Weakly Supervised Deep Learning for Brain Disease Prognosis Using MRI and Incomplete Clinical Scores

Mingxia Liu, Jun Zhang, Chunfeng Lian, and Dinggang Shen, *Fellow, IEEE*

***Abstract*—As a hot topic in brain disease prognosis, predicting clinical measures of subjects based on brain magnetic resonance imaging (MRI) data helps to assess the stage of pathology and predict future development of the disease. Due to incomplete clinical labels/scores, previous learning-based studies often simply discard subjects without ground-truth scores. This would result in limited training data for learning reliable and robust models. Also, existing methods focus only on using hand-crafted features (e.g., image intensity or tissue volume) of MRI data, and these features may not be well coordinated with prediction models. In this paper, we propose a weakly supervised densely connected neural network (wiseDNN) for brain disease prognosis using baseline MRI data and incomplete clinical scores. Specifically, we first extract multiscale image patches (located by anatomical landmarks) from MRI to capture local-to-global structural information of images, and then develop a weakly supervised densely connected network for task-oriented extraction of imaging features and joint prediction of multiple clinical measures. A weighted loss function is further employed to make full use of all available subjects (even those without ground-truth scores at certain time-points) for network training. The experimental results on 1469 subjects from both ADNI-1 and ADNI-2 datasets demonstrate that our proposed method can efficiently predict future clinical measures of subjects.**

***Index Terms*—Alzheimer's disease (AD), clinical score, disease prognosis, neural network, weakly supervised learning.**

I. Introduction

AS THE most sensitive imaging test of the head (particularly the brain) in routine clinical practice, magnetic resonance imaging (MRI) allows physicians to evaluate medical conditions of the brain and determine the presence of certain diseases. Currently, MRI has been widely used in computer-aided diagnosis of Alzheimer's disease (AD) and its prodromal stage, that is, mild cognitive impairment (MCI) [1]–[6]. In particular, structural MRI provides a non-invasive solution to potentially identify abnormal structural changes of the brain, and helps identify AD-related imaging biomarkers for clinical applications [1], [4], [7]–[9]. Recently, it has remained a hot topic to assess the stage of pathology and predict future advances of AD and MCI, by estimating clinical scores of subjects in future time using baseline (BL) MRI data.

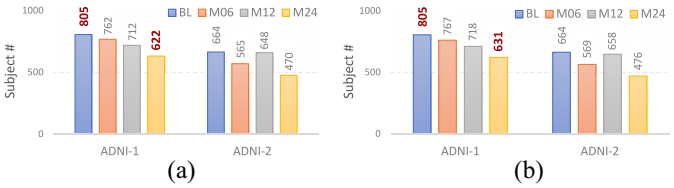

Fig. 1. Distribution of subjects with two types of clinical measures (i.e., CDR-SB and MMSE) from ADNI-1 and ADNI-2 datasets [10] at four time-points: 1) BL; 2) 6th month (M06); 3) 12th month (M12); and 4) M24 after the BL time.

Although several machine-learning methods have been proposed for predicting clinical scores using BL MRI [11], a common challenge of existing methods is the *weakly labeled data problem*, that is, subjects may miss ground-truth clinical scores/labels at certain time-points. As shown in Fig. 1, among 805 subjects at BL time in the AD Neuroimaging Initiative-1 (ADNI-1) dataset [10], only 622 subjects and 631 subjects have complete scores of clinical dementia rating sum of boxes (CDR-SB) and mini-mental state examination (MMSE) at the 24th month (M24) after the BL time, respectively. Due to the nature of supervised learning, previous studies merely discard subjects with missing clinical scores. For instance, Zhang and Shen [11] estimated 2-year changes of two clinical measures from BL MRI, using only 186 subjects with complete ground-truth clinical scores from ADNI-1. It is worth noting that removing subjects with missing scores will significantly reduce the number of training samples, thus affecting the accuracy and robustness of prediction models. Also, previous machine-learning methods usually feed predefined representations [e.g., image intensity and tissue volume within regions-of-interest (ROIs)] of MR images to

Manuscript received November 15, 2018; revised February 3, 2019; accepted March 7, 2019. This work was supported in part by NIH under Grant EB006733, Grant EB008374, Grant EB009634, Grant MH100217, Grant AG041721, Grant AG042599, Grant AG010129, and Grant AG030514. This paper was recommended by Associate Editor D. Goldgof. *(Mingxia Liu and Jun Zhang contributed equally to this work.) (Corresponding author: Dinggang Shen.)*

M. Liu, J. Zhang, and C. Lian are with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA.

D. Shen is with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA, and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea (e-mail: dgshen@med.unc.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2019.2904186

Fig. 2. Flowchart of the proposed weakly supervised deep learning method for brain disease prognosis, containing three main steps, that is, MR image preprocessing; multiscale patch extraction from MRI; and weakly supervised neural network for clinical score regression.

subsequent prediction models, while these features may not be optimal for prediction models, thus degrading the prognosis performance.

Inspired by the recent success of deep learning techniques in medical image analysis [7], [12], [13], several studies resort to deep neural networks to extract MRI features for AD/MCI diagnosis in a data-driven manner [14]–[16]. However, these methods are generally within the supervised learning framework and, thus, cannot directly employ subjects with incomplete ground-truth clinical scores for network training. Therefore, using all available weakly labeled data (i.e., training subjects with incomplete ground-truth scores at certain time-points) becomes an essential problem in MRI-based brain disease prognosis.

In this paper, we propose a weakly supervised deep neural network (wiseDNN) for brain disease prognosis using subjects with BL MRI and incomplete ground-truth clinical scores at multiple time-points. The schematic of our method is shown in Fig. 2. We first preprocess all MR images, and then extract multiscale image patches based on multiple AD-related anatomical landmarks. A deep convolutional neural network (CNN) is finally developed for joint prediction of multiple clinical scores at multiple time-points, with a unique weighted loss function that allows the network to learn from weakly labeled training data. Compared with the previous MRI-based studies, our proposed wiseDNN method can employ all available subjects (even though some of them may lack clinical scores at certain time-points) for model training. Also, our anatomical landmark-based multiscale patch extraction strategy can partly alleviate the problem of limited data, by using image patches other than entire 3-D MR images as the training samples.

The main contributions of this paper can be summarized as follows. First, we develop a neural network with a weighted loss function that can employ all available weakly labeled subjects (i.e., with incomplete ground-truth clinical scores), without discarding subjects having missing scores as in the previous studies [11], [17]. This may help improve the robustness of the learned network by including more subjects in the training process. Second, based on AD-related anatomical landmarks, we propose to extract multiscale (rather than fixed-sized) image patches from MRI, where both small-scale and large-scale patches centered at each landmark are extracted. This kind of strategy helps to capture both local and global structural information of brain MRIs. Third, we develop a joint prediction strategy for estimating multiple clinical scores at multiple time-points simultaneously. Such joint learning strategy is expected to model the inherent relationship among different scores at/across different time-points, thus helping to improve the prediction performance. Finally, based on an MR image of a new test subject, the proposed method can simultaneously predict four types of clinical scores at four time-points within 12 s, which is close to real time.

The remainder of this paper is organized as follows. We present the most relevant studies in Section II. In Section III, we introduce the materials used in this paper and elaborate the details of our proposed method. In Section IV, we first introduce experimental settings and methods for comparison, and then present experimental results of clinical score prediction on both ADNI-1 and ADNI-2 datasets. In Section V, we compare our method with previous studies and discuss limitations of this paper as well as possible future research directions. We finally conclude this paper in Section VI.

II. Related Work

In this section, we first introduce the conventional representations of structural brain MRIs, and then present recent MRI-based deep learning studies for brain disease prognosis/diagnosis.

A. Representations of Structural Brain MR Images

Thanks to the development of neuroimaging techniques, we can directly access brain structures provided by MRI to understand the neurodegenerative underpinnings of AD and MCI [1], [4]. In the literature, many types of feature representations of brain MRI have been developed for automatic AD/MCI diagnosis and prognosis. These representations can be roughly categorized into three classes, that is, voxel-based representation; ROI-based representation; and patch-based representation, with details given below.

1) Voxel-Based Representation: In general, voxel-based approaches [18]–[20] compare brain MR images by directly measuring local tissue [e.g., gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF)] density of a brain via voxel-wise analysis, after deformable registration of individual brain images. For instance, Fan *et al.* [21], [22] proposed to extract volumetric features from brain regions from MR images and applied them to AD classification and gender classification. However, the voxel-based methods are often based on the assumption of one-to-one anatomical mapping between subjects and Gaussian distributions of focal tissue densities during statistical testing [23]. To make data fit the voxel-based model, tissue densities are blurred with large kernels at the expense of focal accuracy, and therefore may reduce the discriminative power of voxel-based representation for MRIs. Another disadvantage of voxel-based representation is that feature dimension is often very high (e.g., millions), while the number of training subjects is very limited (e.g., hundreds), leading to the small-sample-size problem [24] and degrading the performance of learned models.

