PAPER

Multi-level fusion network for mild cognitive impairment identification using multi-modal neuroimages

To cite this article: Haozhe Xu *et al* 2023 *Phys. Med. Biol.* **68** 095018

IOP Publishing Phys. Med. Biol. 68 (2023) 095018 https://doi.org/10.1088/1361-6560/accac8

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
22 August 2022

REVISED
17 February 2023

ACCEPTED FOR PUBLICATION
5 April 2023

PUBLISHED
26 April 2023

PAPER

Multi-level fusion network for mild cognitive impairment identification using multi-modal neuroimages

Haozhe Xu[1,2,3], **Shengzhou Zhong**[1,2,3] **and Yu Zhang**[1,2,3,*]

[1] School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, People's Republic of China
[2] Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou 510515, People's Republic of China
[3] Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou 510515, People's Republic of China
[*] Author to whom any correspondence should be addressed.

E-mail: yuzhang@smu.edu.cn

Keywords: mild cognitive impairment, multi-modal neuroimages, convolutional neural network, multi-level fusion

Abstract

Objective. Mild cognitive impairment (MCI) is a precursor to Alzheimer's disease (AD) which is an irreversible progressive neurodegenerative disease and its early diagnosis and intervention are of great significance. Recently, many deep learning methods have demonstrated the advantages of multi-modal neuroimages in MCI identification task. However, previous studies frequently simply concatenate patch-level features for prediction without modeling the dependencies among local features. Also, many methods only focus on modality-sharable information or modality-specific features and ignore their incorporation. This work aims to address above-mentioned issues and construct a model for accurate MCI identification. *Approach*. In this paper, we propose a multi-level fusion network for MCI identification using multi-modal neuroimages, which consists of local representation learning and dependency-aware global representation learning stages. Specifically, for each patient, we first extract multi-pair of patches from multiple same position in multi-modal neuroimages. After that, in the local representation learning stage, multiple dual-channel sub-networks, each of which consists of two modality-specific feature extraction branches and three sine-cosine fusion modules, are constructed to learn local features that preserve modality-sharable and modality specific representations simultaneously. In the dependency-aware global representation learning stage, we further capture long-range dependencies among local representations and integrate them into global ones for MCI identification. *Main results*. Experiments on ADNI-1/ADNI-2 datasets demonstrate the superior performance of the proposed method in MCI identification tasks (Accuracy: 0.802, sensitivity: 0.821, specificity: 0.767 in MCI diagnosis task; accuracy: 0.849, sensitivity: 0.841, specificity: 0.856 in MCI conversion task) when compared with state-of-the-art methods. The proposed classification model has demonstrated a promising potential to predict MCI conversion and identify the disease-related regions in the brain. *Significance*. We propose a multi-level fusion network for MCI identification using multi-modal neuroimage. The results on ADNI datasets have demonstrated its feasibility and superiority.

1. Introduction

Mild cognitive impairment (MCI) is a condition in which an individual has mild but measurable changes in thinking abilities that do not affect daily activities (Association *et al* 2016). Partial old people with MCI (especially those with progressive MCI) are likely to suffer from Alzheimer's disease (AD) in the future, which is an irreversible disease (Kantarci *et al* 2009, Mitchell and Shiri-Feshki 2009). Timely medical intervention can help delay the deterioration process by discovering significant biomarkers and structural changes in the early

stage. In clinical practices, neuroimages, such as magnetic resonance imaging (MRI) and positron emission tomography (PET), atrophy status and brain function information (Hosseini-Asl *et al* 2016, Singh *et al* 2017, Zhang *et al* 2019a). Therefore, various neuroimage-based computer-aided diagnosis (CAD) methods have been developed for MCI status identification (Leandrou *et al* 2018). In general, exiting MCI identification approaches can be roughly classified into two categories, including traditional machine learning methods (Escudero *et al* 2012, Cheng *et al* 2015, Liu *et al* 2015, Liu *et al* 2016, Nie *et al* 2016, Tong *et al* 2016, Liu *et al* 2017a, Zhou *et al* 2019, Ansart *et al* 2021) and deep learning methods (Cui and Liu 2018, Liu *et al* 2018a, Li and Fan 2019, Spasov *et al* 2019, Zhang *et al* 2019b, Fang *et al* 2020, Lian *et al* 2020, Zhang and Shi 2020). The traditional mhachine learning methods refer to three main steps: (1) identifying the regions of interest (ROIs), (2) extracting features from ROIs, and (3) constructing the classifier. Recently, deep learning methods have shown promising potential in the field of brain disease identification. Different from traditional machine learning methods, they can not only learn features in a data-driven manner but also jointly conduct discriminative feature learning and classifier modeling (Cui and Liu 2018, Li and Fan 2019, Lian *et al* 2020). With the development of new neuroimaging technologies, many studies have demonstrated that multi-modal neuroimages can advance brain disease diagnosis, especially in cognitive impairment evaluation. For example, Zhang *et al* (2019b) proposed a deep learning network to combine multi-modal neuroimage in which two convolutional neural networks (CNNs) in conjunction with clinical neuropsychological information are conducted to distinguish AD from NC. Liu *et al* (2018a) constructed cascaded CNNs to learn multi-level and multi-modal features from MRI and PET images for AD classification. Fang *et al* (2020) employed three CNNs to generate a probabilistic score for the input slices of each modality and fused the probabilistic scores to train an ensemble classifier for prediction. Zhang and Shi (2020) proposed a deep multi-modal fusion model based on MRI and PET images for the early diagnosis of MCI conversion by learning the synergy between the multi-modal data. Researchers developed a graph-based deep neural network (Zhang *et al* 2021) to simultaneously model brain structure and function information using structural MRI and functional MRI to maximize the capability of differentiating MCI patients from elderly normal controls (NC). Although these methods significantly improve diagnostic performance, they merely focus on either modality-sharable information or modality-specific representations, or the simple concatenation of both, thereby the complementary information within multi-modal data is still not fully exploited.

However, deep learning methods for MCI diagnosis are usually hindered by the overfitting issue due to the limited sample size, to tackle which, several studies have proposed to extract local patches from the whole volume as a data augmentation strategy. For example, Liu *et al* (2019) proposed a weakly supervised densely connected neural network (wiseDNN) based on multi-scale patches centered on disease-related anatomical landmarks for predicting multiple types of clinical measures. Lian *et al* (2018) constructed a hierarchical fully convolutional network to automatically identify discriminative locations in MRI images, on which multi-scale representations are generated to construct the hierarchical classifier for MCI conversion prediction. Liu *et al* (2018b) proposed a deep multi-task multi-channel learning framework for joint brain disease classification and clinical score regression using both landmark-around patches and demographic information of subjects. Despite the fact that those patch-based methods have shown impressive accuracy, they simply concatenate patch-level features for diagnosis without modeling the dependencies among local features and ignore the correlation of brain regions, which may result in sub-optimal performance. In this paper, we propose a multi-level fusion network (MFN) for MCI identification using multi-modal neuroimages, which consists of local representation learning and dependency-aware global representation learning stages. The framework of our proposed networks is illustrated in figure 1. Specifically, for each patient, we extract multi-pair of patches from multiple same position in the multi-modal neuroimages. Then, in the local representation learning stage, we construct multiple dual-channel sub-networks (DCSs), each of which consists of two branches of modality-specific feature extraction (MFE) units and three sine-cosine fusion (SCF) modules, to learn local representations from multi pair of patches. Three MFE units in each branch are designed to extract multi-level modality-specific features while three SCF modules are devised to simultaneously learn modality-specific and modality-sharable representations along spatial and channel directions from multi-modal features of two branches. In the dependency-aware global representation learning stage, we additionally employ the long-range dependency capture (LRDC) module to model the correlations among local representations and integrate them into global ones for MCI identification. The main contributions of this paper are summarized as follows:

- We propose a multi-level fusion network for MCI identification with multi-modal neuroimages, and extensive experiments on public datasets demonstrate its superior abilities of generalization and biomarker localization.

