



Unsupervised Conditional Consensus Adversarial Network for Brain Disease Identification with Structural MRI

Jing Zhang^{1,2}, Mingxia Liu^{2(✉)}, Yongsheng Pan^{2,3}, and Dinggang Shen^{2(✉)}

¹ Department of Mathematics, Zhejiang A & F University, Hangzhou 311300, China

² Department of Radiology and BRIC,
University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
{mxliu, dgshen}@med.unc.edu

³ School of Computer Science and Engineering,
Northwestern Polytechnical University, Xi'an 710072, China

Abstract. Effective utilization of multi-domain data for brain disease identification has recently attracted increasing attention since a large number of subjects from multiple domains could be beneficial for investigating the pathological changes of disease-affected brains. Previous machine learning methods often suffer from inter-domain data heterogeneity caused by different scanning parameters. Although several deep learning methods have been developed, they usually assume that the source classifier can be directly transferred to the target (i.e., to-be-analyzed) domain upon the learned domain-invariant features, thus ignoring the shift in data distributions across different domains. Also, most of them rely on fully-labeled data in both target and source domains for model training, while labeled target data are generally unavailable. To this end, we present an Unsupervised Conditional consensus Adversarial Network (UCAN) for deep domain adaptation, which can learn the disease classifier from the labeled source domain and adapt to a different target domain (without any label information). The UCAN model contains three major components: (1) a *feature extraction module* for learning discriminate representations from the input MRI, (2) a *cycle feature adaptation module* to assist feature and classifier adaptation between the source and target domains, and (3) a *classification module* for disease identification. Experimental results on 1,506 subjects from ADNI1 (with 1.5 T structural MRI) and ADNI2 (with 3.0 T structural MRI) have demonstrated the effectiveness of the proposed UCAN method in brain disease identification, compared with state-of-the-art approaches.

1 Introduction

Alzheimer's disease (AD) is a slow fatal neurodegenerative disease affecting people over the age of 65 years. Thus, the identification and analysis of AD and its prodromal phase, *i.e.*, mild cognitive impairment (MCI), are essential for early treatment and possible delays in disease progression. Significant advances in neuroimaging have provided opportunities to study brain-related diseases, bringing

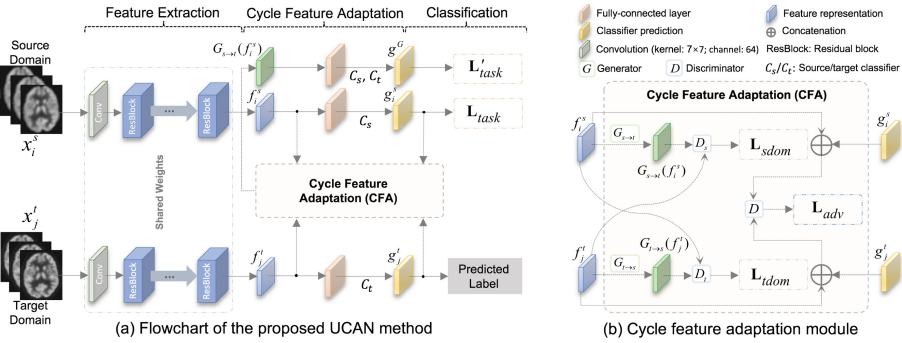


Fig. 1. Illustration of the proposed Unsupervised Conditional consensus Adversarial Network (UCAN), including (1) a *feature extraction* module with a convolutional (Conv) layer and four type of residual blocks as the backbone, (2) a *cycle feature adaptation* (CFA) module to harmonize learned features and classifiers of the source and target domains, and (3) a *classification* module. The input are a pair of images, *i.e.*, x_i^s from the labeled source domain and x_j^t from the unlabeled target domain (with learned features as f_i^s and f_j^t , respectively). Here, $g^G = C_s(G_{s \rightarrow t}(f_i^s)) + C_t(G_{s \rightarrow t}(f_i^s))$.

improvements in early detection of AD and MCI. Structural magnetic resonance imaging (MRI) is widely used in AD studies because it provides an excellent spatial resolution and a non-invasive method to study a patient’s brain. MRI-based methods help predict the progression of MCI to AD, by describing brain atrophy and change in the size of brain tissues explicitly [1, 2].

