Dual Attention Multi-Instance Deep Learning for Alzheimer's Disease Diagnosis With Structural MRI

Wenyong Zhu, Liang Sun, Jiashuang Huang, Liangxiu Han, and Daoqiang Zhang, *Member, IEEE*

***Abstract*— Structural magnetic resonance imaging (sMRI) is widely used for the brain neurological disease diagnosis, which could reflect the variations of brain. However, due to the local brain atrophy, only a few regions in sMRI scans have obvious structural changes, which are highly correlative with pathological features. Hence, the key challenge of sMRI-based brain disease diagnosis is to enhance the identification of discriminative features. To address this issue, we propose a dual attention multi-instance deep learning network (DA-MIDL) for the early diagnosis of Alzheimer's disease (AD) and its prodromal stage mild cognitive impairment (MCI). Specifically, DA-MIDL consists of three primary components: 1) the Patch-Nets with spatial attention blocks for extracting discriminative features within each sMRI patch whilst enhancing the features of abnormally changed micro-structures in the cerebrum, 2) an attention multi-instance learning (MIL) pooling operation for balancing the relative contribution of each patch and yield a global different weighted representation for the whole brain structure, and 3) an attention-aware global classifier for further learning the integral features and making the AD-related classification decisions. Our proposed DA-MIDL model is evaluated on the baseline sMRI scans of 1689 subjects from two independent datasets (i.e., ADNI and AIBL). The experimental results show that our DA-MIDL model can identify discriminative pathological locations and achieve better classification performance in terms of accuracy and generalizability, compared with several state-of-the-art methods.**

***Index Terms*— Alzheimer's disease diagnosis, discriminative pathological location, multi-instance learning, attention mechanism, convolutional neural network, sMRI.**

I. Introduction

ALZHEIMER'S disease (AD) is one of the most prevalent neurological diseases with a significant growth rate in incidence [1]. The progression of AD gradually results in memory deterioration and impairment of cognitive functions, ultimately leading to irreversible neuron injury [2]. Although no treatment has been proven to be effective in preventing the progression of AD [3], the early diagnosis of AD still remains important to subsequent treatments to delay the onset of cognitive symptoms [4]. Specifically considering that the atrophic process occurs even earlier than the appearance of amnestic symptoms [5], many studies [6]–[13] based on machine learning methods are developed to identify anatomical differences between Alzheimer's disease (AD) patients and normal controls (NC), and predict the progression of mild cognitive impairment (MCI) using structural magnetic resonance imaging (sMRI), which are sensitive to morphological changes caused by brain atrophy [14].

The conventional sMRI-based AD diagnosis methods usually partition the entire MR image into multiple regions with different scales for better feature extraction of local abnormal brain structural changes [15]–[18]. Based on the partition with different scales, most of the existing sMRI-based studies can be roughly divided into three categories, including 1) voxel-level, 2) region-level and 3) patch-level. In voxel-level methods [19]–[22], the tissue features (e.g., gray matter densities) extracted from sMRI scans compose high-dimensional voxel-wise structural features for AD diagnosis. However, compared with the dimensionality of features, the number of training images for AD classification is too small, which often leads to the curse of dimensionality. To alleviate this problem, region-level methods [7], [11], [12], [23] are proposed to identify the AD patients from normal controls with the handcrafted features (e.g. gray matter, cerebrospinal fluid and cortical thickness) derived from segmented regions of interest (ROIs). However, these methods are resource-intensive for segmenting ROIs. In contrast, patch-level (an intermediate scale between voxel-level and region-level) feature representations [24]–[27] are proposed for more effectively characterizing the local structural changes in MR images. Specifically, the centers

Manuscript received March 29, 2021; revised April 26, 2021; accepted April 28, 2021. Date of publication May 3, 2021; date of current version August 31, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61861130366, Grant 61876082, Grant 61732006, and Grant 62006115; in part by the National Key Research and Development Program of China under Grant 2018YFC2001600, Grant 2018YFC2001602, and Grant 2018ZX10201002; and in part by the Royal Society-Academy of Medical Sciences Newton Advanced Fellowship under Grant NAF\R1\180371. *(Wenyong Zhu and Liang Sun are co-first authors.) (Corresponding author: Daoqiang Zhang.)*

Wenyong Zhu, Liang Sun, Jiashuang Huang, and Daoqiang Zhang are with the MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: dqzhang@nuaa.edu.cn).

Liangxiu Han is with the Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M1 5GD, U.K.

This article has supplementary downloadable material available at https://doi.org/10.1109/TMI.2021.3077079, provided by the authors.

Digital Object Identifier 10.1109/TMI.2021.3077079

of patches can be located by certain anatomical landmark detectors [24] or statistics methods [25]. However, how to combine the local patches into a global feature representation for the whole brain structure is still a challenge in patch-level methods.

In recent years, deep learning methods have shown great success in image classification tasks such as medical imaging analysis. For instance, deep convolutional neural networks (CNNs) are empirically verified to have the excellent ability to learn high-level features from sMRI data, and greatly improve the performance of brain disease diagnosis with the efforts of many researchers [28]–[35]. However, most existing deep learning methods for AD diagnosis still rely on the manual pre-defined ROIs with experts' experience to build diagnosis models based on CNNs, which leads to insufficient consideration of individual differences using the same template space and may not include the entire disease-related atrophy features distributed in the whole brain. Moreover, due to the black box characteristics of neural networks, few deep learning methods have specific output for pathological locations, which neglects the issues of interpretability in medical practice. Since brain atrophy usually occurs locally, only a few regions in sMRI scans have obvious structural changes which are highly correlative with pathological features, while the rest of regions have little useful information for distinction. Therefore, the key challenge of deep learning-based diagnosis with sMRI is to enhance the identification of discriminative features, including 1) informative micro-structures within local regions and 2) relatively important regions in a global image.

To address aforementioned challenges, we propose a dual attention multi-instance deep learning model (DA-MIDL) to identify discriminative pathological locations for AD diagnosis. Specifically, as illustrated in Fig. 1, DA-MIDL consists of three major components, i.e., the Patch-Nets, the attention multi-instance learning (MIL) pooling module and the attention-aware global classifier. Through the Patch-Nets with spatial attention blocks, DA-MIDL could learn discriminative structural features from multiple local sMRI patches distributed in the brain. Then through the attention MIL pooling, all the patch-level features are given different weights and combined into a global feature representation for the whole brain structure information, based on which finally a global classifier is constructed for AD diagnosis. We have evaluated the proposed method on two public datasets (i.e., ADNI and AIBL) and the experimental results on multiple AD-related classification tasks (e.g., AD classification and MCI conversion prediction) demonstrate that our DA-MIDL method outperforms several state-of-the-art methods in terms of accuracy performance and generalizability. Different from the existing approaches, our major contributions can be summarized as follows.

1) A dual attention multi-instance deep learning model (DA-MIDL) is proposed for improving AD diagnosis performance, which can automatically capture local and global structural features from sMRI scans and make AD-related classification decisions in a unified framework.
2) The Patch-Nets with spatial attention blocks are designed to extract discriminative features within each patch and to enhance the local features of abnormally changed micro-structures caused by atrophy in the brain.
3) An attention multi-instance learning (MIL) pooling operation is proposed to balance the relative contribution of each patch and yield a global different weighted feature representation for the whole brain structure.

The rest of the paper is organized as follows: Section II introduces the related works; Section III describes the studied materials and our proposed DA-MIDL method; Section IV shows the experimental settings and results for multiple AD diagnosis tasks compared with several state-of-the-art methods; Section V presents the discussion on the effectiveness of our attention modules, identified pathological locations and limitations; Section VI concludes the work.

II. Related Work

In this section, we briefly introduce previous studies on computer-aided AD diagnosis methods with sMRI data. Then we respectively review multi-instance learning and attention mechanism related works in medical imaging analysis.

A. Alzheimer's Disease Diagnosis With sMRI

According to the partition of ROIs from sMRI scans, the previous brain disease diagnosis studies could be roughly divided into three categories, including voxel-level, region-level, and patch-level methods.

The voxel-level methods [19]–[22] aimed at distinguishing disease-related microstructures in MR images of the patients and normal controls. In a voxel-wise manner, the tissue (e.g., gray matter and white matter) densities were generally measured as features for the classification algorithms. However, only analyzing features on isolated voxels would lead to the ignorance of the high correlation between voxels. Another limitation of voxel-level methods was the overfitting problem, because the voxel-level feature representation always had a high dimensionality compared with the small number of subjects for model training. Therefore, feature dimension reduction was the main challenge of voxel-level methods for improving the performance of AD classification. In [36], a sparse coding method with a hierarchical tree-guided regularization was adopted to identify the relevant biomarkers (i.e., voxel-wise gray matter density) with structured sparsity from MR images for brain disease classification. In [37], an incremental learning-based method for AD diagnosis was proposed to effectively reduce the dimension of data and achieve robustness to noises by filtering out high frequency components of the voxel-wise cortical thickness data.