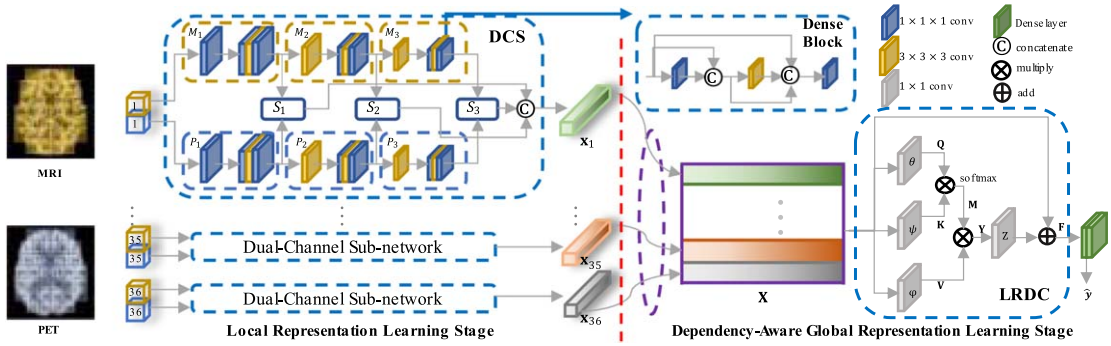


Figure 1. The framework of our proposed network for dementia identification. The proposed network consists of the local representation learning stage and dependency-aware global representation learning stage. The former stage contains multiple DCSs, each of which consists of two branches of three MFE units and three SCF modules, denoted as $\{M_i\}_{i=1}^{3}$, $\{P_i\}_{i=1}^{3}$, and $\{S_i\}_{i=1}^{3}$, respectively. Specifically, the ith MFE units, i.e. M_i and P_i, are designed to extract ith level local features from multi-modal patches. And the ith SCF module, i.e. S_i, is applied to simultaneously learn modality-sharable and modality-specific representations from ith level multi-modal features along the channel and spatial directions. Notably, multiple DCSs share the same network parameters across different patch locations. In the latter stage, we employ the LRDC module to model long-range dependencies among local representations. After integrating these local representation into the global ones, we feed them into dense layers for MCI identification.

Table 1. Demographic information of the subjects included in the studied datasets (i.e. ADNI-1 and ADNI-2).

Dataset	Category	Female/Male	Age	Education	MMSE
ADNI-1	NC	38/60	75.7 ± 4.7	15.9 ± 3.1	28.9 ± 1.1
	sMCI	35/86	74.9 ± 7.5	15.8 ± 2.9	27.4 ± 1.6
	pMCI	31/48	75.0 ± 6.7	15.8 ± 2.7	26.8 ± 1.7
ADNI-2	NC	62/28	71.9 ± 5.8	16.2 ± 2.5	29.2 ± 1.1
	sMCI	44/60	70.2 ± 6.3	16.6 ± 2.6	28.3 ± 1.7
	pMCI	29/40	73.1 ± 7.0	16.5 ± 2.6	27.3 ± 1.8

- Multiple DCSs based on multi-pair of patches, each of which consists of two branches of three MFE units and three SCF modules, are constructed to learn local features that preserve both modality-sharable and modality-specific representations along spatial and channel directions.
- We employ the LRDC module to model the long-range dependencies among local representations, based on which global representations are learned for MCI identification.

The rest of this paper is organized as follows. In section 2, we introduce the studied data and preprocessing steps. The proposed method is described in section 3. Subsequently, we present the experimental setting and results in section 4. Discussion and conclusion are provided in sections 5 and 6, respectively.

2. Materials

Two datasets from Alzheimer's disease neuroimaging initiative (ADNI) database (Jack *et al* 2008), including ADNI-1 and ADNI-2, were enrolled to evaluate the proposed method. According to standard clinical criteria, such as mini-mental state assessment scores (MMSE) and clinical dementia rating, these subjects were divided into two groups (NC and MCI). MCI subjects were further classified into stable MCI (sMCI) and progressive MCI (pMCI) based on whether they would convert to AD within 36 months after the baseline evaluation. Note that subjects who appeared in both datasets were retained in ADNI-1 but removed from ADNI-2. Totally, the ADNI-1 dataset consists of 98 NC, 121 sMCI and 79 pMCI subjects, while the ADNI-2 dataset contains 78 NC, 138 sMCI and 65 pMCI subjects. More demographic information can be found in table 1.

All MRI images were processed following a standard pipeline: (1) anterior commissure-posterior commissure (AC-PC) correction, (2) intensity inhomogeneity correction using N3 algorithm (Sled *et al* 1998), (3) skull stripping and cerebellum removal with aBEAT[4], (4) image registration to the Colin27 template

[4] https://www.nitrc.org/projects/abeat/

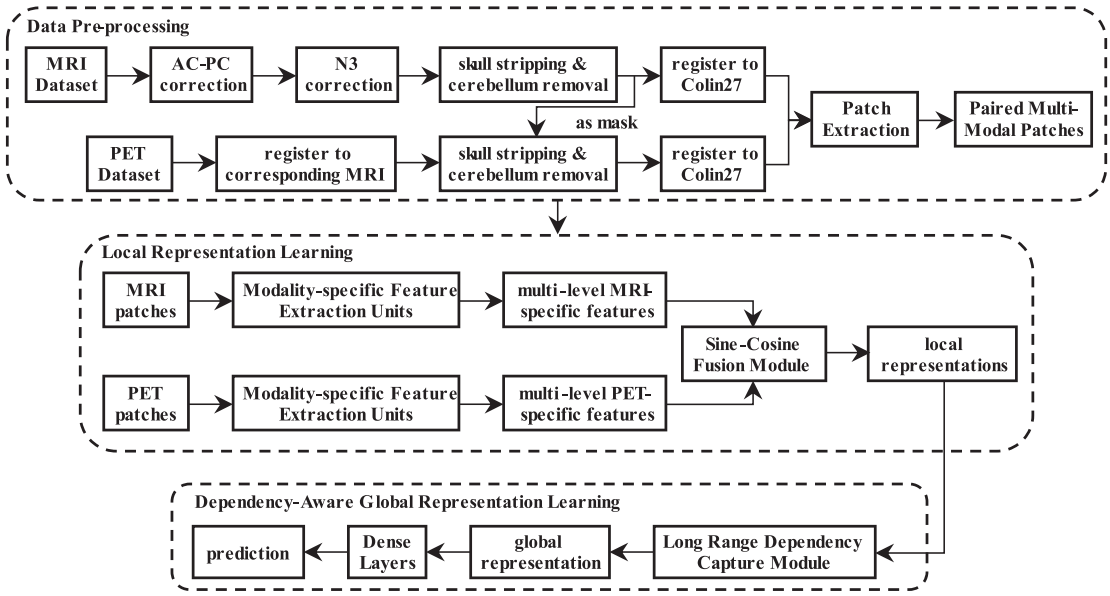


Figure 2. The flowchart of the proposed MFN mainly consists of three parts: data pre-processing, local representation learning, and dependency-aware global representation learning. We first feed the original MR image data into the aBEAT for AC-PC correction, N3 correction, skull stripping, and cerebellum removal. Meanwhile, the original PET image is registered to the corresponding MR image, further performing skull stripping and cerebellum removal using the processed MRI as the mask. Then, the cerebellum-removed multi-modal images are registered to the Colin27 template and also extracted patches to construct the paired multi-modal patches dataset. In the local representation learning stage, the multi-modal patches are fed into the corresponding modality-specific feature extraction units to extract multi-level and -modal features. Using the proposed sine-cosine fusion module, we can learn the local representation based on the features learned from multi-modal patches extracted from the same position. Finally, in the dependency-aware global representation learning stage, we fed the local representations into the long-range dependency capture module to learn the dependencies among patches to construct the global representation, which is further fed into the dense layers to make disease prediction.

(Holmes *et al* 1998) via SPM (Penny *et al* 2011). All PET images preprocessing contains following procedures: (1) registering to the MRI image of the same subject, (2) using skull-stripped and cerebellum-removed MRI image as mask to yield skull-stripped and cerebellum-removed PET image by multiplication, (3) registering the above PET image to the Colin27 template by using the deformation field between the corresponding MRI image and the template. Finally, all processed images were splitted into 36 non-overlapping patches with a size of $32 \times 32 \times 32$.

3. Method

In this section, we introduce the proposed multi-level fusion network in detail. For ease of understanding, we also drew a flowchart of MFN, as shown in figure 2. Specifically, we first pre-processed the multi-modal neuroimages and extracted patches to build the multi-modal patch dataset. Subsequently, we utilized multiple DCS and SCF modules to learn local representation from paired multi-modal patches. Furthermore, the local representation of patches extracted from multiple locations will be input into the LRDC module to learn the global representation to perform the final classification. The detailed architecture of our proposed networks is illustrated in figure 1 and we further elobrate on each components in the following sections.