Conventional MRI-based methods often suffer from inter-domain data heterogeneity caused by different scanning parameters, because models learned on source domains are directly applied to problems in the to-be-analyzed target domain [3–5]. For example, Cheng *et al.* [3] proposed the sharing domain transfer learning method for MCI conversion prediction, which directly use the source domain to be auxiliary for the target domain to select a subset of common features. Several deep learning methods have been recently developed to alleviate the issue of domain shift [6–10]. For example, Motiiian *et al.* [6] provided a unified framework to learn an embedding subspace for addressing the problem of supervised domain adaptation. A deep domain confusion network [7] was developed to map data from both domains into a common feature space to reduce the domain shift, measured by the maximum mean discrepancy (MMD). This method was further extended to a deep adaptation network (DAN) via a multi-layer multi-kernel selection technique [8] and a joint adaptation network (JAN) [9] for natural image classification. Long *et al.* [10] presented a conditional adversarial domain adaptation network (CDAN) by integrating adversarial learning and domain adaptation to a unified framework.

However, existing methods typically suffer from two limitations. (1) They are usually based on the assumption that the source classifier and target classifier can be shared directly. Unfortunately, such an assumption is too strong and could not always hold in practical applications. (2) Many of them are supervised,

relying on fully-labeled data in both target and source domains, and hence, they cannot be applied to general problems, where there are labeled data in the source domain and unlabeled data in the target domain.

As shown in Fig. 1, we propose an Unsupervised Conditional consensus Adversarial Network (UCAN) for brain disease identification, where classifiers learned from labeled source domain can be robustly adapt to a different target domain (without any label information). Three major components are included: (1) a *feature extraction module* for learning discriminate representations from the input MRI, (2) a *cycle feature adaptation module* to harmonize learned MRI features and classifiers of the source and target domains, and (3) a *classification module*. Experimental results on two public datasets with structural MRIs (acquired using different scanning parameters) suggest the efficacy of our method in both tasks of AD/MCI identification and MCI conversion prediction.

2 Materials and Method

2.1 Subjects and Structural MR Image Pre-processing

Two datasets from the ADNI database [11] were employed in this work, including ADNI1 and ADNI2. Since several subjects participated in both ADNI1 and ADNI2, we remove these subjects from ADNI2 to ensure that these two datasets are independent. Subjects in these datasets were divided into four categories: (1) AD, (2) cognitively normal (CN), (3) progressive MCI (pMCI) that would progress to AD within 36 months after baseline, and (4) static MCI (sMCI) that would not progress to AD. The baseline ADNI1 dataset consists of 1.5 T T1-weighted MR images acquired from a total of 785 subjects, including 231 NC, 246 sMCI, 103 pMCI, and 205 AD subjects. The baseline ADNI2 dataset includes 3.0 T T1-weighted sMRI data acquired from 721 subjects, including 205 NC, 312 sMCI, 42 pMCI, and 162 AD subjects. A standard pipeline was used to pre-process structural MR images, including (1) anterior commissure (AC)-posterior commissure (PC) alignment; (2) skull stripping; (3) intensity correction; (4) cerebellum removal; (5) linear alignment to the Colin27 template [12]; (6) re-sampling all MR images to have the same size of $142 \times 142 \times 178 \text{ mm}^3$ (with a spatial resolution of $1 \times 1 \times 1 \text{ mm}^3$) corresponding to the coronal-plane view, sagittal-plane view, and axial-plane view, respectively; and (7) intensity inhomogeneity correction using the N3 algorithm [13].

