Disentangling Site Effects with Cycle-Consistent Adversarial Autoencoder for Multi-site Cortical Data Harmonization

Fenqiang Zhao[1], Zhengwang Wu[1], Dajiang Zhu[2], Tianming Liu[3], John Gilmore[4], Weili Lin[1], Li Wang[1], and Gang Li[1(✉)]

[1] Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
gang_li@med.unc.edu
[2] Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA
[3] Department of Computer Science, University of Georgia, Athens, GA, USA
[4] Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Abstract. Modern multi-site neuroimaging studies are known to be biased by significant site effects observed in imaging data and their derived structural and functional features. Although many statistical models and deep learning methods have been proposed to eliminate the site effects while maintaining biological characteristics, they have two major drawbacks. *First*, statistical models are applicable for harmonizing regional-level data but are inherently not suitable to represent the complex non-linear mapping of vertex-wise cortical property maps. *Second*, existing deep learning methods can only harmonize data between two sites, which are practically less useful in multi-site data harmonization scenario and also ignore the rich information in the whole dataset. To address these issues, we develop a novel, flexible deep learning method to harmonize multi-site cortical surface property maps. Specifically, to detect and remove site effects, we employ a surface-based autoencoder and decompose the encoded cortical features into site-related and site-unrelated components and use an adversarial strategy to encourage the disentanglement. Then decoding the site-unrelated features with other site features can generate mappings across different sites. To learn more controllable and meaningful mappings, we enforce the cycle consistency between forward and backward mappings. Our method can thus efficiently learn rich information from the whole dataset and generate realistic harmonized surface maps at the target site. Experiments on harmonizing infant cortical thickness maps of 2,342 scans from four sites with different scanners and imaging protocols validate the superior performance of our method on both site effects removal and biological variability preservation compared to other methods. To the best of our knowledge, this is the largest validation of different methods on infant cortical data harmonization.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14227, pp. 369–379, 2023.
https://doi.org/10.1007/978-3-031-43993-3_36

1 Introduction

Modern large multi-site neuroimaging studies have shown increasing power to detect biological variability of interest and provided invaluable insights into the changes underlying neurodevelopmental and neurodegenerative disorders [8,12]. However, the aggregation of neuroimaging data across different sites and scanners typically introduces non-biological variability, also known as site effects [7]. Many harmonization methods are thus proposed to remove such unwanted site effects while preserving biological variability [14,20].

Inspired by image-to-image translation techniques in the computer vision field [30], many methods harmonized neuroimaging data in the image domain by synthesizing brain images among different sites [2,4,11]. However, image-based harmonization techniques cannot guarantee that the final derived structural or functional features are free of site effects, due to the huge complexity in neuroimaging data processing pipelines [23]. Alternatively, feature-based harmonization techniques have been proposed to directly mitigate the site effects in the final derived volumetric [19,21], structural [7,19,21,28], functional [26], or diffusion magnetic resonance imaging (dMRI) features [18,25]. Most of these methods are developed based on linear statistical models, e.g., Combat [7] and its variants [21], for harmonizing summarized cortical properties in each region of interest (ROI), e.g., cortical thickness [7] and surface area [20]. Although achieving promising results for long ROI-wise data (number of samples $n >$ number of features p), statistical models have inherent problems when applied to wide vertex-wise data ($n \ll p$) with spatially fine-grained cortical information [1,15]. This is because the growth in data dimension and possible associations in wide vertex-wise data makes the model more complex and consequently statistical inference becomes less tractable and precise [15]. Moreover, since the sources and underlying mechanisms of site effects are heterogeneous and not fully uncovered, linear models and their hypothesis on the model's parameter distribution may not sufficiently represent the complex non-linear mapping of vertex-wise data. Therefore, deep learning methods, that make minimal assumptions about the data-generating mechanisms and thus can automatically learn to fit the complex non-linear mappings, are more favored [15]. For example, Zhao et al. [28] proposed a surface-to-surface CycleGAN to harmonize cortical thickness maps between two sites, which, however, is inefficient and inconvenient in practice for harmonizing multi-site data, because a model needs to be re-trained between any two sites and ignores rich global information in the whole multi-site data.