In contrast, region-level methods were based on the pre-segmented ROIs, which had much lower feature dimensionality than voxel-level methods. For instance, the volumetric features were extracted from 93 ROIs automatically labeled by an atlas warping algorithm and a linear support vector machine (SVM) was used for AD classification [11]. The hippocampal features were segmented from sMRI scans for AD diagnosis and MCI conversion prediction [20], since

the hippocampus is usually affected at the earliest stage of AD. Ensemble classification models were constructed based on multiple sets of regional gray matter density features from multiple spatially normalized template spaces for AD and MCI diagnosis [12]. In [38], a multi-kernel-based method combined with Marginal Fisher Analysis was proposed to achieve the sparsity of ROIs for dimensionality reduction and capture the complicated relationship between MRI features and the disease status. However, the definition and segment of ROIs were resource-intensive due to the requirement of experts' experience. Furthermore, most region-level methods [7], [11], [12], [23] only used part of the handcrafted features (e.g., gray matter, white matter, cerebrospinal fluid and cortical thickness) extracted from the ROIs, which may not include the entire disease-related features.

As an intermediate scale between voxel-level and region-level, patch-level methods [24]–[27] were proposed for more effectively capturing the local structural changes in MR images. For instance, many weak classifiers were constructed based on the features extracted from randomly sampled patches in MR images and were combined to make a final decision for AD diagnosis [26]. The graph representations were measured based on squared Euclidean distance between intensity features of patches, and then the SVM was used for classification [8]. In [27], a fully convolutional network (FCN) was trained on the randomly sampled patches from the full MRI volumes. Based on the trained FCN, the voxels of high-risk were selected and fed to the multilayer perceptron (MLP) for individual-level AD classification. However, the spatial correlation among patches were processed inadequately in these works. A hierarchical full convolutional neural network was proposed for AD diagnosis [6], which could learn multi-scale feature representations (e.g., patch-level, region-level and subject-level) from sMRI scans. Then a pruning strategy was used to remove uninformative patches and cut down the learnable parameters. However, it may lead to the loss of potential spatial correlation between the removed patches and left patches. Therefore, highlighting the discriminative features while retaining the spatial correlation among patches is still a challenge in patch-level methods.

B. Multi-Instance Learning

In multi-instance learning (MIL) [39], one sample consists of multiple observed instances and is only annotated with a general category. That is, the training set can be regarded as a set of labeled bags, where each bag contains multiple unlabeled instances. Specifically, one positive labeled bag contains at least one positive instance. In addition, the positive labeled bag may contain negative instances or useless instances which are irrelevant to the label of the bag. While, all the instances in the negative bags are negative. The main task of MIL is to predict the labels of unseen bags.

Multi-instance learning performs well in the computer-aided medical diagnosis domain [40]–[45]. For example, a novel MIL framework MIS-Boost was employed for the identification of cerebral small vessel disease, using the intensity patches from regions with high probability of containing lesions in CT images [43]. A new method as MIL pooling was proposed based on the quantile function to aggregate the predictions from smaller regions into an image-level classification for breast tumor histology [40]. A deep MIL model was proposed for AD diagnosis, which simply concatenated the local features learned from sub-CNNs for the global feature representation of whole brain structure [10]. However, how to combine the instance-level features into a global bag-level feature representation is still a challenge in MIL.

C. Attention Mechanism

Since the features of different parts make different contributions to the overall classification performance, the attention mechanism has been proposed to automatically find and highlight the most informative points on feature maps for boosting the performance of image classification [46], [47]. Specifically, the attention modules can learn task-oriented different weighted feature maps for subsequent representation learning and classification.

Recently, the attention mechanism has been widely used in the medical imaging analysis domain [48]–[51]. Different task-oriented attention modules were proposed to help classification or segmentation models to enhance the features of disease-related regions in images. For instance, a weakly-supervised attention network was proposed for dementia status prediction [48], which consisted of a fully convolutional network, a trainable dementia attention block and a multi-task regression block. A cross-attention model was designed to find the areas with high pathogenic chances and eliminate noises for thoracic disease diagnosis [49]. A channel attention module was integrated into the conventional residual block to extract more informative features for improving tissue quantification in fingerprinting [50].

Although both attention mechanism and multi-instance learning have good performance in the field of medical imaging analysis, there are few studies to combine these two methods. In MIL, the key stage is the combination of instance-level features into a global bag-level feature representation. It may be unreasonable to combine the instance-level features equally, since different instances contain different amounts of information. Therefore, the attention mechanism can be used to estimate the weight of each instance. To this end, we propose a dual attention multi-instance deep learning model (DA-MIDL) for identifying discriminative pathological locations and AD diagnosis with structural MRI data.

III. Materials and Method

In this section, we first present materials used in our study. Then we introduce the proposed DA-MIDL method, including the overall architecture, key components and loss function based on multi-instance learning and attention mechanisms. Finally, we provide the implementation details.

A. Subjects and Image Pre-Processing

Two datasets (i.e., ADNI and AIBL) used in our study are acquired from the public Alzheimer's Disease Neuroimaging

TABLE I
DEMOGRAPHIC DETAIL OF THE STUDIED SUBJECTS INCLUDING DATASET, GROUP TYPE, GENDER, AGE, MINI-MENTAL STATE EXAMINATION (MMSE) AND CLINICAL DEMENTIA RATING (CDR)

Dataset	Group Type	Gender (Male/Female)	Age (Mean±Std)	MMSE (Mean±Std)	CDR (Mean±Std)
ADNI	AD	202/187	75.13±7.86	23.28±2.03	0.75±0.25
	pMCI	105/67	75.73±7.05	26.59±1.71	0.50±0.00
	sMCI	155/77	76.40±7.94	27.27±1.78	0.49±0.04
	NC	202/198	73.85±6.38	29.10±1.01	0.00±0.00
AIBL	AD	33/46	73.34±7.77	20.42±5.46	0.95±0.51
	pMCI	9/8	75.29±6.16	26.24±2.04	0.47±0.13
	sMCI	48/45	74.67±7.21	27.23±2.08	0.46±0.12
	NC	134/173	73.12±6.19	28.77±1.25	0.02±0.20

Initiative (ADNI) database (http://adni.loni.usc.edu) and Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL) database (https://aibl.csiro.au). In the ADNI dataset, there are totally 1193 1.5T/3T T1-weighted structural MRI (sMRI) scans from subjects at their own baseline/screening visit (i.e., the first examination) across three ADNI phases (i.e., ADNI-1, ADNI-2 and ADNI-3). These subjects can be divided into three categories: AD (Alzheimer's disease), MCI (mild cognitive impairment) and NC (normal control) in accordance with the standard clinic criteria, such as Mini-Mental State Examination (MMSE) scores and Clinical Dementia Rating (CDR). For MCI conversion prediction, MCI subjects can be further categorized into two classes: pMCI (progressive MCI subjects who had converted to AD within 36 months after baseline visit) and sMCI (stable MCI subjects who were continuously diagnosed as MCI for 36 months after baseline visit). The studied ADNI dataset contains 389 AD, 172 pMCI, 232 sMCI and 400 NC subjects. The AIBL dataset consists of baseline sMRI scans from 496 different subjects, including 79 AD, 17 pMCI, 93 sMCI and 307 NC subjects. The demographic detail of these 1689 subjects from the ADNI and AIBL datasets is shown in Table I.

The original structural MRI data downloaded from ADNI are pre-processed for subsequent better feature learning and classification. First, the original images in 3D Neuroimaging Informatics Technology Initiative (NIfTI) format are standardized through geometry correction for gradient nonlinearity by 3D gradwarp correction [52] and intensity correction for non-uniformity by B1 non-uniformity correction [53]. Then, we perform linear registration to the Colin27 template [54] to remove global linear differences (including global translation, scale, and rotation differences) and skull-stripping on all the structural MR images respectively using 'flirt' instruction with default parameters (e.g., DOF (degrees of freedom) as 12 and Correlation Ratio as cost function) and 'bet' instruction with default fractional intensity threshold (0.5) in FSL toolbox [55]. After image normalization to the Colin27 standard space, MR images have a size of $181 \times 217 \times 181$ voxels.

B. Overall Architecture Based on Multi-Instance Learning

We regard the patch-level brain morphometric pattern analysis for AD diagnosis as a multi-instance problem and construct our model based on multi-instance learning. In MIL, the training data is a set of bags, i.e., $D = \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{N}$, where $\boldsymbol{X}_i$ is the i-th sample/bag, Y_i is the bag-level label of $\boldsymbol{X}_i$, and N is the number of bags. Each bag contains multiple unlabeled instances, i.e., $\boldsymbol{X}_i = \{\boldsymbol{I}_{i,j}\}_{j=1}^{N_i}$, where $\boldsymbol{I}_{i,j}$ is the j-th instance, N_i is the number of instances in $\boldsymbol{X}_i$. Besides, in a positive bag there is at least one positive instance while all the instances in a negative bag are negative. We denote $Y_i = 0$ only when $\sum_{j=0}^{N_i} y_{i,j} = 0$, where $y_{i,j}$ represents instance-level label of $\boldsymbol{I}_{i,j}$, otherwise $Y_i = 1$.

