2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871061

'Harmless' adversarial network harmonization approach for removing site effects and improving reproducibility in neuroimaging studies

Weizheng Yan, Zening Fu, Jing Sui, *Senior Member, IEEE,* Vince D. Calhoun, *Fellow, IEEE*

***Abstract*—Multi-site collaboration, which gathers together samples from multiple sites, is a powerful way to overcome the small-sample problem in the neuroimaging field and has the potential to discover more robust and reproducible biomarkers. However, confounds among the datasets caused by various site-specific factors may dramatically reduce the cross-site reproducibility performance. To properly remove confounds while improving cross-site task performances, we propose a maximum classifier discrepancy generative adversarial network (MCD-GAN) that combines the advantages of generative models and maximum discrepancy theory. The mechanisms of MCD-GAN and how it harmonizes the dataset are visualized using simulated data. The performance of MCD-GAN was also compared with state-of-the-art methods (e.g., ComBat, cycle-GAN) within Adolescent Brain Cognitive Development (ABCD) dataset. Result demonstrates that the proposed MCD-GAN can effectively improve the cross-site gender classification performance by harmonizing site effects. Our proposed framework is also suitable for various classification/prediction tasks and is promising to facilitate the cross-site reproducibility of neuroimaging studies.**

***Clinical Relevance*— This work provides an efficient method for removing sites effects and improving reproducibility in large-cohort neuroimaging studies.**

I. Introduction

Multi-site neuroimaging collaboration is a powerful way to overcome the small-sample problems by gathering samples from multiple datasets. However, data from different sites are often acquired using different vendors, protocols, or software versions. Inconsistencies can arise from the magnetic resonance imaging (MRI) machine's field strength, gradient non-linearity, time-of-day, head motion [1], and other factors such as population and recruitment difference, thus compromising consistency and reproducibility of the downstream analysis across studies. Meta-regression studies demonstrate that sample size significantly moderated the pooled classification performances, with lower accuracy associated with a larger sample size [1]. Hence, properly removing the task-irrelevant confounds is essential for improving the outcomes of large-cohort studies.

Non-biological confounds often have a prior unpredictable distribution, making it challenge to be properly removed. The popular harmonizing methods can be divided into two categories: **dataset harmonization** and **domain adaptation**. The core difference between the two categories lies in whether the specific task (e.g., classification, regression) is taken into account when harmonizing. Data harmonization approaches, such as residual harmonization, ComBat [2], Neuroharmony [3], usually estimate the distribution of confounds and remove them while ignoring the subsequent tasks such as group classification. For example, ComBat performs a Bayesian regression that corrects the measurements from different samples with additive and multiplicative terms. Therefore, dataset harmonization is flexible to multiple downstream tasks. However, if the confound to remove was not accurately estimated, or the confound has a strong correlation with the downstream task, data harmonization may even degrade the performance of downstream cross-site performances. Domain adaptation trains task-specific models while harmonizing the feature by mapping the features into a shared task-specific subspace [4, 5]. Domain adaptation methods can theoretically guarantee that the model trained on the source domain will achieve improved performance on the target domain. However, adapted domains are often task-specific and not intuitive for interpretation because of the mapping (mostly non-linearly) from original space to task-related low-dimensional subspace.

In addition, compared to conventional harmonizing methods which usually use linear models for confound estimation, deep learning-based models are often more efficient in capturing high-level feature relations using multiple non-linear layers. Deep generative adversarial model is theoretically an ideal solution for multi-site harmonizing due to its adversarial training strategy. The idea has been applied to both the data harmonization and domain adaptation fields [4, 6]. For example, Bashyam et al., applied cycle-GAN, first proposed to solve style transfer problems[7], to MRI dataset for aligning distributions between the source domain and target domain. By learning an unsupervised image to image canonical mapping from diverse datasets to a reference domain using generative deep learning methods, the cycle-GAN model can reduce confounding data variation while preserving semantic information. Guan et al. designed a deep adversarial neural network to harmonize the MRI scans from different Alzheimer's disease studies[4]. However, cycle-GAN cannot guarantee better performance on a specific task. Integrating the concept within the maximum classifier discrepancy method proposed by Saito et al., [5] can be a remedy.

*This work is supported by National Institute of Health grants and National Science Foundation.

W. Yan (wyan3@gsu.edu), Z. Fu, J. Sui (kittysj@gmail.com) and V. D. Calhoun (vcalhoun@gsu.edu) are with the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS) Center, Georgia State University, Georgia Institute of Technology, Emory University, Atlanta 30303, GA, USA. J. Sui is also with the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, 100875, China.