2.2 Unsupervised Conditional Consensus Adversarial Network

We now present our conditional consensus adversarial domain adaptation framework (see Fig. 1). We consider the problem of unsupervised adaptation in this work. Denote x_i^s (with its label y_i^s) and x_j^t (without label) as the i -th and the j -th subjects from the source and target domains, respectively. Given n_s subjects from the source domain and n_t subjects from the target domain, we represent the labeled source domain as $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ and the unlabeled target domain as $\mathcal{D}_t = \{x_j^t\}_{j=1}^{n_t}$. The goal is to learn a model that can correctly predict the label of subjects from the target domain, based on labeled source data.

Feature Extraction Module: Each input image is first fed into a convolution (Conv) layer with the kernel/filter size of 7×7 and the channel size of 64 (stride: 2), followed by a max pooling with 3×3 filter (stride: 2). Then, we use the ResNet-50 model [14] to extract MRI features of images from both the source and target domains, including four types of residual blocks (with each block containing 3 Conv layers). The numbers of blocks are 3, 4, 6, and 3 for four types of residual blocks, respectively. And the different parameters (*i.e.*, [filters, channels]) for these four types of blocks are listed below:

$$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix}, \begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix}, \begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix}, \begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix}.$$

Cycle Feature Adaptation Module: Using learned features (*i.e.*, f_i^s for x_i^s and f_j^t for x_j^t) via the feature extraction module, we further perform domain adaptation to harmonize data from the source and target domains. Let C_s and C_t be the task-specific classification models trained on the source and target domains, respectively. We can begin by simply learning a source classifier C_s with the cross-entropy loss as follows

$$\mathcal{L}_{task}(C_s, x_i^s, y_i^s) = \mathbb{E}_{(x_i^s, y_i^s) \sim \mathcal{D}^s} \frac{1}{n_s} \sum_{i=1}^{n_s} L(C_s(x_i^s), y_i^s), \quad (1)$$

where $L(\cdot, \cdot)$ is the cross-entropy function.

As shown in Fig. 1 (b), our proposed cycle feature adaptation module has a symmetric network design, based on which we augment the cycle generative adversarial structure to promote the efficacy of feature adaptation. By directly mapping the source feature (*e.g.*, f_i^s) to the target domain via a generator $G_{s \rightarrow t}$ and mapping the target feature (*e.g.*, f_j^t) to the source domain via a generator $G_{t \rightarrow s}$, we aim to remove the low-level differences between features from two domains, ensuring that our learned model is well-conditioned on target domain alignment. We assume that having cycles in both directions helps perform *global domain alignment* by learning features in the adaptation process, and employ the following source domain loss \mathcal{L}_{sdom} and the target domain loss \mathcal{L}_{tdom} :

$$\mathcal{L}_{sdom}(D_s, f_i^s, f_j^t) = -\mathbb{E}_{x_i^s \sim \mathcal{D}^s} \log D_s(f_i^s) - \mathbb{E}_{x_j^t \sim \mathcal{D}^t} \log(1 - D_s(G_{t \rightarrow s}(f_j^t))), \quad (2)$$

$$\mathcal{L}_{tdom}(D_t, f_i^s, f_j^t) = -\mathbb{E}_{x_j^t \sim \mathcal{D}^t} \log D_t(f_j^t) - \mathbb{E}_{x_i^s \sim \mathcal{D}^s} \log(1 - D_t(G_{s \rightarrow t}(f_i^s))), \quad (3)$$

where D_s and D_t denote the discriminators corresponding to the source and target domains, respectively. And $G_{s \rightarrow t}$ denotes the generator to map source features to the target domain, while $G_{t \rightarrow s}$ is the generator to map target features to the source domain.

To distinguish features from different domains, we first concatenate the feature representation (*e.g.*, f_i^s) and classifier prediction (*e.g.*, g_i^s) from each domain.