3.1. Local representation learning stage

The local representation learning stage, which contains multiple DCSs, is constructed to learn local representations. Each DCS consisting of two branches of MFE units and three SCF modules (denoted as $\{S_i\}_{i=1}^3$) is applied to extract local representation for the specific position. Specifically, the multi-modal patches sampled from the same position of MRI and PET images are first fed into two branches to learn modality-specific features. With multi-modal features of two branches, we devise a SCF module to integrate them into comprehensive representations that preserve modality-sharable information and modality-specific characteristic simultaneously.

3.1.1. Modality-specific feature extraction units

As shown in figure 1, both branches contain three MFE units, which are denoted as $\{M_i\}_{i=1}^3$ and $\{P_i\}_{i=1}^3$, respectively, for extracting multi-level features from the input patches. Each MFE unit contains a convolutional

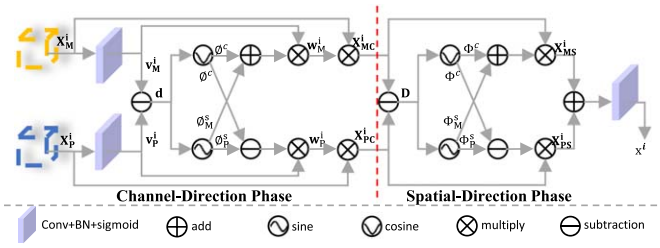


Figure 3. Architecture of the sine-cosine fusion (SCF) module for merging multi-modal feature with the same level into a comprehensive pattern. Specifically, the SCF module consists of channel-direction and spatial-direction phases, which learn modality-sharable and modality-specific representations along channel and spatial directions, respectively.

layer and a dense block that are constructed to extract level-specific features. Specifically, the convolutional layer of the first MFE unit of two branches is a $1 \times 1 \times 1$ convolutional layer with the stride of 1 to avoid the problem of information dropout caused by early down-sampling, while the other two adopt a $3 \times 3 \times 3$ convolutional layer with the stride of 2 to halve the size of spatial resolution and yield higher-level features. Meanwhile, each dense block in MFE units contains three convolutional layers, which are connected densely to avoid gradient vanishing and maximize information flow. Notably, each convolutional layer is followed by batch normalization and ReLU activation operators to prevent gradient explosion and enhance network sparsity, respectively. Moreover, all convolutional layers in the same unit share the same number of channel which is set to 16×2^i for ith MFE unit. Finally, feature maps $\mathbf{X}_{\mathrm{M}}^{\mathrm{i}} \in \mathbb{R}^{D_i \times H_i \times W_i \times C_i}$ and $\mathbf{X}_{\mathrm{P}}^{\mathrm{i}} \in \mathbb{R}^{D_i \times H_i \times W_i \times C_i}$ of M_i and P_i will be fed into ith SCF module, S_i, for exploring modality-sharable and modality-specific representations, where D_i, H_i, W_i and C_i are the depth, height, width and channel of the feature maps, respectively.

3.1.2. Sine-cosine fusion module

Various strategies can be used to fuse features from different modalities (Liu *et al* 2018a, Fang *et al* 2020, Zhang and Shi 2020). However, most of these strategies might not preserve both modality-sharable information and modality-specific representations at the same time. Additionally, different channels and spatial positions in features maps may contribute unequally to the modality-sharable and modality-specific representations. Therefore, we propose a SCF module that consists of channel-direction fusion and spatial-direction fusion phases to learn features that simultaneously preserve modality-sharable and modality-specific representations. As shown in figure 3, each module contains two blocks to learn features in the channel and spatial directions, respectively.

We first introduce the details of learning modality-sharable and modality-specific representations along channel direction. Given the input features $\mathbf{X}_{\mathrm{M}}^{\mathrm{i}}$ and $\mathbf{X}_{\mathrm{P}}^{\mathrm{i}}$ for S_i, we use two $D_i \times H_i \times W_i$ convolutional layers with channel number of C_i to generate two modality-specific representation vectors (i.e. $\mathbf{v}_{\mathrm{M}}^{\mathrm{i}}$, $\mathbf{v}_{\mathrm{P}}^{\mathrm{i}} \in \mathbb{R}^{1 \times 1 \times 1 \times C_i}$) for cross-modality feature learning. Notably, each convolutional layer is followed by batch normalization and sigmoid activation operators to map the feature into the range of [0, 1]. Then, we compute the difference vector $\mathbf{d} = \mathbf{v}_{\mathrm{M}}^{\mathrm{i}} - \mathbf{v}_{\mathrm{P}}^{\mathrm{i}} = [d^1, \cdots, d^{C_i}] \in \mathbb{R}^{1 \times 1 \times 1 \times C_i}$, where $\{d^j\}_{j=1}^{3}$ reflects the heterogeneity of the jth channel of multi-modal features. Subsequently, we apply the cosine function to the difference vector $\mathbf{d}$ to yield a modality-sharable coefficient vector ϕ^c, which is defined as follows

$$\phi^c = \cos(\mathbf{d}) = [\cos(d^1), \cdots, \cos(d^{C_i})]. \tag{1}$$

It's worth noting that $\cos(d^j)$ has a large value when multi-modal representations in the jth channel are close to each other. Similarly, we also apply the sine function on the difference vector $\mathbf{d}$ to construct the modality-specific coefficient vectors $\phi_{\mathrm{M}}^{\mathrm{s}}$ and $\phi_{\mathrm{P}}^{\mathrm{s}}$, which are computed by

$$\begin{aligned} \phi_{\mathrm{M}}^{\mathrm{s}} &= \sin(d) = [\sin(d^1), \cdots, \sin(d^{C_i})], \\ \phi_{\mathrm{P}}^{\mathrm{s}} &= \sin(-d) = [\sin(-d^1), \cdots, \sin(-d^{C_i})]. \end{aligned}$$

Intuitively, $\phi_{\mathrm{M}}^{\mathrm{s}}$ is opposite to $\phi_{\mathrm{P}}^{\mathrm{s}}$ for highlighting modality-specific characteristics. Subsequently, we yield the weighted vectors that preserve the modality-sharable and modality-specific characteristics by

$$\begin{aligned} \mathbf{w}_M^i &= \mathbf{v}_M^i \times (\phi^c + \phi_{\mathrm{M}}^{\mathrm{s}}), \\ \mathbf{w}_P^i &= \mathbf{v}_P^i \times (\phi^c + \phi_{\mathrm{P}}^{\mathrm{s}}). \end{aligned} \tag{2}$$

Finally, the channel weighted multi-modal features with modality-sharable and modality-specific representations along channel direction can be computed as follows:

$$\mathbf{X}^i_{MC} = \mathbf{X}^i_M \times \mathbf{w}^i_M,$$
$$\mathbf{X}^i_{PC} = \mathbf{X}^i_P \times \mathbf{w}^i_P. \tag{3}$$

In the spatial-direction phase, we first calculate the difference matrix $\mathbf{D}$ between the multi-modal features $\mathbf{X}^i_{MC}$ and $\mathbf{X}^i_{PC}$. Then, we compute the modality-sharable coefficient matrix (i.e. Φ^c), modality-specific coefficient matrices (i.e. Φ^s_M and Φ^s_P) and the spatial weighted multi-modal features $\mathbf{X}^i_{MS}$ and $\mathbf{X}^i_{PS}$ according to equations (1)–(3), where $\mathbf{X}^i_{MS}$ and $\mathbf{X}^i_{PS}$ can be formulated as:

$$\mathbf{X}^i_{MS} = \mathbf{X}^i_{MC} \times (\Phi^c + \Phi^s_{\mathrm{M}}),$$
$$\mathbf{X}^i_{PS} = \mathbf{X}^i_{PC} \times (\Phi^c + \Phi^s_{\mathrm{P}}). \tag{4}$$

We further integrate the multi-modal features $\mathbf{X}^i_{MS}$ and $\mathbf{X}^i_{PS}$ into $\mathbf{X}^{\mathrm{i}} \in \mathbb{R}^{D_i \times H_i \times W_i \times C_i}$ by

$$\mathbf{X}^i = \mathbf{X}^i_{MS} + \mathbf{X}^i_{PS}. \tag{5}$$

Then, the feature map $\mathbf{X}^i$ is fed into a $D_i \times H_i \times W_i$ convolutional layer with channel number of 12 to construct a ith level feature vector $\mathbf{x}^i \in \mathbb{R}^{1\times1\times1\times12}$. Finally, multi-level feature vectors from S_1, S_2 and S_3 are connected as local representations (i.e. $\mathbf{x} = [\mathbf{x}^1, \ \mathbf{x}^2, \ \mathbf{x}^3] \in \mathbb{R}^{1\times1\times1\times36}$) for subsequent global representations learning.