2) ROI-Based Representation: Different from voxel-based features, ROI-based representations focus on measuring regionally anatomical volumes in predefined regions in the brain. In particular, previous ROI-based studies often employ tissue volume [11], [25]–[27], cortical

thickness [28]–[33], hippocampal volume [34]–[36], and tissue density [9], [17], [31], [32], [37], [38] in specific brain regions as feature representation of MR images. However, this type of representation requires *a priori* hypothesis on abnormal regions from a structural/functional perspective to define regions in the brain [39], while such a hypothesis may not always hold in practice. For example, an abnormal brain region might span over multiple ROIs or just a small part of an ROI, so using a fixed partition for the brain could produce suboptimal learning performance.

3) Patch-Based Representation: Patch-based morphometry was developed to detect fine anatomical changes in brain MRIs, by taking advantage of nonlocal analysis to model the one-to-many mapping between brain anatomies [23]. As reported in [16], [23], [37], and [40]–[42] neurodegenerative patterns can be presented through a patch-based analysis for assisting AD diagnosis and also evaluating the progression of MCI. Liu *et al.* [37], [41] employed GM density within local patches as the representation of MRI for AD diagnosis, using randomly selected patches (i.e., without localizing AD-related micro-structures in the brain). Zhang *et al.* [6] proposed to extract morphometric features (i.e., local energy pattern [43]) from image patches located by AD-related anatomical landmarks. These hand-crafted features of MRI are usually fed to predefined models (e.g., support vector machine [6], [41] and sparse representation [37]) for disease diagnosis and prognosis. However, since feature extraction and model training are performed independently in these methods, those pre-extracted MRI features may not be optimal for prediction models.

B. Deep Learning for Brain Disease Prognosis/Diagnosis

Recently, several deep learning methods have been proposed to automatically learn MRI features in a task-oriented manner. However, to directly feed the whole MR image into a CNN could not generate robust models, since there are millions of voxels in an MR image and many brain regions may be not affected by dementia. Hence, a common challenge in MRI-based deep learning is determining how to precisely locate informative (e.g., discriminative between different groups) regions in brain MRIs.

To address this challenge, Zhang *et al.* [15] proposed to focus on three ROIs (i.e., hippocampal, ventricular, and cortical thickness surface) in brain MRI, and developed a deep CNN for predicting clinical measures of subjects using 2-D image patches extracted from three ROIs. Khvostikov *et al.* [44] employed only the hippocampal ROI and surrounding regions in brain scans (i.e., both structural MRI and diffusion tensor imaging data) for learning a CNN. Similarly, Li *et al.* [14] presented a deep ordinal ranking model for AD classification, using the hippocampal ROI in MRIs. However, these studies use empirically defined regions in MRI, without considering other potentially important brain regions that may be affected by brain diseases. Besides, Sarraf *et al.* [45] developed a 2-D CNN for identifying AD patients from healthy controls (HCs) using both structural and functional MRI (fMRI) data. However, they simply convert 3-D MR images and 4-D fMR images into 2-D slices as the input of their networks, ignoring the important spatial information of slices in MRIs. More recently, Liu *et al.* [16], [46] proposed an anatomical landmark-based deep learning framework for AD diagnosis and MCI conversion prediction. Specifically, they first locate 3-D image patches via AD-related anatomical landmarks distributed throughout the brain, and then develop a CNN for joint feature extraction from MRI and disease classification. However, a fixed size of image patches is used in these studies, without considering that structural changes caused by dementia could vary largely among different brain regions.

Besides, most of the existing deep learning methods are performed in a fully supervised manner, by merely discarding subjects with missing ground-truth scores at certain time-points. To adequately employ all available subjects (even those without ground-truth scores at multiple time-points) for model training, we propose a weakly supervised CNN for predicting clinical measures based on BL MRI data. The proposed method is different from the previous studies in [46]. Specifically, in this paper, we focus on making use of weakly labeled training subjects by developing a unique weighted loss function in the proposed neural network, while previous methods [46] can only use fully labeled (i.e., with complete ground-truth scores) training subjects. Also, this paper proposes to extract multiscale image patches centered at each landmark location to model multiscale structural information of each brain MRI, while those in [46] only use fixed-sized image patches.

III. Materials and Methods

In this section, we first introduce studied subjects and the procedure of MR image preprocessing, and then present the proposed method in detail.

A. Subjects and Image Preprocessing

We perform experiments on 1469 subjects from two subsets of the public ADNI database [10], including ADNI-1 and ADNI-2. Specifically, there are 805 subjects with BL structural MRI data from ADNI-1, and 664 subjects with BL structural MRI data from ADNI-2. Note that subjects that appear in both ADNI-1 and ADNI-2 are directly removed from ADNI-2. Also, different from subjects with 1.5 T T1-weighted MRI in ADNI-1, the studied subjects in ADNI-2 have 3.0 T T1-weighted MRI. That is, ADNI-1 and ADNI-2 are two independent datasets in our experiments. According to several criteria,[1] these subjects can be categorized into three classes, that is, AD, MCI, and HC.

For each subject, four types of clinical measures/scores are used in the experiments: 1) CDR-SB; 2) classic AD assessment scale cognitive (ADAS-Cog) subscale with 11 items (ADAS-Cog11); 3) modified ADAS-Cog with 13 items (ADAS-Cog13); and 4) MMSE. The date when subjects were scheduled to perform the screening becomes the BL time after approval. Also, the time-points for follow-up visits are denoted by the duration starting from the BL time. Specifically, we

[1] http://adni.loni.usc.edu

TABLE I
NUMBER OF STUDIED SUBJECTS HAVING FOUR TYPES OF CLINICAL SCORES (i.e., CDR-SB, ADAS-COG11, ADAS-COG13, AND MMSE) AT FOUR TIME-POINTS, INCLUDING BL, M06, M12, AND M24 AFTER THE BL TIME

Dataset	Category	CDR-SB				ADAS-Cog11				ADAS-Cog13				MMSE			
		BL	M06	M12	M24	BL	M06	M12	M24	BL	M06	M12	M24	BL	M06	M12	M24
ADNI-1	AD	186	173	155	134	185	174	156	134	181	167	152	123	186	174	157	135
	HC	226	214	204	194	226	218	207	199	226	218	205	198	226	218	208	200
	MCI	393	375	353	294	393	374	352	295	390	371	361	293	393	375	353	296
	Subject #	805	762	712	622	804	766	715	628	797	756	718	614	805	767	718	631
ADNI-2	AD	146	99	96	23	145	97	95	23	143	96	94	23	146	99	95	23
	HC	183	170	162	132	183	171	167	132	183	167	165	132	183	171	167	133
	MCI	335	296	390	315	332	299	395	320	332	299	393	315	335	299	396	320
	Subject #	664	565	648	470	660	567	657	475	658	562	652	470	664	569	658	476

denote M06, M12, and M24 as the 6th month, 12th month, and 24th month after BL, respectively. All studied subjects have MRI data at BL, while many have missing ground-truth scores at certain time-points regarding a specific clinical measure. The detailed information about the studied subjects is shown in Table I.

For each structural MR image corresponding to a specific subject, we first perform anterior commissure (AC)–posterior commissure (PC) correction, followed by skull stripping and cerebellum removal. We then linearly align each image to a common Colin27 template [47], and further resample all MR images to have the same size (i.e., $152 \times 186 \times 144$ with a spatial resolution of $1 \times 1 \times 1$ mm^3). Using the N3 algorithm [48], we finally perform intensity inhomogeneity correction for each MR image.

B. Proposed Method

In this paper, we attempt to deal with two challenging problems in MRI-based brain disease prognosis, that is, how to make full use of weakly labeled training data (i.e., subjects with incomplete ground-truth clinical scores) and how to learn informative features of structural MR images. To this end, we develop a weakly supervised CNN to integrate feature extraction and model learning into a unified framework, where a unique weighted loss function is used to employ all available weakly labeled subjects for network training. Specifically, there are two main steps in the proposed wiseDNN method, that is, extraction of multiscale image patches and weakly supervised neural network, with details given below.

1) Extraction of Multiscale Image Patches: While there are millions of voxels in each brain MR image, the structural changes caused by dementia could be subtle, especially in the early stage of AD (e.g., MCI). If we directly feed the whole MR image into a deep neural network, the input data will include too much noisy/irrelevant information, bringing difficulty in network training based on only a limited (i.e., hundreds) number of training subjects. Therefore, to facilitate the network training for accurate disease prognosis, we would like to first locate informative brain regions in each MRI, rather than using the whole image.

Following [42], we resort to anatomical landmarks to locate AD-related regions in brain MRIs. To be specific, we apply a landmark detection algorithm [42] to generate a total of 1741 anatomical landmarks defined in the Colin27 template [47].

