2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782176

CGDM-GAN: An Adversarial Network Approach with Self-supervised Learning for Site Effect Removal

1st Xiangxiang Cui
The State Key Lab of Cognitive Neuroscience and Learning
Beijing Normal University
Beijing, China
abner.cxx@mail.bnu.edu.cn

2nd Dongmei Zhi
The State Key Lab of Cognitive Neuroscience and Learning
Beijing Normal University
Beijing, China
dmzhi@bnu.edu.cn

3rd Weizheng Yan
National Institute on Alcohol Abuse and Alcoholism, Lab of Neuroimaging
National Institutes of Health
Bethesda, United States
weizheng.yan@nih.gov

4th Vince D. Calhoun
Tri-Institutional Center for Translational Research in Neuroimaging and Data Science
Georgia State University,
Atlanta, United States
vince.calhoun@ece.gatech.edu

5th Chuanjun Zhuo
The Department of Psychiatric-Neuroimaging-Genetics and Morbidity Laboratory (PNGC-Lab), Tianjin Mental Health Center
Nankai University
Tianjin, China
chuanjunzhuotjmg@163.com

6th Jing Sui
IDG/McGovern Institute for Brain Research, State Key Laboratory of Cognitive Neuroscience and Learning,
Beijing Normal University,
Beijing, China
jsui@bnu.edu.cn

***Abstract*—Imaging data collected from different sites is difficult to pool together due to unwarranted variations introduced by different acquisition protocols or scanners. Data harmonization is an effective way to mitigate site-specific bias while preserving the intrinsic image properties, thereby increasing the sample size and enhancing the generalization of models. Although various harmonization methods exist, their performance on specific tasks is often unsatisfactory. Here, we proposed a novel approach, CGDM-GAN, by combining the advantages of generative models, maximum discrepancy theory, and gradient discrepancy minimization with self-supervised learning to harmonize site effects and improve cross-site classification performance. The proposed CGDM-GAN was successfully conducted on synthetic dataset, and further validated on in-house and ABCD datasets, outperforming three data harmonization methods, including ComBat, CycleGAN, and MCD-GAN, suggesting its potential for removing site effects and improving cross-site neuroimaging classification.**

Keywords—Harmonization, Maximum Classifier Discrepancy, Generative Adversarial Network，Self-supervised Learning

I. Introduction

Multi-site neuroimaging collaboration is a powerful strategy for overcoming the small-sample problem . However, magnetic resonance imaging (MRI) acquired using different scanners, protocols, head motion and recruitment disparities may introduce significant heterogeneity, thereby reducing the accuracy and reproducibility across studies [1-3]. However, Meta-regression studies have shown that the accuracy of pooled classification performances is significantly affected by sample size [1]. Consequently, it is essential to remove task-irrelevant confounds to improve the outcomes of large-cohort studies.

Non-biological confounds tend to have unpredictable distributions, making their proper removal challenging. To address this issue, researchers have developed a variety of harmonization methods designed to mitigate the effects of non-biological confounds, which can be categorized into dataset harmonization and domain adaptation. Dataset harmonization approaches, such as ComBat [4], are commonly used to handle non-biological variance. Domain adaptation trains task-specific models while harmonizing the feature by mapping the features into a shared task-specific subspace [5-7].

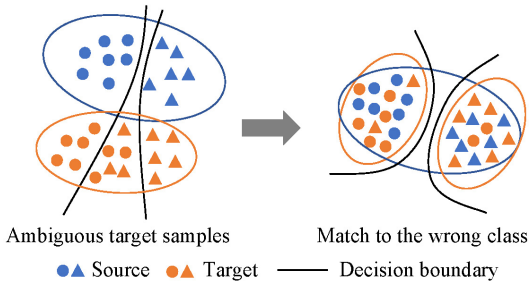

Fig 1. Previous methods only focus on classifier discrepancy and overlook the accuracy of target samples.