3.2. Dependency-aware global representation learning stage

Although modality-sharable and modality-specific representations can be learned by the dual-channel backbones, they are local patterns and focus on region-specific information and ignore the dependencies among brain regions (Lian *et al* 2018, Liu *et al* 2018b, Liu *et al* 2019). Inspired by (Wang *et al* 2018), we employ the LRDC module to construct the global representations that model long-range dependencies among local regions.

Given local representations of N regions, we first reshape and concatenate them as a feature matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{1\times N\times36}$, where $\{\mathbf{x}_k\}^N_{k=1}$ is the local representations of kth pair of multi-modal patches. Subsequently, we apply three 1×1 convolutional layers θ, ψ, and φ to map the feature matrix $\mathbf{X}$ into weight matrices $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ that are denoted as

$$\mathrm{Q} = f(\mathbf{X}; W_\theta),$$
$$\mathrm{K} = f(\mathbf{X}; W_\psi),$$
$$\mathrm{V} = f(\mathbf{X}; W_\varphi). \tag{6}$$

in which W_θ, W_ψ, and W_φ are learnable weight of convolutional layers θ, ψ, and φ, respectively. After that, these three weight matrixes will explore the dependencies among local positions in the self-attention manner (Vaswani *et al* 2017). Specifically, we first give $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ a new shape $\mathbb{R}^{36N\times1}$ through a resize operation, which contains $36N$ local features. Furthermore, we compute the correlation coefficient matrix $\mathbf{M} \in \mathbb{R}^{36N\times36N}$ by

$$\mathrm{M} = \mathrm{softmax}(\mathbf{Q} \otimes \mathbf{K}^T), \tag{7}$$

where $\otimes$ denotes matrix multiplication. Hence, the correlation coefficient m_{ij} in $\mathbf{M}$ reflects the dependency between local features in ith and jth positions. Following that, we multiply the correlation coefficient matrix $\mathbf{M}$ times $\mathbf{V}$ to get the correlation matrix $\mathbf{Y}$, which can be denoted as

$$\mathrm{Y} = \mathrm{M} \otimes \mathrm{V}. \tag{8}$$

Furthermore, we yield the global representation F by

$$\mathrm{F} = f(\mathbf{Y}; \boldsymbol{W_Z}) + \mathrm{X}, \tag{9}$$

where $\boldsymbol{W_Z}$ is the learnable weight of 1×1 convolutional layer Z. Finally, the global representation $\mathbf{F}$ is fed into two dense layers with 36 and 1 neurons which are followed by a sigmoid function to produce a subject-level diagnosis.

3.3. Loss function

We apply the binary cross-entropy as the loss function of our proposed method. Specifically, let I, y be the input neuroimages and corresponding class label. The learnable parameters for the local representation learning and dependency-aware global representation learning stages are denoted as $\boldsymbol{W_l}$ and $\boldsymbol{W_g}$, respectively. After getting the prediction output $\hat{\mathrm{y}}$, the binary cross-entropy loss can be defined as

$$\mathrm{Loss}(I; \boldsymbol{W_l}, \boldsymbol{W_g}) = y \cdot \log(\hat{\mathrm{y}}) + (1 - y) \cdot \log(1 - \hat{\mathrm{y}}). \tag{10}$$

The Loss will be propagated backwards to optimize the network parameters $\boldsymbol{W_l}$ and $\boldsymbol{W_g}$.

3.4. Implementation details

Similar to (Lian *et al* 2020), to evaluate the generalization capability of different methods, all models were trained on ADNI-1 and evaluated on ADNI-2. Additionally, we randomly selected 30% samples in ADNI-1 as the validation dataset. The diagnostic performance was quantitatively evaluated in terms of four criteria, that is, accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the receiver operating characteristic curve (AUC). We trained the proposed MFN by setting the mini-batch size and learning rate as 36 and 0.001, respectively, and applying dropout with the rate of 0.5 to multiple convolutional layers. We also applied the early stopping strategy to select the best model for test, i.e. when the loss reduction on validation dataset is less than 0.001 for 10 consecutive epochs, the training stopped and the model was selected. All parameters were initialized with 'kaiming_normal' (He *et al* 2015) and the model was trained with the SGD optimizer. In addition, all DCSs share the same weights to limit the number of learnable parameters. For competing methods, we followed the parameter settings in the literatures. We also performed delong-test for comparison between the proposed method and other competing methods. We used Pytorch package[5] based on Python3.7 to implement all comparison methods. All computationally intensive calculations were offloaded to a 12 GB NVIDIA RTX 2080Ti.

4. Experiments and results

In this section, we first introduce the experimental settings, including comparison methods, parameter settings, and evaluation strategy. Subsequently, we compare our proposed method with state-of-the-art methods and validate the effectiveness of the proposed method by ablation experiments. Finally, we also conduct discriminative localization analysis for the proposed method.

4.1. Experimental settings

We compare our proposed method with two conventional machine learning methods, including landmark-based morphological method (Zhang *et al* 2016) and voxel-based morphological method (Ashburner and Friston 2000). Besides, we further compare the proposed method with four recent deep learning methods, including two mono-modal approaches (i.e. 2.5D network (Kim *et al* 2021), multi-view separable pyramid network (Pan *et al* 2020a)) and two multi-modal methods (i.e. deep multi-modal fusion network (Zhang and Shi 2020), path-wise transfer dense convolution network (Gao *et al* 2021)). The details of these six competing methods are introduced as follows.

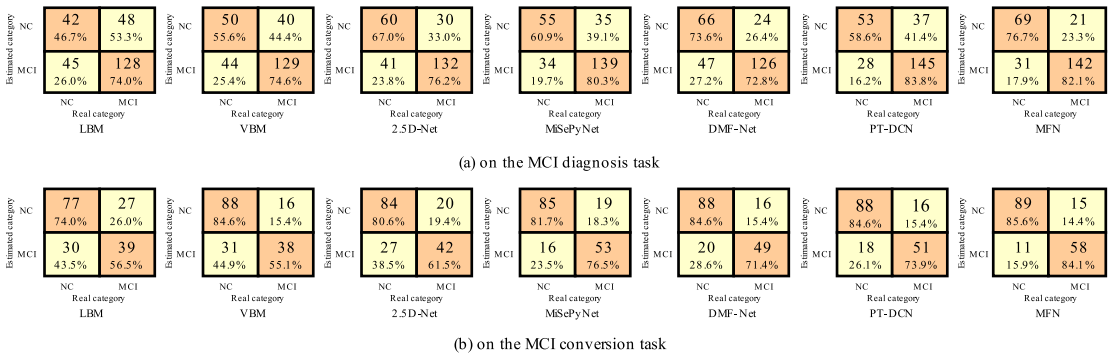
(1) Landmark-based morphological method (LBM): in the LBM method, $K = 50$ anatomical landmarks were used to locate $2K$ 3D patches from the PET and MR images. From each patch, 100D local energy patterns as features were extracted, and the features for different patches were concatenated as a $200K$-D vector. Finally, using these patch-level features, support vector regression (SVR) classifiers were trained to predict disease progress.

(2) 2.5D network (2.5D-Net): the 2.5D-Net method used slices of three views of PET data to construct 2D CNN classifiers. Following (Kim *et al* 2021), we separated 3D PET data into 2D slices from three views as the input of 2D sub-networks to yield slice-wise predictions. Notably, 2D sub-networks shared the same architecture but different parameters. Finally, we concatenated all slice-wise prediction values and then fed them into the fully connected layer for the final prediction.

(3) Multi-view separable pyramid network (MiSePyNet): the MiSePyNet method constructed a 3D CNN-based multi-scale model on the whole-brain MR images. In line with (Pan *et al* 2020a), we first constructed a slice-wise CNN for each view at the starting layer in a multi-scale manner to learn representations among slices. After that, we fed the intermediate feature into spatial-wise CNN which is also with different scales of convolutional kernels, to yield distinguishing spatial patterns for prediction tasks. Finally, we combined feature maps of different views and then fed them into fully-connected layers followed by a softmax function for classification.

