Computers in Biology and Medicine 154 (2023) 106570

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Prototype-guided multi-scale domain adaptation for Alzheimer's disease detection

Hongshun Cai, Qiongmin Zhang *, Ying Long

College of Computer Science and Engineering, Chongqing University of Technology, Chongqing, 400054, China

ARTICLE INFO

Keywords:
Alzheimer's disease
Structural MRI
Multi-scale
Domain adaptation
Feature alignment
Prototype MDD

ABSTRACT

Alzheimer's disease (AD) is the most common form of dementia and there is no effective treatment currently. Using artificial intelligence technology to assist the diagnosis and intervention as early as possible is of great significance to delay the development of AD. Structural Magnetic Resonance Imaging (sMRI) has shown great practical values on computer-aided AD diagnosis. Affected by data from different sources or acquisition domains in realistic scenarios, MRI data often suffer from domain shift problem. In this paper, we propose a deep Prototype-Guided Multi-Scale Domain Adaptation (PMDA) framework to handle MRI data with domain shift problem, and realize automatic auxiliary diagnosis of AD, Mild Cognitive Impairment (MCI) and Cognitively Normal (CN). PMDA is composed of three modules: (1) MRI multi-scale feature extraction module combines the advantages of 3D convolution and self-attention to effectively extract multi-scale features in high-dimensional space, (2) Prototype Maximum Density Divergence (Pro-MDD) module adopts prototype learning to constrain the feature outlier samples in a mini-batch when MDD is used to align source domain and target domain, and (3) Adversarial Domain Adaptation module is applied to achieve global feature alignment of the source domain and target domain and co-training two distinctive discriminators to mitigate the over-fitting issue. Experiments have been performed on 3T and 1.5T sMRI with domain shift in ADNI dataset. The experimental results demonstrated that the proposed framework PMDA outperforms supervised learning methods and several state-of-the-art domain adaptation methods and achieves a superior accuracy of 92.11%, 76.01% and 82.37% on AD vs. CN, AD vs. MCI, and MCI vs. CN tasks, respectively.

1. Introduction

Alzheimer's disease (AD) is one of the most common chronic diseases in the elderly. It is mainly characterized by progressive cognitive impairment and behavioral impairment. The cause of the disease is still unclear [1], and there are no effective clinical methods or medicines can cure it. According to the World Health Organization (WHO) [2], more than 55 million people live with dementia worldwide and there are nearly 10 million new cases every year. AD is the most common form of dementia and may contribute to 60–70% of cases. The general pathological manifestations of AD are the reduction of the brain volume, followed by deepening and widening of the sulci or atrophy of the temporal lobe, especially the hippocampus [3]. Mild Cognitive Impairment (MCI) is a syndrome that occurs in the early stage of AD [4]. MCI is more subtle compared to AD in terms of the characteristics of the brain lesion and is therefore more difficult to identify. Unlike AD, which is irreversible, some MCI is reversible and some patients have a chance of improving or surviving with MCI status after early intervention. So Early detection of MCI and AD patients, and effective intervention treatment are of great significance for preventing or delaying the disease.

Structural Magnetic Resonance Imaging (sMRI) can provide relative values for different tissue types. So the morphological changes caused by AD, such as hippocampal atrophy, could be observed in sMRI. It has shown significant clinical and practical values in computer-aided AD diagnosis. However, such a situation is often encountered when using MRI data for auxiliary diagnosis, the relative values for different tissue types are often affected by the scanner manufacturer, the scan protocol or even the software version. This issue is known as the scanning bias [5]. When use model to extract features, it may contain two parts of features, one is the feature related to the actual task, and the other is the feature related to the scanner bias. If the features that related to the scanner bias are more prominent and distract compare with features related to the actual task, e.g., disease-related features, it will seriously affect the generalization performance of the model. Therefore, scanner

* Corresponding author.
E-mail address: zqm@cqut.edu.cn (Q. Zhang).

https://doi.org/10.1016/j.compbiomed.2023.106570
Received 6 September 2022; Received in revised form 2 January 2023; Accepted 22 January 2023
Available online 23 January 2023

bias could result in inconsistent distribution of data features, also known as "domain shift" [6]. This has become a prospective problem to be solved when faced with MRI data from different scanners and institutions.

Due to deep learning has made breakthroughs in the field of computer vision [7], more and more researchers apply deep learning to the classification and recognition of AD [8]. Labeled imaging data is often used for supervised learning, and assume that train data and test data satisfy the independent and identical feature distributions. However, when the feature distributions are inconsistent, the performance of the trained classifier on test data will degrade seriously.

Domain adaptation techniques is a promising approach to handle the domain shift problem of MRI data analysis. It can accomplish feature alignment by imposing constraints on data from different domains. Then the classifier which trained on the labeled dataset (source domain) could be directly applied to the unlabeled dataset (target domain), and get excellent generalization effect. More and more researchers apply domain adaptation methods to medical imaging analysis, including detection, classification, and segmentation [9] due to its great performance.

In this paper, we propose a deep Prototype-Guided Multi-Scale Domain Adaptation (PMDA) framework and apply it to the auxiliary diagnosis of AD, MCI and CN. As shown in Fig. 2, there are three modules in the proposed method: (1) MRI multi-scale feature extraction module, which is used to extract features from source domain and target domain, (2) Prototype Maximum Density Divergence (Pro-MDD) module, which is used to alleviate the equilibrium challenge of adversarial learning and constrain feature outlier samples to enhance feature alignment, and (3) Adversarial Domain Adaptation module is applied to achieve global feature alignment of the source domain and the target domain and co-training two different domain discriminators for training more generalizable models. We validated the proposed framework PMDA on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and demonstrated the superiority of the method in terms of various medical metrics and algorithm performance.

The major contributions of the study are as follows.

1. The proposed PMDA method is an unsupervised domain adaptation framework for MRI-based AD and MCI detection without requiring any label information of target domain.
2. We design a multi-scale feature extraction module based on 3D convolution and self-attention for MRI, and combine the advantages of metric learning and adversarial learning to achieve feature alignment of source domain and target domain.
3. We propose Pro-MDD which introduces prototype learning in Maximum Density Divergence to strengthen the constraints on outlier samples, thus enhancing the feature alignment effect and making the model more generalizable and robust.
4. Extensive experiments have been performed on sMRI data from ADNI dataset with domain shift demonstrate the significant superiority of PMDA framework in AD and MCI classification and detection.

This paper is organized as follows. We first review the relevant works in Section 2. The details of MRI data used in this work and PMDA method are introduced in section 3. In Section 4, we present various comparative experiments in detail, as well as a discussion of the experimental results. Section 5 we further analyze the influence of several key components of the proposed method and discuss the limitations of the current work and future work. The paper is finally concluded in Section 6.

2. Related work

2.1. MRI-based AD and MCI analysis

The sMRI have been widely used in computer-aided systems for AD and MCI identification. Ahmed et al. [10] extracted local features from hippocampus and posterior cingulate cortex, and then apply a Support Vector Machine (SVM) for AD and MCI classification. Beheshti et al. [11] proposed a histogram-based feature generation framework which based on patient-specific brain connectivity networks for AD diagnosis and MCI transformation prediction. In recent years, deep learning has achieved promising results in neuroimaging analysis. Aderghal et al. [12] used convolution neural networks (CNNs) with a data augmentation strategy adapted to the specificity of sMRI scans for AD classification. Oh et al. [13] proposed to use convolutional autoencoder (CAE)-based unsupervised learning and supervised transfer learning for AD and MCI classification. Korolev et al. [14] proposed deep 3D VGG-like CNN architectures for classification of brain MRI scans.

In these methods, it is common to consider the testing data has the same or similar feature distribution with the training data. However, the collected imaging data usually come from different scanners or institutions actually, the feature distributions of the test and train data can be even completely different. If the feature distribution of train and test data is quite different, the transferability of the model will be significantly affected.