Brain abnormal atrophy occurs at few local regions, especially at the early stage of AD [56]–[58]. To this end, we regard the bag of patches from a certain patient's MR image as a positive bag. Correspondingly, we group the patches from a normal control into a negative bag. Thus, the bags of multiple patches with bag-level labels take the place of whole large images as the training data for AD-related diagnosis.

Our proposed DA-MIDL model (shown in Fig. 1) contains four key steps, including the selection of instances for composing a bag $\boldsymbol{X}$ (i.e., Patch Location Proposals described in Section III-C), a transform f of instance-level features (i.e, Patch-Net described in Section III-D), a combination ϕ of transformed instances (i.e, Attention MIL Pooling described in Section III-E), a classification g based on the combined bag-level features (i.e, Attention-Aware Global Classifier described in Section III-F). The probability Θ of positive category is expressed as: $\Theta(\boldsymbol{X}) = g\phi f(\boldsymbol{X})$.

C. Patch Location Proposals

Patch location proposals are used to initially select patches from sMRI scans as input to our model. Inspired by the patch extraction in [25], we propose a novel method, considering the group comparison on patch-level features instead of voxel-wise features. In our method, we first uniformly divide the MR images into multiple cubic patches with a fixed size (e.g., $W \times W \times W$) without overlapping in order to simplify calculations and avoid redundant information. However, not all the partitioned patches are related to abnormal atrophy caused by AD. The t-test is a method to identify the significance of the difference between the experimental group and the control group. In our experiment, the patch locations with more significant differences between AD group and NC group are more likely to be the brain regions with abnormal atrophy. Thus, we apply the t-tests to sort the informativeness in all patches. We calculate the average of the voxel-wise features in one patch as its patch-level feature. Then we make a group comparison on two groups of patch-level features at one patch location respectively from the same amount of AD patients and normal controls in training set using a t-test. So we can obtain a p-value at this patch location, which can represent the informativeness of this location. All the p-values respectively calculated on all the locations are normalized by $\frac{pvalue-MIN}{MAX-MIN}$ and form a p-value map covering the whole brain MR image. Additionally, the locations with smaller p-values are roughly considered to have higher discrimination. According to the p-value map, we select a number of patches in one image

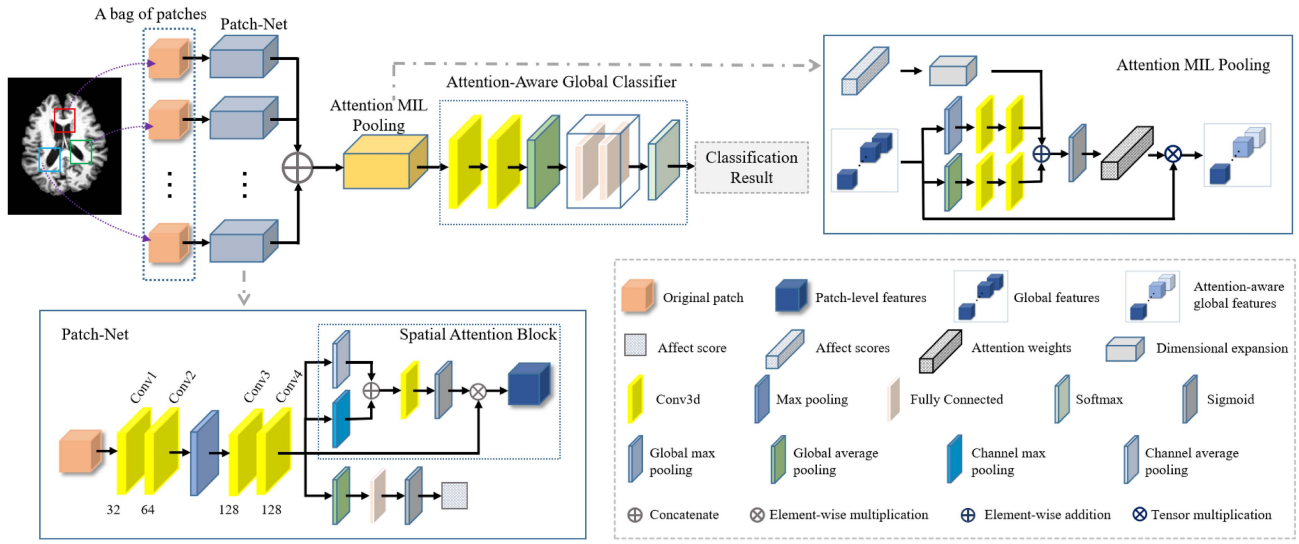

Fig. 1. Illustration of our dual attention multi-instance deep learning network (DA-MIDL), which consists of Patch-Nets with spatial attention blocks, attention MIL pooling and attention-aware global classifier.

at the locations with the smallest p-values to compose a bag (e.g., $\boldsymbol{X} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, \cdots, \boldsymbol{I}_k\}$, where $\boldsymbol{I}_i \in \mathbb{R}^{W\times W\times W}$ and k is the number of selected patches) as input to our model.

D. Patch-Net With Spatial Attention Block

Fig. 1 shows the structure of Patch-Net with spatial attention block. There are two tasks in Patch-Net, including 1) learning a spatial attention-aware patch-level feature representation, and 2) outputting an affect score which indicates the ability of triggering the bag label. Spatial attention blocks are used for feature enhancement of discriminative parts in fixed-size patches. Specifically, all the Patch-Nets in our DA-MIDL method have the same architecture.

1) Patch-Net: The former part as a backbone of Patch-Net aims to learn more abstract feature representations from original patches and reduce the size of feature maps. It consists of four 3D convolutional layers and a max pooling in the middle for adapting the size of input patches. The first convolutional layer has a kernel size of $4\times4\times4$. The last three convolutional layers have the same filter size of $3 \times 3 \times 3$. The max pooling has a filter size of $2 \times 2 \times 2$ with 2-unit-length stride for down-sampling. In detail, the number of channels from conv1 to conv4 is 32, 64, 128 and 128 sequentially. All the convolutional layers are trained in a unit stride with non-zero-padding feature maps. Each convolutional layer is followed by batch normalization (BN) and rectified linear unit (ReLU) activations. Based on the feature maps output from conv4, the Patch-Net extends to two branching modules. One is the spatial attention block (Section III-D2) for learning a spatial attention-aware patch-level representation (whose size is $C \times w \times w \times w$, where C is the number of channels and $w \times w \times w$ is the size of the feature maps). The other module (including a global average pooling, a fully connected layer and a sigmoid function) aims to produce an affect score which is used to identify the discriminative pathological locations. Instead of generating one-dimensional feature vectors in most existing instance-level transforms in MIL, the local patch-level features output from Patch-Nets maintain the three-dimensional shape for the better combination of patches and further learning of the spatial relationship among patches.

2) Spatial Attention Block: Inspired by the spatial attention module proposed in [47], we design our spatial attention block embedded into our Patch-Net to adapt the local structural feature extraction from 3D image patches. The architecture of the spatial attention block is also shown in Fig. 1. Two different pooling along the channel axis (i.e., channel max pooling and channel average pooling) are adopted to generate two feature maps in the name of the max features and average features respectively. Then the two feature maps are concatenated with a size of $2 \times w \times w \times w$ as the input of the subsequent convolutional layer (stride: 1, kernel size: $3\times3\times3$, padding: 1 for maintaining the size of feature maps). The output of the convolutional layer can be regarded as a spatial attention map ($\mathbb{A}_{spatial} \in \mathbb{R}^{w\times w\times w}$, the same size as the feature maps from conv4) where the attention score at each location is limited to the range of 0 to 1 through the sigmoid layer. The spatial attention map describes the spatially-varying contributions of different parts in a patch, which reveals which part to emphasize or suppress in feature representations. Each feature map in the output of conv4 is multiplied at element wise with the calculated attention map $\mathbb{A}_{spatial}$ so that the local spatial attention-aware structural representations are generated ultimately. Then we explain the proposed spatial attention block with several formulas.

We denote the output of conv4 as $\boldsymbol{F} = \{\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_C\}$, where $\boldsymbol{F}_i \in \mathbb{R}^{w\times w\times w}$ and C is the number of channels. Max pooling along channel axis can be expressed as

$$\boldsymbol{F}_{max} = ChannelMaxPooling(\boldsymbol{F}), \tag{1}$$

where $\boldsymbol{F}_{max}^{w,h,l} = max\{\boldsymbol{F}_1^{w,h,l}, \boldsymbol{F}_2^{w,h,l}, \cdots, \boldsymbol{F}_C^{w,h,l}\}$. Average pooling along channel axis is denoted as

$$\boldsymbol{F}_{average} = ChannelAveragePooling(\boldsymbol{F}), \tag{2}$$

where $\boldsymbol{F}_{average}^{w,h,l} = \frac{1}{C}\sum_{c=1}^{C} \boldsymbol{F}_{c}^{w,h,l}$. Then we concatenate the two feature maps and calculate a spatial attention map.

$$\mathbb{A}_{spatial} = \sigma(\mathbb{W}([\boldsymbol{F}_{max}; \boldsymbol{F}_{average}])), \tag{3}$$

where σ is sigmoid activation, $\mathbb{W}$ is the weight of the convolutional layer and [;] is concatenation. The patch-level spatial-attention-aware feature representation $\mathbf{F}$ is denoted as

$$\mathbf{F} = [\boldsymbol{F}_1 \otimes \mathbb{A}_{spatial}; \cdots; \boldsymbol{F}_C \otimes \mathbb{A}_{spatial}], \tag{4}$$

where $\otimes$ is element-wise multiplication.