(4) Deep multi-modal fusion network (DMF-Net): the DMF-Net method used slices filtered by AlexNet to construct a three-branch 2D network for diagnosis. Specifically, in line with (Zhang and Shi 2020), we first applied AlexNet to select slices of three views (i.e. Axial, Coronal and Sagittal) according to the classification accuracy corresponding to each slice-dataset, where the slice-dataset means the slices located in the same position of multi-modal neuroimages of all samples. Finally, we constructed a three-branch network, in

[5] https://pytorch.org/



(a) on the MCI diagnosis task

(b) on the MCI conversion task

Figure 4. The confusion matrices achieved by seven different methods in the MCI diagnosis task (MCI versus NC) and MCI conversion task (pMCI versus sMCI). The corresponding models are trained on ADNI-1 and tested on ADNI-2.

Table 2. Results of MCI diagnosis and MCI conversion prediction tasks, which are obtained by different methods trained on ADNI-1 and tested on ADNI-2, respectively. The best results are marked in boldface. The term denoted by * represents that the results of MFN are statistically significantly better than other comparison methods ($p < 0.05$) using delong-test.

Method	MCI versus NC				sMCI versus pMCI			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
LBM	0.646	0.740	0.467	0.583*	0.671	0.565	0.740	0.697*
VBM	0.681	0.746	0.556	0.542*	0.728	0.551	0.846	0.553*
2.5D-Net	0.730	0.762	0.670	0.791*	0.732	0.615	0.806	0.784*
MiSePyNet	0.738	0.803	0.609	0.739*	0.797	0.765	0.817	0.832*
DMF-Net	0.731	0.728	0.736	0.786*	0.792	0.714	0.846	0.761*
PT-DCN	0.754	**0.838**	0.586	0.776*	0.803	0.739	0.846	0.818*
MFN	**0.802**	**0.821**	**0.767**	**0.842**	**0.849**	**0.841**	**0.856**	**0.887**

which two branches are utilized to extract modality-specific features and the other is used to fuse multi-modal features for diagnosis.

(5) Pathwise transfer dense convolution network (PTDCN): the PT-DCN (Gao *et al* 2021) method gradually learned and combined the multi-level and multi-modal features of 3D MRI and PET data through a path-wise transfer deep convolution network for brain disease diagnosis. Specifically, the network contained two paths of dense convolutional networks and three transfer blocks. Each path learned features from specific single-modality data. Three transfer blocks fused multi-level intermediate features of two paths and then fed them back into two paths for further analysis. Finally, last outputs of two paths were concatenated and then fed into two convolution layers and a fully connected layer for diagnosis.

4.2. Results of MCI Diagnosis and conversion prediction task

Table 2 and figure 4 report the experimental results and confusion matrices of all comparison methods on MCI diagnosis and MCI conversion prediction tasks (i.e. MCI versus NC and sMCI versus pMCI), respectively. From the figure and table, we can observe four key points. First of all, compared with the conventional machine learning approaches (i.e. LBM and VBM), the deep-learning approaches (i.e. 2.5D-Net, MiSePyNet, DMF-Net, PT-DCN, and our MFN) largely improve the diagnostic performance, which demonstrates the significance of learning task-oriented features for brain disease diagnosis. Second, generally speaking, 3D-based methods (i.e. MiSePyNet, PT-DCN, and MFN) obtain better performance than 2D-based approaches (i.e. 2.5D-Net and DMF-Net) on two tasks. Actually, ADNI images contain highly complex patterns and individual-specific information. In 3D-based methods, the 3D convolutional kernel can learn representations across slices simultaneously, which can aid in the construction of spatial structural information to improve classification performance while the 2D convolutional kernel is insufficient to capture it, thereby limiting the performance. Moreover, the complexity (i.e. the sub-network number) of 2.5D-Net is higher than DMF-Net, which may fall into over-fitting when dealing with small-size datasets, whereas DMF-Net pre-selects informative slices to improve diagnosis performance. Third, multi-modal representations learning and fusion are crucial for improving diagnostic accuracy. For example, PT-DCN and MFN, both of which have the built-in multi-modal fusion mechanisms, consistently outperform mono-modal methods on two diagnostic tasks. As a 2D-based multi-modal method, DMF-Net yields better performance

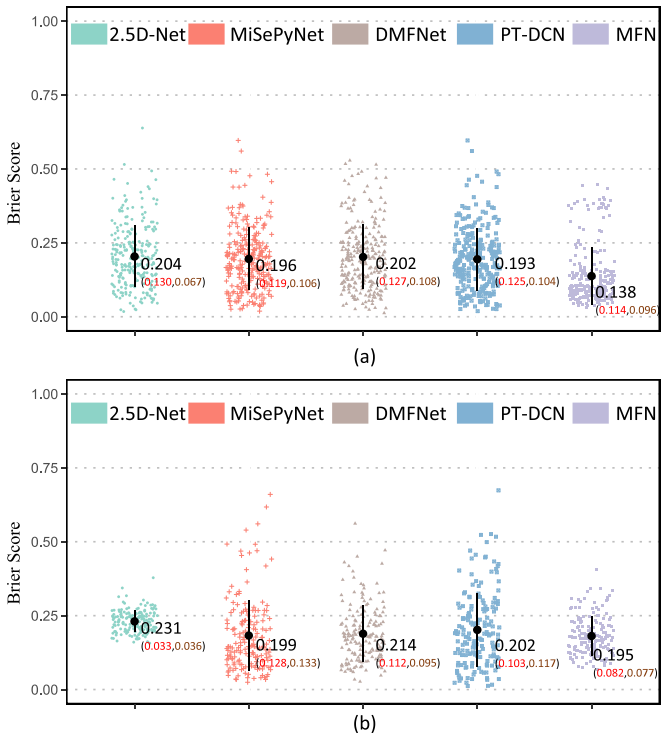


Figure 5. Brier score scatter charts (along with the mean value and intra-group standard deviations) of the deep learning methods on two tasks: (a) NC versus MCI; and (b) sMCI versus pMCI. The mean value, positive-group, and negative-group standard deviations are shown in black, red, and brown, respectively. And the positive and negative groups in (a) and (b) refer to the MCI and NC groups and the pMCI and sMCI groups, respectively.

than 2.5D-Net but is just comparable with MiSePyNet, which demonstrates the effectiveness of multi-modal fusion and also the significance of 3D structural information in the classification tasks. Last but not least, regardless of the MCI versus NC task or the sMCI versus pMCI task, our proposed method obtains the best performance with the highest metric values in most cases. Compared with other methods, several potential advantages exist in the proposed method: (1) The input 3D patches enable the network to learn spatial structural information and to make more robust decisions. (2) The SCF modules learn modality-sharable and modality-specific representations from multi-modal features of two branches along spatial and channel directions. (3) The LRDC module captures the dependencies among position-specific patterns to construct global features for final prediction. We also computed the Brier score (Brier *et al* 1950) and intra-group standard deviation (STD) to evaluate the performance of four deep learning methods on two tasks by providing the gap between the prediction value and the realistic target as well as the predictive variance of the same group. The Brier score scatter charts and intra-group STDs are provided in figure 5, in which the smaller mean value and STD value indicate better performance and robustness, respectively. From the figure, we can observe that MFN can achieve the best performance in terms of STD and Brier score metrics on the NC versus MCI task. Moreover, although the 2.5D-Net can yield the smallest STD value on the sMCI versus pMCI task, the mean Brier score is the higher than others, indicating the enormous gap between the prediction value and the realistic target. Meanwhile, the proposed MFN yields reasonable intra-group STD values and a mean Brier score on the sMCI versus pMCI task, which verifies its robustness and superiority again.