For more and more studies have become aware of the domain shift issue, domain adaption techniques begin to be gradually employed in the auxiliary diagnosis of AD [9]. Guan et al. [15] proposed an attention-guided deep domain adaptation (AD^2A) framework for Multi-site MRI harmonization and apply it to automated brain disorder identification. Li et al. [16] proposed an effective knowledge transfer method to diminish the disparity among different datasets. However, most of them either using adversarial learning or metric learning to align the marginal (global) distributions, or align conditional (local) distributions based on domain adaptation methods. In practice, marginal distributions often accompany with conditional distributions. The marginal distribution is always more important when two domains are very dissimilar. When the marginal distributions are similar, the conditional distribution should be given more attention. Existing methods rarely consider the joint alignment of marginal and conditional distributions. Our approach takes into account the marginal and conditional distributions at the same time, which can effectively improve the model performance when feature alignment is achieved. Disregarding the category information and outlier samples may bring about incorrect matching of categories, which may eventually lead to negative transfer.

2.2. Attention enhanced convolution

By using of convolution kernel, CNNs can automatically extract features from image conveniently and efficiently. Therefore, CNNs have become one of the most powerful techniques in various vision tasks. Previous researches showed that, applying attention mechanism over images can overcome locality limitation for CNNs [17]. Many researchers have explored the possibility of employing attention modules to enhance the functionality of convolutional networks. Squeeze-and-Excitation (SE) [18] module and Gather-Excite (GE) [19] module implement the attention mechanism by reweighing each channel in feature map which focuses on correlations across channels. Bottleneck Attention Module (BAM) [20] and Convolutional Block Attention Module (CBAM) [21] independently reweigh both channels and spatial locations to optimize the feature map. Inspired by the powerful ability of self-attention in long-range dependencies, some research works try to combine self-attention and convolution together to further explore and fuse image features. BoTNet [22] is a hybrid model which achieves great performance. It uses both convolution and self-attention for visual recognition. Pan et al. [17] proposed a hybrid

operator to integrate self-attention and convolution modules by sharing the same heavy operations. Results on image classification and object detection benchmarks demonstrate the effectiveness of the proposed operator. Existing methods have proved that attention mechanism can enhance the extraction effect of convolution on image features.

2.3. Prototype learning

Prototypes can provide useful insight into the inner workings of the network, the relationship between classes, and the important aspects of the latent space [23]. Prototype learning is considered as a robust method when handling open-set recognition [24] and few-shot learning [25]. Wang et al. [26] proposed a novel prototype alignment network to learn class-specific prototype representations from a few support images within an embedding space for image segmentation. Tanwisuth et al. [27] proposed prototype-oriented alignment method that works well under various application scenarios of unsupervised domain adaptation. It is robust against class imbalance and data-privacy concerns. Yang et al. [28] proposed convolutional prototype network, which keeps CNN for representation learning but replaces the closed-world assumed softmax with an open-world oriented and human-like prototype model. Experiments demonstrate the efficiency and effectiveness of the method for both closed-set recognition and open-set recognition tasks. Therefore, existing studies support the notion that domain adaptation and few-shot learning are significantly benefited by prototype learning.

3. Materials and methodology

3.1. Materials and image preprocessing

Data used in this work is collected from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and AD.

All sMRI scans are performed anterior commissure (AC)–posterior commissure (PC) correction, skull and dura strip, and align to the Colin27 template to obtain the standard image data ($1 \times 1 \times 1$ mm^3). The data details are shown in Table 1.

3.2. Domain shift assessment

Software and hardware setting of MRI equipment has a great impact on the MRI imaging. One of the elements is the magnetic field strength. The MRI data in the ADNI database were primarily obtained by scanning at two magnetic field strengths: 3T and 1.5T. 3T MRI has more advantages compared with 1.5T, such as higher spatial resolution, higher signal-to-noise ratio, and better image enhancement effect. In this paper, 3T and 1.5T MRI data were used as the source domain and the target domain, respectively. Data preprocessing can match normalized intensities of all data from different data sources [29]. However, the impact of scanner effects remains and the scanner-specific bias is unable to remove [30]. When using sMRI for the identification of AD and MCI, even after careful preprocessing, the domain shift introduced by scanner effects remains and the scanner-specific bias is the main factor affecting the generalization performance of deep learning model.

Table 1
The details of dataset.

Category	3T	1.5T	Subject
AD	128	144	272
MCI	160	152	312
CN	152	160	312

To evaluate the domain shift of the used MRI dataset, we use T-SNE [31] to perform dimensionality reduction and show the feature distributions on source domain and target domain data before feed them into model for supervised learning or domain adaptation. Fig. 1(a) shows that after data preprocessing, although a large amount of data clustered together, there are still some outliers. So it shows the differences in marginal (global) distribution of all data. Fig. 1(b) shows the zoomed in result of the clustered area. We can see that although the marginal distributions of source domain and target domain aligned, all the categories are mixed together and difficult to separate from each other. Therefore, differences in conditional (local) distributions are existed in the used dataset as well.

By observing the feature distribution of all data, it can be found that despite both our source domain and target domain data have been preprocessed in the same way, marginal and conditional distributions differences of the two domains are still existed. It is conceivable that the difference in data distribution will be even greater if the data comes from different sources or acquisition domains.

3.3. Problem definition

Experiments focus on the problem of Unsupervised Domain Adaptation (UDA) for AD and MCI classification. In UDA setting, the source and target data are sampled from different distributions $P_S(x,y)$ and $P_T(x,y)$ respectively, where $P_S(x,y) \neq P_T(x,y)$. The source domain is denoted as $\mathscr{D}_S = \{(x_i^S, y_i^S)\}_{i=1}^{n_S}$, where x_i^S is the i-th source sample, y_i^S is the corresponding category label and n_S indicate the number of source data. Also, we have n_t samples of unlabeled sMRI data in target domain which denoted as $\mathscr{D}_T = \{x_i^T\}_{i=1}^{n_T}$, where x_i^T is the i-th target sample. The source and target domain are assumed to share the same set of category labels $C = \{1, 2, ..., K\}$ and K indicate the number of classes in both domains. Our work is to design an unsupervised learning model, which is trained on the labeled source domain data and unlabeled target domain data. Then by using this model to classify the target domain data.

3.4. Method

As shown in Fig. 2, the proposed PMDA framework consists of three parts: (1) MRI Multi-Scale Feature Extraction module, (2) Prototype Maximum Density Divergence (Pro-MDD) module, and (3) Adversarial Domain Adaptation module. The details of each component are as follows.

3.4.1. MRI multi-scale feature extraction module

Considering the high dimension of MRI data, we adopt 3D CNNs as the backbone to extract the spatial feature. The convolution layers of 3D CNN could get the high-level semantic information easily in feature extraction module.

In order to locate the disease-related regions, we generate a spatial attention map which can detect disease-related brain areas automatically by utilizing the inter-spatial relationship of features. To compute the spatial attention, the convolution results of the first layer are denoted as the input feature map M, then average-pooling and max-pooling operations along the channel axis are performed on M respectively. The results of the two pooling results are concatenated and a standard 3D convolution layer is applied to obtain the spatial attention map. In the spatial attention map, larger weight is given to the identified important regions and relatively smaller weight to the unimportant regions. The attention map A is defined as follows:

$$A = \sigma\left(f^{3\times3\times3}\left(\left[M_{max}, M_{avg}\right]\right)\right) \tag{1}$$

where σ represents the sigmoid function and $f^{3\times3\times3}$ denotes a convolution operator with $3 \times 3 \times 3$ kernel.

The source domain and target domain data use the same feature extraction network and share weights. In order to encourage the atten-

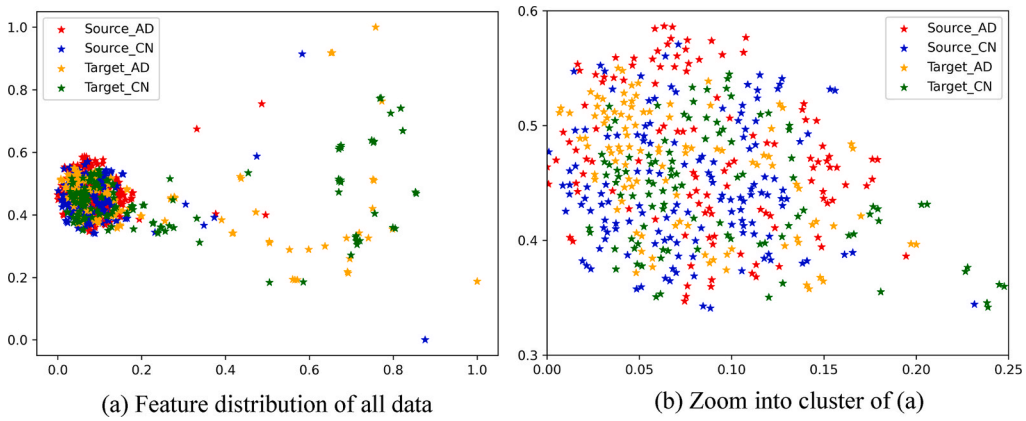

Fig. 1. Data distribution visualization of source domain and target domain.