The deep generative adversarial model with the superiority of adversarial training strategy has been successfully applied to both the data harmonization and domain adaptation fields [6, 8]. For example, [8] proposed CycleGAN to solve style transfer problems and [9] proposed MCD-GAN by combining CycleGAN (data harmonization) with maximum classifiers discrepancy (domain adaptation) to harmonize the datasets

This work was supported by National key research and development program (2021YFE0202500), 2030 BSBIIT program (2021ZD0200500)，NSF of China (62373062,82022035) and China Postdoctoral Foundation (2022M710434).

from different scanners without mapping the original features to a lower-dimensional subspace. However, maximum classifiers discrepancy [7] only focused on the similarity of the outputs between two distinct classifiers and cannot guarantee the accuracy of target labels, resulting in less robust and discriminative features. Therefore, the cross-domain gradient discrepancy minimization (CGDM) approach can be a remedy by incorporating self-supervised labels into models, which can enhance the classification accuracy of the target domain label [4].

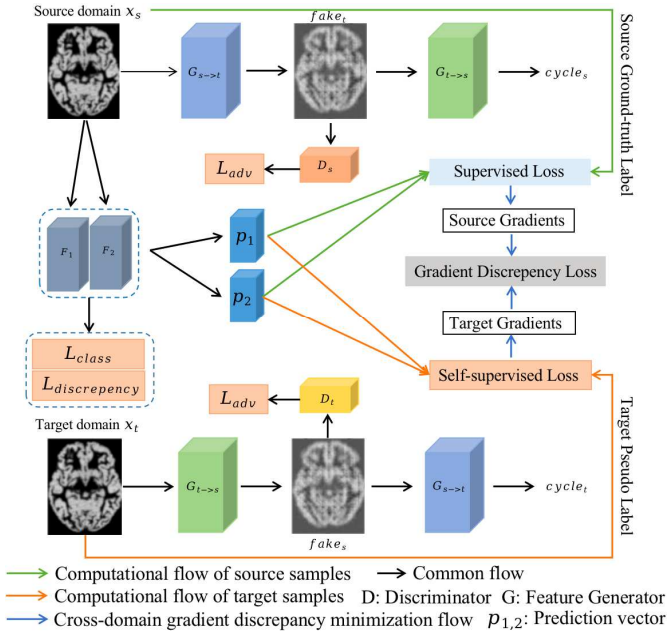

Fig 2. The framework of the proposed CGDM-GAN. 1) MCD and CycleGAN (black line) for data harmonization and domain adaptation. 2) Source samples supervised (green line) and Target samples self-supervised (orange line) for optimizing $classifier_1$ and $classifier_2$. 3) Cross-domain gradient discrepancy minimization (blue line) for improving the accuracy of the target domain.

In this study, we propose a CGDM-GAN harmonization method, which combines a gradient difference minimization method to explicitly reduce the discrepancy between the gradient vectors generated from source and target domain samples. Pseudo-labels are used for self-supervised learning to reduce ambiguous target samples and achieve accurate class-level distribution alignment through gradient vector alignment. The proposed CGDM-GAN has three main highlights: 1) The gradient discrepancy minimization is utilized as a supervised signal to improve the accuracy of the labels for samples in the target domain. 2) Improving task-specific performance across multiple neuroimaging datasets. The proposed CGDM-GAN outperforms current state-of-the-art harmonization methods using both the Adolescent Brain Cognitive Development (ABCD) dataset and our in-house datasets.

II. Methods and Materials

A. CGDM-GAN network

As shown in Fig. 2, CGDM-GAN framework proposed consists of three modules: CycleGAN for data harmonization, Maximum Classifier Discrepancy (MCD) for domain adaptation, and Cross-domain Gradient Difference Minimization (CGDM) for improved accuracy in the target domain. This architecture builds upon our previous integration MCDGAN (As shown in the black line in Fig. 2) [9] of CycleGAN and MCD, introducing CGDM to enhance the removal of site effects. Specifically, CycleGAN is primarily designed for style transfer between different domains, ensuring the transformation maintains stylistic coherence. The MCD leverages two classifiers to analyze the target domain, aiming to maximize their prediction differences. This discrepancy is then minimized by training the generator. CGDM aligns gradient across source and target domains, enhancing domain adaptation by precisely minimizing gradient discrepancies. This approach significantly improves target domain accuracy by ensuring that the learning process for both domains is closely harmonized, thus facilitating a more effective generalization to the target domain. Our training strategy employs a phased approach, initially leveraging CycleGAN and MCD to improve classification in the target domain. CGDM is introduced in later stages, optimizing its effectiveness once pseudo-label accuracy is enhanced.