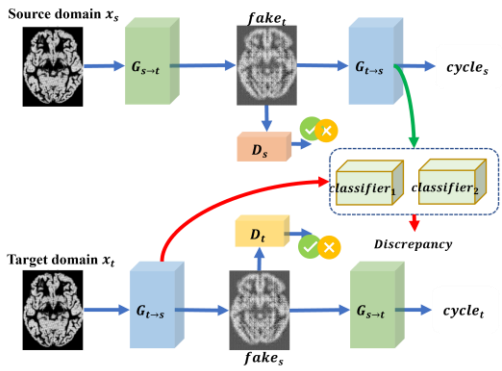

Figure 1. Framework of the proposed MCD-GAN. The model contains two generative functions $G_{s\to t}$, $G_{t\to s}$, and associated adversarial discriminators D_s and D_t . D_s encourages $G_{s\to t}$ to translate source domain, and vice versa for $D_t$and $G_{t\to s}$. Two classifiers $classifier_1$ and $classifier_2$ are trained on the source domain.

In this work, we aim to combine the advantage of cycle-GAN (data harmonization) with those of maximum discrepancy classifiers (domain adaptation) techniques and propose a new "harmless" harmonization method MCD-GAN. The proposed MCD-GAN has two main highlights: 1) harmonizing the datasets from different scanners without mapping the original features to a lower-dimensional subspace; 2) improving task-specific performance. We first systematically introduce and compare the current popular harmonizing methods including ComBat, cycle-GAN, and maximum discrepancy classifiers. Then the architecture and mechanisms of the proposed MCD-GAN were described in detail. We simulated data to visualize the mechanism of the proposed MCD-GAN. After that, the performance of MCD-GAN was validated within the Adolescent Brain Cognitive Development (ABCD) dataset.

II. Methods and Materials

A. MCD-GAN network

Fig. 1 is an overview of the proposed MCD-GAN framework which has two main modules: cycle-GAN and two classifiers. The cycle-GAN consist of two generators $G_{s\to t}$, $G_{t\to s}$, and respective adversarial discriminators D_s and D_t. D_s encourages $G_{s\to t}$ to generate data from the source domain, and vice versa for D_t and $G_{t\to s}$. Two classifiers F_1 and F_2 are then trained on the source domain. As for maximum classifier discrepancy, the unlabeled target domain samples are sent to $classifier_1$ and $classifier_2$ to get respective predicted labels. The discrepancy of $classifier_1$ and $classifier_2$ are then obtained. Two main steps are iteratively conducted to optimize the MCD-GAN model. In the first step, two classifiers are trained to maximize the discrepancy of the $G_{t\to s}$ outputs (red lines) while maintaining the classification performance on the source domain (green line). Second, the generator which maps the samples from the target domain to the source domain was optimized to minimize the discrepancy of the classifiers on the target dataset.

B. MCD-GAN loss functions

Loss functions play a key role in optimizing the deep learning model. The loss functions of MCD-GAN consist of cycle-GAN loss, classification loss, and max-discrepancy loss. Details of the losses are as follows:

Adversarial loss: We assume that cycles in both directions help perform global domain alignment by learning features in the adaptation process, and employ the following source domain loss $L_s(G_{s\to t}, G_{t\to s}, D_s)$ and the target domain loss $L_t(G_{s\to t}, G_{t\to s}, D_t)$:

$$L_s(G_{s\to t}, G_{t\to s}, D_s) = -E_{x_s\sim D^s}\log D_s(x_s) - E_{x_t\sim D^t}\log\left(1 - D_s(G_{t\to s}(x_t))\right) \quad (1)$$

$$L_t(G_{s\to t}, G_{t\to s}, D_t) = -E_{x_t\sim D^t}\log D_t(x_t) - E_{x_s\sim D^s}\log\left(1 - D_t(G_{s\to t}(x_s))\right) \quad (2)$$

where D_s and D_t are discriminators corresponding to the source and target domains. $G_{s\to t}$ is the generator mapping source features to the target domain, $G_{s\to t}$ is the generator to map target features to the source domain.

Cycle-consistency loss: The cycle consistency losses were also applied to regularize the two generators. The intuitive explanation is that if we translate from one domain to the other and back again, we should arrive at where we started. Therefore, the loss for cycle consistency is as follows:

$$L_{cyc} = E_{x_s\sim D^s}\|G_{s\to t}(G_{t\to s}(x_s) - x_s)\|_1 + E_{x_t\sim D^t}\|G_{t\to s}(G_{s\to t}(x_s) - x_t)\|_1 \quad (3)$$

Hereafter, the cycle-GAN loss is the weighted sum of the adversarial loss and cycle-consistency loss:

$$L(G_{s\to t}, G_{t\to s}, D_s, D_t) = L_s(G_{s\to t}, G_{t\to s}, D_s) + L_t(G_{s\to t}, G_{t\to s}, D_t) + \boldsymbol{\lambda} L_{cyc} \quad (4)$$

where λ is the hyperparameter to control the ratio between adversarial loss and cycle-consistency loss.