E. Attention MIL Pooling

We also propose an attention MIL pooling operation for learning a patch-attention map which indicates relative contribution of each patch. The architecture of attention MIL pooling is also shown in Fig. 1.

Each patch-level structural representation $\mathbf{F} \in \mathbb{R}^{C\times w\times w\times w}$ output from Patch-Net are firstly compressed by average-pooling along channel axis to $\bar{\mathbf{F}} \in \mathbb{R}^{1\times w\times w\times w}$. Then, the compressed patch-level feature representations are concatenated to the global feature representation as $\mathbf{F}_{global} = \{\bar{\mathbf{F}}_1, \bar{\mathbf{F}}_2, \cdots, \bar{\mathbf{F}}_C\}$, where C is also the number of patches and $\bar{\mathbf{F}}_i$ represents the patch-level features of the i-th input patch. The global average pooling (GAP) and global max pooling (GMP) are constructed in parallel for generating two different feature descriptors, since it is empirically confirmed that exploiting both above feature descriptors can improve representation power of networks rather than only adopting one of them [47]. Then the two descriptors are respectively further learned by corresponding two $1 \times 1 \times 1$ convolutional layers to produce two patch-attention maps.

$$\boldsymbol{A}_{average} = \mathbb{W}_1 ReLU(\mathbb{W}_0 GAP(\mathbf{F}_{global})) \tag{5}$$

$$\boldsymbol{A}_{max} = \mathbb{W}_1 ReLU(\mathbb{W}_0 GMP(\mathbf{F}_{global})) \tag{6}$$

We respectively use $\mathbb{W}_0$, $\mathbb{W}_1$ as the weights of the convolutional layers. Herein, the convolutional layers in processing the average feature descriptor share the parameters with the convolutional layers in processing the max feature descriptor. Apart from the two patch-attention maps learned from inter-patch relationships, the affect score learned from each intra-patch feature is also considered to estimate the contribution of each patch. The affect scores from all the Patch-Nets form an affect vector $\boldsymbol{a} = \{a_1, a_2, \cdots, a_C\}$, where C is the number of patches. The affect vector is extended to the same size as the patch-attention maps. Thus, the three different attention maps can be merged into a comprehensive patch-attention map $\mathbb{A}_{patch}$ by element-wise summation, which is activated by sigmoid function σ afterwards.

$$\mathbb{A}_{patch} = \sigma(\boldsymbol{A}_{average} + \boldsymbol{A}_{max} + \boldsymbol{a}) \tag{7}$$

Finally, the previous global representations are multiplied with the patch-attention map to produce the attention-aware global feature representation $\mathcal{F}_{global}$.

$$\mathcal{F}_{global} = \mathbf{F}_{global} \otimes \mathbb{A}_{patch}, \tag{8}$$

where $\otimes$ represents tensor multiplication.

Different from the conventional max MIL pooling and average MIL pooling, our attention MIL pooling not only considers all patch features instead of only depending on the most probably discriminative patch, but also gives each patch a different weight instead of combining all the patches equally. Therefore, the attention MIL pooling can emphasize the feature representations for crucial patches to lighten the noise interference and meanwhile remain the connection between unimportant patches and key patches to avoid the loss of potential relevant features, so that it can improve the classification performance and reduce the misdiagnosis rate of special subjects. Specifically, the patch-attention map can be a reference to identify pathological locations.

F. Attention-Aware Global Classifier

Attention-aware global classifier (shown in Fig. 1) continues to process the bag-level representations $\mathcal{F}_{global}$ by considering the high correlations among patches and makes a final diagnosis. Compared with directly using fully connected layers to explore the correlation among patch-level features, the convolutional layers show a superior high-level feature extraction capability for deep learning on inter-patch features. Thus, the two-layer convolutional network in the front of the global classifier are used to further learn the attention-aware feature representation from the MIL pooling for extracting more structural information among patches and squeezing the feature maps along channels. The two convolutional layers respectively have 128 filters and 64 filters with the same size of $2\times2\times2$ and unit stride, followed by batch normalization (BN) and rectified linear unit (ReLU) activations. Then an adaptive 3D average pooling is adopted to downsample the feature maps to $\mathcal{F} \in \mathbb{R}^{64\times1\times1\times1}$. Then the feature representation is flattened as the input of subsequent two fully connected layers with 32 and 2 units to generate two scores (normalized by softmax function) representing the negative and positive probability respectively.

Based on the different weighted feature maps output from the previous attention MIL pooling, the attention-aware global classifier is designed to further learn the integral feature representations for the whole brain structural information in MRI scans and output classification results for AD classification or MCI conversion prediction.

G. Loss Function

Since only image-level labels are given while patch-level labels are ambiguous, the image-level label is regarded as the unique guidance used in back propagation for updating our network weights $\mathbf{W}$. The loss function we use in model training based on the cross entropy loss is described as:

$$\mathcal{L}(\mathbf{W}) = -\frac{1}{N}\sum_{n=1}^{N} log(P(Y_n|\boldsymbol{X}_n; \mathbf{W})), \tag{9}$$

where N is the number of images, $P(Y_n|\boldsymbol{X}_n; \mathbf{W})$ is the probability of correct prediction for $\boldsymbol{X}_n$. As an end-to-end network, the training losses are backpropagated from the global classifier to the MIL pooling and Patch-Nets for assisting in

updating the parameters of the network with an optimization algorithm (e.g., Adam). By minimizing the loss function, our network finally learns a map: $\boldsymbol{X}$ to Y.

H. Implementation

Our proposed DA-MIDL network (whose framework is shown in Fig. 1 and detailed in Table SI of the *Supplementary Materials*) is implemented using Python based on the Pytorch packages.[1] To alleviate the overfitting issue, we use the batch normalization activation after the convolutional layers. We make all the Patch-Nets share the weights, which reduces the number of trainable parameters especially when a large cohort of patches are inputted. Besides, the input image patches to Patch-Nets are extracted from different brain locations with various anatomical structures, which effectively augments the diversity of training data.

Two datasets (i.e., ADNI and AIBL) are used to evaluate the performance and generalizability of our DA-MIDL method. Specifically, we divide the samples from the ADNI dataset into training and test datasets, in which 80% samples are used for model training while the remaining 20% samples held out as a test dataset. A five-fold cross validation strategy is used for choosing the hyper-parameters and model training on the ADNI training dataset. Then the trained model with optimized hyper-parameters is tested on the held out ADNI test dataset. To further verify the robustness and generalizability of our model, we have also evaluated our model on an independent dataset (AIBL).

In the training stage, we first calculate the p-value map covering the whole MR image by group comparison on the training set (i.e., 4 subsets each round in 5-fold) to initialize the input patch locations. Then we feed the patches extracted from the selected locations in MR images to the corresponding Patch-Nets, respectively. The proposed DA-MIDL is trained using the Adam optimizer for 100 epoch, which requires ~5.5 hours on one NVIDIA GTX Titan X GPU, and evaluated on the remaining 1 validation subset. The architecture (e.g., the number of channels) of DA-MIDL and its hyper-parameters (e.g., learning rate $= 0.001$, batch size $= 10$, patch size $= 25 \times 25 \times 25$ and patch number $= 60$) are chosen by the mean validation performance across all folds. In addition, the validation performances of our method with different parameters are shown in Section II-A of the *Supplementary Materials*.

In the test stage, we feed the patches extracted at the same locations used in the training stage from an unseen MR image to the trained network for AD diagnosis, which takes ~0.25 seconds for one subject based on the pre-processed MR image.

IV. Experiments

In this section, we present the experimental settings and the performance and generalizability of our DA-MIDL method on multiple AD-related diagnosis tasks compared with several state-of-the-art methods.

[1] https://github.com/WyZhuNUAA/DA-MIDL

A. Experimental Settings

Our DA-MIDL method is verified on multiple AD-related diagnosis tasks, such as AD classification (AD vs. NC), MCI conversion prediction (pMCI vs. sMCI) and MCI classifications (pMCI vs. NC and sMCI vs. NC). We apply four metrics to evaluate the classification performance, including accuracy (ACC), sensitivity (SEN), specificity (SPE), and the area under receiver operating characteristic curve (AUC). These metrics are defined as: $ACC = \frac{TP+TN}{TP+TN+FP+FN}$, $SEN = \frac{TP}{TP+FN}$, $SPE = \frac{TN}{TN+FP}$, where TP, TN, FP and FN are denoted as true positive, true negative, false positive and false negative value respectively. ACC, SEN and SPE are calculated using the default threshold of 0.5. AUC is calculated on all possible pairs of true positive rate ($TPR = SEN$) and false positive rate ($FPR = 1 - SPE$) by changing the thresholds performed on the prediction results from our trained DA-MIDL network.