B. CGDM-GAN loss functions

Adversarial loss: The adaptation process helps in domain alignment by learning features in both directions, and employ the following definition:

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

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

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

Cycle consistency loss: The cycle consistency loss was also applied to regularize the two generators. The loss for cycle consistency is as follows:

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

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

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

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 follows:

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

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

Max classification discrepancy loss: We utilizes the absolute value of the probabilistic output differences between two deep learning classifiers as the basis for calculating discrepancy loss:

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

where the $classifier_1(x_s^k)$ and $classifier_2(x_s^k)$ denote probability output of $classifier_1(x_s)$ and $classifier_2(x_s)$ for class k respectively.

Identify mapping loss: To ensure that the identity of the target domain is the same as the source domain and further stabilize the training procedure, as used in [10], we require the generator to be the identity mapping if the real samples of the target domain are provided as the input to the generator. The loss function is as follows:

$$L_{identity}(G_{s\to t}, G_{t\to s}) = E_{x_s\sim D^{x_s}} \| G_{s\to t}(x_s) - x_s \|_1 + E_{x_t\sim D^{x_t}} \| G_{t\to s}(x_t) - x_t \|_1 \tag{7}$$

The identity mapping loss acts as an effective stabilizer at the early stage of training.

Gradient Discrepancy loss: Inspired by [5], to learn a classifier that can correctly classify all samples from both domains, the gradient vectors produced by source and target samples should be similar. We denote expected gradients over source and target examples as g_s and g_t respectively, and formulate appropriate gradient for source samples as follows:

$$g_s = \frac{1}{2}\sum_{n=1}^{2} \underset{(x_i^s, y_i^s)\sim(x^s, y^s)}{E} [\nabla_{\theta_{f_n}} L_{ce}(F_n(G(x_i^s)), y_i^s)] \tag{8}$$

Where the $E, F, n, x^s, y^s, x_i^s, y_i^s$ and $\nabla_{\theta_{f_n}}$ denote expecteds value, classifier, classifier number, source training sets, training sets label, i^{th} source training sample, i^{th} source training sample label and derivative of f_n concerning θ respectively. For computing the gradient generated by the target samples, we assign pseudo-labels to target samples, denoted by y^* To mitigate the impact of potentially incorrect pseudo-labels on ambiguous target samples. The formulation for the gradient vector of the target samples is as follows:

$$g_t = \frac{1}{2}\sum_{n=1}^{2} \underset{(x_i^t, y_i^*)\sim(x^t, y^*)}{E} [\nabla_{\theta_{f_n}} L_{ce}^{w}(F_n(G(x_i^t)), y_i^*)] \tag{9}$$

Here L_{ce}^{w} is the weighted cross entropy loss function which is formulated as follows:

$$L_{ce}^{w}(F_n(G(x_i^t)), y_i^*) = w_j(x_i^t) L_{ce}(F_n(G(x_i^t)), y_i^*), \quad w_j(x_i^t) = 1 + e^{-E(\delta(F_n(G(x_i^t)), y_i^*))} \tag{10}$$

where δ represents the softmax output and E denotes the standard information entropy. Currently, we have obtained gradient vectors for both source and target samples. To align the domain distribution, we formulate the discrepancy as follows:

$$L_{GD} = 1 - \frac{g_s^T g_t}{\| g_s \|_2 \| g_t \|_2} \tag{11}$$

L_{GD} employs cosine similarity to capture the discrepancy between source and target domains, while also incorporating semantic information for distribution alignment.