Such a concatenation operation could capture the inherent relationship of task-specific features and classifier prediction results, formulated as follows:

$$h_i^s = f_i^s \oplus g_i^s, \quad h_j^t = f_j^t \oplus g_j^t.$$

We then employ the above inherent relationship between features and classifier prediction as discriminative representation conditioning on the adversarial learning, encouraging the distributions of source data and target data to be similar. Thus, the adversarial loss based on the discriminative representations (*e.g.*, h_i^s and h_j^t) of the source and target domains can be formulated as follows:

$$\mathcal{L}_{adv}(D, h_i^s, h_j^t) = -\mathbb{E}_{x_i^s \sim \mathcal{D}^s} \log D(h_i^s) - \mathbb{E}_{x_j^t \sim \mathcal{D}^t} \log(1 - D(h_j^t)), \quad (4)$$

where D is the discriminator to tell the domain labels (*e.g.*, source or target domain) of the input paired images based on h_i^s and h_j^t .

Classification Module: Since data in the target domain are unlabeled, we propose to employ the generated feature representation (via the generator $G_{s \rightarrow t}$) and its label in the source domain to improve the robustness of the target classifier. Given a feature vector f_i^s , we denote $g^G = C_s(G_{s \rightarrow t}(f_i^s)) + C_t(G_{s \rightarrow t}(f_i^s))$ as the prediction for our generated feature $G_{s \rightarrow t}(f_i^s)$, and such predictions are achieved by the source classifier C_s and the target classifier C_t . Then, we develop a task-specific loss to assist the classification tasks in both the source and target domains, which is formulated as follows:

$$\begin{aligned} & \mathcal{L}'_{task}(C_s, C_t, G_{s \rightarrow t}(f_i^s), y_i^s) \\ &= \mathbb{E}_{(G_{s \rightarrow t}(f_i^s), y_i^s) \sim \mathcal{D}^s} \frac{1}{n_s} \sum_{i=1}^{n_s} (L(C_s(G_{s \rightarrow t}(f_i^s)), y_i^s) + L(C_t(G_{s \rightarrow t}(f_i^s)), y_i^s)), \end{aligned} \quad (5)$$

through which we can explicitly learn the target classifier C_t , even though there are no labeled data in the target domain.

By combining Eqs. (1)–(5), we obtain the objective function of our UCAN model as follows:

$$\begin{aligned} & \min_{C_s, C_t} \min_{G_{t \rightarrow s}, G_{s \rightarrow t}} \min_{D_s, D_t, D} \mathcal{L}_{total}(C_s, C_t, x_i^s, y_i^s, x_j^t, G_{t \rightarrow s}, G_{s \rightarrow t}, D_s, D_t, D) \\ &= \mathcal{L}_{task}(C_s, x_i^s, y_i^s) + \mathcal{L}'_{task}(C_s, C_t, G_{s \rightarrow t}(f_i^s), y_i^s) \\ & \quad + \mathcal{L}_{sdom}(D_s, f_i^s, f_j^t) + \mathcal{L}_{tdom}(D_t, f_i^s, f_j^t) \\ & \quad + \mathcal{L}_{adv}(D, h_i^s, h_j^t). \end{aligned} \quad (6)$$

Implementation: The proposed network is implemented in PyTorch. Specifically, in the *training* stage, we first pre-train the feature extraction module using ResNet-50 [14] on ImageNet (with the Pytorch source code provided in <https://github.com/KaimingHe/deep-residual-networks>), and then train the entire network in an end-to-end manner for 500 epochs. The Adam solver is used with a



Fig. 2. Results of five different methods in both tasks of (a) disease identification and (b) MCI conversion prediction.

batch size of 36 and a learning rate of 3×10^{-3} . In the *testing* stage, we feed the testing MRI into the bottom part of the trained UCAN model (only the target domain) to predict its class label. For each 3D structural MR image, we extract its slices from three views (*i.e.*, coronal-plane, sagittal-plane and axial-plane views), followed by feeding slides of each view to the proposed network. Finally, the results based on slides of three views for each subject are fused using the majority voting strategy to get the final result. Each slice is re-sized to 256×256 with zero-filling before being fed into our network. Besides, we apply both horizontal and vertical flip to MRI slides for data augmentation.