To address these issues, we develop a novel, flexible deep learning-based method to harmonize multi-site cortical surface maps in a vertex-wise manner free of parcellation scheme, thus preserving the spatially detailed cortical measure information after harmonization and enabling comprehensive vertex-wise analysis in further studies. Our approach builds on a surface-based autoencoder and uses an adversarial strategy [5] to encourage the disentanglement of site-related and site-unrelated components. To learn a more controllable and meaningful generative model, we enforce the cycle consistency between forward and backward mappings, inspired by CycleGAN [30]. Our method also shares

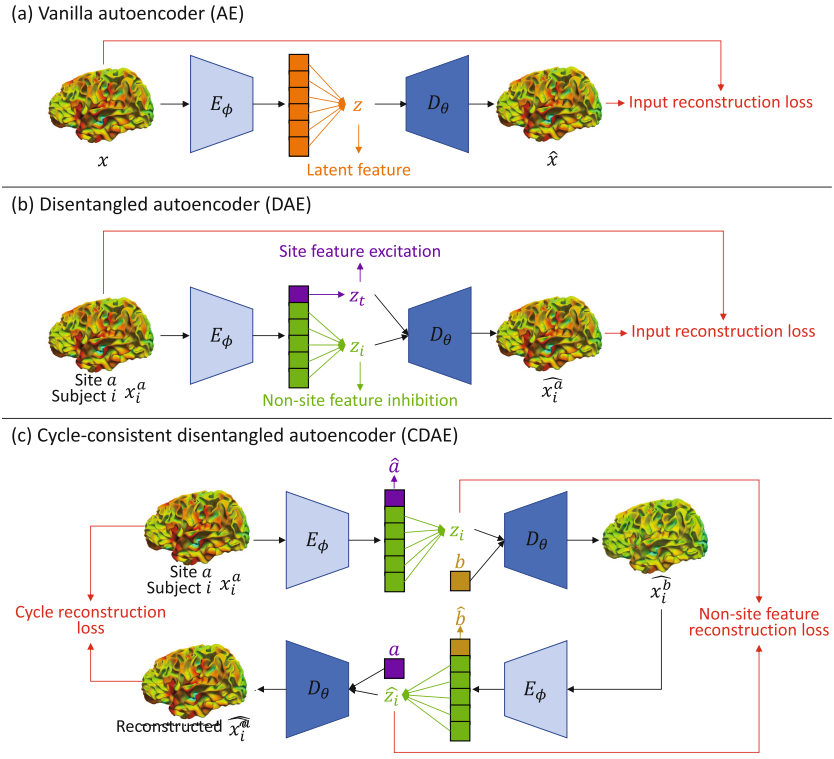

Fig. 1. Illustration of different autoencoder models for cortical surface data representation and harmonization.

certain similarity with the adversarial autoencoder [17], domain-adversarial neural networks [9], and guided variational autoencoder [5], but significantly differs in the operation space, network structure, and most importantly, the aim and application, where we focus on a more meaningful data harmonization task in the neuroimaging field. To sum up, the main contributions of this paper are:

1. We propose a novel method to fill the critical gap for multi-site vertex-wise cortical data harmonization, while there exist statistical models suitable for ROI-wise data harmonization and CycleGAN model for two-site data;
2. Taking advantage of disentanglement learning and adversarial strategies, we successfully learn a transparent, controllable, and meaningful generative model for efficiently mapping cortical surface data across different sites;
3. To the best of our knowledge, we performed the largest validation on infant cortical data harmonization with 2,342 scans from 4 sites and demonstrate the superior performance of our method compared to other methods.