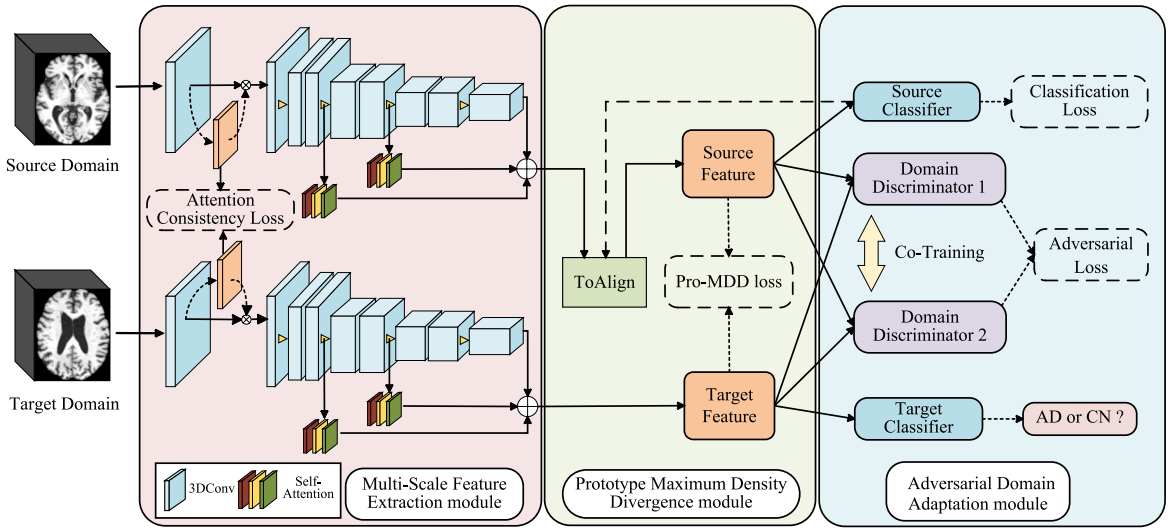

Fig. 2. Illustration of the proposed Prototype-Guided Multi-Scale Domain Adaptation (PMDA) framework for MRI-based AD diagnosis. There are three components: (1) MRI Multi-Scale Feature Extraction module, (2) Prototype Maximum Density Divergence module, and (3) Adversarial Domain Adaptation module.

tion consistency and transfer semantic information from source domain to target domain, an attention consistency loss is added in our framework. We define the spatial attention map of the source domain as A^S, and the target domain as A^T. Calculate the mean square difference of the two spatial attention maps as the attention consistency loss function:

$$\mathscr{L}_{att} = \frac{1}{N \times D \times W \times H} \sum_{i=1}^{N} \left\| A_i^S - A_i^T \right\| \tag{2}$$

where N is the number of samples in a batch, D, W and H represent the depth, width and height of the feature map, respectively.

Self-attention can capture long-range contextual information between feature maps, so we use it to enhance representation capability. Considering that self-attention requires $O(n^2d)$ memory and computation [32] when performed globally across n entities, we incorporate self-attention module after Conv6 and Conv8 layers. To aggregate multi-scale features, we first use self-attention to process the output features of Conv6 and Conv8, then together with the last layer of features to perform feature fusion on them. We describe the detailed operation of self-attention operation below.

The output feature maps of conv6 and conv8 have different scales of semantic information, and we use them as input of self-attention module respectively. Taking the output of Conv6 as an example, named the feature map of Conv6 as $M_{in} \in \mathbb{R}^{C \times H \times W \times D}$, we use 3D-convolution to generate two new feature maps $\mathbf{Q}$ and $\mathbf{K}$ respectively, $\{\mathbf{Q}, \mathbf{K}\} \in \mathbb{R}^{C \times H \times W \times D}$. Then reshape them to $\mathbb{R}^{C \times N}$, where $N = H \times W \times D$ is the number of voxels. Finally perform matrix multiplication on the $\mathbf{Q}$ and $\mathbf{K}$, and apply a softmax layer to calculate the spatial attention map $\mathbf{S} \in \mathbb{R}^{N \times N}$:

$$s_{ji} = \frac{\exp\left(Q_i \cdot K_j\right)}{\sum_{i=1}^{N} \exp\left(Q_i \cdot K_j\right)} \tag{3}$$

Where s_{ji} measures the i^{th} position's impact on j^{th} position. The more similar feature representations of the two position contributes to greater correlation between them.

Meanwhile, we feed feature M_{in} into a 3D-convolution layer to generate a new feature map $\mathbf{V} \in \mathbb{R}^{C \times H \times W \times D}$ and reshape it to $\mathbb{R}^{C \times N}$. Then perform a matrix multiplication on $\mathbf{V}$ and $\mathbf{S}$ and multiply a scale parameter α. Finally perform a skip-connection operation with features M_{in} to obtain the final output $\mathbf{E} \in \mathbb{R}^{C \times H \times W \times D}$:

$$E_j = \alpha \sum_{i=1}^{N} \left(s_{ji} V_i\right) + M_{in} \tag{4}$$

where α is initialized as 0 and gradually learns to assign more weight.

It can be inferred from Eq. (4) that $\mathbf{E}$ at each position is a weighted sum of the features across all positions and features of Conv6. Therefore,

E selectively aggregates contexts based on the spatial attention map and has a global contextual view. It could capture long-range contextual information more effectively and learn better feature representation.

3.4.2. Prototype Maximum Density Divergence module

In order to alleviate the negative impact of domain shift and outlier samples, we use Pro-MDD module in the proposed framework. We integrate the advantages of both prototype learning and metric learning, and both of them has a mutually beneficial effect with the subsequent adversarial domain adaptation module.

MDD [33] manages to align the source domain and target domain by simultaneously minimizing the inter-domain divergence and maximizing the intra-class density and it was verified could alleviate the equilibrium challenge of adversarial learning in domain adaptation. Suppose we have n_s samples $\{x_{s,1}, x_{s,2}, \cdots x_{s,n_s}\}$ in the source domain and n_t samples $\{x_{t,1}, x_{t,2}, \cdots x_{t,n_t}\}$ in the target domain, and their labels are denoted as y_s and y_t, respectively. Our purpose is to minimize MDD(P_S, P_T) and the formula is defined as follows:

$$\begin{aligned}\mathrm{MDD}(P_S, P_T) = &\frac{1}{n_b}\sum_{i}^{n_b}\left\|x_{s,i} - x_{t,i}\right\|_2^2 \\ +&\frac{1}{m_s}\sum_{y_{s,i}=y'_{s,j}}\left\|x_{s,i} - x'_{s,j}\right\|_2^2 + \frac{1}{m_t}\sum_{y_{t,i}=y'_{t,j}}\left\|x_{t,i} - x'_{t,j}\right\|_2^2\end{aligned} \tag{5}$$

where n_b is equal to the half of the batch size, $y_{s,i} = y'_{s,j}$ indicates that $x_{s,i}$ and $x'_{s,j}$ have the same label. Since label information is unavailable for the target domain, we use pseudo labels. Due to the number of samples in batch which satisfies $y_{s,i} = y'_{s,j}$ is uncertain before training, we use m_s and m_t to represent the appropriate number. In experiments, it calculates the pair-wise distance at the relative position for the reason that trained by batch.