Classification loss: The classifiers are trained on source domain samples. The loss function is as follow:

$$L_{class}(x_s) = 0.5 * \frac{1}{K}\sum_{k=1}^{K} L(classifier_1(x_s^k), y_s) + 0.5 * \frac{1}{K}\sum_{k=1}^{K} L(classifier_2(x_s^k), y_s) \quad (5)$$

where L denotes the cross-entropy loss and k denote the number of classes.

Max classification discrepancy loss: Two deep learning classifiers are trained on the original datasets. Similar to Saito et al's work [5], we utilize the absolute values of the difference between two classifiers' probabilistic outputs as discrepancy loss:

$$d(classifier_1(x_s), classifier_2(x_s)) = \frac{1}{K}\sum_{k=1}^{K}|classifier_1(x_s^k) - classifier_2(x_s^k)| \quad (6)$$

where the p_{1k} and p_{2k} denote probability output of p_1 and p_2 for class k respectively.

C. MCD-GAN training steps

Step A: First, we pre-train cycle-GAN by minimizing Eq.4 to harmonize the source domain and target domain. The objective function is as follow:

$$\min_{G_{s\to t}, G_{t\to s}, D_s, D_t} L(G_{s\to t}, G_{t\to s}, D_s, D_t) \quad (7)$$

Step B: We train classifiers ($classifier_1$ and $classifier_2$) using the features generated from the cycle generator $G_{t\to s}G_{s\to t}(x_s)$. This is equivalent to training the classifiers on the source domain. The optimized classifiers are tested on the unlabeled target domain samples which are mapped to the source domain using $G_{t\to s}$. The discrepancy of the two classifiers on unlabeled target datasets are maximized by maximizing $d(classifier_1, classifier_2)$. At the same time,

TABLE I. Demographic information

	GE (Discovery)	SIEMENS (Prisma)
Cortical thickness (68 features)	2708 subjects	3072 subjects
sMRI (121*145*121 voxels)	2708 subjects	3072 subjects
Gender(F/M)	1291/1417	1431/1641
Months(mean±std)	118.2±7.6	119.3±7.5

the classification performance on the source domain should also be maintained. The objective is as follows:

$$\min_{classifier_1, classifier_2} \{L_{class}(G_{t\to s}G_{s\to t}(x_s)) - d\left(classifier_1(G_{t\to s}(x_t)), classifier_2(G_{t\to s}(x_t))\right)\} \quad (8)$$

Step C: We optimize the generator $G_{t\to s}$ to minimize the discrepancy for fixed classifiers. The term denotes the trade-off between the generator and classifiers. The objective is as follows:

$$\min_{G_{t\to s}} d\left(classifier_1(G_{t\to s}(x_t)), classifier_2(G_{t\to s}(x_t))\right) \quad (9)$$

The above three steps are repeated until convergence.

C. Model implementation

The proposed models are implemented on the platform of Tensorflow (https://www.tensorflow.org/) and ScikitLearn (https://scikit-learn.org/). Adam was used as the optimizer with an initial learning rate of 10^{-4}. The batch size was set to 4. The parameter λ was empirically set to 10. All the above models were implemented on the cluster (Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz, 20 CPU cores) with a GPU card (Tesla V100-SXM2-32GB).

D. Data and preprocessing

Simulated data (Double moon): As shown in Fig. 2(a). Double moon data is simulated using Sklearn (https://scikit-learn.org/stable/). The dataset contains 2000 samples from two sites (site 1 and site 2). Each site consists of two categories (class 1 and class 2). Site2 is obtained by counterclockwise rotating site 1 by 45 degrees.

ABCD MRI cortical thickness features: As shown in Table 2, cortical thickness features are included in this work. The features are downloaded from the ABCD (https://abcdstudy.org/) official website.

ABCD MRI volumes: As shown in Table I, T1 MRI volumes collected using GE and SIEMENS scanners are included. The sMRI data were segmented into tissue probability maps for gray matter, white matter, and cerebral spinal fluid using SPM12. The gray matter images were then warped to standard space, modulated, and smoothed using a Gaussian kernel with an FWHM = 10 mm. The preprocessed gray matter volume images had a dimensionality of 121 × 145 × 121 in the voxel space, with the voxel size of 1.5 × 1.5 × 1.5 mm^3.