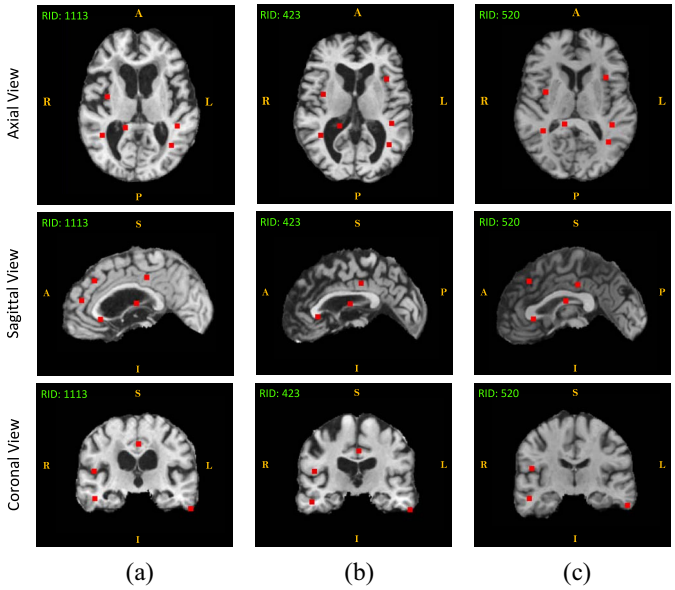

Fig. 3. Illustration of $K = 40$ anatomical landmarks of three typical subjects in their original image spaces: (a) AD; (b) MCI; and (c) HC, shown in three views (i.e., axial, sagittal, and coronal views). Each row denotes a particular view, and each column corresponds to a specific subject. RID: Roster ID.

As shown in Fig. S1 of the supplementary material, many landmarks are spatially close to each other. To reduce the information redundancy and computational burden, we select $K = 40$ anatomical landmarks from the original landmark pool via the following steps. We first rank these landmarks in the ascending order according to their p-values, where such p-values are generated by the landmark detection algorithm [42] by group comparison between AD and HC subjects. We then use a spatial Euclidean distance threshold (i.e., 20) as a criterion to control the distance between landmarks, and the top $K = 40$ landmarks are finally chosen to be used in this paper. As an illustration, we show the identified landmarks on three typical subjects in Fig. 3, while these landmarks shown in the template space can be found in Fig. S2 of the supplementary material.

Based on these identified landmarks, we extract multiscale image patches located by each landmark from an input MRI, to capture richer structural information of brain MRIs. Specifically, centered at each landmark, we extract both small-scale (i.e., with the size of $24 \times 24 \times 24$) and large-scale (i.e., with the size of $48 \times 48 \times 48$) patches from each MRI. Hence,

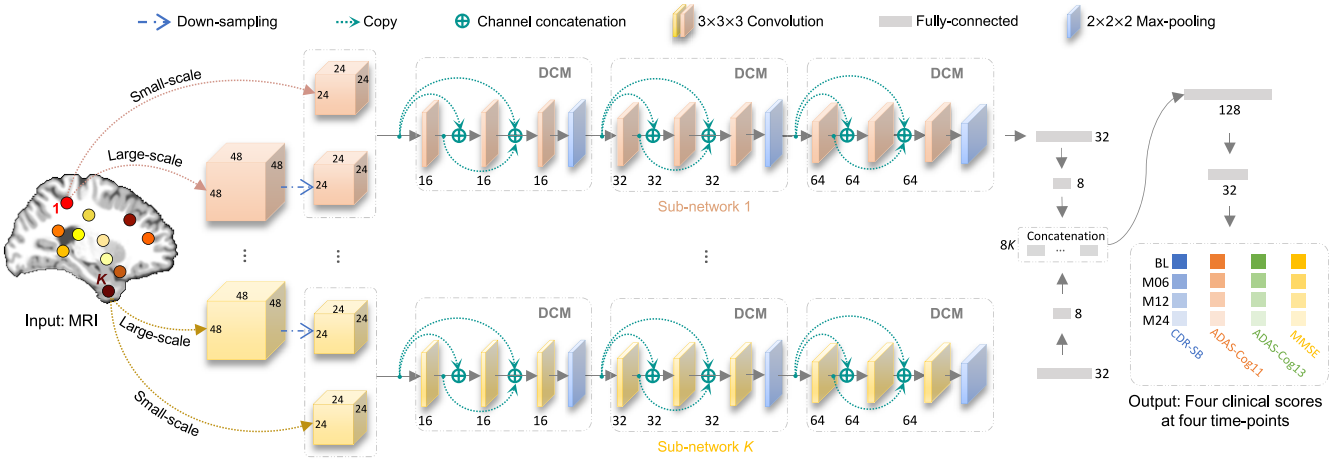

Fig. 4. Illustration of the proposed network using BL MRI data, containing K subnetworks. The input data are $2K$ image patches from each MR image, and the outputs are four types of clinical scores (i.e., CDR-SB, ADAS-Cog11, ADAS-Cog13, and MMSE) at four time-points (i.e., BL, M06, M12, and M24). DCM: Densely connected module. Each DCM contains a sequence of three convolutional layers, followed by max-pooling layer for image down-sampling.

given K landmarks, we can obtain $2K$ image patches from each subject (corresponding to a particular MRI). These multiscale image patches will be used as the input data of our proposed neural network.

2) Weakly Supervised Neural Network: Using multiscale image patches from each MRI, we jointly perform feature learning of MRIs and regression of multiple clinical scores at four time-points via the proposed neural network (with the architecture given in Fig. 4). As shown in Fig. 4, the input of the proposed network includes $2K$ image patches from each subject, and the output contains four types of clinical measures (i.e., CDR-SB, ADAS-Cog11, ADAS-Cog13, and MMSE) at four time-points (i.e., BL, M06, M12, and M24).

We first focus on modeling relatively local structural information contained in multiscale image patches via K parallel subnetworks, with each subnetwork corresponding to a specific landmark location. In each subnetwork, we first down-sample the large-scale (i.e., $48 \times 48 \times 48$) patch to have the same size as that of the small-scale (i.e., $24 \times 24 \times 24$) patch. Then, these two small-scale patches are treated as the two-channel input and fed into each subnetwork that contains a sequence of three densely connected modules (DCMs) and two fully connected (FC) layers. There are three convolutional layers in each DCM, followed by a $2 \times 2 \times 2$ max-pooling layer for feature map down-sampling. Particularly, for a specific convolutional layer in each DCM, feature maps (i.e., output images of each convolutional layer) of all preceding layers are used as inputs, and its own feature maps are used as inputs for all subsequent layers. All convolutional layers are followed by batch normalization and rectified linear unit (ReLU) activation. It has been proven that such densely connected architecture is useful in strengthening feature propagation, encouraging feature reuse, as well as substantially reducing the number of parameters to be optimized in the network [49]. Even though the K parallel subnetworks share the same architecture, their parameter weights are optimized independently. The motivation is that we would like to learn landmark-specific local features from image patches via K subnetworks to keep the unique local structural information provided by each landmark location. If those subnetworks share parameters, we would not be able to capture the landmark-specific local structural information of brain MRIs via shallow subnetworks.

It is worth noting that using only each local patch individually would not be able to capture the global structure of an MRI. To this end, the feature maps learned from the last K FC layers in K subnetworks are further concatenated, followed by two additional FC layers for learning local-to-global feature representations of the input MR image. The final FC layer (with 32 neurons) is employed to predict four types of clinical scores at four time-points.

Inspired by [50], we design a weighted loss function in the proposed network, to make full use of all available weakly labeled training subjects (with missing ground-truth clinical scores at certain time-points). We denote $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_N]$ as the training set containing N subjects, where $\mathbf{W}$ is the network coefficient. For the nth ($n = 1, \ldots, N$) subject $\mathbf{x}_n$, its sth ($s = 1, \ldots, S$) ground-truth clinical score at the t-th ($t = 1, \ldots, T$) time-point is denoted as $y_n^{s,t}$. The proposed objective function aims to minimize the difference between the predicted score $f^{s,t}(\mathbf{x}_n; \mathbf{W})$ and the ground truth $y_n^{s,t}$ in the following:

$$\arg\min_{\mathbf{W}} \frac{1}{N} \sum_{n=1}^{N} \frac{1}{\sum_{t=1}^{T} \sum_{s=1}^{S} \gamma_n^{s,t}} \sum_{t=1}^{T} \sum_{s=1}^{S} \gamma_n^{s,t} \times \left(y_n^{s,t} - f^{s,t}(\mathbf{x}_n; \mathbf{W}) \right)^2 \quad (1)$$

where $\gamma_n^{s,t}$ is an indicator to denote whether $\mathbf{x}_n$ is labeled with the sth clinical score at the t-th time-point. Specifically, $\gamma_n^{s,t} = 1$ if the ground-truth score $y_n^{s,t}$ is available for $\mathbf{x}_n$; and $\gamma_n^{s,t} = 0$, otherwise. To be specific, even when a training subject has missing scores at certain time-points and does not contribute to the loss computation (i.e., $\gamma_n^{s,t} = 0$), it can still contribute to the prediction tasks at the *remaining* time-points during the network training. Hence, more samples can be employed at different time-points. Using (1), we can *not only* automatically learn feature representation of MR images in a data-driven manner *but also* utilize all available subjects (even though some of them may lack ground-truth clinical scores

at several time-points) for model training. This is different from the conventional supervised methods [14], [16], [46] that simply discard subjects with incomplete ground-truth scores.