However, the use of MDD loss will introduce the following problems in our experiments. First, MDD may sensitive to outlier samples in a mini-batch when used to align the source and target marginal distributions. We attribute this problem to the sampling variability of both the source and target samples. Second, limited by the high dimensionality of MRI data and the impact of computing power, we can only use small batches for model training. Even if the two domains share the same set of category labels, we cannot guarantee that the samples drawn from the source and target domains will cover the same set of classes in each mini-batch. Especially, if the label proportions shift between domains, learning a domain-invariant representation may not perform better than using the source data alone to train the model. Therefore, in order to deal with both of these problems, we propose prototype MDD (Pro-MDD), which introduce the theory of prototype learning in MDD loss.

Prototype learning by constructing class prototypes which can represent the samples with the same label in the latent space, integrate this idea with MDD can improve generalization performance of the model. We compute the mean vector over the source data features with true labels as class prototypes. The influence of outlier samples can be alleviated to a certain extent by calculating the Euclidean norm distance between the source domain and target domain data with the corresponding prototype instead of the data with the same label in a batch. The class prototype is defined as the following formula:

$$c_k = \frac{1}{|X_S|}\sum_{(x_i, y_i)\in X_S} f_\varphi(x_i) \tag{6}$$

where c_k represents the prototype of class k, $f_\varnothing$ represents the feature extraction module with learnable parameters φ. Each prototype is the mean vector of the source data belonging to its class.

Meanwhile, if we directly compute the mean vector of all data of a category in the source domain is computational-intensive during training. To address this problem, we estimate the class prototypes as the moving average of the cluster centroids in mini-batches, so that we can track the prototypes that slowly move. Specifically, in each iteration, the prototype is estimated as:

$$c_k = \gamma c_k + (1-\gamma)c'_k \tag{7}$$

where c'_k is the mean vector of class k calculated within the current training batch from the feature extraction module, and γ is the momentum coefficient.

At this point the Pro-MDD loss calculated by the following equation:

$$\begin{aligned}\mathscr{L}_{pro-MDD} = &\frac{1}{n_b}\sum_{i}^{n_b}\left\|x_{s,i} - x_{t,i}\right\|_2^2 \\ +&\frac{1}{m_s}\sum_{n=1}^{k}\sum_{y_{s,i}=y_n}\left\|x_{s,i} - c_n\right\|_2^2 + \frac{1}{m_t}\sum_{n=1}^{k}\sum_{y_{t,i}=y_n}\left\|x_{t,i} - c_n\right\|_2^2\end{aligned} \tag{8}$$

where k represents the number of class prototypes.

We can see that the $\frac{1}{n_b}\sum_{i}^{n_b}\left\|x_{s,i} - x_{t,i}\right\|_2^2$ of Eq. (8) considers the inter-domain divergence which can bring the feature distribution of the source domain and the target domain closer, the $\frac{1}{m_s}\sum_{n=1}^{k}\sum_{y_{s,i}=y_n}\left\|x_{s,i} - c_n\right\|_2^2$ and $\frac{1}{m_t}\sum_{n=1}^{k}\sum_{y_{t,i}=y_n}\left\|x_{t,i} - c_n\right\|_2^2$ of Eq. (8) consider the intra-domain density, which can make the features of which have same class label closer and different class label farther. The use of prototype learning addresses the impact of sampling variability and negative transfer caused by outlier samples.

To achieve task-oriented alignment, we introduce ToAlign [34] which motivated by Grad-CAM [35] in this module. It uses the gradients of the predicted score corresponding to the ground-truth class as the attention weights to obtain the task-discriminative features. So we utilize it to decompose a holistic feature of each source sample into a task-discriminative feature and a task-irrelevant feature to enable task-oriented alignment with the target features.

3.4.3. Adversarial domain adaptation module

The adversarial domain adaptation algorithm is used to reduce the feature distribution differences between source domain and target domain. Feature extractor and domain discriminator are the keys of this module, and both of them are trained in an adversarial manner. Domain discriminator is used to distinguish whether the data comes from source domain or target domain, and feature extractor is applied to confuse the domain discriminator by extracting domain-invariant features. Once the domain discriminator cannot distinguish whether a learned feature from the source domain or the target domain, it is considered that the learned representations are domain-invariant.

The Source Classifier is trained to discriminate the labels of the input MRI samples from the source domain. Thus, the supervised information on source domain can be utilized. Using the feature map generated by the Multi-Scale Feature Extraction module as input, we apply three fully-connected layers in the source classifier for classification. Its training objective is a cross-entropy loss, which can be formulated as:

$$\mathscr{L}_{cls} = -\frac{1}{N_S}\sum_{c=1}^{C}\sum_{i=1}^{N_S} y_{x_i}\log G_y\big(G_f(x_i)\big) \tag{9}$$

where N_S is the number of samples on source domain, C is the number of classes, G_y is the source classifier and G_f is the feature extractor. The target classifier shares weights with the source classifier and achieve the final classification of the target domain.

The domain discriminator here is based on the Domain-adversarial Neural Network (DANN), and parameters can be trained efficiently with the Gradient Reversal Layer (GRL) [36]. Due to data limitation, one discriminator over-fitting is easy to occur in domain adaptation, which leads to degradation. Specifically, over-fitting discriminator produces very high prediction scores and very small discriminator loss. It will lead to training divergence and degraded generation [37]. Inspired by co-training with multiple tasks [38], we utilize two domain

discriminators to co-train and calculate the loss of the domain discriminator as:

$$\mathscr{L}_{D_1} = \frac{1}{n_s + n_t} \sum_{x_i \in D_s \cup D_t} \mathscr{L}_{d_1}\left(G_{d_1}\left(G_f(x_i)\right), d_i\right) \tag{10}$$

$$\mathscr{L}_{D_2} = \frac{1}{n_s + n_t} \sum_{x_i \in D_s \cup D_t} \mathscr{L}_{d_2}\left(G_{d_2}\left(G_f(x_i)\right), d_i\right) \tag{11}$$

where d_i is the domain label of the input data x_i. G_{d_1} and G_{d_2} are two distinctive domain discriminators. $\mathscr{L}_{d_1}$ and $\mathscr{L}_{d_2}$ are the domain discriminator losses calculated using the cross-entropy loss.

With the co-training design, although one discriminator may overfit and focuses on learning simple patterns or structures, the other discriminator will be encouraged to learn different information such as complex patterns and structures. Inspired by GenCo [39], Weight-Discrepancy Co-training (WeCo) which co-trains multiple distinctive discriminators by diversifying their parameters with a weight discrepancy loss is used in this module. The two discriminators thus complement each other to focus on different types of information, which helps mitigate the over-fitting issue effectively.

Define the weight discrepancy loss $\mathscr{L}_{wd}$, it minimizes the cosine distance between the weights of G_{d_1} and G_{d_2}.

$$\mathscr{L}_{wd}(G_{d_1}, G_{d_2}) = \frac{\overrightarrow{W_{G_{d_1}}}\overrightarrow{W_{G_{d_2}}}}{\left|\overrightarrow{W_{G_{d_1}}}\right|\left|\overrightarrow{W_{G_{d_2}}}\right|} \tag{12}$$

where $\overrightarrow{W_{G_{d_1}}}$ and $\overrightarrow{W_{G_{d_2}}}$ are the weights of G_{d_1} and G_{d_2}.

Apply $\mathscr{L}_{wd}$ on only one discriminator for simplicity. The overall adversarial loss can thus be shown as:

$$\mathscr{L}_{Adv} = \lambda(\mathscr{L}_{D_1} + \mathscr{L}_{D_2}) + (1-\lambda)\mathscr{L}_{wd} \tag{13}$$

where λ is a trade-off parameter.

To summarize, the objective of our proposed method is to jointly optimize four components including source classification loss $\mathscr{L}_{cls}$, attention consistency loss $\mathscr{L}_{att}$, adversarial loss $\mathscr{L}_{Adv}$ and prototype MDD loss $\mathscr{L}_{pro-MDD}$. The overall optimization problem can be written as follows:

$$\min_F \mathscr{L}_{total} = \mathscr{L}_{cls} + \alpha\mathscr{L}_{att} + \beta\mathscr{L}_{Adv} + \omega\mathscr{L}_{pro-MDD} \tag{14}$$

where the parameters α、 β and ω weight the relative importance of these loss terms.