III. Results

As shown in Fig. 2, the performance of the proposed MCD-GAN was compared with ComBat and cycle-GAN using a simulated 'double moon' dataset.

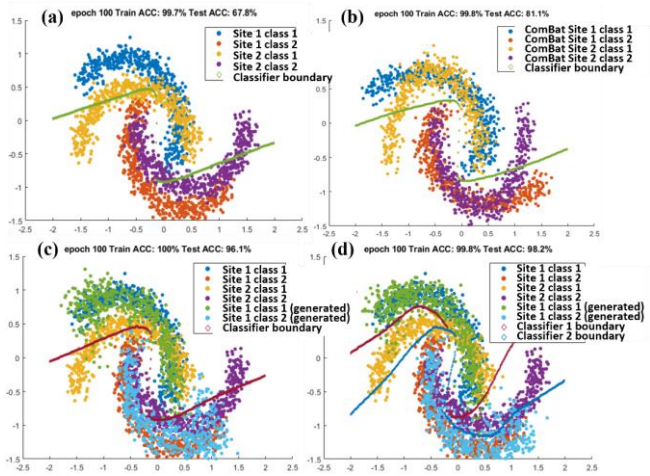

Figure 2. (a) Double moon simulated dataset and the classifier's boundary (green curves) before harmonization. Two sites of data (site 1 and site 2) are simulated. Each site consists of two classes (class 1 and class 2). Site2 is obtained by counterclockwise rotating Site1. (b) After ComBat harmonization; (c) After cycle-GAN harmonization; (d) After MCD-GAN harmonization. *Notes: "Train ACC" represents the classification performance on training dataset (source domain). "Test ACC" represents the classification performance on testing dataset (target domain).*

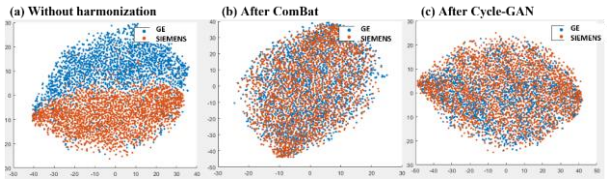

Figure 3. Visualization and comparison of data harmonization method on ABCD cortical thickness features. (a) before harmonization. (b) after ComBat; (c) after cycle-GAN.

Before harmonization, the classifier trained on Site 1 (accuracy=99.7%) cannot perform well on Site 2 (accuracy=67.8%). ComBat approach changes the scale of both source and target domains for harmonization. However, the ComBat cannot harmonize the non-linear confounds. Cycle-GAN performed better on harmonizing non-linear site effects than ComBat. MCD-GAN performed even better than cycle-GAN by constraining the samples which are not close to decision boundaries.

Fig. 3 shows the effect of harmonizing methods. Cortical thickness features of each ABCD subject were used for tSNE unsupervised clustering and visualization. Before harmonization, the samples from GE and SIEMENS can be clustered into two groups with a clear boundary. After applying harmonization methods (ComBat and cycle-GAN) to the cortical thickness features, the site effects are removed. In addition, compared to ComBat results, the scale of the harmonized features obtained by cycle-GAN is more similar to the original features. In other words, the cycle-GAN method can better maintain features within the original feature space.

TABLE II. Comparison of methods on datasets.

	Double moon simulated data		ABCD cortical thickness		ABCD 3D MRI	
	Train	Test	Train	Test	Train	Test
No harmony	99.7%	67.8%	67.7%	63.2%	99.2%	67.5%
ComBat	99.8%	81.1%	67.2%	65.7%	98.4%	86.0%
Cycle-GAN	100%	96.1%	66.5%	65.7%	98.4%	86.6%
MCD-GAN	99.8%	**98.2%**	66.8%	**66.0%**	98.2%	**87.1%**

Table II shows the classification performance on the simulated double moon and ABCD dataset. The results show that the proposed MCD-GAN outperforms the compared state-of-the-art harmonization methods on the cross-site gender-classification task.

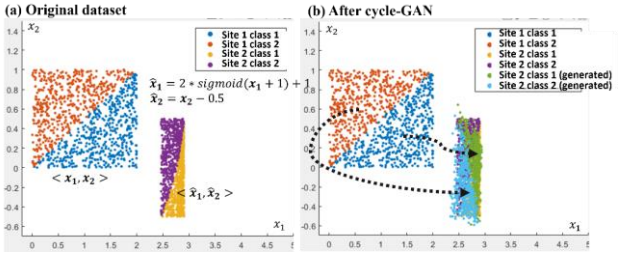

Figure 4. Simulated dataset and cycle-GAN. (a) The simulated "GE/SIEMENS dataset before harmonization. The SIEMENS samples are generated from GE samples with a non-linear transformation. (b) The learned mappings from GE domain to SIEMENS domain. In this example, the harmonized dataset will not improve the cross-site classification performance because of the rotation problem.