Pseudo-label classification loss: We use softmax outputs from two classifiers to assign pseudo-label y_i^* for each target sample i . To improve model classification performance through self-supervised learning, we utilize pseudo-labeled target samples to encourage correct decision boundary towards the discriminative target distribution. The weighted classification loss $L_{cls}^{W}(x^t, y^*)$ for self-supervised learning can be formulated as follows:

$$L_{cls}^{W}(x^t, y^*) = \frac{1}{2n_t}\sum_{i=1}^{n_t}\sum_{n=1}^{2} L_{ce}^{W}(F_n(G(x_i^t)), y_i^*) \tag{12}$$

Therefore, by improving the discriminability of the target distribution, we can further align samples between two domains at a category level.

C. CGDM-GAN training steps

Step-1: First, to pre-train CycleGAN. Specifically, train the generator $(G_{s\to t}, G_{t\to s})$ and discriminator $(D_{s\to t}, D_{t\to s})$ on source domain X_s , target domain X_t and fake data X_{fake}^t generated by $G_{s\to t}$ compute adversarial loss, and cycle loss, and identify loss. Update $G_{s\to t}, G_{t\to s}, D_{s\to t}, D_{t\to s}$ parameters. All the above processes are conducted through the Eq. (1) ~ (4) and Eq. (7).

Step-2: Second, to train two classifiers (F_1 and F_2). Specifically, train classifier (F_1, F_2) on X_{same}^s generated by $G_{t\to s}$ on X^s . 2) Get the pseudo-label of the target sample through two classifiers. 3) Compute classification loss, and max classification discrepancy loss through Eq. (5), Eq. (6), and Eq. (12). 4) Update classifier (F_1, F_2) parameters.

Step-3: Third, use self-supervised learning and cross-domain gradient discrepancy minimization to optimal generator model. Specifically, 1) Compute gradient discrepancy loss, max classification discrepancy loss, and adversarial loss through Eq. (8) ~ (11), Eq. (1), Eq. (2), and Eq. (6). 2) Update $G_{t\to s}$ parameters.

D. Data and preprocessing

ABCD MRI cortical thickness features: In GE, the mean months are 118.2±7.6., there are 1417 Male and 1,291 Female subjects. In SIEMENS, the mean months are 119.3±7.5, there are 1641 Male and 1431 Female subjects.

In-house Dataset: The dataset is from two sites. Site 1 has 158 schizophrenia patients and 176 normal controls. Site 2 has 69 schizophrenia patients and 49 normal controls.

ABCD MRI volumes: T1 MRI volumes collected using GE and SIEMENS scanners. In GE scanner, there are 1362 Male and 1,281 Female subjects. In SIEMENS scanner, there are 1585 Male and 1370 Female subjects. 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.

Simulated data (Double moon): The dataset comprises 2000 samples, with 1000 subjects per site (Site1 and Site2). Each site consists of two categories. Site2 is derived by rotating Site1 counterclockwise by an angle of 45 degrees.

III. Resutls

As shown in Fig. 3(a), before harmonization, the distribution of site1 and site2 are decentralized based on t-SNE visualization of double moon simulated data. However, Fig. 3(b) shows that the two sites are well harmonized after the proposed CGDM-GAN harmonization.

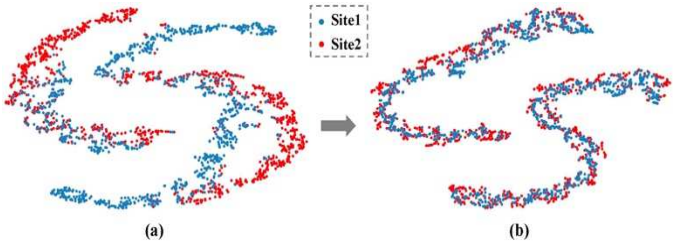

Fig 3. (a) Before harmonizing the double moon simulated data. (b) After harmonizing the double moon simulated data using the proposed CGDM-GAN.

As shown in TABLE I, the performance of the proposed CGDM-GAN was compared with ComBat and CycleGAN, MCD-GAN using multiple datasets. Table I shows the classification performance on the simulated double moon, In-house, and ABCD datasets. The results show that the proposed CGDM-GAN outperforms the three compared state-of-the-art harmonization methods on the cross-site classification task.