3 Experiment

Experimental Setup: The proposed UCAN method was compared with three state-of-the-art methods for deep domain adaptation using structural MR images, including (1) Deep Adaptation Network (**DAN**) [8], and (2) Joint Adaptation Network (**JAN**) [9], and (3) Conditional Domain Adaptation Network (**CDAN**) [10]. To evaluate the efficacy of the proposed cycle feature adaptation (CFA) module (see Fig. 1), we further compare our UCAN with its variant without using the CFA module (called **UCAN-c**). Except for the CFA module, UCAN-c and UCAN share the same network architecture. For a fair comparison, both DAN and JAN methods employ the suggested network architecture and parameters provided by the authors (see <https://github.com/thuml/Xlearn>), while CDAN uses the same settings provided by the respective paper (see <https://github.com/thuml/CDAN>). Two groups of experiments were performed, including (1) brain disease identification (*i.e.*, AD vs. pMCI vs. sMCI vs. NC classification), and (2) MCI conversion prediction (*i.e.*, pMCI vs. sMCI classification). To validate the robustness of a specific method, a two-fold cross-validation was used in the experiments. In the 1st fold (*i.e.*, “ADNI1→ADNI2”), we treated ADNI1 as the source domain and ADNI2 as the target domain. In the 2nd fold (*i.e.*, “ADNI2→ADNI1”), we regarded ADNI2 as the source domain and ADNI1 as the target domain. The classification accuracy was employed as the evaluation metric in the task of pMCI vs. sMCI classification, while the averaged accuracy among four categories was used in the task of brain disease identification.

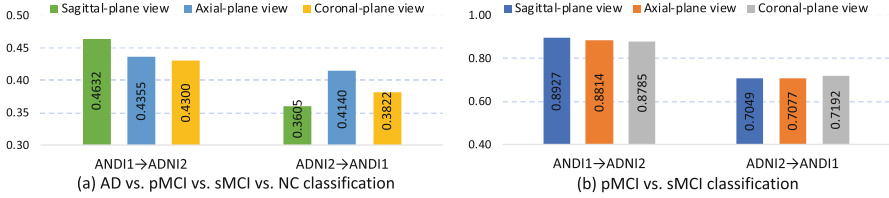


Fig. 3. Results of the proposed UCAN model using MRI slices from three different views in (a) disease identification and (b) MCI conversion prediction.

Results of Disease Identification: Figure 2(a) reports the results achieved by five different methods in the task of brain disease identification (*i.e.*, AD vs. pMCI vs. sMCI vs. NC classification), from which one may have the following observations. *First*, our UCAN method achieved the best performance in two-fold cross-validation, compared with three deep learning methods (*i.e.*, DAN, JAN, and CDAN). *Second*, our UCAN consistently outperforms its degenerated variant (*i.e.*, UCAN-c). For instance, using ADNI1 as the source domain (*i.e.*, “ADNI1→ADNI2”), the average accuracy of four-category classification achieved by UCAN is 0.4429 for subjects in the target ADNI2 domain, while UCAN-c only yields an average accuracy of 0.4351. This clearly suggests the effectiveness of our proposed cycle feature adaptation module to alleviate the data distribution shift among source and target domains. *Besides*, the overall performance achieved by five methods in the 2nd fold (*i.e.*, “ADNI2→ADNI1”) is worse than that of different methods in the 1st fold (*i.e.*, “ADNI1→ADNI2”). The possible reason is that ADNI2 is a relatively unbalanced dataset (*e.g.*, the number of pMCI subjects is largely less than the other three categories, compared with ADNI1. In such a case, models trained on ADNI2 are less robust than those trained on the balanced ADNI1 dataset.