4. Experiments

4.1. Experimental setup

We conduct three groups of experiments, including: (1) AD vs. CN classification, (2) AD vs. MCI classification, and (3) MCI vs. CN classification. Each task is trained with labeled source domain data and unlabeled target domain data, and finally compare the recognition effect of PMDA and other methods on the target domain.

The proposed model PMDA was implemented in PyTorch. The network was trained for 100 epochs. The Adam [40] was used as the optimizer with a learning rate of 1×10^{-3} and use a small batch of 8 samples per domain. The dropout operation with a rate of 0.5 was used in classifier to prevent over-fitting. We empirically set the parameter λ、α、 β and ω to be 0.9、0.1、0.1 and 0.2 respectively.

At the beginning of the training work, we pretrain the MRI Multi-Scale Feature Extraction module on source domain for 30 epochs to get the initial weight. Then, the proposed method with three modules were further fine-tuned and co-trained via Eq. (14).

Four metrics were employed for performance evaluation in the experiments, i.e., classification accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the receiver operating characteristic curve (AUC). Denote TP, TN, FP, FN as the true positive, true negative, false positive and false negative, respectively. Then, the first three evaluation metrics can be defined as ACC $= \frac{TP+TN}{TP+TN+FP+FN}$, SEN $= \frac{TP}{TP+FN}$, SPE $= \frac{TN}{TN+FP}$. For each metric, a higher value indicates better classification performance.

4.2. Comparison with state-of-the-art methods

To acquire a broad perspective, the proposed PMDA framework is compared with several recent literatures which applied their methods on the ADNI database as well. It should be emphasized that, although all methods we compare were trained and tested on the ADNI, the exact dataset parameters such as scanning parameters, imaging equipment, sample size, etc., and the model with training parameters used in each literature are different. Therefore, the results are only used for the comparison of relative levels among methods, and the numbers do not represent the absolute superiority or inferiority.

From Table 2, it can be found that our method outperforms most of the listed machine learning methods and deep learning methods. By performing feature alignment on the data of source domain and target domain, the proposed PMDA effectively improved the model performance on AD detection. Compared with above supervised learning methods that assume train data and test data are identically distributed, our method takes into account the inconsistency of feature distribution and weaken the effect of domain shift.

4.3. Results of classification in target domain

We evaluate the proposed PMDA method and the competing methods in cross-domain problems. Take 3T MRI with label as the source domain and 1.5T MRI without label as the target domain and the results we show in the following tables are classification results on the target domain. We compared our method with 3DResNet50 [43] and five domain adaptation methods, including DANN [36], DAAN [44], CDAN [45], AD^{2}A [15] and ATM [33]. These competing methods are briefly introduced as follows.

(1) 3DResNet50. ResNet is a ubiquitously used architecture which enabled efficient implementation of deeper and bigger networks. ResNet50 is used as baseline model of supervised learning on source domain and make prediction on target domain. Thus, no feature alignment is performed on it. Since MRI data we used are 3D data, the 2D convolution blocks in ResNet50 are replaced with 3D convolution blocks.

Table 2
Classification accuracy of different methods.

Method	AD vs. CN	AD vs. MCI	MCI vs. CN	Subject
Ahmed et al. [10]	83.77	62.07	69.45	137AD, 210MCI, 162CN
Beheshti et al. [11]	84.17	67.59	70.38	83AD, 87MCI, 61CN
Aderghal et al. [12]	82.80	62.50	66.00	188AD, 399MCI, 228CN
Oh et al. [13]	86.60	60.97	63.04	198AD, 267MCI, 230CN
Korolev et al. [14]	79.00	62.00	63.00	50AD, 120MCI, 61CN
Prajapati et al. [41]	87.50	**83.33**	79.17	58AD, 60MCI, 60CN
Xia et al. [42]	88.30	–	79.02	154AD, 346MCI, 207CN
PMDA	**92.11**	76.01	**82.37**	128AD, 160MCI, 152CN

(2) DANN. DANN is a classic adversarial learning-based domain adaptation method that has been widely used in medical imaging tasks. Unlike PMDA, it adopts one single domain classifier for domain adaptation.
(3) DAAN. The model can dynamically learn domain-invariant representations while quantitatively evaluate the relative importance of global and local domain distributions. It is the first attempt to perform dynamic adversarial distribution adaptation for deep adversarial learning.
(4) CDAN. CDAN is a principled framework that conditions the adversarial adaptation models on discriminative information conveyed in the classifier predictions. It can improve the discriminability and entropy conditioning which could control the uncertainty of classifier predictions to guarantee the transferability.
(5) AD^{2}A. AD^{2}A is an attention-guided deep domain adaptation framework which used for Multi-site MRI harmonization. The method is trained in an adversarial learning manner using a domain discriminator and feature extractor. In the area of domain adaptation, this method is effectively in identifying brain diseases.
(6) ATM. ATM enjoys the benefits of both adversarial training and metric learning. The MDD loss make ATM can align two domains by simultaneously minimizing the inter-domain divergence and maximizing the intra-class density. Experiments show that ATM can outperform previous state-of-the-arts methods with significant advantages.

For fairness comparison, the Multi-Scale Feature Extraction module of PMDA is used in each method. Meanwhile, the training strategies are exactly the same, including the learning rate and the number of training epochs.

From Table 3, it is found that the proposed PMDA method achieves the highest accuracy on the classification tasks of AD vs CN, which is 9.54% higher than the baseline 3DResNet50. Among the compared domain adaptation methods, all the ACC, SEN and AUC of PMDA are the best.

From Table 4, the four indicators of PMDA on this task show excellent results than 3DResNet50. Compared to the listed domain adaptation methods, the ACC, SPE and AUC of PMDA is the best, the SEN is 9.72%, 10.41% lower than CDAN and AD^{2}A respectively. However, PMDA has a better balance between SEN and SPE. Compare with Table 3, AD vs. MCI is much more difficult to identify. The reason may be that MCI is an early developmental stage of AD, and its features are not obvious in MRI, which makes identification difficult.

From Table 5, PMDA achieves the highest accuracy of 82.37% on this classification task, and the SEN, SPE and AUC values are the best as well. Compared with AD, the brain structure lesions of MCI are more subtle, and the structural lesions of the brain are not exactly the same for different patients, so the identification of MCI and CN is difficult than AD and CN. The classification accuracy of MCI and CN was 6.36% higher than that of AD and MCI, indicating that our model PMDA has more advantages in predicting MCI than distinguishing early and late stages of dementia.

From the above 3 tables, we came to the following conclusions.

Table 3
AD vs. CN classification results.

Method	ACC(%)	SEN(%)	SPE(%)	AUC(%)
3DResNet50	82.57	77.08	87.50	82.29
DANN	86.51	79.17	**93.13**	86.14
DAAN	86.51	84.03	88.75	86.39
CDAN	85.86	87.50	84.38	85.94
AD^{2}A	88.49	86.81	90.00	88.40
ATM	89.14	88.89	89.38	89.13
PMDA	**92.11**	**91.67**	92.50	**92.08**

Table 4
AD vs. MCI classification results.

Method	ACC(%)	SEN(%)	SPE(%)	AUC(%)
3DResNet50	63.51	63.89	63.16	63.52
DANN	67.91	59.72	75.66	67.69
DAAN	64.19	65.97	62.50	64.24
CDAN	66.55	82.64	51.32	66.98
AD^{2}A	67.91	**83.33**	53.29	68.31
ATM	70.95	71.53	70.39	70.96
PMDA	**76.01**	72.92	**78.95**	**75.93**

Table 5
MCI vs. CN classification results.

Method	ACC(%)	SEN(%)	SPE(%)	AUC(%)
3DResNet50	75.96	75.00	76.88	75.94
DANN	76.60	82.89	70.63	76.76
DAAN	79.17	80.26	78.13	79.19
CDAN	77.89	75.66	80.00	77.83
AD^{2}A	80.13	82.89	77.50	80.20
ATM	80.77	80.26	**81.25**	80.76
PMDA	**82.37**	**83.55**	**81.25**	**82.40**

Compared with DANN, AD^{2}A which only use one domain discriminator, the mentioned four indicators of PMDA in the AD vs. CN and MCI vs. CN classification are improved, indicating the co-training two domain discriminators strategy in adversarial domain adaptation module can effectively improve the generalization performance.