TABLE I. Comparison of methods on datasets.

Data		No harmony	ComBat	Cycle GAN	MCD-GAN	**CGDM-GAN**
Cortical thickness	Train	67.7%	67.2%	66.5%	66.8%	67.5%
	Test	63.2%	65.7%	65.7%	66.0%	**67.1%**
In-house	Train	99.1%	98.3%	99.1%	97.4%	98.3%
	Test	72.5%	75.1%	75.2%	76.0%	**77.2%**
sMRI	Train	99.2%	98.4%	98.4%	98.2%	98.3%
	Test	67.5%	86%	86.6%	87.1%	**88.2%**
Simulated data	Train	99.7%	99.8%	100%	99.8%	99.8%
	Test	67.8%	81.1%	96.1%	98.2%	**98.8%**

IV. Discussion

Properly addressing site-related confounds is crucial for achieving reproducible results in multi-site studies. However, conventional methods of data harmonization do not incorporate gradient discrepancy minimization with pseudo-labels to enhance the accuracys of the target domain label, resulting in poor performance. Our proposed CGDM-GAN has advantages in three aspects: 1) The gradient discrepancy minimization is utilized as a supervised signal to improve the accuracy of the labels for samples in the target domain. 2) Harmonizing the datasets from different scanners without mapping the original features to a lower-dimensional subspace. 3) Improving task-specific performance across multiple neuroimaging datasets.

Obtaining pseudo-labels is a critical issue. In our research, we attempted to use clustering for obtaining pseudo-labels. However, the visualization results indicate that clustering labels decreases the performance of data harmonization across classes. Therefore, we opt to utilize the results from dual classifiers for obtaining pseudo-labels. Additionally, the training strategy of CGDM-GAN requires attention, and the optimal training scheme suggests initiating supervised learning of pseudo-labels and backpropagation of gradient similarity loss in the middle and late stages of training, which will enhance the effectiveness of the training.

V. Conclusions

In summary, we propose a novel approach, CGDM-GAN, which leverages the strengths of generative model, maximum discrepancy classifier, and gradient discrepancy minimization with self-supervised learning, for harmonizing the confounds while training the classifiers to improve the cross-site/scanner classification performance. The proposed approach outperforms ComBat, CycleGAN, and MCD-GAN, on a simulated, an In-house and ABCD datasets, demonstrating its superiority of the cross-site reproducibility of neuroimaging findings.

References

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

[2] R. Jiang, C. C. Abbott, T. Jiang, Y. Du, R. Espinoza, K. L. Narr, B. Wade, Q. Yu, M. Song, and D. Lin, "SMRI biomarkers predict electroconvulsive treatment outcomes: accuracy with independent data sets," *Neuropsychopharmacology,* vol. 43, no. 5, pp. 1078-1087, 2018.

[3] D. Yao, V. D. Calhoun, Z. Fu, Y. Du, and J. Sui, "An ensemble learning system for a 4-way classification of Alzheimer's disease and mild cognitive impairment," *Journal of neuroscience methods,* vol. 302, pp. 75-81, 2018.

[4] 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, Feb 15, 2018.

[5] Z. K. Du, J. J. Li, H. Z. Su, L. Zhu, and K. Lu, "Cross-Domain Gradient Discrepancy Minimization for Unsupervised Domain Adaptation," *2021 Ieee/Cvf Conference on Computer Vision and Pattern Recognition, Cvpr 2021*, pp. 3936-3945, 2021.

[6] 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," *Med Image Anal,* vol. 71, pp. 102076, Jul, 2021.

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

[8] 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 arXiv:2010.05355*, 2020.

[9] W. Yan, Z. Fu, J. Sui, and V. D. Calhoun, "'Harmless' adversarial network harmonization approach for removing site effects and improving reproducibility in neuroimaging studies," *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2022, pp. 1859-1862, Jul, 2022.

[10] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *arXiv preprint arXiv:1611.02200*, 2016.