Results of Disease Progression Prediction: We also report the results of five methods in MCI conversion prediction (*i.e.*, pMCI vs. sMCI classification) in Fig. 2(b). From this figure, a similar trend can be found as that in the task of brain disease identification. That is, our UCAN method is consistently superior to the four competing methods in both folds. The underlying reason is being that our method can capture the relationship between feature representations and classifier predictions as complementary information in the training process. Considering that the domain adaptation between the source and target domains in the task of pMCI vs. sMCI classification is a severe data unbalanced problem, these results further demonstrate the robustness of the proposed UCAN method.

Discussion and Future Work: In the current work, the input data of UCAN are 2D image slices extracted from three views (*i.e.*, sagittal-plane, axial-plane, and coronal-plane views). We now investigate the influence of different views on the performance of our method, with results reported in Fig. 3. From Fig. 3, one

can observe that the overall best performance is achieved by using the sagittal-plane view in 1st fold. In the experiments, we equally treat three views to generate the prediction result for a testing subject, by fusing the results of UCAN using slides of three views via majority voting. Using a weighted voting strategy seems to be more reasonable, which will be our future work. Besides, we plan to employ 3D (rather than 2D) convolution in the proposed network to take advantage of the global structure information of 3D MRIs.

4 Conclusion

We present an unsupervised conditional consensus adversarial network (UCAN) for deep domain adaptation, which can learn the disease classifier from the labeled source domain and adapt to a different target domain. Specifically, we first design a feature extraction module to learn representations from input MRI, followed by a cycle feature adaptation module to harmonize features and classifiers of the source and target domains. Experimental results on 1,506 subjects suggest the efficacy of the proposed method.

References

1. Liu, M., Zhang, D., Shen, D.: Relationship induced multi-template learning for diagnosis of Alzheimer’s disease and mild cognitive impairment. *IEEE Trans. Med. Imaging* **35**(6), 1463–1474 (2016)
2. Lian, C., et al.: Multi-channel multi-scale fully convolutional network for 3D perivascular spaces segmentation in 7T MR images. *Med. Image Anal.* **46**, 106–117 (2018)
3. Cheng, B., Liu, M., Zhang, D., Munsell, B.C., Shen, D.: Domain transfer learning for MCI conversion prediction. *IEEE Trans. Biomed. Eng.* **62**(7), 1805–1817 (2015)
4. Lian, C., Liu, M., Zhang, J., Shen, D.: Hierarchical fully convolutional network for joint atrophy localization and Alzheimer’s disease diagnosis using structural MRI. *IEEE Trans. Pattern Anal. Mach. Intell.* (2018)
5. Zhu, Y., et al.: MRI-based prostate cancer detection with high-level representation and hierarchical classification. *Med. Phys.* **44**(3), 1028–1039 (2017)
6. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. *CoRR*. abs/1709.10190 (2017)
7. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: maximizing for domain invariance. *arXiv preprint [arXiv:1412.3474](https://arxiv.org/abs/1412.3474)* (2014)
8. Long, M., Cao, Y., Cao, Z., Wang, J., Jordan, M.I.: Transferable representation learning with deep adaptation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* (2018)
9. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: *ICML*, pp. 2208–2217 (2017)
10. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: *NIPS*, pp. 1640–1650 (2018)
11. Jack, C., Bernstein, M., Fox, N., et al.: The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging* **27**(4), 685–691 (2008)

12. Holmes, C.J., Hoge, R., Collins, L., Woods, R., Toga, A.W., Evans, A.C.: Enhancement of MR images using registration for signal averaging. *J. Comput. Assist. Tomogr.* **22**(2), 324–333 (1998)
13. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans. Med. Imaging* **17**(1), 87–97 (1998)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *CVPR*, pp. 770–778 (2016)