Compared with DANN, DAAN, CDAN and AD^{2}A, which only use adversarial learning, PMDA combines metric learning and adversarial learning. In three classification tasks both ACC and AUC have been improved in a certain extent, indicating that the Pro-MDD has played a significant role in alleviating problems such as the instability of adversarial learning training and the disappearance of gradients.

Compared with ATM which use the combination of metric learning and adversarial learning methods, PMDA generally improves the four indicators in the three classification tasks, indicating that the use of attention consistency loss for source domain and target domain semantic sharing and the strategy for incorporating prototype learning in MDD can enhance the predictive performance on the target domain.

4.4. Results of 5-fold cross-validation classification

In order to show more convincing and robust results for domain adaptation, 5-fold cross-validation strategy is used here. We mainly use the whole 3T data and four folds of the 1.5T data to serve as the training set (source domain). The rest of the 1.5T data serve as the testing set (target domain). The AD vs. CN classification results of 3DResNet50 and other domain adaptation methods in each fold are listed in Table 6.

In Table 6, we can find that compare with supervised learning (3DResNet50) and other domain adaptation methods, the proposed PMDA method achieves the best results on ACC and AUC. These results show that our method still has great classification performance and robustness when the domain shift is not obvious. Meanwhile, compare with Table 3, all the methods in Table 6 have higher ACC and AUC values. This may be attributed to the fact that the source domain contains both 3T and 1.5T MRI data, while the target domain only contains 1.5T MRI data. Therefore, the domain shift of the source domain and the target domain is significantly decreased compared to the previous setting. Another reason for the better performance of our model is that we use more labeled source domain data for training.

Table 6
AD vs. CN classification results achieved by PMDA and six competing methods on one fold of target domain from ADNI using 5-fold cross-validation.

Method	Fold #1		Fold #2		Fold #3		Fold #4		Fold #5	
	ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)
3DResNet50	87.50	85.42	85.71	85.71	85.71	85.64	91.07	90.74	85.71	87.88
DANN	94.64	94.00	94.64	94.27	87.50	87.55	92.86	93.33	92.86	92.56
DAAN	92.86	92.59	91.07	89.13	89.29	89.29	92.86	93.55	91.07	90.64
CDAN	91.07	89.13	89.29	87.50	94.64	94.23	96.43	96.55	96.43	96.55
AD^2A	94.64	93.75	92.86	92.00	92.86	93.10	94.64	94.64	94.64	94.27
ATM	92.86	92.98	94.64	93.48	94.64	94.64	**100.00**	**100.00**	**98.21**	**98.39**
PMDA	**96.43**	**96.30**	**98.21**	**98.00**	**96.43**	**96.00**	**100.00**	**100.00**	**98.21**	98.33

5. Discussion

5.1. Ablation study

In order to verify the effectiveness of each module of PMDA, we use seven variants to conduct ablation experiments. The effectiveness of self-attention, attention consistency loss, Pro-MDD and co-training on the PMDA is demonstrated through the following comparative experiments.

(1) Apply nine 3D convolution layers with spatial attention as the backbone network of feature extraction (3DCNN) and three fully connection layers as the classifier which on the source domain with supervised learning. When the model reaches convergence, we use 3DCNN to extract features and the classifier on target domain is adopted to make prediction.
(2) Add the Self-Attention (+SA) mechanism to the feature extraction module on the basis of (1). Currently, the feature extraction module implements multi-scale feature fusion, and it is still trained by supervised learning.
(3) Add the attention consistency loss (+ATTC) to the multi-scale feature extraction module on the basis of (2) which can transfer semantic information from the source domain to the target domain.
(4) Add the adversarial domain adaptation module with one single domain discriminator (+Adv1) on the basis of (3).
(5) On the basis of (3), add the adversarial domain adaptation module (+Adv2). Co-train two different domain discriminators and use weight discrepancy loss to mitigate the discriminator over-fitting issue.
(6) Add the MDD module (+MDD) on the basis of (5);
(7) Add the Pro-MDD module (+Pro-MDD) on the basis of (5).

Table 7 reports the results achieved by seven variants in the task of AD vs. CN, AD vs. MCI and MCI vs. CN classification. The baseline 3DCNN has achieved general classification results on the three classification tasks. When self-attention (+SA) is added to achieve multi-scale feature fusion, the evaluation metrics on all three tasks are significantly improved. The inclusion of attention consistency loss (+ATTC) also resulted in different magnitudes of performance improvement for the model on the three tasks, suggesting that it may play a role in sharing semantic information between the source domain and target domain.

Almost all the indicators of the three tasks are all improved after the +Adv1 is added to the multi-scale feature extraction module, where the SEN and SPE performance of three tasks have great difference. It reflects that the adversarial training with single discriminator is always instable in a certain extent. The ACC and AUC have a small range of improvement again on the three tasks after the +Adv2 is added to the multi-scale feature extraction module. At the same time, the gap between the SEN and SPE metrics becomes smaller, and this phenomenon is particularly evident on the MCI vs. CN task. This result verifies that co-training two domain discriminators and add weight discrepancy loss can effectively enhance the generalization performance of model and alleviate instable problems of adversarial training.

Table 7
Ablation experiment results.

Task	model	ACC(%)	SEN(%)	SPE(%)	AUC(%)
AD vs. CN	3DCNN	76.97	65.28	87.50	76.39
	+SA	80.92	77.08	84.38	80.73
	+ATTC	83.55	76.39	90.00	83.19
	+Adv1	86.51	79.17	**93.13**	86.14
	+Adv2	87.83	84.72	90.63	87.67
	+MDD	89.47	88.19	90.63	89.41
	+Pro-MDD	**92.11**	**91.67**	92.50	**92.08**
AD vs. MCI	3DCNN	60.47	49.31	71.05	60.18
	+SA	63.85	67.36	60.53	63.94
	+ATTC	64.53	54.17	74.34	64.25
	+Adv1	67.91	59.72	75.66	67.69
	+Adv2	68.24	**72.92**	63.82	68.37
	+MDD	70.27	65.97	74.34	70.16
	+Pro-MDD	**76.01**	**72.92**	**78.95**	**75.93**
MCI vs. CN	3DCNN	69.87	73.03	66.88	69.95
	+SA	74.36	76.97	71.88	74.42
	+ATTC	75.32	78.29	72.50	75.39
	+Adv1	76.60	82.89	70.63	76.76
	+Adv2	77.56	76.32	78.75	77.53
	+MDD	78.21	69.08	**86.88**	77.98
	+Pro-MDD	**82.37**	**83.55**	81.25	**82.40**

Although adversarial domain adaptation can achieve effective feature alignment when faced with cross-domain data, it only completes the alignment of the marginal feature distribution, does not consider the category feature information. The results of the three classification tasks are getting better + MDD, which demonstrate MDD can effectively alleviate the difference in conditional distributions. Therefore MDD has obvious benefits in minimizing the inter-domain divergence and maximizing the intra-class density.

However, MDD does not take into account the negative impact of outliers samples on model generalization performance. If there are samples with obvious outliers, it is easy to affect the final classification performance. Results of +Pro-MDD show that the effect is more significant than + MDD, which verifies that prototype learning in MDD can effectively enhance feature alignment and improve classification performance of feature outlier samples.

5.2. Visualization via Grad-CAM

Since visualization of salient regions provides important clinical information, we propose to provide interpretable information by localizing brain regions relevant to decision-making using Gradient-weighted Class Activation Mapping (Grad-CAM) [35]. For the target domain, we visualized the results for six subjects of ADNI under PMDA (Fig. 3). We can see the model focuses on the hippocampus [3], ventricles [46], and some areas of the cortex, which are consistent with the important areas for AD diagnosis by physicians. Especially, it can be noticed that the detected key areas of AD are more obvious than those of MCI. In some extent, it verifies that structural changes caused by AD are relatively easier to be detected than MCI.