2 Method

2.1 Vanilla Autoencoder (AE)

As shown in Fig. 1(a), we developed our method based on a vanilla autoencoder (AE) to enjoy its transparency and simplicity. Let $X = (x_1, ..., x_n)$ denote a set of input cortical surface property maps, e.g., cortical thickness map, where n is the number of samples. The encoder network E_ϕ parameterized by ϕ extracts the latent features $z : z = E_\phi(x)$, which is developed based on the fundamental operations of spherical CNN proposed in Spherical U-Net [27]. Spherical U-Net leverages the consistent and regular data structure of resampled spherical cortical surfaces to design the 1-ring filter on cortical surfaces and accordingly extends convolution, pooling, and upsampling operations to the spherical surface. Herein, E_ϕ consists of 5 repeated spherical 1-ring-convolution+Batch Normalization (BN)+ReLU layers with 4 spherical mean pooling layers between them. The feature channel at each resolution is 8, 16, 32, 64, and 128, respectively. To enable more compact feature extraction, we add a flatten and a linear layer with 512 neurons to the end of the encoder and finally obtain the latent vector z with size 1×512. The decoder D_θ first uses a linear layer with 20,736 neurons and then reshapes it into 162×128 to recover the feature map at lowest resolution. Then it gradually upsamples the features and finally reconstructs the input data with its original size, which can be formulated as: $\hat{x} = D_\theta(z)$. The training process of an AE thus tries to minimize the reconstruction loss:

$$\mathcal{L}_{recon}(\phi, \theta) = \sum_{i=1}^{n} \|x - D_\theta(E_\phi(x))\|^2 \tag{1}$$

2.2 Disentangled Autoencoder (DAE)

To detect and remove site effects from multi-site vertex-wise cortical measurements, we introduce a disentanglement learning strategy for its controllability and interpretability [5]. As shown in Fig. 1(b), suppose for training data $X = (x_1, \ldots, x_n)$, there are M sites in total and the corresponding ground-truth site labels are $T = (t_1, ..., t_n)$. Let $z = (z_t, z_i)$, where z_t defines a $1 \times M$ vector representing the site-wise classification probability and z_i represents remaining latent variables. We use the adversarial excitation and inhibition method [5] to encourage the disentanglement of the latent variables:

$$\mathcal{L}_{Excitation}(\phi, W_t) = -\sum_{i=1}^{n}[t_i \cdot log(W_t(z_t)) + (1 - t_i) \cdot log(1 - W_t(z_t))] \tag{2}$$

$$\mathcal{L}_{Inhibition}(\phi, W_i) = \sum_{i=1}^{n}[t_i \cdot log(W_i(z_i)) + (1 - t_i) \cdot log(1 - W_i(z_i))] \tag{3}$$

where W_t and W_i refer to the site classifiers using latent variables z_t and z_i, respectively, t_i is a one-hot vector encoding the ground-truth site label. The multi-class binary cross entropy loss in Eq. (2) thus encourages z_t to be the

same as correct site label. Conversely, Eq. (3) is an inhibition process, as we want the remaining variables z_i to be site-unrelated. W_t is a simple sigmoid layer for outputting probability, while W_i consists of four linear layers with BN+ReLU+dropout layers between them and a sigmoid layer at the end. This is because z_t should be directly correlated with the site-wise predictions, while z_i with more features needs a stronger classifier to detect site-related information in it and adversarially train the encoder to extract site-unrelated features for z_i.

2.3 Cycle-Consistent Disentangled Autoencoder (CDAE)

Previously, a DAE, after successful training, will be directly used for image synthesis [5] or MR image harmonization [31] by combining site-unrelated variables with the target site label. Such image generation process may introduce new and even unseen style patterns in computer vision applications, which, however, are artifacts that are not meaningful and acceptable in medical imaging field [16].

To learn a more controllable and meaningful generative model, we propose to train the decoder with additional constraints on backward mapping. As shown in Fig. 1(c), after generating the surface map from source site a to target site b, denoted as $\widehat{x_i^b}$, we backward map it to the source site: $(\hat{a}, z_i) = E_\phi(x_i^a), \widehat{x_i^b} = D_\theta(b, z_i)$, $(\hat{b}, \hat{z}_i) = E_\phi(\widehat{x_i^b}), \widehat{\widehat{x_i^a}} = D_\theta(a, \hat{z}_i)$, and enforce the cycle-consistency loss to guarantee the generated surface map is meaningful to the original surface map:

$$\mathcal{L}_{cycle}(\phi, \theta) = \sum_{i=1}^{n} \|\widehat{\widehat{x_i^a}} - x_i^a\|^2. \tag{4}$$