IV. Discussion and Conclusion

Properly harmonizing site-effected confounds is vital in achieving reproducible results in multi-site studies especially when data are collected from different sites. Conventional data harmonization methods cannot guarantee improved performance on specific tasks. Our proposed MCD-GAN has advantages in two aspects: 1) compared to conventional data harmonization methods, MCD-GAN can guarantee improved performance in specific classification tasks; 2) compared to conventional domain adaptation methods, MCD-GAN does not map the features into a low-dimensional subspace but keeps the features in their original space.

Data harmonization and domain adaptation are two main categories of methods for removing confounds. Data harmonization methods focus on estimating the distribution of the confounds and then using specific algorithms to remove the confounds from the original features. The confound-removed dataset can be used for a series of subsequent analyses. Domain adaptation methods focus on mapping the original features into a common specific feature space. Therefore, if distributions of the confounds are clear or easy to be estimated, removing the confounds using the data harmonization method should be optimal because it does not affect the downstream analysis, however, since most of the actual confounds are non-linear and difficult to estimate, domain adaptation methods should be a better choice because it can guarantee the improvement of task-specific performance.

The non-linear characteristic of deep learning often makes it not intuitive for interpretation, especially when the input features are high-dimensional. Simulated 2D samples are helpful to simplify the problem. To further explain the reason why data harmonization cannot guarantee improved performance on downstream tasks, we simulated two sites (GE and SIEMENS) of samples. As shown in Fig. 4(a), site 2 can be generated from site 1 with a non-linear transformation. A cycle-GAN model was trained to map the data from GE to the SIEMENS site. The result demonstrates that cycle-GAN can accurately learn the distribution mappings between two sites. However, it cannot guarantee improving performance on subsequent classification tasks because of the rotation problem. This experiment further illustrates the limitations of conventional data harmonization approaches.

In summary, we propose MCD-GAN, which combines the advantages of both generative model and maximum discrepancy classifier approaches, for harmonizing the confounds while training the classifiers to improve the cross-site/scanner classification performance. The performance of MCD-GAN is compared with conventional methods, such as ComBat and Cycle-GAN, on a simulated and ABCD MRI dataset. The result demonstrates the superiority of the proposed method. The mechanisms of MCD-GAN are visualized by applying them to the simulated dataset. The proposed MCD-GAN is promising to facilitate the cross-site reproducibility of neuroimaging studies.

References

[1] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," *NeuroImage,* vol. 180, pp. 68-77, 2018/10/15/, 2018.

[2] J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, M. McInnis, M. L. Phillips, M. H. Trivedi, M. M. Weissman, and R. T. Shinohara, "Harmonization of cortical thickness measurements across scanners and sites," *NeuroImage,* vol. 167, pp. 104-120, 2018/02/15/, 2018.

[3] R. Garcia-Dias, C. Scarpazza, L. Baecker, S. Vieira, W. H. L. Pinaya, A. Corvin, A. Redolfi, B. Nelson, B. Crespo-Facorro, C. McDonald, D. Tordesillas-Gutiérrez, D. Cannon, D. Mothersill, D. Hernaus, D. Morris, E. Setien-Suero, G. Donohoe, G. Frisoni, G. Tronchin, J. Sato, M. Marcelis, M. Kempton, N. E. M. van Haren, O. Gruber, P. McGorry, P. Amminger, P. McGuire, Q. Gong, R. S. Kahn, R. Ayesa-Arriola, T. van Amelsvoort, V. Ortiz-García de la Foz, V. Calhoun, W. Cahn, and A. Mechelli, "Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners," *NeuroImage,* vol. 220, pp. 117127, 2020/10/15/, 2020.

[4] H. Guan, Y. Liu, E. Yang, P.-T. Yap, D. Shen, and M. Liu, "Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification," *Medical Image Analysis,* vol. 71, pp. 102076, 2021/07/01/, 2021.

[5] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation." pp. 3723-3732.

[6] V. M. Bashyam, J. Doshi, G. Erus, D. Srinivasan, A. Abdulkadir, M. Habes, Y. Fan, C. L. Masters, P. Maruff, and C. Zhuo, "Medical Image Harmonization Using Deep Learning Based Canonical Mapping: Toward Robust and Generalizable Learning in Imaging," *arXiv preprint,* vol. arXiv:1409.0473, 2020.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks." pp. 2223-2232.