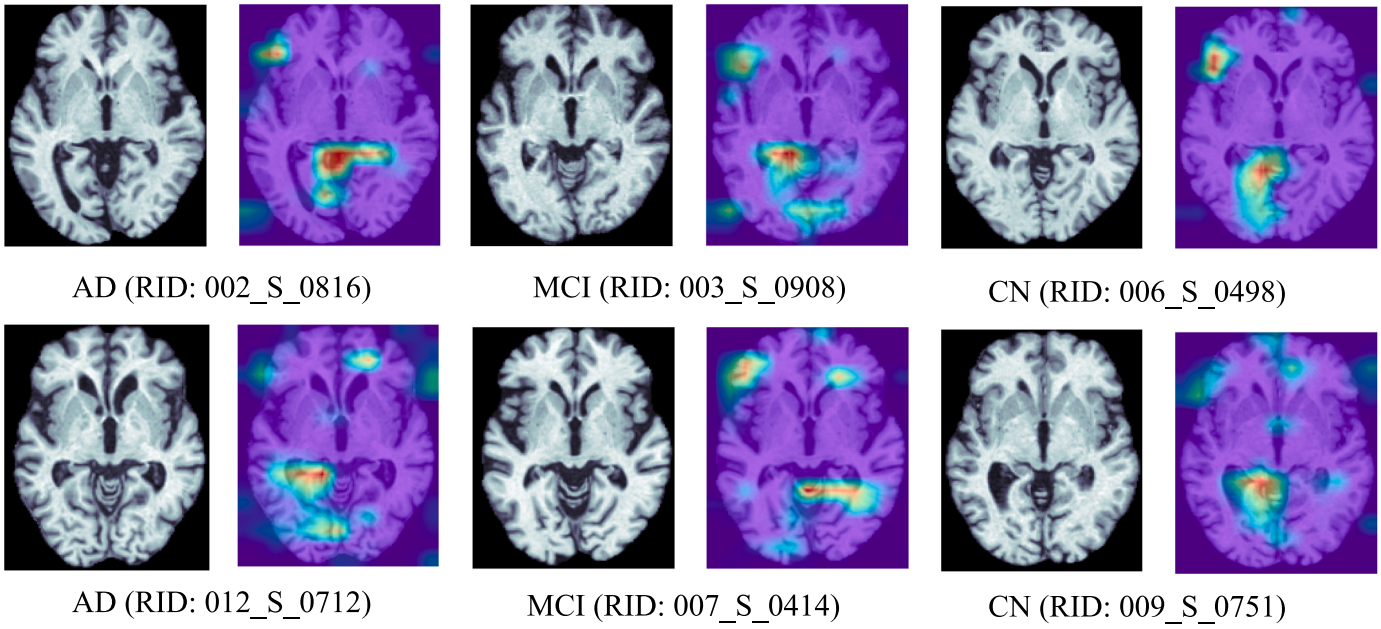

Fig. 3. Visualization via Grad-CAM in target domain. The red and purple denote the greater and lower impact on the model, respectively.

5.3. Visualization of distribution after adaptation

We use AD and CN sMRI data to show the feature distribution of source domain (3T MRI) and target domain (1.5T MRI) after domain adaptation. 584 data (AD and CN) in the source domain and target domain are used as the input of 3DResNet50 and PMDA. T-SNE algorithm [31] is employed for dimensionality reduction visualization. Feature distributions of 3DResNet50 model has performed supervised learning training on the source domain, the dimensionality reduction visualization of the features retrieved from the source domain and target domain is shown in Fig. 4(a). PMDA perform feature alignment between source domain and target domain, the visualization of the feature distribution is shown in Fig. 4(b). Fig. 4(a) show the features of AD and CN in source domain can be clearly distinguished. But for target domain, AD and CN features have mixed in a certain degree, which leads to the poor final classification. At the same time, PMDA performs feature alignment during feature extraction, the features distributions in the target domain (yellow spots and green spots) will close to the same class of the source domain (red spots and blue spots). It verifies our method has good transferability and discriminability.

5.4. Limitations and future work

Although the proposed PMDA method has achieved great performance on AD, MCI and CN classification tasks, whereas limitations of this method need to be concerned as well.

Firstly, PMDA is trained from scratch, and does not consider pretrained weights. Pretrain the existing 3D CNNs or Transformer models on the other large-scale 3D medical image datasets and fine-tune them on our datasets may further improve the classification performance. Secondly, for deep learning work, the size of the used dataset is relatively small. If more MRI data with obvious domain shift could be used, we can develop models with stronger recognition and generalization performance. Thirdly, experiments in this paper set up one source domain and one target domain. In the follow-up work, multi-source domain with single-target domain or multi-source domain with multi-target domain could be extended. Finally, due to privacy preserving policies, the training data in source domain required by most of the existing domain adaptation methods is usually unavailable. How to improve target domain performance without source domain should be in concern in the future.

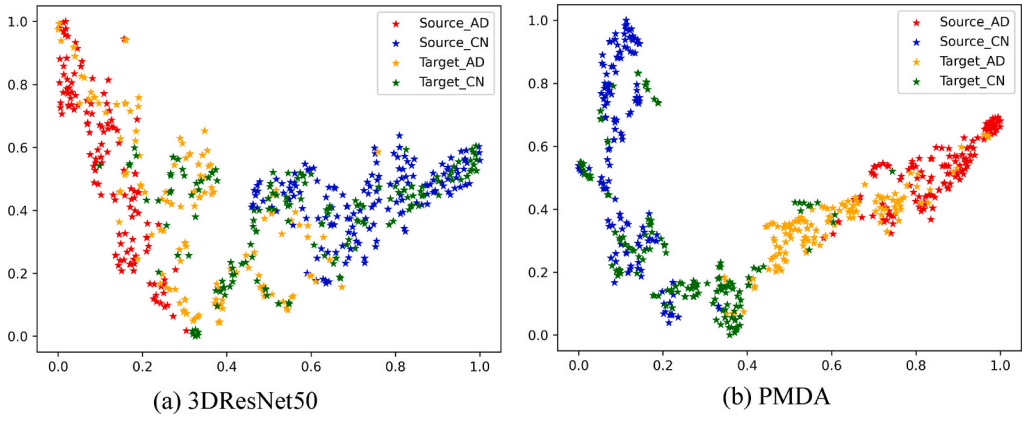

Fig. 4. Visualization of distribution of features extracted from 3DResNet50 and PMDA for source domain and target domain.

6. Conclusion

In this paper, we proposed a PMDA framework for AD, MCI and CN identification. Specifically, our method consists of three modules, i.e., a MRI multi-scale feature extraction module, which integrates convolution and self-attention for feature extraction and multi-scale fusion, a Proto-MDD module which introduces prototype learning in MDD to enhance feature alignment and alleviate the impact of outlier samples on model training, and an Adversarial Domain Adaptation module which aligns the marginal distributions of source domain and target domain and co-training two different domain discriminators to mitigate the discriminator over-fitting issue. We evaluated the proposed PMDA method on 896 sMRIs data acquired from ADNI. Experimental results show that, compared to supervised learning method such as 3DResNet50 and other state-of-the-art domain adaptation methods, the PMDA method is more effective on AD, MCI and CN classification.

Declaration of competing interest

None Declared.

Acknowledgement

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This work was financially supported by grants from Science and Technology Research Program of Chongqing Municipal Education Commission (No. KJQN202101116), Postgraduate research innovation project finance in Chongqing (No. CYS22660) and Chongqing University of Technology Joint Project (No. gzlcx20223193).