B. Competing Methods

We compare our DA-MIDL method with three baseline methods, i.e., a conventional voxel-level method (i.e., VBM), a conventional ROI-level method (i.e., ROI), a conventional patch-level method (i.e., PLM), and two state-of-the-art patch-level deep learning-based methods (i.e., DMIL [10] and HFCN [6]).

1) Voxel-Level Morphometry (VBM): According to the study [22], each MR image is processed by the spatial normalization to a standard stereotactic brain space (i.e., Colin27 template) and the local gray matter density is measured as the voxel-level feature. Due to the high dimension of the voxel-wise features, a t-test is adopted to make a difference comparison on two groups of images at each voxel respectively from AD patients and normal controls for feature selection. Then based on the selected voxel-level feature vectors, a linear SVM is trained for AD-related diagnosis.

2) ROI-Level Method (ROI): Following the work [11], all the registered MR images after a deformable registration algorithm Hammer [59] are segmented into 93 regions according to the template with 93 manually labeled ROIs [60]. Then we calculate the gray matter volume in each ROI as the region-level feature which is further normalized by the total intracranial volume. Based on the feature vectors which consist of 93 ROI features, a linear SVM is constructed for AD-related classification.

3) Patch-Level Method (PLM): Similar to [26], we uniformly divide the tissue density maps into non-overlapping patches. Then, we use t-tests to select the relevant voxels with p-values smaller than 0.05. The selected patches (i.e., contain the relevant voxels) from tissue density maps are used to compose a patch pool as the features for Alzheimer's disease classification. We construct a classifier (i.e., SVM) based on the patch pool to obtain the classification results.

4) Deep Multi-Instance Learning (DMIL): In this work [10], the deep learning framework with patch-wise input data is constructed based on the multi-instance learning. Multiple sub-CNNs with the same structure of 6 convolutional layers generate patch-level feature representations, where each sub-CNN corresponds to a patch and has different parameters.

Then the patch-level feature representations are concatenated into a global feature representation as the input of the subsequent classifier including 5 fully connected layers and a softmax layer for AD diagnosis.

5) Hierarchical Fully Convolutional Network (HFCN): The HFCN model [6] is reproduced as a comparison method, which is implemented by fully convolutional layers and contains three levels of networks including multiple patch-level sub-networks, several region-level sub-networks and a subject-level sub-network. Multi-scale feature representations are jointly learned and fused for the construction of hierarchical classifiers. That is, the outputs from low-level sub-networks are spatially combined to form input features for high-level sub-networks.

Note that the conventional methods are simply implemented by linear SVMs based on selected local features of different scales (i.e., voxel-level, ROI-level and patch-level) and the patch-level deep learning-based models are trained on the patches at the same proposal locations as our DA-MIDL method instead of the landmarks with prior knowledge. Thus, these contrast methods may fail to achieve the first-rate results in their papers. All the methods are trained and evaluated on the same training set and test sets.

TABLE II
RESULTS FOR AD CLASSIFICATION (AD VS. NC) AND MCI CONVERSION PREDICTION (PMCI VS. SMCI) ON THE ADNI TEST SET

Method	AD vs. NC				pMCI vs. sMCI			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
VBM	0.816	0.756	0.875	0.883	0.679	0.629	0.717	0.709
ROI	0.804	0.718	0.888	0.852	0.667	0.571	0.739	0.692
PLM	0.848	0.846	0.850	0.905	0.716	0.657	0.761	0.732
DMIL	0.892	0.859	0.925	0.950	0.765	0.714	0.804	0.790
HFCN	0.905	0.897	0.913	0.942	0.778	0.686	**0.848**	0.812
DA-MIDL	**0.924**	**0.910**	**0.938**	**0.965**	**0.802**	**0.771**	0.826	**0.851**

TABLE III
RESULTS FOR PMCI VS. NC AND SMCI VS. NC CLASSIFICATIONS ON THE ADNI TEST SET

Method	pMCI vs. NC				sMCI vs. NC			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
VBM	0.816	0.647	0.888	0.853	0.698	0.674	0.713	0.742
ROI	0.789	0.618	0.862	0.846	0.675	0.652	0.688	0.698
PLM	0.825	0.765	0.850	0.876	0.738	0.652	0.788	0.756
DMIL	0.868	0.735	**0.925**	0.908	0.794	0.783	0.800	0.808
HFCN	0.877	0.795	0.913	0.910	0.802	0.717	**0.850**	0.832
DA-MIDL	**0.895**	**0.824**	**0.925**	**0.917**	**0.825**	**0.804**	0.838	**0.860**

C. Classification Performance on ADNI

The performances on AD classification and MCI conversion prediction achieved by our DA-MIDL method and the competing methods on the test set from ADNI are shown in Table II. Also, the 5-fold validation performances of our method on the training set from ADNI are shown in Table SII of the *Supplementary Materials*.

As shown from Table II, our DA-MIDL method achieves better performance in both AD classification and MCI conversion prediction tasks in most cases. For example, our DA-MIDL method obtains better results on all four metrics (i.e., ACC = 0.924, SEN = 0.910, SPE = 0.938 and AUC = 0.965) in AD classification. Additionally, in the MCI conversion prediction task, the ACC (0.802), SEN (0.771) and AUC (0.851) yielded by our DA-MIDL method are also much better than the results from the other five methods. Meanwhile, the patch-level methods (i.e., PLM, DMIL, HFCN and DA-MIDL) all outperform the voxel-level and ROI-level methods (i.e., VBM and ROI). The possible reason is that the patch-level feature representation can capture more suitable local discriminative structural features. Furthermore, compared with the conventional patch-level method (i.e., PLM), the deep learning-based methods (i.e., DMIL, HFCN and DA-MIDL) achieve much better results for Alzheimer's disease diagnosis. The main reason could be that using the task-oriented features learned by deep learning methods can mitigate the heterogeneity between features and subsequent classification algorithms. Compared with the two state-of-the-art methods (i.e., DMIL and HFCN), our DA-MIDL method overall achieves better classification performance with the same inputs. The underlying reason could be that the different weighted feature representations learned by our DA-MIDL model are effective for AD detection. Specifically, our DA-MIDL method has a superior improvement on the sensitivity metric, which implies that our DA-MIDL method has much lower missed diagnosis rate in AD classification and MCI conversion predication. It indicates that our DA-MIDL method is more sensitive to the disease-related structural changing features in the brain.

To further evaluate the performance of DA-MIDL, we perform the additional experiments on MCI classification tasks (including pMCI vs. NC and sMCI vs. NC). Specifically, the classification between sMCI and NC is also as challenging as MCI conversion prediction (i.e., pMCI vs. sMCI) due to the slight structural changes in the cerebrum at the early stage of AD. As shown in Table III, our DA-MIDL method also achieves better performance on the both classification tasks. For example, in the classification task of distinguishing pMCI subjects from normal controls, our DA-MIDL method acquires better results (i.e., ACC = 0.895, SEN = 0.824, SPE = 0.925 and AUC = 0.917). In the challenging classification between sMCI subjects and normal controls, our DA-MIDL method also obtains quite better results, especially on ACC (0.825) and AUC (0.860).

D. Generalization on AIBL

To verify the generalizability of our method, we further use an independent AIBL dataset to evaluate our DA-MIDL method and its competing methods trained on the ADNI dataset. The experimental results for AD classification and MCI conversion prediction on the AIBL dataset are shown in Table IV.

As shown in Table IV, our proposed DA-MIDL method generally outperforms the other five competing methods (i.e., VBM, ROI, PLM, DMIL and HFCN) in most metrics in both AD-related diagnosis tasks. For example, the DA-MIDL achieves the best ACC (0.902) for the AD vs. NC classification task on the AIBL dataset by using the model trained on

TABLE IV
RESULTS FOR AD CLASSIFICATION AND MCI CONVERSION PREDICTION ON THE AIBL DATASET

Method	AD vs. NC				pMCI vs. sMCI			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
VBM	0.808	0.582	0.866	0.817	0.673	0.529	0.720	0.717
ROI	0.793	0.519	0.863	0.796	0.664	0.471	0.710	0.671
PLM	0.839	0.722	0.870	0.846	0.709	0.529	0.742	0.725
DMIL	0.868	0.772	0.893	0.901	0.764	0.588	0.796	0.793
HFCN	0.889	0.823	0.906	0.930	0.782	0.647	0.806	0.796
DA-MIDL	**0.902**	**0.848**	**0.915**	**0.939**	**0.809**	**0.706**	**0.828**	**0.824**

TABLE V
RESULTS FOR AD CLASSIFICATION AND MCI CONVERSION PREDICTION ACHIEVED BY DA-MIDL AND ITS COUNTERPARTS (I.E., N-MIDL, S-MIDL AND A-MIDL) ON THE ADNI TEST SET

Method	AD vs. NC				pMCI vs. sMCI			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
N-MIDL	0.886	0.885	0.888	0.944	0.753	0.743	0.761	0.758
S-MIDL	0.899	0.897	0.900	0.951	0.765	0.714	0.804	0.790
A-MIDL	0.905	0.903	0.913	0.960	0.778	**0.800**	0.761	0.805
DA-MIDL	**0.924**	**0.910**	**0.938**	**0.965**	**0.802**	0.771	**0.826**	**0.851**

the ADNI dataset, which is better than VBM (0.808), ROI (0.793), PLM (0.839), DMIL (0.868), and HFCN (0.889). For the pMCI vs. sMCI classification task, our DA-MIDL method also obtains better results (0.809, 0.706, 0.828, and 0.824 for ACC, SEN, SPE and AUC, respectively), which is superior to that of the second-best method (0.782, 0.647, 0.806, and 0.796 for ACC, SEN, SPE and AUC, respectively). These results suggest our DA-MIDL method can achieve a robust performance across different datasets. Furthermore, compared with the results reported in Table II, the performance of our DA-MIDL has no obvious drop for the AD vs. NC task in most of metrics. The performance of our DA-MIDL has a serious drop in terms of sensitivity for the pMCI vs. sMCI task. The possible reason is the small number of pMCI in the AIBL dataset. Even one misclassified sample will lead to a serious drop in terms of sensitivity. Nevertheless, these results overall indicate the good generalization capability of our method for AD diagnosis.