4.3. Ablation study

The ablation experiments are designed to validate the effectiveness of each component in the proposed method. All MFN-related variants can be divided into three categories as follows. (I) Multi-level learning related: Let MFN-M1 and MFN-M2 denote MFN with the DCS that only outputs the lowest-level features (i.e. $\mathbf{x}^1$) and highest-level features (i.e. $\mathbf{x}^3$) rather than comprehensive multi-level features, respectively. (II) Sine-Cosine fusion module related: let MFN-S1, MFN-S2, MFN-S3, MFN-S4, and MFN-S5 denote the MFN without modality-sharable exploration, modality-specificity exploration, spatial-direction fusion, channel- direction fusion, and the SCF module, respectively. (III) Dependency-aware global representation learning related: let MFN-L denotes the MFN without the LRDC module. The experimental results are shown in table 3, from which several points can be found. (I) The efficacy of multi-level learning mechanism. We can observe that MFN outperforms MFN-M1 and MFN-M2, which demonstrates that the multi-level features can help learn a comprehensive pattern including both containing more discriminative information than shallow features (i.e. $\mathbf{x}^1$). (II) The efficacy of the sine-cosine

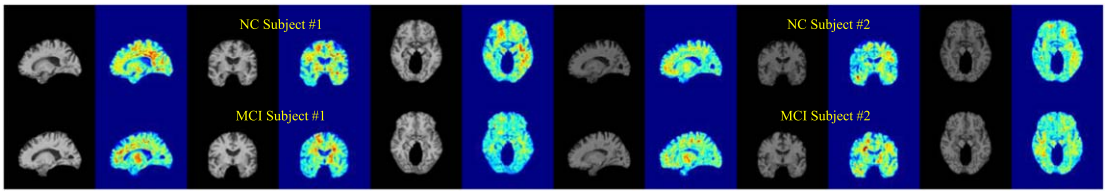


Figure 6. Illustration of the DAMs for four subjects in ADNI-2, including 2 NC subjects (top) and 2 MCI subjects (bottom), which are generated by the S_1 of the model trained on ADNI-1 in the task of MCI diagnosis.

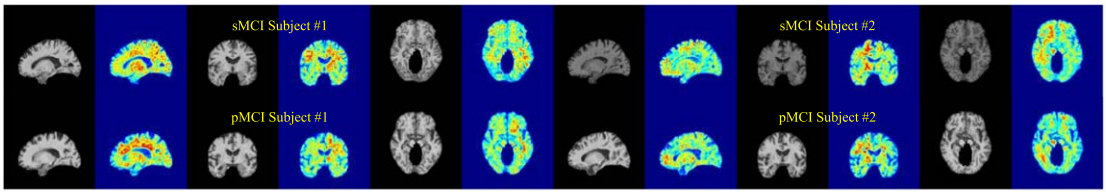


Figure 7. Illustration of the DAMs for four subjects in ADNI-2, including 2 sMCI subjects (top) and 2 pMCI subjects (bottom), which are generated by the S_1 of the model trained on ADNI-1 in the task of MCI conversion prediction.

Table 3. Results of MCI diagnosis and MCI conversion prediction tasks, which are obtained by different MFN-related variants trained on ADNI-1 and tested on ADNI-2, respectively. The best results are marked in boldface.

Method	MCI versus NC				sMCI versus pMCI			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
MFN-M1	0.688	0.786	0.500	0.676	0.659	0.493	0.769	0.648
MFN-M2	0.726	0.798	0.589	0.755	0.763	0.696	0.808	0.822
MFN-S1	0.722	**0.879**	0.422	0.758	0.763	0.594	0.879	0.800
MFN-S2	0.745	0.740	0.756	0.808	0.815	0.725	0.875	0.868
MFN-S3	0.722	0.798	0.578	0.749	0.815	0.696	0.894	0.849
MFN-S4	0.749	0.809	0.633	0.803	0.821	0.826	0.817	0.866
MFN-S5	0.711	0.740	0.656	0.769	0.809	0.739	0.856	0.865
MFN-L	0.730	0.746	0.700	0.761	0.763	0.739	0.779	0.839
MFN	**0.802**	0.821	**0.767**	**0.842**	**0.849**	**0.841**	**0.856**	**0.887**

fusion module. It is not surprising that MFN-S3, MFN-S4, and MFN work better than MFN-S5 since they execute feature enhancement operations in the channel, spatial, and both directions, respectively, to keep low redundancy. Also, the performance achieved by MFN-S1 and MFN-S2 (i.e. with the SCF module merely exploring modality-specific and modality-sharable representations) is superior to that of MFN-S5 (i.e. without the SCF module), which demonstrates the effectiveness of the SCF module in commonality and specificity exploration directions. Furthermore, the performance yielded by MFN-S1 is inferior to that of MFN-S2. A possible reason is that modality-sharable representation may be more significant to identify dementia-induced abnormalities than modality-specific information. (III) The efficacy of dependency-aware global representation learning. As expected, when the local dependency correlations among brain regions are not fully utilized, the diagnostic performance of MFN-L decreases significantly. This might be attributed to the fact that dementia-induced anatomical abnormalities are distributed across many brain regions, so considering multiple local information collaboratively will boost the diagnostic performance.

4.4. Discriminative localization analysis

It's of great significance to identify potential biomarkers associated with the prognosis of dementia. We randomly selected two subjects from two comparison groups in ADNI-2 as the input to a model trained on two tasks to generate the corresponding cross-channel averaged intermediate feature to produce a disease attention map (DAM). The DAMs obtained in the MCI versus NC and sMCI versus pMCI tasks have shown in figures 6 and 7. We depicted each DAM in 2D form from three perspectives (i.e. axial, coronal, and sagittal views). From these figures, we can observe three key points. First of all, the proposed MFN consistently highlights multiple parts at the locations of the hippocampus, frontal lobe, fusiform gyrus, amygdala, and thalamus for different

Table 4. Results of MCI diagnosis and MCI conversion prediction tasks, which are obtained by the MFN model with different patch sizes trained on ADNI-1 and tested on ADNI-2, respectively. The best results are marked in boldface.

size	amount	MCI versus NC				sMCI versus pMCI			
		ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
16	316	0.684	0.659	0.733	0.737	0.694	0.826	0.606	0.742
24	97	0.681	0.699	0.644	0.735	0.723	0.812	0.663	0.771
28	58	0.715	0.763	0.622	0.746	0.809	0.725	0.845	0.849
32	36	**0.802**	**0.821**	**0.767**	**0.842**	**0.849**	**0.841**	**0.856**	**0.887**
36	29	0.703	0.746	0.622	0.757	0.809	0.783	0.827	0.855
48	12	0.719	0.757	0.644	0.790	0.798	0.826	0.779	0.869

subjects with MCI or NC. It implies that the model can automatically focus on disease-related biomarkers to improve diagnostic performance, given that the discriminative power of these brain regions for dementia diagnosis has already been validated in previous studies (Wang *et al* 2007, Frisoni *et al* 2010, Zhang *et al* 2011, Coupé *et al* 2012, Aggleton *et al* 2016, Liu *et al* 2017b). Second, the most discriminative brain regions are consistent but have variant influences on different subjects. For example, most DAMs highlight the thalamus, while to varying degrees. It implies that the proposed MFN method is feasible for individualized diagnosis of brain atrophies associated with MCI, which should be a valuable property in practice. Third, by comparing figure 6 with figure 7, we can see that the concerning regions for the two tasks are partially different, although they are largely consistent. For example, the lingual gyrus regions are highlighted for most subjects in the sagittal view of figure 7, while not for figure 6. It is worth noting that previous studies (Liu *et al* 2017b) have shown that the lingual gyrus is an important biomarker for analyzing the progression of AD, which in some sense implies that our proposed method is effective in task-oriented discriminative localization to predict disease progression.

5. Discussion

In this section, we discuss the influence of the patch size, the comparison results with some previous studies, and limitations and future work in sequence.

5.1. The influence of patch size

The proposed network consists of local representation learning and dependency-aware global representation learning stages, in which the former stage aims to explore multi-level features of patches and the latter one aims to exploit the correlations among patches. Therefore, it is worth investigating that the influence of the input patch size. In this section, we conducted experiments on multiple networks with different input patch sizes, including 16, 24, 28, 32, 36, and 48. We trained these models on ADNI-1 and evaluated them on the ADNI-2 dataset as before. The quantified classification results in terms of four different metrics (i.e. ACC, SEN, SPE, and AUC) and the amount of patch corresponding to each size are listed table 4. From table 4, we can find that the proposed framework yields the best performance when the input size is equal to 32, while smaller or bigger inputs lead to worse results. It can be attributed to two reasons: (1) the patches with a small size contain less structural information that is necessary for the diagnosis. (2) A large patch size will reduce the quantity of extracted patches, which in unbeneficial to constructing location dependency among patches for the LRDC module. Fortunately, the patch size of 32 is a trade-off between enough structural information and a sufficient number of patches.