3) Implementation: To augment the training samples as well as reduce the negative influence of landmark detection errors, we randomly sample different patches centered at each landmark location with displacements within a $5 \times 5 \times 5$ cubic, and the step size is 1. Thus, a total of 125 patches centered at each landmark can be extracted from each MRI at each scale. Given K landmarks, we can obtain 125^K combinations of patches at each scale, with each combination being regarded as a particular sample for the proposed network. In this way, we can theoretically generate 125^K samples for representing each MR image, and these samples will be randomly used as the input data of the proposed network. More details can be found in Fig. S4 of the supplementary material.

At the *training* stage, we train the network based on the training subjects, using their BL MRIs as input and the corresponding ground-truth four clinical scores at four time-points (with missing values) as output. Specifically, based on K anatomical landmarks, we first sample multiscale (i.e., $24 \times 24 \times 24$ and $48 \times 48 \times 48$) image patches from each train MRI, and then feed these patches to the network. In this way, we can learn a nonlinear mapping from each input MRI to its four clinical scores at four time-points. At the *testing* stage, for an unseen test subject with only a BL MR image, we first locate its corresponding landmarks via the deep learning-based landmark detection algorithm [42], and then extract multiscale patches based on these landmarks. Finally, we feed these multiscale image patches into the learned network to predict the clinical scores at four time-points for this test subject.

We implement the proposed network using Keras [51] with Tensorflow backend [52]. The objective function in (1) is optimized by a stochastic gradient descent (SGD) approach [53] combined with a backpropagation algorithm for computing network gradients as well as parameter update. Specifically, we empirically set the momentum coefficient and the learning rate for SGD to 0.9 and 10^{-4}, respectively. The change curves of the training and validation loss function on the ADNI-1 dataset can be found in Fig. S2 of the supplementary material. Particularly, for an unseen testing subject with BL MRI, our method requires approximately 12 s for predicting its four types of clinical measures at four time-points, using a computer with a single GPU (i.e., NVIDIA GTX TITAN 12 GB). This implies that the proposed wiseDNN method is expected to perform real-time brain disease prognosis in real-world applications. For readers' convenience, the code and the pretrained model have been made publicly available online.[2]

IV. Experiments

In this section, we first present the experimental settings and introduce the methods for comparison. We then present and analyze the experimental results achieved by different methods on both ADNI-1 and ADNI-2 datasets, and compare our method with several state-of-the-art methods for MRI-based AD prognosis.

[2]https://github.com/mxliu/wiseDNN

A. Experimental Settings

To investigate the robustness of the proposed method, we perform two groups of experiments in a twofold cross-validation manner. Specifically, in the first group of experiments, we train models on subjects from ADNI-1, and test them on subjects from the independent ADNI-2 dataset. In the second group, we train models on ADNI-2 and test them on ADNI-1, respectively. In the experiments, we aim to predict four types of clinical scores (i.e., CDR-SB, ADAS-Cog11, ADAS-Cog13, and MMSE) at four time-points (i.e., BL, M06, M12, and M24), by using BL MRI data. Two criteria are used to evaluate the performance of our method and those competing approaches: 1) the correlation coefficient (CC) and 2) the root mean square error (RMSE) between the ground-truth and the predicted clinical scores achieved by a particular method. We also perform a paired t-test (with a significance level of 0.05) on prediction results achieved by our wiseDNN method and each specific comparison method.

B. Methods for Comparison

We first compare the proposed **wiseDNN** method with three conventional feature representations of brain MRI: 1) voxel-based tissue density (denoted as **Voxel**) [19]; 2) ROI-based GM volume (denoted as **ROI**) [11]; and 3) landmark-based morphological features (denoted as **LMF**) [42]. The details of these methods are introduced in the following.

1) *Voxel:* In this method, a nonlinear image registration algorithm [54] is first applied to spatially normalize all MR images to the template space. Then, the FAST algorithm in the FSL package [55] is employed to segment each MR image into three tissues (i.e., GM, WM, and CSF). Then, the local GM tissue density in each voxel is used as a feature value. The feature vector (with each element corresponding to a particular voxel) of each MR image is finally fed into a support vector regressor (SVR) for clinical score regression. For each type of clinical scores at a specific time-point, a linear SVR (with a default parameter $C = 1$) will be learned based on training subjects and, thus, models for different scores at different time-points are independently trained in this method.
2) *ROI:* Similar to the Voxel method, the ROI method first segments the GM tissue from the spatially normalized MR image. Using a nonlinear registration algorithm, the subject-labeled image based on the AAL template with 90 manually labeled ROIs can be generated. For each of 90 brain regions in the labeled MR image, the GM tissue volume of that region is computed as a feature. Then, a 90-D feature vector can be generated for each MR image, and such a feature vector is then fed into a linear SVR (with a default parameter $C = 1$) for independent regression of different clinical scores at different time-points.
3) *LMF:* In this method, the same K landmarks as those used in our method are employed to locate AD-related image patches in MRI. Specifically, centered at a particular landmark, LMF first extracts an image patch

TABLE II
RESULTS OF CCS BETWEEN THE GROUND-TRUTH AND PREDICTED SCORES OF FOUR TYPES OF CLINICAL MEASURES, ACHIEVED BY DIFFERENT METHODS AT FOUR TIME-POINTS (i.e., BL, M06, M12, AND M24). HERE, LEARNING MODELS ARE TRAINED AND TESTED ON ADNI-1 AND ADNI-2, RESPECTIVELY. THE BEST RESULTS ARE SHOWN IN BOLD. THE TERM DENOTED BY * REPRESENTS THAT THE RESULTS OF WISEDNN ARE STATISTICALLY SIGNIFICANTLY BETTER THAN OTHER COMPARISON METHODS ($p < 0.05$) USING PAIRWISE t-TEST

Correlation Coefficients	CDR-SB				ADAS-Cog11				ADAS-Cog13				MMSE			
	BL	M06	M12	M24	BL	M06	M12	M24	BL	M06	M12	M24	BL	M06	M12	M24
Voxel	0.238	0.181	0.174	0.134	0.325	0.334	0.151	0.150	0.331	0.253	0.255	0.192	0.309	0.310	0.293	0.146
ROI	0.239	0.223	0.114	0.155	0.309	0.243	0.099	0.165	0.344	0.349	0.309	0.234	0.306	0.254	0.233	0.265
LMF	0.431	0.435	0.452	0.339	0.527	0.539	0.512	0.414	0.554	0.573	0.543	0.445	0.331	0.405	0.423	0.364
wiseDNN-IS	0.487	0.478	0.510	0.424	0.584	0.605	0.567	0.462	0.616	0.625	0.601	0.502	0.538	0.564	0.526	0.474
wiseDNN-I	0.495	0.503	0.512	0.456	0.585	0.616	0.572	0.497	0.618	0.641	0.607	0.524	0.572	0.543	0.505	0.478
wiseDNN-S	0.515	0.506	0.514	0.455	0.593	0.615	0.538	0.493	0.622	0.634	0.575	0.522	0.542	**0.590**	0.530	0.473
wiseDNN-C	0.515	0.520	0.509	0.507	0.597	**0.632**	0.590	0.536	0.626	0.622	**0.627**	0.569	0.546	0.559	0.559	0.523
wiseDNN	**0.539***	**0.538***	**0.551***	**0.527***	**0.626***	0.622	**0.591**	**0.552***	**0.652***	**0.649***	0.616	**0.583***	**0.589***	0.586	**0.579***	**0.537***

$(24 \times 24 \times 24)$, and then compute the 100-D local energy pattern [43] as features for this patch. By concatenating these patch-based features, a $100K$-D feature vector is finally obtained for representing each MR image, followed by a linear SVR for clinical score regression.

Three critical strategies are used in wiseDNN, that is, joint prediction of multiple clinical scores at multiple time-points; multiscale patch extraction; and utilization of weakly supervised subjects (with complete BL MRI data and incomplete ground-truth clinical scores) for model training. To investigate the influence of each strategy, we further compare wiseDNN with its four variants, with details given below.