E. Comparison With Previous Works

For a broad comparison between our method and related studies on the performance of AD diagnosis, in Table VI we sort out several state-of-the-art results reported in the corresponding literature using structural MRI data from the ADNI database on AD classification and MCI conversion prediction tasks, including two voxel-level methods [13], [20], three ROI-level methods [38], [61], [62] and three patch-level methods [8], [27], [30].

As shown in Table VI, we can have several observations as follows. First, our method achieves competing performance in both AD-related classification tasks. Second, compared with voxel-level methods [13], [20], our method has much better performance. The possible reason is that our method can deal with the spatial correlation (i.e., latent non-linear features) of local patch-level brain structures by convolutional neural networks better than the linear voxel-level feature vectors. Third, different from the region-level methods based on empirically predefined ROIs [38], [62], our method attempts to extract structural features from multiple patches distributed in the whole brain, which is much more difficult. However, our method has competing performance, which implies the effectiveness of our DA-MIDL model for identifying the pathological locations. Finally, our method outperforms the other patch-level methods [8], [27], [30], which demonstrates that our method has a good feature extraction ability of local-to-global representations for AD diagnosis by specifically balancing the relative contribution of each patch.

V. DISCUSSION

In this section, we first evaluate the effectiveness of attention modules (i.e. spatial attention block and attention MIL pooling) and the influence of Attention MIL Pooling in our method. Then we present the discriminative pathological locations identified by our method and the potential of clinical translation. Finally, we analyze the limitations of our work and possible future research directions.

A. Effectiveness of Attention Modules

To evaluate the effectiveness of the attention modules used in our study, we further compare the proposed DA-MIDL method with its counterparts, i.e., the model with neither attention modules (N-MIDL), the model only with spatial attention blocks (S-MIDL), and the model only with attention MIL pooling (A-MIDL). We evaluate these four methods on two AD-related diagnosis tasks (e.g., AD vs. NC and pMCI vs. sMCI), with results reported in Table V.

As shown in Table V, our proposed attention modules can overall improve the classification performance. For instance, our DA-MIDL method with dual attention modules has higher accuracy than its counterparts (i.e., N-MIL, S-MIL and A-MIL) in AD classification and MCI conversion prediction. These results imply that using both attention modules could achieve better classification performance. The possible reason is that the attention modules are effective in enhancing the discriminative features for AD-related classification.

B. Influence of Attention MIL Pooling

We further compare the proposed attention MIL pooling with several common MIL pooling which are widely used for aggregating the instance-level representations in processing MIL problems such as average MIL pooling and max MIL pooling [63]. Average MIL pooling is used to calculate the average feature representation of instance-level features as the bag representation (i.e, $B = \frac{1}{K}\sum_{k=1}^{K} I_k$). In contrast, max MIL pooling only focuses on discriminative instance-level features (i.e., $B = max_{k=1,\cdots,K}\{I_k\}$). We replace the attention MIL pooling in DA-MIDL with average MIL pooling and max MIL pooling as two control methods, which are implemented respectively by element-wise average and maximum operators. As shown in Fig. 2, the proposed attention MIL pooling

TABLE VI
REFERENTIAL COMPARISON ON SMRI-BASED STUDIES FOR AD CLASSIFICATION AND MCI CONVERSION PREDICTION

Reference	Method	Subjects	AD vs. NC				pMCI vs. sMCI			
			ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
Salvatore et al. [13]	Voxel-level features, PCA + SVM	137 AD + 162 NC + 76 pMCI + 134 sMCI	0.760	-	-	-	0.660	-	-	-
Cuingnet et al. [20]	Voxel-level features, SVM	137 AD + 162 NC + 76 pMCI + 134 sMCI	0.886	0.810	**0.950**	-	0.704	0.570	0.780	-
Eskildsen et al. [61]	ROI-level features, mRMR + LDA	194 AD + 226 NC + 61 pMCI + 134 sMCI	0.867	0.804	0.920	0.917	0.773	0.690	0.791	0.835
Cao et al. [38]	ROI-level features, multi-kernel + KNN	192 AD + 229 NC + 168 pMCI + 229 sMCI	0.886	0.857	0.904	0.898	0.704	0.677	0.718	0.705
Lin et al. [62]	ROI-level features, CNN + Lasso + ELM	188 AD + 229 NC + 169 pMCI + 139 sMCI	0.888	-	-	-	0.799	**0.840**	0.748	**0.861**
Tong et al. [8]	Patch-level Features, mi-Graph + SVM	198 AD + 231 NC + 167 pMCI + 238 sMCI	0.900	0.860	0.930	-	0.720	0.690	0.740	-
Li et al. [30]	Path-level features, K-means + DenseNet	199 AD + 229 NC	0.895	0.879	0.908	0.924	-	-	-	-
Qiu et al. [27]	Patch-level features, FCN + MLP	188 AD + 229 NC	0.834	0.767	0.889	-	-	-	-	-
Proposed	Patch-level features, Attention + MIL + CNN	398 AD + 400 NC + 172 pMCI + 232 sMCI	**0.924**	**0.910**	0.938	**0.965**	**0.802**	0.771	**0.826**	0.851

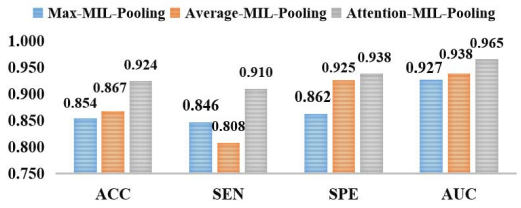

Fig. 2. AD classification performance of our DA-MIDL model with different MIL pooling including max-mil-pooling, average-MIL-pooling and attention-MIL-pooling on the ADNI test set.

achieves much better performance on AD classification than the other two MIL pooling in all cases. It implies that only focusing on one single discriminative patch (as max MIL pooling does) may not adequately represent the disease-related features for classification, since the atrophy occurs at multiple patches distributed in the brain. In addition, the average MIL pooling may relatively lack of identifying discriminative features with a lower sensitivity, due to aggregating all patch-level features equally. In contrast, attention MIL pooling may balance the contribution of each patch for AD classification by learning a relative weight for each patch-level features with different informativeness, which could be effective for improving the classification performance.

C. Discriminative Pathological Locations and the Potential of Clinical Translation

The potential of clinical translation is of importance to the computer-aided diagnosis. One of the keys to the clinical diagnosis of AD is to observe the morphological changes of the brain (i.e., to find abnormally atrophy areas of the brain). As an auxiliary diagnostic approach, our proposed DA-MIDL method can automatically identify the possible pathological locations in the whole MR images for doctors to find the regions of interest for diagnosis easily. That is, our method can identify the subject-specific discriminative pathological locations, including relative discriminative patches in global images and discriminative micro-structures in local patches.

1) Discriminative Patch Locations: The upper part of Fig. 3 shows several discriminative patch locations in sMRI scans identified by DA-MIDL. The discriminative patch locations are marked at the perspective direction of one view (e.g., coronal, axial or sagittal view) in 3D images. In total 12 most discriminative patches are marked for one subject, which cover ~10.61% of non-zero voxels in the whole image. Also, the left and right panels respectively correspond to the probable pathological locations for AD classification and MCI conversion prediction. Besides, the marked patch locations in the first and second rows are respectively suggested by the affect scores and attention weights yielded from DA-MIDL. Compared with the discrete locations identified by affect scores, the patches identified by attention weights gather in certain regions. The possible reason is that the affect scores are calculated depending on isolated patches, while the attention weights take account of the correlations among patches. Furthermore, the probable pathological locations in AD classification and MCI conversion prediction are very similar, which is in line with the high correlation between the two classification tasks according to the progression of Alzheimer's disease.