5.2. Comparison with previous studies

In table 5, we roughly summarize and compare our results with those of several state-of-the-art methods (Suk *et al* 2014, Liu *et al* 2017a, Cui and Liu 2018, Lian *et al* 2018, Pan *et al* 2018, Li *et al* 2019, Spasov *et al* 2019, Zhou *et al* 2019, Aderghal *et al* 2020, Hao *et al* 2020, Pan *et al* 2020b, Gao *et al* 2021) reported in the literature for AD diagnosis using baseline MRI or multi-modal data. The results presented in table 5 demonstrate that our proposed method achieves comparable performance to state-of-the-art methods in most cases. Despite making a direct comparison may not be entirely fair due to the varying number of subjects and inconsistent dataset partitions, we can still draw some conjectures from the results. First, the multi-modal fusion methods can learn more discriminative information by exploring the comprehensive characteristics inherent in multi-modal data for MCI diagnosis. As shown in table 5, the multi-modal data-based models (Suk *et al* 2014, Liu *et al* 2017a, Pan *et al* 2018, Spasov *et al* 2019, Aderghal *et al* 2020, Hao *et al* 2020, Pan *et al* 2020b, Gao *et al* 2021) achieved better performances on two tasks than single-modality data-based methods (Cui and Liu 2018, Lian *et al* 2018,

Table 5. A brief summary of the state-of-the-art studies based on ADNI dataset for MCI diagnosis. The best results are marked in boldface.

Method	Modality	Subjects	MCI versus NC				sMCI versus pMCI			
			ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
(Lian *et al* 2018)	MRI	429 NC + 465 sMCI + 205 pMCI	—	—	—	—	0.809	0.526	0.854	0.781
(Cui and Liu 2018)	MRI	223 NC +231 sMCI + 165 pMCI	0.746	0.773	0.700	0.777	0.750	0.733	0.762	0.797
(Li *et al* 2019)	MRI	216 NC + 233 sMCI + 164 pMCI	0.750	0.819	0.622	0.758	0.725	0.610	0.825	0.746
(Gao *et al* 2021)	MRI+PET	427 NC + 332 sMCI + 234 pMCI	—	—	—	—	0.753	0.708	0.784	0.778
(Pan *et al* 2018)	MRI+PET	429 NC + 465 sMCI + 205 pMCI	—	—	—	—	0.791	0.553	0.829	0.758
(Pan *et al* 2020b)	MRI+PET	420 NC + 562 sMCI + 146 pMCI	—	—	—	—	0.774	0.791	0.772	0.825
(Hao *et al* 2020)	MRI+PET	52 NC + 56 sMCI + 43 pMCI	0.845	0.940	0.662	0.810	0.778	0.674	0.855	0.760
(Suk *et al* 2014)	MRI+PET	101 NC + 128 sMCI + 76 pMCI	**0.857**	**0.954**	0.659	**0.881**	0.759	0.480	**0.952**	0.747
(Aderghal *et al* 2020)	MRI+DTI	399 NC + 273 MCI	0.785	0.777	**0.814**	0.796	—	—	—	—
(Zhou *et al* 2019)	MRI+PET+Genetic	211 NC + 205 sMCI + 157 pMCI	—	—	—	—	0.743	—	—	0.755
(Liu *et al* 2017a)	MRI+PET+CSF	226 NC + 395 MCI	0.800	0.862	0.688	0.805	0.790	0.608	0.925	0.797
(Spasov *et al* 2019)	MRI+DTI+clinical	184 NC + 228 sMCI + 181 pMCI	—	—	—	—	0.830	**0.875**	0.810	**0.917**
Proposed	MRI+PET	188 NC + 225 sMCI + 148 pMCI	0.802	0.821	0.767	0.842	**0.849**	0.841	0.856	0.887

Li *et al* 2019) in most situations, which demonstrates the significance of multi-modality neuroimages in MCI diagnosis. Second, the increases of data modalities might boost classification performance by providing informative specific views for MCI. For instance, compared with two-modality data-based models (Pan *et al* 2018, 2020b, Gao *et al* 2021), three-modality data-based methods (Liu *et al* 2017a, Spasov *et al* 2019) further improved the diagnosis performance in the MCI conversion task. Notably, the significant improvement of Hao *et al* (2020) and Suk *et al* (2014) on the MCI diagnosis task might benefit from the different data partitioning strategies, i.e. they are evaluated by cross-validation on a relatively small dataset. Third, comparing three-modality data-based models (Liu *et al* 2017a, Spasov *et al* 2019, Zhou *et al* 2019), the deep-learning method (Spasov *et al* 2019) can achieve better performance than the conventional machine-learning methods (Liu *et al* 2017a, Zhou *et al* 2019). A possible reason might be that the deep learning methods integrate feature extraction and model construction into a unified framework, while the machine learning methods are based on hand-crafted features that may mismatch with models and lead to suboptimal performance. Though only using two modalities of data, our MFN still performs better than most methods, which can be attributed to: (1) extracting multi-level feature representations for multi-modal neuroimages to enhance the feature robustness; (2) learning the underlying correlation between multi-modal neuroimages for feature fusion; (3) exploring the dependency correlation among local patches to construct robust global features to boost the diagnostic performance.

5.3. Limitations and future work

Although our proposed method achieves superior results in automatic MCI diagnostic tasks, its performance and generalization capacity could be further improved in the future by carefully dealing with the following limitations or challenges. First, in our current implementation, the input patches are sampled across the entire brain in a non-overlapping manner. Considering the abnormities caused by dementia are subtle and concentrated in pathological positions, it might be reasonable to extend our proposed method by using clinical prior knowledge to extract discriminative patches centered on disease-related landmarks. Second, the network pruning strategy can be used into our method to purify informative patches after pre-training the network. Third, it is worth noting that the datasets studied in this paper have different imaging data distributions due to different scanners (i.e. 1.5 T and 3 T scanners for ADNI-1 and ADNI-2, respectively). Hence, incorporating the domain adaptation strategy into our current framework might enhance its generalization capability. Finally, due to the expensive cost and radiation injury of multi-modality neuroimages, not all subjects have complete data, which has remained a common challenge in AD diagnosis based on multi-modality neuroimages. In this paper, we discard the modality-incomplete subjects, which could decrease the dataset's size and degrade the generality and accuracy of the classification model. Inspired by Liu *et al* (2022), a promising direction is to evaluate a virtual PET image based on its relevance with the corresponding MR image and extract multi-modal features in the generation process to ensure the feature's reliability.

6. Conclusion

In this paper, we propose a multi-level fusion network for mild cognitive impairment identification using multi-modal neuroimages that consists of local representation learning and dependency-aware global representation learning stages. Specifically, in the local representation learning stage, we construct multiple DCSs, each of which consists of two branches of MFE units and three SCF modules, to learn local representations which preserve both modality-sharable and modality-specific representations. Three MFE units in each branch are designed to extract multi-level modality-specific features while the SCF modules are devised to learn modality-sharable and modality-specific representations from multi-modal features of two branches. In the dependency-aware global representation learning stage, we employ the LRDC module to model the correlations among local representations and integrate them into global ones for MCI identification. On the ADNI public dataset with 561 subjects, the effectiveness of our proposed method on MCI versus NC and sMCI versus pMCI tasks has been extensively evaluated. Compared with several state-of-the-art CAD methods, our proposed method achieves better or at least comparable classification performance, especially in the relatively challenging task of MCI conversion prediction.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [grant numbers 61971213, 61671230]; the Basic and Applied Basic Research Foundation of Guangdong Province [grant number 2019A1515010417]; and the Guangdong Provincial Key Laboratory of Medical Image Processing [grant number No.2020B1212060039]. The approvals of the Ethics Committee are not available, in that the datasets used are

obtained from public databases, and we have cited the required references. We would like to express our gratitude to Dr Zhenyuan Ning for his constructive idea support and writing help.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://adni.loni.usc.edu/. Data will be available from 1 October 2004.