1) *wiseDNN-IS*, that independently trains a model for each clinical measure using only small-scale (i.e., $24 \times 24 \times 24$) patches. In wiseDNN-IS, we simply train an independent network for each type of clinical measure at each time-point, with the input of small-scale image patches centered at K landmark locations.
2) *wiseDNN-I*, that learns an independent network for each type of clinical scores using multiscale patches. Specifically, in wiseDNN-I, we separately train a specific network for each type of clinical scores at each time-point, without exploiting the underlying relationship among different clinical measure and that among different time-points.
3) *wiseDNN-S*, that uses only small-scale patches for learning a joint network for four clinical measures. In this method, we only extract a small-scale $(24 \times 24 \times 24)$ patch from each landmark location in an MR image, without accounting for more global information of that image captured by large-scale $(48 \times 48 \times 48)$ patches.
4) *wiseDNN-C*, that uses only labeled subjects (i.e., with complete ground-truth clinical scores at four time-points) for network training. In other words, in wiseDNN-C, subjects with clinical scores missed at least at one time-point will be discarded. In contrast, wiseDNN can use all weakly labeled subjects for network training via a weighted loss function in (1). For a fair comparison, similar to wiseDNN, wiseDNN-C uses multiscale image patches for joint prediction of four types of clinical scores at four time-points.

In summary, the proposed wiseDNN method is compared with seven methods in the experiments. Among these methods, five approaches (i.e., Voxel, ROI, LMF, wiseDNN-IS, and wiseDNN-I) independently train models for different clinical measures, while three approaches (i.e., wiseDNN-S, wiseDNN-C, and wiseDNN) jointly predict multiple clinical measures. In three conventional feature-based methods (i.e., Voxel, ROI, and LMF), a linear SVR (with a default parameter $C = 1$) will be learned independently for each type of clinical scores at each time-point. Six landmark-based methods (i.e., LMF, wiseDNN-IS, wiseDNN-I, wiseDNN-S, wiseDNN-C, and wiseDNN) share the same landmark pool as shown in Fig. 3. Also, four methods (i.e., Voxel, ROI, LMF, and wiseDNN-C) can only employ subjects with complete ground-truth scores at four time-points, while the remaining ones (i.e., wiseDNN-IS, wiseDNN-I, wiseDNN-S, and wiseDNN) can take advantage of all available training subjects (even though some of them may lack clinical scores at certain time-points).

C. Prognosis Results on ADNI-2

In this group of experiments, we train models on ADNI-1, and test them on ADNI-2. In Table II, we report the CC values achieved by eight different methods, with * denoting that the results of wiseDNN are statistically significantly better than other compared methods ($p < 0.05$) using pairwise t-test. Besides, the RMSE values obtained by different methods are reported in Fig. 5. We further show scatter plots of the ground-truth *versus* predicted scores achieved by our proposed wiseDNN method in Fig. 6. From Table II and Figs. 5 and 6, one could have the following observations.

1) Methods (i.e., wiseDNN-IS, wiseDNN-I, wiseDNN-S, wiseDNN-C, and wiseDNN) using task-oriented features via neural networks usually yield better results (regarding both CC and RMSE), compared with methods (i.e., Voxel, ROI, and LMF) using hand-crafted features of MRI. Also, even though LMF employs the same landmarks as wiseDNN to locate patches in MRI, its performance is worse than that of wiseDNN. For instance, LMF achieves a CC value of 0.431, which is worse than that (i.e., 0.539) of wiseDNN in predicting CDR-SB at BL. This may occur due to that LMF uses expert defined features of MRI for disease prognosis, where these features may not be well coordinated with subsequent prediction models. These findings imply that the integration of feature extraction into model learning (as we do in this paper) helps improve the prognosis performance.

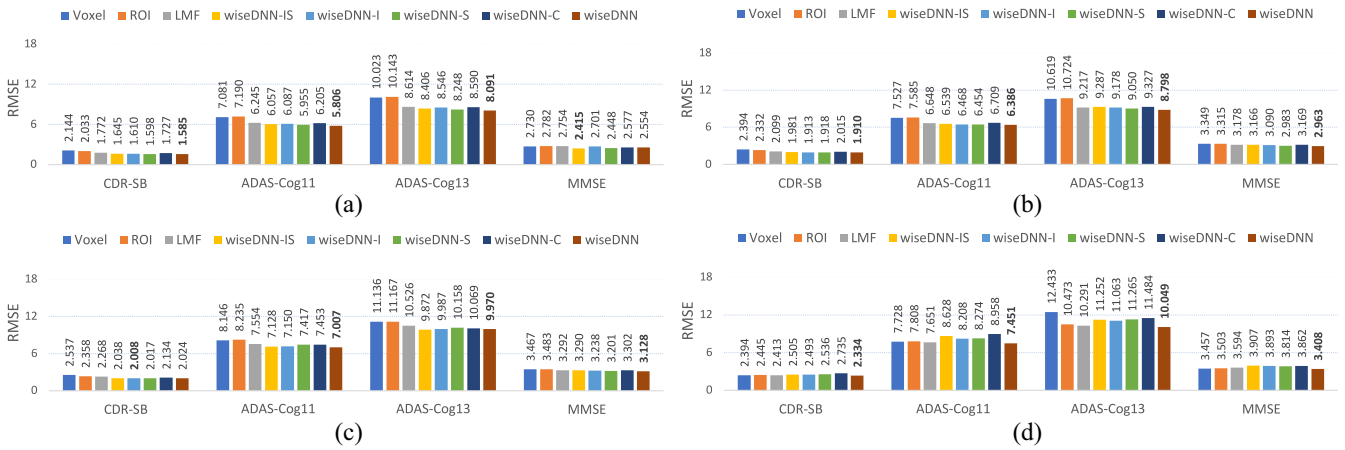

Fig. 5. Results of RMSE between the ground-truth and predicted clinical scores achieved by eight different methods at four time-points: (a) BL; (b) M06; (c) M12; and (d) M24. Here, learning models are trained and tested on ADNI-1 and ADNI-2, respectively.

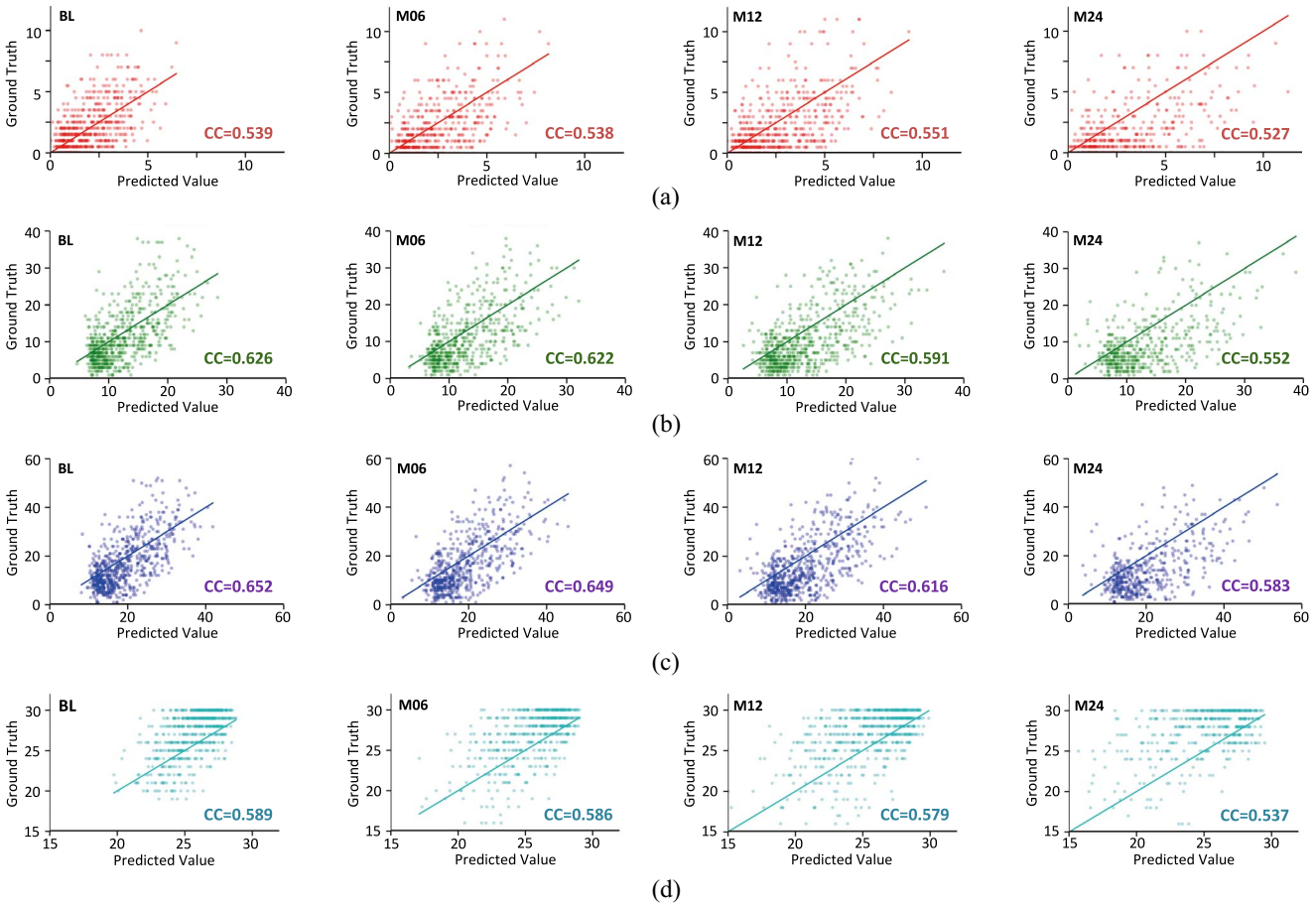

Fig. 6. Scatter plots of ground-truth versus predicted scores of: (a) CDR-SB; (b) ADAS-Cog11; (c) ADAS-Cog13; and (d) MMSE at four time-points, achieved by the proposed wiseDNN method. Each row denotes a particular clinical measure, and each column denotes a time-point. Here, the proposed neural network is trained and tested on ADNI-1 and ADNI-2, respectively.