We also add the cycle-consistency loss to the latent site-unrelated variable z_i to reinforce the correlation between the generated and original surface map:

$$\mathcal{L}_{latent}(\phi, \theta) = \sum_{i=1}^{n} \|\hat{z}_i - z_i\|^2, \tag{5}$$

and an explicit correlation loss to further reduce the ambiguity of indirect cycle-consistency losses for better preserving global structural information:

$$\mathcal{L}_{cc}(\phi, \theta) = \sum_{i=1}^{n} cov(\widehat{x_i^b}, x_i^a)/(\sigma_{\widehat{x_i^b}} \sigma_{x_i^a}) \tag{6}$$

where *cov* denotes the covariance, σ is the standard deviation. Besides, we also use the losses in AE and DAE to train the backward mapping. This means we reuse the site classifier W_t to adversarially train the decoder to generate fake surface maps at the target site that cannot be distinguished from the true maps, which is a standard GAN training process [10]. Finally, our model can automatically detect and disentangle the site-related feature from the input data using DAE losses, and generate more meaningful mappings across sites and better preserve individual variability using CDAE losses, thus fulfilling the requirement of data harmonization task.

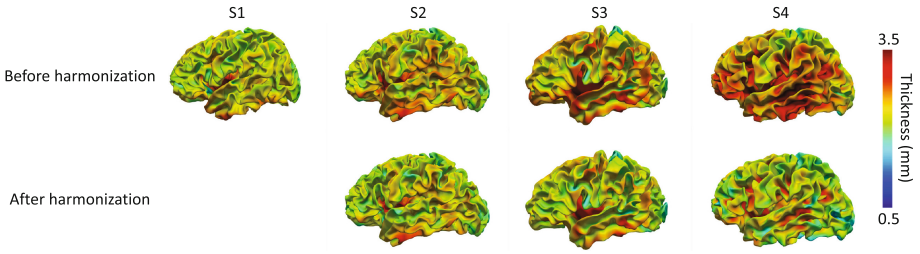

Fig. 2. Representative cortical thickness map harmonization results of four age-matched subjects from four different sites.

3 Experiments and Results

3.1 Experimental Setting

We evaluated our method using 4 large infant datasets (S1, S2, S3, and S4) acquired by different scanners and imaging protocols with resolutions ranging from isotropic 0.8 mm^3 (S1) to $1.25 \times 1.25 \times 1.95$ mm^3 (S3). S1 [13] and S4 [12] are two public datasets with 609 and 779 scans, respectively. S2 [22] and S3 [24] are two in-house datasets containing 335 and 619 scans, respectively. To the best of our knowledge, this is the largest collection of infant MRI datasets for harmonization and joint analysis purposes. All images were preprocessed using the infant-dedicated pipeline iBEAT V2.0 (http://www.ibeat.cloud/) [23]. Then the reconstructed cortical surfaces were mapped onto the sphere, nonlinearly aligned based on geometric features, and further resampled with 40,962 vertices [6]. Figure 2 shows typical reconstructed surfaces color-coded by cortical thickness in the first row. As can be seen, the site effects introduced by different acquisition methods largely dominate the data variance and will inevitably mislead the joint analysis of the four sites if without performing harmonization.