We further mark out three major brain regions where more patches are gathered with a visualization tool [64] in the rightmost panel of Fig. 3, including hippocampus, amygdala and thalamus. The marked regions are consistent with many previous works [6], [7], [23] and are considered as related regions for AD diagnosis. Specifically, the hippocampus is highly correlative with long-term memory. The influence of

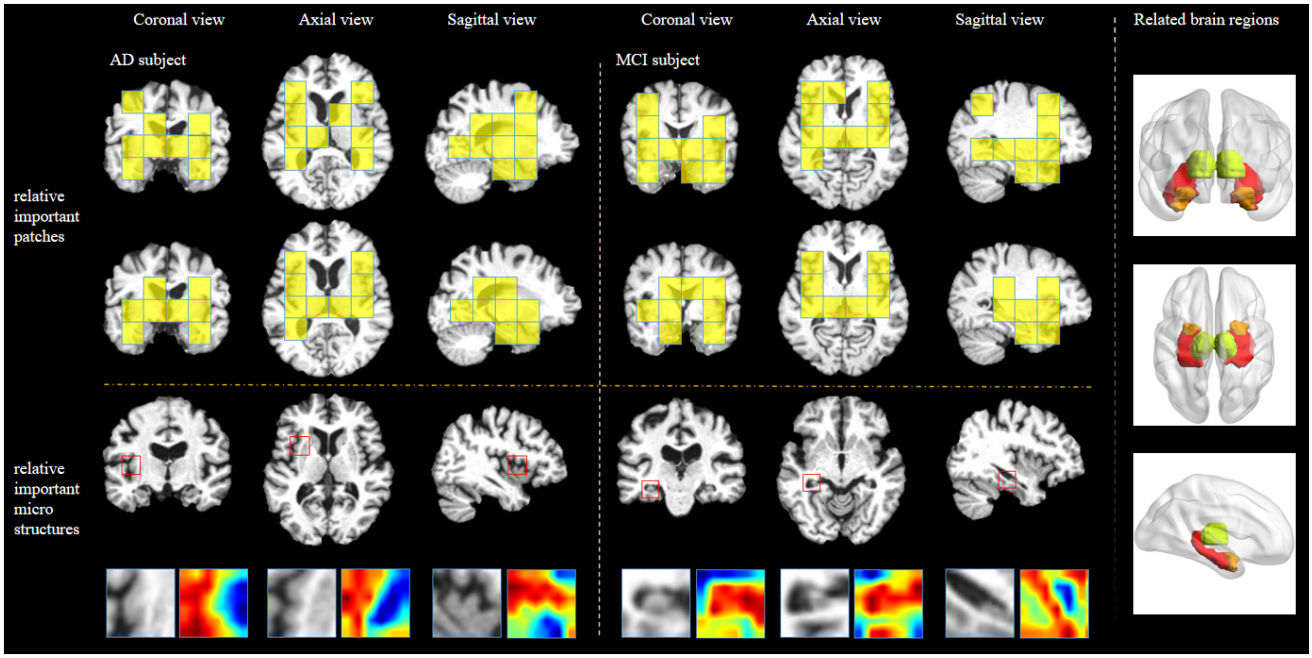

Fig. 3. Discriminative pathological locations identified by DA-MIDL on AD classification (i.e., the left panel) and MCI conversion prediction (i.e., the right panel). The first and second rows show the informative patch locations respectively suggested by affect scores and attention weights. The rightmost panel shows the marked brain regions where more patches are gathered. The last two rows show the discriminative micro structures in several fixed-size informative patches identified by spatial attention blocks.

brain atrophy caused by AD on the hippocampus has been biologically verified [65]. The amygdala has effect on emotion functions and control of learning and memory [66], which is also relevant to AD. In addition, the thalamus is thought to be related to cognition and information processing speed [7].

2) Discriminative Parts Within Patches: The last two rows of Fig. 3 show the spatially-varying contributions of different parts within the corresponding discriminative patches with relatively high attention weights in sMRI scans produced from spatial attention blocks. The spatial heat maps demonstrate the discriminative micro-structures in fixed-size patches for AD diagnosis. We can observe that most of informative parts with red colors are located at the edges of sulcus gyrus and gray matter, which may effectively reflect the local structural changes by brain atrophy.

D. Limitations and Future Work

Although our proposed DA-MIDL method achieves good performance in AD-related diagnosis and identifying discriminative pathological locations, there are still several limitations which may influence the generalization capability of our model. We summarize the limitations and potential solutions as follows. 1) The size of input patches is fixed and equivalent. However, the structural changes in the cerebrum caused by brain atrophy may occur across multiple regions with different scales. Using the fixed size could not represent various local features. It's more reasonable to use multi-scale patches as inputs, while it may increase the difficulty of constructing the networks. In addition, ROI pooling [67] may be adopted for settling the inputs with non-uniform sizes. 2) The patch location proposals based on the group comparison are isolated from the subsequent network. This means that our proposed method is not strictly an end-to-end analysis procedure, which may affect the optimal performance of our model. Therefore, it is important to combine the generator of patch location proposals and the network into a unified framework. In the future works, we can embed a weakly-supervised network for detecting informative landmarks in a whole brain and based on the detected patches at the discriminative landmarks we construct our DA-MIDL model for AD-related diagnosis. The parameters in the detection network and DA-MIDL model can be optimized jointly as an end-to-end model.

VI. Conclusion

In this study, we propose a dual attention multi-instance deep learning network (DA-MIDL) for computer-aided AD diagnosis, which includes three major components: 1) Patch-Nets with spatial attention blocks for extracting discriminative features from local patches, 2) an attention MIL pooling operation for balancing the relative contribution of each patch, and 3) an attention-aware global classifier for making the AD-related diagnosis decisions based on the combined feature representation for the whole brain structure. Our proposed DA-MIDL method is evaluated on 1689 subjects from two independent datasets (i.e., ADNI and AIBL) in multiple AD-related diagnosis tasks. Experimental results demonstrate that our method can not only identify discriminative pathological locations in sMRI scans, but also achieve better diagnosis performance than several state-of-the-art methods.

References

[1] A. Association, W. Thies, and L. Bleiler, "2013 Alzheimer's disease facts and figures," *Alzheimer's Dementia*, vol. 9, no. 2, pp. 208–245, Mar. 2013.

[2] W. Jagust, "Vulnerable neural systems and the borderland of brain aging and neurodegeneration," *Neuron*, vol. 77, no. 2, pp. 219–234, Jan. 2013.

[3] J. Cummings, G. Lee, A. Ritter, M. Sabbagh, and K. Zhong, "Alzheimer's disease drug development pipeline: 2019," *Alzheimer's Dementia: Transl. Res. Clin. Interventions*, vol. 5, no. 1, pp. 272–293, Jan. 2019.

[4] A. Atri, "Current and future treatments in Alzheimer's disease," in *Seminars Neurol.*, vol. 39, no. 2. New York, NY, USA: Thieme Medical Publishers, 2019, pp. 227–240.

[5] R. L. Buckner, "Memory and executive function in aging and AD: Multiple factors that cause decline and reserve factors that compensate," *Neuron*, vol. 44, no. 1, pp. 195–208, 2004.

[6] C. Lian, M. Liu, J. Zhang, and D. Shen, "Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 880–893, Apr. 2020.

[7] W. Shao, Y. Peng, C. Zu, M. Wang, and D. Zhang, "Hypergraph based multi-task feature selection for multimodal classification of Alzheimer's disease," *Computerized Med. Imag. Graph.*, vol. 80, Mar. 2020, Art. no. 101663.

[8] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J. V. Hajnal, and D. Rueckert, "Multiple instance learning for classification of dementia in brain MRI," *Med. Image Anal.*, vol. 18, no. 5, pp. 808–818, Jul. 2014.

[9] D. Dai, H. He, J. T. Vogelstein, and Z. Hou, "Accurate prediction of AD patients using cortical thickness networks," *Mach. Vis. Appl.*, vol. 24, no. 7, pp. 1445–1457, Oct. 2013.

[10] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Landmark-based deep multi-instance learning for brain disease diagnosis," *Med. Image Anal.*, vol. 43, pp. 157–168, Jan. 2018.

[11] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, Apr. 2011.

[12] M. Liu, D. Zhang, and D. Shen, "Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment," *IEEE Trans. Med. Imag.*, vol. 35, no. 6, pp. 1463–1474, Jun. 2016.

[13] C. Salvatore, A. Cerasa, P. Battista, M. C. Gilardi, A. Quattrone, and I. Castiglioni, "Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: A machine learning approach," *Frontiers Neurosci.*, vol. 9, p. 307, Sep. 2015.

[14] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in alzheimer disease," *Nature Rev. Neurol.*, vol. 6, no. 2, pp. 67–77, Feb. 2010.

[15] S. Leandrou, S. Petroudi, P. A. Kyriacou, C. C. Reyes-Aldasoro, and C. S. Pattichis, "Quantitative MRI brain studies in mild cognitive impairment and Alzheimer's disease: A methodological review," *IEEE Rev. Biomed. Eng.*, vol. 11, pp. 97–111, 2018.

[16] S. Rathore, M. Habes, M. A. Iftikhar, A. Shacklett, and C. Davatzikos, "A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages," *NeuroImage*, vol. 155, pp. 530–548, Jul. 2017.