2) Compared with those methods (i.e., Voxel, ROI, wiseDNN-IS, and wiseDNN-I) that independently learn models for different clinical scores, the proposed methods (i.e., wiseDNN-S, wiseDNN-C, and wiseDNN) that jointly predicting multiple scores generally yield higher CC and lower RMSE values. It suggests that our joint learning strategy could boost the learning performance, by implicitly exploiting the inherent relationship among different clinical measures.
3) Methods (i.e., wiseDNN-C and wiseDNN) using multiscale image patches consistently outperform their counterparts (i.e., wiseDNN-IS and wiseDNN-S) using only single-scale patches, in terms of both CC and RMSE values. Thanks to both small-scale and

TABLE III
RESULTS OF CCS BETWEEN THE GROUND-TRUTH AND PREDICTED SCORES OF FOUR TYPES OF CLINICAL MEASURES, ACHIEVED BY EIGHT DIFFERENT METHODS AT FOUR TIME-POINTS (i.e., BL, M06, M12, AND M24). HERE, LEARNING MODELS ARE TRAINED AND TESTED ON ADNI-2 AND ADNI-1, RESPECTIVELY. THE BEST RESULTS ARE SHOWN IN BOLD. THE TERM DENOTED BY * REPRESENTS THAT THE RESULTS OF WISEDNN ARE STATISTICALLY SIGNIFICANTLY BETTER THAN OTHER COMPARISON METHODS ($p < 0.05$) USING PAIRWISE t-TEST

Correlation Correlations	CDR-SB				ADAS-Cog11				ADAS-Cog13				MMSE			
	BL	M06	M12	M24	BL	M06	M12	M24	BL	M06	M12	M24	BL	M06	M12	M24
Voxel	0.197	0.210	0.247	0.245	0.146	0.281	0.283	0.232	0.222	0.261	0.291	0.224	0.208	0.216	0.292	0.251
ROI	0.190	0.286	0.213	0.227	0.205	0.267	0.250	0.277	0.203	0.265	0.293	0.337	0.211	0.230	0.199	0.218
LMF	0.417	0.420	0.426	0.459	0.512	0.513	0.503	0.510	0.541	0.525	0.514	0.493	0.435	0.416	0.426	0.457
wiseDNN-IS	0.436	0.449	0.448	0.496	0.575	0.601	0.585	0.588	0.596	0.615	0.603	0.597	0.489	0.475	0.512	0.511
wiseDNN-I	0.454	0.447	0.454	0.497	0.573	0.585	0.584	0.567	0.608	0.607	0.596	0.577	**0.528**	0.505	0.527	0.558
wiseDNN-S	0.468	0.474	0.478	0.496	0.580	0.574	0.565	0.557	0.604	0.582	0.579	0.554	0.502	0.485	0.521	0.526
wiseDNN-C	0.446	0.431	0.447	0.488	0.578	0.576	0.581	0.577	0.599	0.594	0.595	0.581	0.523	0.501	0.529	0.543
wiseDNN	**0.486***	**0.489***	**0.489***	**0.541***	**0.595***	**0.612***	**0.615***	**0.620***	**0.626***	**0.637***	**0.622***	**0.627***	0.525	**0.515***	**0.537***	**0.559**

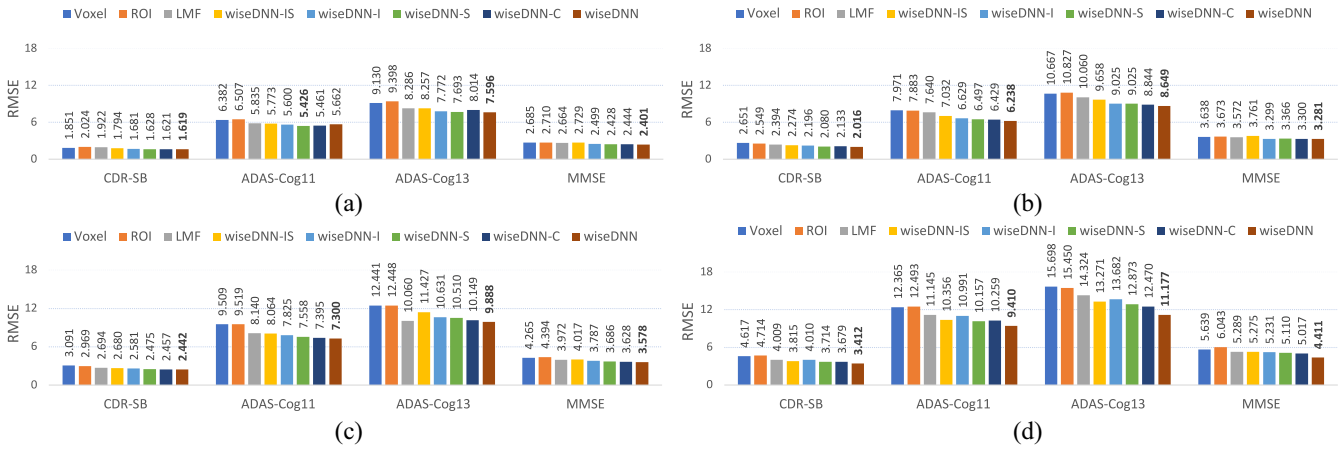

Fig. 7. Results of RMSE between the ground-truth and predicted clinical scores achieved by eight different methods at four time-points: (a) BL; (b) M06; (c) M12; and (d) M24. Here, learning models are trained and tested on ADNI-2 and ADNI-1, respectively.

large-scale image patches, wiseDNN-C and wiseDNN can model more global structure information of MRI, while wiseDNN-IS and wiseDNN-S merely focus on local information of MRI via single-scale patches. This could partly explain why using multiscale image patches can generate better prediction results.

4) The performance of each method slightly decreases over time in predicting four types of clinical scores. As an example, wiseDNN obtains an RMSE value of 3.408 at M24, which is worse than that (i.e., 2.554) at BL in predicting MMSE. This result may be caused by the fact that we only use BL MRI data to predict clinical scores at four time-points, while the brain structure may slightly change along time after BL time. A more reasonable solution is to use MRI at multiple time-points for predicting clinical scores at different time-points, which is expected to generate better performance.

D. Prognosis Results on ADNI-1

In the second group of experiments, we train and test learning models on ADNI-2 and ADNI-1, respectively. The CC and RMSE values achieved by different methods are shown in Table III and Fig. 7, respectively. From Table III and Fig. 7, we can see that, in most cases, our proposed wiseDNN method achieves better performance than the competing methods, regarding both CC and RMSE values.

Besides, it can be observed from Tables II and III and Figs. 5 and 7 that wiseDNN is superior to wiseDNN-C in predicting four types of clinical scores at four time-points. Note that wiseDNN uses all subjects with incomplete ground-truth scores for model training, while wiseDNN-C employs only subjects with complete scores. These results imply that using all available weakly labeled subjects for model learning, as we do in wiseDNN, provides a good solution to improve the prognosis performance. Also, Tables II and III suggest that the proposed wiseDNN method achieves comparable results using independent models trained on ADNI-1 and ADNI-2, respectively. This implies that our method has a good generalization ability because MRIs in ADNI-1 and ADNI-2 were acquired by 1.5 T and 3.0 T scanners, respectively.

In addition, we can see from Tables II and III that the overall correlation between the estimated scores achieved by all eight methods and the ground-truth scores are not high, and hence these methods are not yet ready to be used in a clinical setting. The underlying reason could be that the brain MRIs of subjects within four categories (AD, pMCI, sMCI, and HC) are combined for network training. Note that it is challenging to accurately distinguish between four categories, because the AD-related structural changes of the brain could be very subtle. Besides, MCI is the early stage of AD, and many MCI subjects (such as sMCI) will not necessarily convert to AD, leading to the complex data distribution of four categories.

E. Comparison With the State-of-the-Art Methods

In the literature, several methods have been proposed to predict clinical scores at multiple time-points [11], [46].