References

[1] E. Gerardin, G. Chételat, M. Chupin, et al., Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging[J], Neuroimage 47 (4) (2009) 1476–1486.
[2] W.H. Organization, Global Action Plan on the Public Health Response to Dementia 2017–2025[R], 2017.
[3] F. Li, M. Liu, A.S.D.N. Initiative, A hybrid convolutional and recurrent neural network for hippocampus analysis in Alzheimer's disease[J], J. Neurosci. Methods 323 (2019) 108–118.
[4] L. Kang, J. Jiang, J. Huang, et al., Identifying early mild cognitive impairment by multi-modality MRI-based deep learning[J], Front. Aging Neurosci. 12 (2020), 206-206.
[5] J. Wolleb, R. Sandkühler, M. Barakovic, N. Hadjikhani, A. Papadopoulou, Ö. Yaldizli, J. Kuhle, C. Granziera, P.C. Cattin, Learn to ignore: domain adaptation for multi-site MRI analysis[C], in: In Medical Image Computing and Computer Assisted Intervention (MICCAI 2022): 25th International Conference, 2022, pp. 725–735.
[6] S.G. Finlayson, A. Subbaswamy, K. Singh, et al., The clinician and dataset shift in artificial intelligence[J], N. Engl. J. Med. (2020) 283–286.
[7] T. Fernando, H. Gammulle, S. Denman, et al., Deep learning for medical anomaly detection–a survey[J], ACM Comput. Surv. 54 (7) (2021) 1–37.
[8] W. Zhu, L. Sun, J. Huang, et al., Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI[J], IEEE Trans. Med. Imag. 40 (9) (2021) 2354–2366.
[9] H. Guan, M. Liu, Domain adaptation for medical image analysis: a survey[J], IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 69 (3) (2022) 1173–1185.
[10] O.B. Ahmed, M. Mizotin, J. Benois-Pineau, et al., Alzheimer's disease diagnosis on structural MR images using circular harmonic functions descriptors on hippocampus and posterior cingulate cortex[J], Comput. Med. Imag. Graph. 44 (2015) 13–25.
[11] I. Beheshti, N. Maikusa, M. Daneshmand, et al., Classification of Alzheimer's disease and prediction of mild cognitive impairment conversion using histogram-based analysis of patient-specific anatomical brain connectivity networks[J], J. Alzheim. Dis. 60 (1) (2017) 295–304.
[12] K. Aderghal, M. Boissenin, J. Benois-Pineau, et al., Classification of sMRI for AD Diagnosis with Convolutional Neuronal Networks: A Pilot 2-D+ε Study on ADNI [C], MultiMedia Modeling (2017) 690–701.
[13] K. Oh, Y.-C. Chung, K.W. Kim, et al., Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning[J], Sci. Rep. 9 (1) (2019) 1–16.
[14] S. Korolev, A. Safiullin, M. Belyaev, et al., Residual and Plain Convolutional Neural Networks for 3D Brain MRI classification[C], 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017), 2017, pp. 835–838.
[15] H. Guan, Y. Liu, E. Yang, et al., Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification[J], Med. Image Anal. 71 (2021), 102076.
[16] W. Li, Y. Zhao, X. Chen, et al., Detecting Alzheimer's disease on small dataset: a knowledge transfer perspective[J], IEEE J. Biomed. Health Inf. 23 (3) (2018) 1234–1242.
[17] X. Pan, C. Ge, R. Lu, et al., On the integration of self-attention and convolution[C], Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (2022) 815–825.
[18] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks[C], Proc. IEEE Conf. Comput. Vision Pattern Recognit. (2018) 7132–7141.
[19] J. Hu, L. Shen, S. Albanie, et al., Gather-excite: exploiting feature context in convolutional neural networks[J], Adv. Neural Inf. Process. Syst. (2018) 31.
[20] J. Park, S. Woo, J.-Y. Lee, I.S. Kweon, BAM: Bottleneck attention module[J], arXiv (2018) preprint arXiv:1807.06514.
[21] S. Woo, J. Park, J.-Y. Lee, et al., Cbam: convolutional block attention module[C], Proc. Eur. conf. Comput. Vis. (ECCV) (2018) 3–19.
[22] A. Srinivas, T.-Y. Lin, N. Parmar, et al., Bottleneck transformers for visual recognition[C], Proc. IEEE/CVF conf. Comput. Vis. Pattern Recognit. (2021) 16519–16529.
[23] O. Li, H. Liu, C. Chen, et al., Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains its predictions[C], in: Proceedings of the AAAI Conference on Artificial Intelligence, 32, 2018, pp. 3530–3537 (1).
[24] Y. Shu, Y. Shi, Y. Wang, et al., p-oDn: prototype-based open Deep network for open Set Recognition[J], Sci. Rep. 10 (1) (2020) 1–13.
[25] T. Gao, X. Han, Z. Liu, et al., Hybrid attention-based prototypical networks for noisy few-shot relation classification[C], Proc. AAAI Conf. Artif. Intell. (2019) 6407–6414.
[26] K. Wang, J.H. Liew, Y. Zou, et al., Panet: few-shot image semantic segmentation with prototype alignment[C], Proc. IEEE/CVF Int. Conf. Comput. Vis. (2019) 9197–9206.
[27] K. Tanwisuth, X. Fan, H. Zheng, et al., A prototype-oriented framework for unsupervised domain adaptation[J], Adv. Neural Inf. Process. Syst. 34 (2021) 17194–17208.
[28] H.-M. Yang, X.-Y. Zhang, F. Yin, et al., Convolutional prototype network for open set recognition[J], IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2022), 2358-2370.
[29] E. Kondrateva, M. Pominova, E. Popova, et al., Domain shift in computer vision models for mri data analysis: an overview[C], Thirteenth Int. Conf. Machine Vis. 11605, (2021) 126–133.
[30] B. Glocker, R. Robinson, D.C. Castro, et al., Machine Learning with Multi-Site Imaging Data: an Empirical Study on the Impact of Scanner effects[J], 2019 arXiv preprint arXiv:1910.04597.
[31] L. Van Der Maaten, G. Hinton, Visualizing Data using t-SNE[J], J. Mach. Learn. Res. 9 (2008) 2579–2605.
[32] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need[C], Proc. 31st Int. Conf. Neural Inf. Process. Syst. (2017) 6000–6010.
[33] J. Li, E. Chen, Z. Ding, et al., Maximum density divergence for domain adaptation [J], IEEE Trans. Pattern Anal. Mach. Intell. 43 (11) (2021) 3918–3930.
[34] G. Wei, C. Lan, W. Zeng, et al., ToAlign: task-oriented alignment for unsupervised domain adaptation[J], Adv. Neural Inf. Process. Syst. 34 (2021) 13834–13846.
[35] R.R. Selvaraju, M. Cogswell, A. Das, et al., Grad-cam: visual explanations from deep networks via gradient-based localization[C], Proc. IEEE Int. Conf Comput. Vis. (2017) 618–626.
[36] Y. Ganin, E. Ustinova, H. Ajakan, et al., Domain-adversarial training of neural networks[J], J. Mach. Learn. Res. 17 (1) (2016), 2096-2030.
[37] R. Pascanu, T. Mikolov, Y. Bengio, Understanding the exploding gradient problem [J], CoRR (2012) 1, abs/1211.5063 2.417.

[38] J. Huang, D. Guan, A. Xiao, et al., Cross-view regularization for domain adaptive panoptic segmentation[C], Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (2021) 10133–10144.
[39] K. Cui, J. Huang, Z. Luo, et al., GenCo: Generative Co-training for Generative Adversarial Networks with Limited Data[C], in: Proceedings of the AAAI Conference on Artificial Intelligence, 36, 2021, pp. 499–507 (1).
[40] D.P. Kingma, J. Ba, Adam: A Method for Stochastic optimization[J], 2014 arXiv preprint arXiv:1412.6980.
[41] R. Prajapati, G.-R. Kwon, A binary classifier using fully connected neural network for Alzheimer's disease classification[J], J. Multimedia Inf. Syst. 9 (1) (2022) 21–32.
[42] Z. Xia, et al., A Novel End-to-End Hybrid Network for Alzheimer's Disease Detection Using 3D CNN and 3D CLSTM[C], in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 2020, pp. 1–4, https://doi.org/10.1109/ISBI45749.2020.9098621.
[43] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition[C], Proc. IEEE Conf. Comput. Vision Pattern Recognit. (2016) 770–778.
[44] C. Yu, J. Wang, Y. Chen, M. Huang, Transfer Learning with Dynamic Adversarial Adaptation Network[C], in: 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, 2019, pp. 778–786, https://doi.org/10.1109/ICDM.2019.00088.
[45] M. Long, Z. Cao, J. Wang, et al., Conditional adversarial domain adaptation[C], Proc. 32nd Int. Conf. Neural Inf. Process. Syst. (2018) 1647–1657.
[46] A. Bartos, D. Gregus, I. Ibrahim, et al., Brain volumes and their ratios in Alzheimer s disease on magnetic resonance imaging segmented using Freesurfer 6.0[J], Psychiatr. Res. Neuroimaging 287 (2019) 70–74.