We implemented our method using PyTorch and public Spherical U-Net code [27,29]. We trained the models in an easy-to-hard manner by gradually adding losses from Eq. (1) to Eq. (6) using Adam optimizer with a fixed learning rate 5e-4. All surfaces were randomly split into training and testing sets with the proportion of 7:3. The weights of different loss terms are empirically set as 1.0, 1.0, 0.5, 0.5, 4.0, 5.0 for $\mathcal{L}_{recon}$, $\mathcal{L}_{Excitation}$, $\mathcal{L}_{Inhibition}$, $\mathcal{L}_{cycle}$, $\mathcal{L}_{latent}$, and $\mathcal{L}_{cc}$, respectively. All the experiments were run on a PC with an Nvidia RTX 3080-Ti GPU and an Intel Core i7-9700K CPU. We adopted the popular statistical model, Combat [7], as the baseline method for comparison. We used its public code with age and sex as biological covariates for harmonizing the four sites' cortical thickness in a vertex-wise manner and ROI-wise manner, referred to *Vertex-wise Combat* and *ROI-wise Combat*, respectively. Of note, Combat harmonizes multiple sites into one intermediate site, while our method maps less reliable sites (with low-quality images) to the more reliable site (with high-quality images), i.e., S1, in this work.

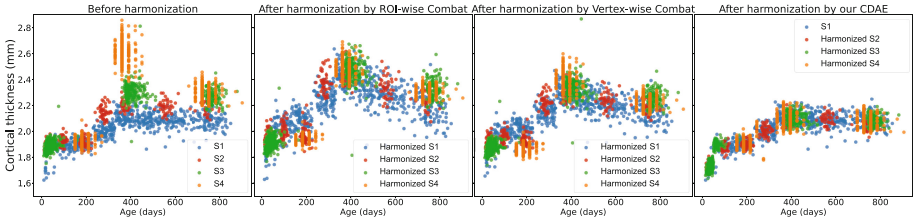

Fig. 3. Scatter plots of population-level developmental trajectories of average cortical thickness in four datasets after harmonization by different methods.

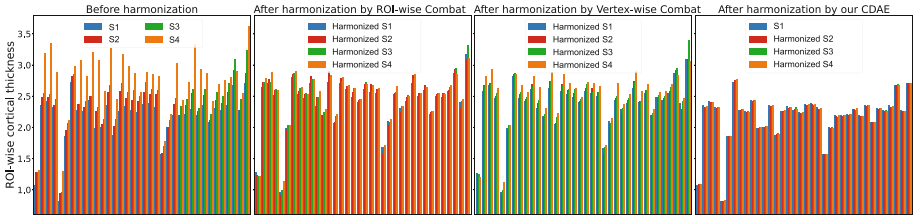

Fig. 4. ROI-wise average cortical thickness of 1-year-old infants at different sites. In each subplot, the x-axis represents the 36 ROIs of Desikan-Killiany parcellation [3].

3.2 Results

Validation on Removing Site Effects. To validate if site effects are successfully removed, we first show the population-level developmental trajectory of average cortical thickness in Fig. 3. To fairly compare with statistical models that use all data and deliver more reliable comparison with larger sample size, we draw the figures based on the whole dataset. As shown, the developmental trajectories from different sites are not consistent and comparable before harmonization and after harmonization using Combat, but are well harmonized to a common space and enable joint analysis using our method. We also compare the ROI-level thickness values before and after harmonization as performed in [7,28]. To be less biased by the data distribution of ages, we show the average thickness of 1-year-old infants for each ROI in Desikan-Killiany parcellation [3]. As shown in Fig. 4, the differences of ROI-wise thickness of different sites are significant before harmonization (all $p < 0.05$) and still remain after harmonization by Combat, but are not significant after harmonization by our method (all $p > 0.1$). Of note, although our method deals with vertex-wise data, compared to ROI-wise Combat, Fig. 3 and Fig. 4 still show better performance of our method on population-level and region-level site effects removal. These results indicate that Combat and other linear models indeed suffer from limited modeling ability and thus are not able to capture the complex non-linear mapping in heterogeneous multi-site data, while our method based on deep learning can effectively solve this problem with the proposed loss functions.

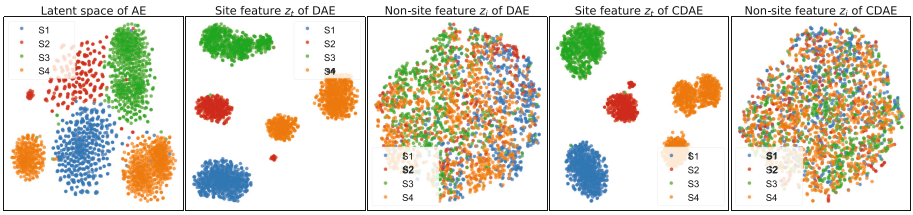

Fig. 5. t-SNE visualization of the latent variables of different autoencoders.