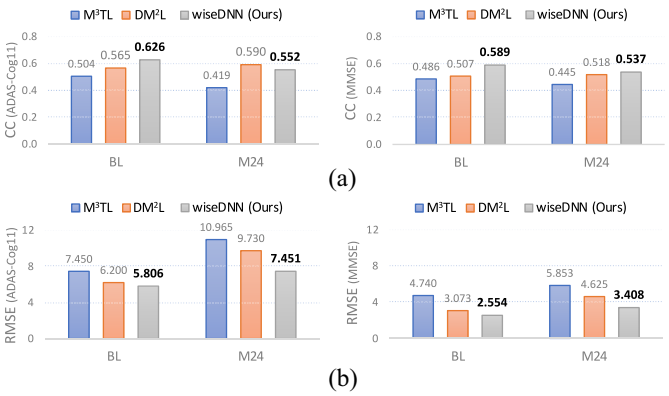

Fig. 8. Comparison between the proposed wiseDNN method and two state-of-the-art methods (i.e., M^{3}TL in [11] and DM^{2}L in [46]) in estimating two types of clinical scores (i.e., ADAS-Cog11 and MMSE) at two time-points (i.e., BL and M24), in terms of (a) CC and (b) RMSE. The models are trained on ADNI-1 and tested on ADNI-2, respectively.

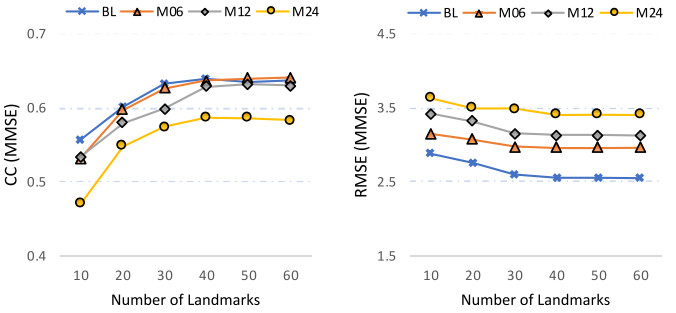

Fig. 9. Results achieved by the proposed wiseDNN method using different numbers of landmarks in MMSE score regression at four time-points, in terms of (left) CC and (right) RMSE. In this group of experiments, we train the network on ADNI-1 and test it on ADNI-2.

However, there are at least two major differences between our wiseDNN method and the conventional studies [11], [46]. Specifically, our wiseDNN method can automatically learn discriminative feature representations for MR images via a deep CNN, rather than using hand-crafted features (ROI-based GM tissue volume) [11]. Also, different from [11] and [46] that use only training subjects with complete ground-truth labels/scores, wiseDNN can utilize all weakly labeled (i.e., with incomplete ground-truth clinical scores) subjects for model learning, thus significantly increasing the number of training subjects and potentially boosting the diagnosis performance.

For comparison, we further report the prognosis results achieved by wiseDNN and these two previous methods, that is, multimodal multitask learning (M^{3}TL) [11] and deep multitask multichannel learning (DM^{2}L) [46]. Note that, M^{3}TL relies on an independent linear SVR for clinical score regression at each time-point, while both DM^{2}L and our wiseDNN methods perform the joint regression of multiple clinical scores via CNNs. For predicting two types of clinical scores (i.e., ADAS-Cog11 and MMSE) at two time-points (i.e., BL and M24), models in this group of experiments are trained on ADNI-1 and tested on ADNI-2, respectively. Both M^{3}TL and DM^{2}L can only employ subjects with complete ground-truth clinical scores for model training, while our wiseDNN method is capable of using all available subjects.

In Fig. 8, we report the CC and RMSE values in terms of the ground-truth and estimated clinical scores achieved by three different methods. It can be observed from Fig. 8 that our wiseDNN method outperforms M^{3}TL and DM^{2}L in most cases, further suggesting the effectiveness of our proposed method. Besides, compared with M^{3}TL that uses conventional hand-crafted features of MRI, two deep learning methods (i.e., DM^{2}L and wiseDNN) consistently yield higher CC and lower RMSE values, further suggesting that incorporating task-oriented feature learning into the process of training prognosis models can further improve the learning performance.

V. Discussion

In this paper, we propose a wiseDNN for simultaneous prediction of multiple clinical scores at multiple time-points, using subjects with BL MRI data and incomplete ground-truth scores. This method is potentially useful in clinical practice for disease prognosis in a fast and objective way. For instance, given a subject who is suspected to have AD or MCI evaluated by a physician, we can acquire the BL brain MR image. Using the proposed framework, we can feed this BL MRI to the trained network to predict how the clinical status changes over time for this patient, requiring only about 12 s. In the following, we analyze the influence of parameters and the effect of different network architectures, and then present the limitation of this paper and possible future research direction.

A. Parameter Analysis

We now evaluate the effect of the number of landmarks on the performance of the proposed wiseDNN method. In this group of experiments, we vary the number of landmarks in the range of [10, 20, . . . , 60], and record the CC and RMSE results achieved by wiseDNN in MMSE regression. The experimental results are shown in Fig. 9, with models trained and tested on ADNI-1 and ADNI-2, respectively.

From Fig. 9, we can observe that wiseDNN with less than 20 landmarks does not produce satisfactory results. The underlying reason could be that using a limited number of landmarks are not able to effectively capture global structural information of the brain MRI, thus leading to suboptimal learning performance. Besides, using more landmarks (e.g., 30), the results of our method are stable. Considering the computational burden, the optimal number of landmarks for our method can be selected in the range of [30, 50].

B. Effect of Densely Connected Module

In the proposed network shown in Fig. 4, we employ three DCMs in each subnetwork to learn patch-level local features. To study the influence of these DCMs, we further perform a group of experiments by replacing three DCMs with three residual learning modules (RLMs) [56] and denote this new network as **wiseDNN-R**. Note that wiseDNN-R and wiseDNN share the same input and objective functions. In Fig. S5 of the supplementary material, we show the network architecture of wiseDNN-R. The experimental results produced by both wiseDNN and wiseDNN-R in estimating MMSE scores at four time-points are reported in Fig. 10. It can be seen from Fig. 10 that wiseDNN-R achieves similar results with wiseDNN. This

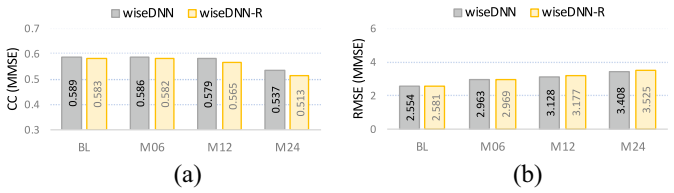

Fig. 10. Comparison between the proposed wiseDNN method and its variant (i.e., wiseDNN-R) in estimating MMSE at four time-points (i.e., BL, M06, M12, and M24), in terms of (a) CC and (b) RMSE. The models are trained on ADNI-1 and tested on ADNI-2, respectively.

demonstrates that our proposed landmark-based weakly supervised framework has strong compatibility with different types of network structures.

C. Limitation and Future Work

This paper still has several limitations, as listed below.

1) The proposed method was only validated on predicting clinical scores using only BL MRI scans, although there exist longitudinal MRI scans for subjects in both ADNI-1 and ADNI-2 datasets. The challenge of using longitudinal MRI scans for clinical score prediction is that many subjects have missing MRI scans at later time-points.
2) The current network primarily works for estimating multiple types of clinical scores, without considering the underlying association between clinical scores and class labels (e.g., AD or HC) of subjects.
3) The preselection of local patches based on anatomical landmarks is still independent of feature extraction and classifier construction, which may hamper the prognostic performance.
4) We did not consider the difference of data distribution of subjects in ADNI-1 and ADNI-2. This may negatively affect the generalization capability of our method.

Accordingly, we will continue this paper in the following directions.

1) We will employ both BL and longitudinal MRI scans to predict clinical scores, by imputing those missing MR scans via deep learning algorithms (e.g., generative adversarial networks [57], [58]) for reliable prediction. Also, it is possible to determine which time-point is the most important in the disease progression, based on the complete (after computation) MRI scans for predicting clinical scores at all time-points.
2) Since clinical scores and class labels for a particular subject are highly associated, it is reasonable to develop a unified deep learning model for joint regression and classification.
3) It is desired to automatically identify both patch- and region-level discriminative locations in whole brain MRI, upon which both patch- and region-level feature representations can be jointly learned and fused in a data-driven manner to construct disease classification models.
4) We plan to design a domain adaptation method for dealing with the problem of different data distribution [59], [60], which is expected to further improve the generalization capability of the proposed network further.