[17] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," *NeuroImage*, vol. 145, pp. 137–165, Jan. 2017.

[18] F. Falahati, E. Westman, and A. Simmons, "Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging," *J. Alzheimer's Disease*, vol. 41, no. 3, pp. 685–708, Jul. 2014.

[19] M. Vounou *et al.*, "Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease," *NeuroImage*, vol. 60, no. 1, pp. 700–716, Mar. 2012.

[20] R. Cuingnet *et al.*, "Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database," *NeuroImage*, vol. 56, no. 2, pp. 766–781, May 2011.

[21] J. C. Baron *et al.*, "*In vivo* mapping of gray matter loss with voxel-based morphometry in mild Alzheimer's disease," *NeuroImage*, vol. 14, no. 2, pp. 298–309, Aug. 2001.

[22] J. Ashburner and K. J. Friston, "Voxel-based morphometry—The methods," *NeuroImage*, vol. 11, no. 6, pp. 805–821, 2000.

[23] D. Zhang and D. Shen, "Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease," *NeuroImage*, vol. 59, no. 2, pp. 895–907, Jan. 2012.

[24] J. Zhang, Y. Gao, Y. Gao, B. C. Munsell, and D. Shen, "Detecting anatomical landmarks for fast Alzheimer's disease diagnosis," *IEEE Trans. Med. Imag.*, vol. 35, no. 12, pp. 2524–2533, 2016.

[25] M. Liu, D. Zhang, P. T. Yap, and D. Shen, "Hierarchical ensemble of multi-level classifiers for diagnosis of Alzheimer's disease," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Berlin, Germany: Springer, Oct. 2012, pp. 27–35.

[26] M. Liu, D. Zhang, and D. Shen, "Ensemble sparse classification of Alzheimer's disease," *NeuroImage*, vol. 60, no. 2, pp. 1106–1116, Apr. 2012.

[27] S. Qiu *et al.*, "Development and validation of an interpretable deep learning framework for Alzheimer's disease classification," *Brain*, vol. 143, no. 6, pp. 1920–1933, 2020.

[28] J. Wen *et al.*, "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101694.

[29] K. Aderghal, A. Khvostikov, A. Krylov, J. Benois-Pineau, K. Afdel, and G. Catheline, "Classification of alzheimer disease on imaging modalities with deep CNNs using cross-modal transfer learning," in *Proc. IEEE 31st Int. Symp. Computer-Based Med. Syst. (CBMS)*, Jun. 2018, pp. 345–350.

[30] F. Li and M. Liu, "Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks," *Computerized Med. Imag. Graph.*, vol. 70, pp. 101–110, Dec. 2018.

[31] M. Liu *et al.*, "A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease," *NeuroImage*, vol. 208, Mar. 2020, Art. no. 116459.

[32] L. Fang *et al.*, "Automatic brain labeling via multi-atlas guided fully convolutional networks," *Med. Image Anal.*, vol. 51, pp. 157–168, Jan. 2019.

[33] Y. Huang *et al.*, "Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network," *Frontiers Neurosci.*, vol. 13, p. 509, 2019.

[34] H. Wang *et al.*, "Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease," *Neurocomputing*, vol. 333, pp. 145–156, Mar. 2019.

[35] H.-I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, Nov. 2014.

[36] M. Liu, D. Zhang, P. T. Yap, and D. Shen, "Tree-guided sparse coding for brain disease classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, Oct. 2012, pp. 239–247.

[37] Y. Cho, J.-K. Seong, Y. Jeong, and S. Y. Shin, "Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data," *NeuroImage*, vol. 59, no. 3, pp. 2217–2230, Feb. 2012.

[38] P. Cao *et al.*, "Nonlinearity-aware based dimensionality reduction and over-sampling for AD/MCI classification from MRI measures," *Comput. Biol. Med.*, vol. 91, pp. 21–37, Dec. 2017.

[39] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013.

[40] H. D. Couture, J. S. Marron, C. M. Perou, M. A. Troester, and M. Niethammer, "Multiple instance learning for heterogeneous images: Training a CNN for histopathology," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Sep. 2018, pp. 254–262.

[41] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," 2018, *arXiv:1802.04712*. [Online]. Available: http://arxiv.org/abs/1802.04712

[42] Y. Xu, J. Zhang, I. Eric, C. Chang, M. Lai, and Z. Tu, "Context-constrained multiple instance learning for histopathology image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, Oct. 2012, pp. 623–630.

[43] L. Chen *et al.*, "Identification of cerebral small vessel disease using multiple instance learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Oct. 2015, pp. 523–530.

[44] G. Chen, J. Zhao, R. Zhang, T. Wang, G. Zhang, and B. Lei, "Automated stage analysis of retinopathy of prematurity using joint segmentation and multi-instance learning," in *Proc. Int. Workshop Ophthalmic Med. Image Anal.* Cham, Switzerland: Springer, Oct. 2019, pp. 173–181.

[45] M. Combalia and V. Vilaplana, "Monte-Carlo sampling applied to multiple instance learning for histological image classification," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 274–281.

[46] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," *Comput. Sci.*, vol. 2015, pp. 2048–2057, Feb. 2015.

[47] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[48] C. Lian, M. Liu, L. Wang, and D. Shen, "End-to-end dementia status prediction from brain mri using multi-task weakly-supervised attention network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Oct. 2019, pp. 158–167.

[49] C. Ma, H. Wang, and S. C. Hoi, "Multi-label thoracic disease image classification with cross-attention networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Oct. 2019, pp. 730–738.
[50] Z. Fang, Y. Chen, D. Nie, W. Lin, and D. Shen, "RCA-U-Net: Residual channel attention U-Net for fast tissue quantification in magnetic resonance fingerprinting," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Oct. 2019, pp. 101–109.
[51] T. D. Bui, L. Wang, J. Chen, W. Lin, G. Li, and D. Shen, "Multi-task learning for neonatal brain segmentation using 3D dense-unet with dense attention guided by geodesic distance," in *Domain Adaptation and Representation Transfer and Medical Image Learning With Less Labels and Imperfect Data*. Cham, Switzerland: Springer, 2019, pp. 243–251.
[52] J. Jovicich *et al.*, "Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data," *NeuroImage*, vol. 30, no. 2, pp. 436–443, Apr. 2006.
[53] P. A. Narayana, W. W. Brey, M. V. Kulkarni, and C. L. Sievenpiper, "Compensation for surface coil sensitivity variation in magnetic resonance imaging," *Magn. Reson. Imag.*, vol. 6, no. 3, pp. 271–274, May 1988.
[54] C. J. Holmes, R. Hoge, L. Collins, R. Woods, A. W. Toga, and A. C. Evans, "Enhancement of MR images using registration for signal averaging," *J. Comput. Assist. Tomogr.*, vol. 22, no. 2, pp. 324–333, Mar. 1998.
[55] M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782–790, Aug. 2012.
[56] S. Frenzel *et al.*, "A biomarker for Alzheimer's disease based on patterns of regional brain atrophy," *Frontiers Psychiatry*, vol. 10, p. 953, Jan. 2020.
[57] N. C. Fox and J. M. Schott, "Imaging cerebral atrophy: Normal ageing to Alzheimer's disease," *Lancet*, vol. 363, no. 9406, pp. 392–394, Jan. 2004.
[58] R. I. Scahill, J. M. Schott, J. M. Stevens, M. N. Rossor, and N. C. Fox, "Mapping the evolution of regional atrophy in Alzheimer's disease: Unbiased analysis of fluid-registered serial MRI," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 7, pp. 4703–4707, Apr. 2002.
[59] D. Shen and C. Davatzikos, "HAMMER: Hierarchical attribute matching mechanism for elastic registration," *IEEE Trans. Med. Imag.*, vol. 21, no. 11, pp. 1421–1439, Nov. 2002.
[60] N. J. Kabani, D. J. MacDonald, C. J. Holmes, and A. C. Evans, "3D anatomical atlas of the human brain," *NeuroImage*, vol. 7, no. 4, p. S717, May 1998.
[61] S. F. Eskildsen, P. Coupé, D. García-Lorenzo, V. Fonov, J. C. Pruessner, and D. L. Collins, "Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning," *NeuroImage*, vol. 65, pp. 511–521, Jan. 2013.
[62] W. Lin *et al.*, "Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment," *Frontiers Neurosci.*, vol. 12, p. 777, Nov. 2018.
[63] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognit.*, vol. 74, pp. 15–24, Feb. 2018.
[64] M. Xia, J. Wang, and Y. He, "BrainNet viewer: A network visualization tool for human brain connectomics," *PLoS ONE*, vol. 8, no. 7, 2013, Art. no. e68910.
[65] C. Pennanen *et al.*, "Hippocampus and entorhinal cortex in mild cognitive impairment and early AD," *Neurobiol. Aging*, vol. 25, no. 3, pp. 303–310, 2004.
[66] L. Goldstein, "The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction," *Yale J. Biol. Med.*, vol. 65, no. 5, p. 540, 1992.
[67] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.