$	$4.04 \pm 3.86$	52 (50)
<b>15 Years</b>	7. Standard	$8.91 \pm 5.31$	$10.4 \pm 5.55$	100 (50)
	8. Naïve Unlearning	$3.82 \pm 2.84$	$4.39 \pm 4.07$	57(50)
	9. Unlearning on Overlap	$3.75 \pm 2.78$	$3.99 \pm 3.52$	50 (50)

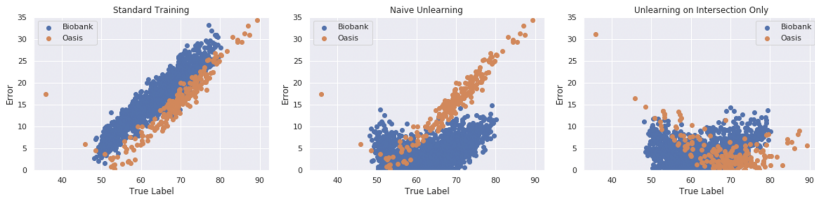
With naïve unlearning, the network is not able to correct for both scanners and the results for the OASIS dataset are poor, whereas by using unlearning on just the overlap subjects, the error is reduced across both datasets. The only region which performs worse is the lower end of the OASIS dataset, probably because when the network was being driven only by the Biobank data, the network generalised to OASIS testing points from the same range. Naïve unlearning also performs slightly less well at removing scanner information on the testing domain. This probably indicates that the features learned also encode some age information and so generalise less well across the whole age range.

These results show the power of the network to remove strong bias from the data, with a large reduction in error across the dataset compared to standard learning. There is also a clear improvement compared to unlearning on the whole dataset, showing that information which was key to the age prediction task was being removed when the whole dataset was used for unlearning, and this is lessened by training on only the overlap dataset.

## 5 Discussion

We have shown that, through using an iterative training scheme to ‘unlearn’ scanner information, we can create features from which most scanner information has been removed and, thus, harmonise the data for a given task. The training regime is flexible and could be implemented with any feedforward network and likely data scenarios such as biased datasets, meaning that unlearning should be applicable for most real-world MRI harmonisation problems.





**Fig. 4.** Errors for the three different training regimes: standard training, naïve unlearning and unlearning only on the overlap data (10 year case). It can be seen that unlearning only on the overlap dataset leads to much lower losses across both datasets.

**Acknowledgements.** ND is supported by the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC) [grant number EP/L016052/1]. MJ is supported by the National Institute for Health Research (NIHR), Oxford Biomedical Research Centre (BRC), and this research was funded by the Wellcome Trust [215573/Z/19/Z]. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust [203139/Z/16/Z]. AN is grateful for support from the UK Royal Academy of Engineering under the Engineering for Development Research Fellowships scheme. The computational aspects of this research were supported by the Wellcome Trust Core Award [Grant Number 203141/Z/16/Z] and the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

## References

1. Alfaro-Almagro, F., et al.: Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *bioRxiv* **166**, 130385 (2017)
2. Alvi, M.S., Zisserman, A., Nellåker, C.: Turning a blind eye: explicit removal of biases and variation from deep neural network embeddings. In: *ECCV Workshops* (2018)
3. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.: A theory of learning from different domains. *Mach. Learn.* **79**, 151–175 (2010)
4. Dewey, B., et al.: DeepHarmony: a deep learning approach to contrast harmonization across scanner changes. *Magn. Reson. Imaging* **64**, 160–170 (2019)
5. Filippini, N., et al.: Study protocol: the Whitehall II imaging sub-study. *BMC psychiatry* **14**, 159 (2014). <https://doi.org/10.1186/1471-244X-14-159>
6. Fortin, J.P., et al.: Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage* **167**, 104–120 (2017). <https://doi.org/10.1016/j.neuroimage.2017.11.024>
7. Ganin, Y., Lempitsky, V.S.: Unsupervised domain adaptation by back propagation. *ArXiv* (2014)
8. Ganin, Y., et al.: Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **17**, 59:1–59:35 (2015)
9. Hoffman, J., Tzeng, E., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4068–4076 (2015)

10. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008)
11. Marcus, D., Wang, T., Parker, J., Csernansky, J., Morris, J., Buckner, R.: Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* **19**, 1498–507 (2007). <https://doi.org/10.1162/jocn.2007.19.9.1498>
12. Pomponio, R., et al.: Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage* **208**, 116450 (2019). <https://doi.org/10.1016/j.neuroimage.2019.116450>
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, September 2014
14. Sudlow, C., et al.: UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos Med.* **12**, e1001779 (2015)
15. Wachinger, C., Rieckmann, A., Pölsterl, S.: Detect and correct bias in multi-site neuroimaging datasets. *bioRxiv* (2020)
16. Wilson, G., Cook, D.J.: A survey of unsupervised deep domain adaptation (2018)
17. Zhao, F., Wu, Z., Wang, L., Lin, W., Xia, S., Li, G.: Harmonization of infant cortical thickness using surface-to-surface cycle-consistent adversarial networks, pp. 475–483 (2019)
18. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251 (2017)