VI. Conclusion

In this paper, we proposed a wiseDNN for predicting multiple types of clinical measures, based on subjects with BL MRI data and incomplete ground-truth clinical scores. Specifically, we first preprocessed all MR images and identified disease-related anatomical landmarks via a landmark detection algorithm. Based on each landmark location, we extracted multiscale patches centered at each landmark. Using image patches as input data, we developed a densely connected neural network to simultaneously learn discriminative features of MRI and predict multiple clinical scores at four time-points. In our proposed network, a weighted loss function was developed to employ all available training subjects, even though some may lack complete ground-truth clinical scores at certain time-points. Experiments on 1469 subjects from the ADNI-1 and ADNI-2 datasets suggest that the proposed wiseDNN method can effectively predict clinical scores at future time-points using BL MRI data.

Acknowledgment

Data used in preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).

References

[1] N. Fox *et al.*, "Presymptomatic hippocampal atrophy in Alzheimer's disease. A longitudinal MRI study," *Brain*, vol. 119, no. 6, pp. 2001–2007, 1996.
[2] 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, 2002.
[3] C. R. Jack, Jr., *et al.*, "Serial PIB and MRI in normal, mild cognitive impairment and Alzheimer's disease: Mplications for sequence of pathological events in Alzheimer's disease," *Brain*, vol. 132, no. 5, pp. 1355–1365, 2009.
[4] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in Alzheimer disease," *Nat. Rev. Neurol.*, vol. 6, no. 2, pp. 67–77, 2010.
[5] B. Jie, M. Liu, D. Zhang, and D. Shen, "Sub-network kernels for measuring similarity of brain connectivity networks in disease diagnosis," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2340–2353, May 2018.
[6] J. Zhang, M. Liu, L. An, Y. Gao, and D. Shen, "Alzheimer's disease diagnosis using landmark-based features from longitudinal structural MR images," *IEEE J. Biomed. Health Inf.*, vol. 21, no. 6, pp. 1607–1616, Nov. 2017.
[7] 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.*, to be published.
[8] X. Liu, D. Tosun, M. W. Weiner, and N. Schuff, "Locally linear embedding (LLE) for MRI based Alzheimer's disease classification," *NeuroImage*, vol. 83, pp. 148–157, Dec. 2013.
[9] 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.
[10] C. R. Jack *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J. Magn. Reson. Imag.*, vol. 27, no. 4, pp. 685–691, 2008.
[11] 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, 2012.

[12] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, Jan. 2017.

[13] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1240–1251, May 2016.

[14] H. Li, M. Habes, and Y. Fan, "Deep ordinal ranking for multi-category diagnosis of Alzheimer's disease using hippocampal MRI data," *arXiv preprint arXiv:1709.01599*, 2017.

[15] J. Zhang, Q. Li, R. J. Caselli, J. Ye, and Y. Wang, "Multi-task dictionary learning based convolutional neural network for computer aided diagnosis with longitudinal images," *arXiv preprint arXiv:1709.00042*, 2017.

[16] 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.

[17] X. Zhu, H.-I. Suk, and D. Shen, "A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis," *NeuroImage*, vol. 100, pp. 91–105, Oct. 2014.

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

[19] 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, 2001.

[20] R. Honea, T. J. Crow, D. Passingham, and C. E. Mackay, "Regional deficits in brain volume in schizophrenia: A meta-analysis of voxel-based morphometry studies," *Amer. J. Psychiatry*, vol. 162, no. 12, pp. 2233–2245, 2005.

[21] Y. Fan, D. Shen, R. C. Gur, R. E. Gur, and C. Davatzikos, "COMPARE: Classification of morphological patterns using adaptive regional elements," *IEEE Trans. Med. Imag.*, vol. 26, no. 1, pp. 93–105, Jan. 2007.

[22] Y. Fan *et al.*, "Unaffected family members and schizophrenia patients share brain structure patterns: A high-dimensional pattern classification study," *Biol. Psychiatry*, vol. 63, no. 1, pp. 118–124, 2008.

[23] P. Coupé, J. Manjón, V. Fonov, S. F. Eskildsen, and D. L. Collins, "Patch-based morphometry: Application to Alzheimer's disease," in *Proc. Alzheimer's Assoc. Int. Conf.*, 2012, p. 1.

[24] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning* (Springer Series in Statistics), vol. 1. New York, NY, USA: Springer, 2001.

[25] H. Yamasue *et al.*, "Voxel-based analysis of MRI reveals anterior cingulate gray-matter volume reduction in posttraumatic stress disorder due to terrorism," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 15, pp. 9039–9043, 2003.

[26] E. A. Maguire *et al.*, "Navigation-related structural change in the hippocampi of taxi drivers," *Proc. Nat. Acad. Sci. USA*, vol. 97, no. 8, pp. 4398–4403, 2000.

[27] M. Liu and D. Zhang, "Sparsity score: A novel graph-preserving feature selection method," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 28, no. 4, 2014, Art. no. 1450009.

[28] B. Fischl and A. M. Dale, "Measuring the thickness of the human cerebral cortex from magnetic resonance images," *Proc. Nat. Acad. Sci. USA*, vol. 97, no. 20, pp. 11050–11055, 2000.

[29] 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, 2011.

[30] J. Lötjönen *et al.*, "Fast and robust extraction of hippocampus from MR images for diagnostics of Alzheimer's disease," *NeuroImage*, vol. 56, no. 1, pp. 185–196, 2011.

[31] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye, "Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data," *NeuroImage*, vol. 61, no. 3, pp. 622–632, 2012.

[32] S. Xiang *et al.*, "Bi-level multi-source learning for heterogeneous block-wise missing data," *NeuroImage*, vol. 102, pp. 192–206, Nov. 2014.

[33] L. Nie *et al.*, "Modeling disease progression via multisource multitask learners: A case study with Alzheimer's disease," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1508–1519, Jul. 2017.

[34] C. R. Jack, R. C. Petersen, P. C. O'Brien, and E. G. Tangalos, "MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease," *Neurology*, vol. 42, no. 1, p. 183, 1992.

[35] C. Jack *et al.*, "Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment," *Neurology*, vol. 52, no. 7, p. 1397, 1999.

[36] M. Atiya, B. T. Hyman, M. S. Albert, and R. Killiany, "Structural magnetic resonance imaging in established and prodromal Alzheimer's disease: A review," *Alzheimer's Disease Assoc. Disorders*, vol. 17, no. 3, pp. 177–195, 2003.

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

[38] B. Lei, P. Yang, T. Wang, S. Chen, and D. Ni, "Relational-regularized discriminative sparse learning for Alzheimer's disease diagnosis," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1102–1113, Apr. 2017.

[39] G. W. Small *et al.*, "Cerebral metabolic and cognitive decline in persons at genetic risk for Alzheimer's disease," *Proc. Nat. Acad. Sci. USA*, vol. 97, no. 11, pp. 6037–6042, 2000.

[40] P. Coupé *et al.*, "Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease," *NeuroImage Clin.*, vol. 1, no. 1, pp. 141–152, 2012.

[41] M. Liu, D. Zhang, and D. Shen, "Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis," *Human Brain Mapping*, vol. 35, no. 4, pp. 1305–1319, 2014.

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

[43] J. Zhang, J. Liang, and H. Zhao, "Local energy pattern for texture classification using self-adaptive quantization thresholds," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 31–42, Jan. 2013.

[44] A. Khvostikov, K. Aderghal, J. Benois-Pineau, A. Krylov, and G. Catheline, "3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies," *arXiv preprint arXiv:1801.05968*, 2018.

[45] S. Sarraf, D. D. DeSouza, J. Anderson, and G. Tofighi, "DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI," *BioRxiv*, 2017, Art. no. 070441.

[46] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis," *IEEE Trans. Biomed. Eng.*, to be published. doi: 10.1109/TBME.2018.2869989.

[47] C. J. Holmes *et al.*, "Enhancement of MR images using registration for signal averaging," *J. Comput. Assisted Tomography*, vol. 22, no. 2, pp. 324–333, 1998.

[48] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Trans. Med. Imag.*, vol. 17, no. 1, pp. 87–97, Feb. 1998.

[49] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993*, 2016.

[50] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. MICCAI*, 2016, pp. 424–432.

[51] F. Chollet *et al.* (2015). *Keras*. [Online]. Available: https://keras.io

[52] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Design Implementation*, 2016, pp. 265–283.

[53] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[54] Z. Xue, D. Shen, and C. Davatzikos, "CLASSIC: Consistent longitudinal alignment and segmentation for serial image computing," *NeuroImage*, vol. 30, no. 2, pp. 388–399, 2006.

[55] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, Jan. 2001.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[57] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[58] Y. Pan *et al.*, "Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2018, pp. 455–463.

[59] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 136–144.

[60] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," *arXiv preprint arXiv:1702.05374*, 2017.

Mingxia Liu, photograph and biography not available at the time of publication.

Jun Zhang, photograph and biography not available at the time of publication.

Chunfeng Lian, photograph and biography not available at the time of publication.

Dinggang Shen (F'18), photograph and biography not available at the time of publication.