Moreover, we also perform the common practice to visualize the latent space of autoencoders, which is based on t-SNE embedding of the extracted latent features z. As shown in Fig. 5, the latent features extracted by AE demonstrate clear clusters of different sites, indicating that the data variance in the original data is dominated by site effects. However, after disentanglement learning using DAE, the site effects are successfully disentangled and removed, and CDAE further improves the results thanks to more constraints on backward mapping. Finally, to quantitatively evaluate site effects removal, we also attempted to predict the site from harmonized cortical thickness features [7]. Following the settings in [7], we used a support vector machine (SVM) model with radial basis kernel and the hyper-parameters of the model are selected using grid search based on 10-fold validation on the training set. After fitting on the training set, the SVM classification accuracies on the testing set are 74.3%, 92.2%, 51.4%, **40.2%** for ROI-wise Combat, Vertex-wise Combat, DAE, and CDAE, respectively. A lower accuracy of DAE and CDAE indicates that our method successfully removes site effects in the cortical thickness measurements, while Combat still preserves certain site-identifiable information.

Validation on Preserving Individual Variability. While it is important to validate if a harmonization method removes site effects, it is equally important to show that the method preserves biological variability after harmonization. Following [28], we computed the Euclidean distances between any two surface maps in the same site, thus forming a distance matrix, denoted as $H_{i,j}^{n\times n} = \|x_i - x_j\|^2$. We computed the Pearson correlation coefficient (PCC) of the distance matrices before and after harmonization to estimate if relative distances of different scans in the same site are preserved after harmonization. The average PCC values across sites are 0.853, 0.921, 0.842, **0.959** for ROI-wise Combat, Vertex-wise Combat, DAE, and CDAE, respectively, suggesting that our method preserves the most individual differences. From Fig. 2, we can also see that individual variability is well preserved while the site-related information is successfully removed with our method.

Validation on Downstream Task. We further investigate how different harmonization methods affect a downstream task. Same as in [7], we used a support vector regression (SVR) model with radial basis function to predict the scan age using cortical thickness features. After hyper-parameters selection and training, the mean absolute error on the testing set is 81.1, 59.6, 50.5, 56.3, and **45.1**

days for no harmonization, ROI-wise Combat, Vertex-wise Combat, DAE and CDAE, respectively. The coefficient of determination, or R^2, are 0.823, 0.865, 0.892, 0.884, **0.936** for no harmonization, ROI-wise Combat, Vertex-wise Combat, DAE, and CDAE, respectively. Note that Vertex-wise Combat improves the prediction accuracy of ROI-wise Combat possibly due to more spatial details preservation. All these results indicate that our CDAE model substantially increases the accuracy and trustworthiness of joint analysis of multi-site data compared to no harmonization and Combat-based harmonization methods.

4 Conclusion

In this paper, to address the important yet unsolved problem of multi-site vertex-wise cortical data harmonization, we proposed a novel, flexible deep learning-based method, named Cycle-consistent Disentangled Autoencoder (CDAE). Our CDAE takes advantage of the simplicity and transparency of autoencoder, the controllability and interpretability of disentanglement learning, and reinforced meaningfulness of cycle-consistency to successfully learn the complex non-linear mapping among heterogeneous multi-site data, which is inherently difficult and unsuitable for existing methods. Both visual and quantitative results on four datasets with 2,342 scans show the effectiveness of our method on both site effects removal and biological variability preservation. Our method will not only facilitate multi-site vertex-wise neuroimaging data analysis but also inspire novel directions in learning-based data harmonization.

Acknowledgements. This work was supported in part by the National Institutes of Health (NIH) under Grants MH116225, MH117943, MH123202, AG075582, and NS128534.