Conflict of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aderghal K *et al* 2020 Improving Alzheimer's stage categorization with convolutional neural network using transfer learning and different magnetic resonance imaging modalities *Heliyon* **6** e05652

Aggleton J P, Pralus A, Nelson A J and Hornberger M 2016 Thalamic pathology and memory loss in early Alzheimer's disease: moving the focus from the medial temporal lobe to papez circuit *Brain* **139** 1877–90

Ansart M *et al* 2021 Predicting the progression of mild cognitive impairment using machine learning: a systematic, quantitative and critical review *Med. Image Anal.* **67** 101848

Ashburner J and Friston K J 2000 Voxel-based morphometry the methods *NeuroImage* **11** 805–21

Association A *et al* 2016 2016 Alzheimer's disease facts and figures *Alzheimer»s & Dementia* **12** 459–509

Brier G W *et al* 1950 Verification of forecasts expressed in terms of probability *Mon. Weather Rev.* **78** 1–3

Cheng B, Liu M, Zhang D, Munsell B C and Shen D 2015 Domain transfer learning for mci conversion prediction *IEEE Trans. Biomed. Eng.* **62** 1805–17

Coupé P *et al* 2012 Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease *NeuroImage* **59** 3736–47

Cui R and Liu M 2018 Hippocampus analysis by combination of 3-d densenet and shapes for Alzheimer's disease diagnosis *IEEE J. Biomed. Health Inform.* **23** 2099–107

Escudero J *et al* 2012 Machine learning-based method for personalized and cost-effective detection of Alzheimer's disease *IEEE Trans. Biomed. Eng.* **60** 164–8

Fang X, Liu Z and Xu M 2020 Ensemble of deep convolutional neural networks based multi-modality images for Alzheimer's disease diagnosis *IET Image Proc.* **14** 318–26

Frisoni G B, Fox N C, Jack C R, Scheltens P and Thompson P M 2010 The clinical use of structural mri in alzheimer disease *Nat. Rev. Neurol.* **6** 67–77

Gao X, Shi F, Shen D and Liu M 2021 Task-induced pyramid and attention gan for multimodal brain image imputation and classification in alzheimers disease *IEEE J. Biomed. Health Inform.* **26** 36–43

Hao X *et al* 2020 Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of Alzheimer's disease *Med. Image Anal.* **60** 101625

He K, Zhang X, Ren S and Sun J 2015 Delving deep into rectifiers: surpassing human-level performance on imagenet classification *Proc. of the IEEE Int. Conf. on Computer Vision* pp 1026–34

Holmes C J, Hoge R, Collins L, Woods R, Toga A W and Evans A C 1998 Enhancement of Mr images using registration for signal averaging *J. Comput. Assist. Tomogr.* **22** 324–33

Hosseini-Asl E, Keynton R and El-Baz A 2016 Alzheimer's disease diagnostics by adaptation of 3d convolutional network *2016 IEEE Int. Conf. on Image Processing (ICIP)* (IEEE) 126–30

Jack C R Jr *et al* 2008 The Alzheimer's disease neuroimaging initiative (adni): Mri methods *J. Magn. Reson. Imaging* **27** 685–91

Kantarci K *et al* 2009 Risk of dementia in mci: combined effect of cerebrovascular disease, volumetric mri, and 1h mrs *Neurology* **72** 1519–25

Kim S *et al* 2021 Deep learning-based amyloid pet positivity classification model in the Alzheimer's disease continuum by using 2-[18f] fdg pet *EJNMMI Res.* **11** 1–14

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

Li F *et al* 2019 A hybrid convolutional and recurrent neural network for hippocampus analysis in Alzheimer's disease *J. Neurosci. Methods* **323** 108–18

Li H and Fan Y 2019 Early prediction of Alzheimer's disease dementia based on baseline hippocampal mri and 1-year follow-up cognitive measures using deep recurrent neural networks *2019 IEEE 16th Int. Symp. on Biomedical Imaging (ISBI 2019)* (IEEE) pp 368–71

Lian C, Liu M, Pan Y and Shen D 2020 Attention-guided hybrid network for dementia diagnosis with structural Mr images *IEEE Trans. Cybern.* **52** 1992–2003

Lian C, Liu M, Zhang J and Shen D 2018 Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural mri *IEEE Trans. Pattern Anal. Mach. Intell.* **42** 880–93

Liu M *et al* 2017a View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data *Med. Image Anal.* **36** 123–34

Liu M, Cheng D, Wang K and Wang Y 2018a Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis *Neuroinformatics* **16** 295–308

Liu M, Zhang D, Adeli E and Shen D 2015 Inherent structure-based multiview learning with multitemplate feature representation for Alzheimer's disease diagnosis *IEEE Trans. Biomed. Eng.* **63** 1473–82

Liu M, Zhang D and Shen D 2016 Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment *IEEE Trans. Med. Imaging* **35** 1463–74

Liu M, Zhang J, Adeli E and Shen D 2018b Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis *IEEE Trans. Biomed. Eng.* **66** 1195–206
Liu M, Zhang J, Lian C and Shen D 2019 Weakly supervised deep learning for brain disease prognosis using mri and incomplete clinical scores *IEEE Trans. Cybern.* **50** 3381–92
Liu X *et al* 2017b Decreased functional connectivity between the dorsal anterior cingulate cortex and lingual gyrus in Alzheimer's disease patients with depression *Behavioural Brain Res.* **326** 132–8
Liu Y *et al* 2022 Assessing clinical progression from subjective cognitive decline to mild cognitive impairment with incomplete multi-modal neuroimages *Med. Image Anal.* **75** 102266
Mitchell A J and Shiri-Feshki M 2009 Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies *Acta Psychiatrica Scandinavica* **119** 252–65
Nie L, Zhang L, Meng L, Song X, Chang X and Li X 2016 Modeling disease progression via multisource multitask learners: a case study with Alzheimer's disease *IEEE Trans Neural Netw. Learn. Syst.* **28** 1508–19
Pan X *et al* 2020a Multi-view separable pyramid network for ad prediction at mci stage by 18 f-fdg brain pet imaging *IEEE Trans. Med. Imaging* **40** 81–92
Pan Y, Liu M, Lian C, Xia Y and Shen D 2020b Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages *IEEE Trans. Med. Imaging* **39** 2965–75
Pan Y, Liu M, Lian C, Zhou T, Xia Y and Shen D 2018 Synthesizing missing pet from mri with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Springer) 455–63
Penny W D, Friston K J, Ashburner J T, Kiebel S J and Nichols T E 2011 *Statistical Parametric Mapping: The Analysis of Functional Brain Images* (Amsterdam, New York: Elsevier) (https://doi.org/10.1016/B978-0-12-372560-8.X5000-1)
Singh S *et al* 2017 Deep-learning-based classification of fdg-pet data for Alzheimer's disease categories *13th Int. Conf. on Medical Information Processing and Analysis. Int. Society for Optics and Photonics* 10572
Sled J G, Zijdenbos A P and Evans A C 1998 A nonparametric method for automatic correction of intensity nonuniformity in mri data *IEEE Trans. Med. Imaging* **17** 87–97
Spasov S *et al* 2019 A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease *Neuroimage* **189** 276–87
Suk H-I *et al* 2014 Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis *NeuroImage* **101** 569–82
Tong T *et al* 2016 A novel grading biomarker for the prediction of conversion from mild cognitive impairment to Alzheimer's disease *IEEE Trans. Biomed. Eng.* **64** 155–65
Vaswani A *et al* 2017 Attention is all you need *Adv. Neural Inf. Process. Syst.* **30** 6000–10
Wang L *et al* 2007 Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the alzheimer type *IEEE Trans. Med. Imaging* **26** 462–70
Wang X, Girshick R, Gupta A and He K 2018 Non-local neural networks *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* pp 7794–803
Zhang D *et al* 2011 Multimodal classification of Alzheimer's disease and mild cognitive impairment *NeuroImage* **55** 856–67
Zhang F, Li Z, Zhang B, Du H, Wang B and Zhang X 2019b Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease *Neurocomputing* **361** 185–95
Zhang J, Gao Y, Gao Y, Munsell B C and Shen D 2016 Detecting anatomical landmarks for fast Alzheimer's disease diagnosis *IEEE Trans. Med. Imaging* **35** 2524–33
Zhang L *et al* 2021 Deep fusion of brain structure-function in mild cognitive impairment *Med. Image Anal.* **72** 102082
Zhang T and Shi M 2020 Multi-modal neuroimaging feature fusion for diagnosis of Alzheimer's disease *J. Neurosci. Methods* **341** 108795
Zhang Y *et al* 2019a Strength and similarity guided group-level brain functional network construction for mci diagnosis *Pattern Recognit.* **88** 421–30
Zhou T *et al* 2019 Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data *IEEE Trans. Med. Imaging* **38** 2411–22