References

1. Bzdok, D.: Classical statistics and statistical learning in imaging neuroscience. Front. Neurosci. **11**, 543 (2017)
2. Cackowski, S., Barbier, E.L., Dojat, M., Christen, T.: ImUnity: a generalizable VAE-GAN solution for multicenter MR image harmonization. Med. Image Anal. **88**, 102799 (2023)
3. Desikan, R.S., et al.: An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage **31**(3), 968–980 (2006)
4. Dewey, B.E., et al.: DeepHarmony: a deep learning approach to contrast harmonization across scanner changes. Magn. Reson. Imaging **64**, 160–170 (2019)
5. Ding, Z., et al.: Guided variational autoencoder for disentanglement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7920–7929 (2020)
6. Fischl, B.: FreeSurfer. Neuroimage **62**(2), 774–781 (2012)
7. Fortin, J.P., et al.: Harmonization of cortical thickness measurements across scanners and sites. Neuroimage **167**, 104–120 (2018)

8. Frisoni, G.B., Fox, N.C., Jack, C.R., Jr., Scheltens, P., Thompson, P.M.: The clinical use of structural MRI in Alzheimer disease. Nat. Rev. Neurol. **6**(2), 67–77 (2010)
9. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 1–35 (2016)
10. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
11. Guan, H., Liu, M.: DomainATM: domain adaptation toolbox for medical data analysis. NeuroImage **268**, 119863 (2023)
12. Hazlett, H.C., et al.: Early brain development in infants at high risk for autism spectrum disorder. Nature **542**(7641), 348–351 (2017)
13. Howell, B.R., et al.: The UNC/UMN baby connectome project (BCP): an overview of the study design and protocol development. Neuroimage **185**, 891–905 (2019)
14. Hu, F., et al.: Image harmonization: a review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. NeuroImage **274**, 120125 (2023)
15. Ij, H.: Statistics versus machine learning. Nat. Meth. **15**(4), 233 (2018)
16. Kazeminia, S., et al.: GANs for medical image analysis. Artif. Intell. Med. **109**, 101938 (2020)
17. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. arXiv preprint arXiv:1511.05644 (2015)
18. Moyer, D., Ver Steeg, G., Tax, C.M., Thompson, P.M.: Scanner invariant representations for diffusion MRI harmonization. Magn. Reson. Med. **84**(4), 2174–2189 (2020)
19. Pomponio, R.: Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. Neuroimage **208**, 116450 (2020)
20. Solanes, A., et al.: Removing the effects of the site in brain imaging machine-learning-measurement and extendable benchmark. Neuroimage **265**, 119800 (2023)
21. Torbati, M.E., et al.: A multi-scanner neuroimaging data harmonization using ravel and combat. Neuroimage **245**, 118703 (2021)
22. Wang, F., et al.: Developmental topography of cortical thickness during infancy. Proc. Natl. Acad. Sci. **116**(32), 15855–15860 (2019)
23. Wang, L., Wu, Z., Chen, L., Sun, Y., Lin, W., Li, G.: iBEAT V2. 0: a multisite-applicable, deep learning-based pipeline for infant cerebral cortical surface reconstruction. Nat. Protoc. 18, 1488–1509 (2023)
24. Xia, K.: Genetic influences on longitudinal trajectories of cortical thickness and surface area during the first 2 years of life. Cereb. Cortex **32**(2), 367–379 (2022)
25. Xia, Y., Shi, Y.: Personalized DMRI harmonization on cortical surface. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13436, pp. 717–725. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_68
26. Yamashita, A., et al.: Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. PLoS Biol. **17**(4), e3000042 (2019)
27. Zhao, F., et al.: Spherical deformable U-Net: application to cortical surface parcellation and development prediction. IEEE Trans. Med. Imaging **40**(4), 1217–1228 (2021)
28. Zhao, F., et al.: Harmonization of infant cortical thickness using surface-to-surface cycle-consistent adversarial networks. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 475–483. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_52

29. Zhao, F.: Spherical U-Net on cortical surfaces: methods and applications. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 855–866. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_67
30. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
31. Zuo, L., et al.: Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. Neuroimage **243**, 118569 (2021)