Medical Image Analysis 91 (2024) 103032

Contents lists available at ScienceDirect

Medical Image Analysis

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

Joint learning framework of cross-modal synthesis and diagnosis for Alzheimer's disease by mining underlying shared modality information

Chenhui Wang [a], Sirong Piao [b], Zhizhong Huang [c,d], Qi Gao [a], Junping Zhang [c,d], Yuxin Li [b,*], Hongming Shan [a,e,f,g,**], the Alzheimer's Disease Neuroimaging Initiative[1], the Australian Imaging Biomarkers and Lifestyle flagship study of aging[2]

[a] Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai 200433, China
[b] Department of Radiology, Huashan Hospital, Fudan University, Shanghai 200040, China
[c] Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China
[d] School of Computer Science, Fudan University, Shanghai 200433, China
[e] MOE Frontiers Center for Brain Science, Fudan University, Shanghai, 200433, China
[f] MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Fudan University, Shanghai, 200433, China
[g] Shanghai Center for Brain Science and Brain-inspired Technology, Shanghai 201210, China

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Alzheimer's disease (AD) diagnosis
Joint learning
Cross-modal synthesis
Magnetic resonance imaging (MRI)
Positron emission tomography (PET)

ABSTRACT

Alzheimer's disease (AD) is one of the most common neurodegenerative disorders presenting irreversible progression of cognitive impairment. How to identify AD as early as possible is critical for intervention with potential preventive measures. Among various neuroimaging modalities used to diagnose AD, functional positron emission tomography (PET) has higher sensitivity than structural magnetic resonance imaging (MRI), but it is also costlier and often not available in many hospitals. How to leverage massive unpaired unlabeled PET to improve the diagnosis performance of AD from MRI becomes rather important. To address this challenge, this paper proposes a novel joint learning framework of unsupervised cross-modal synthesis and AD diagnosis by mining underlying shared modality information, improving the AD diagnosis from MRI while synthesizing more discriminative PET images. We mine underlying shared modality information in two aspects: diversifying modality information through the cross-modal synthesis network and locating critical diagnosis-related patterns through the AD diagnosis network. First, to diversify the modality information, we propose a novel unsupervised cross-modal synthesis network, which implements the inter-conversion between 3D PET and MRI in a single model modulated by the AdaIN module. Second, to locate shared critical diagnosis-related patterns, we propose an interpretable diagnosis network based on fully 2D convolutions, which takes either 3D synthesized PET or original MRI as input. Extensive experimental results on the ADNI dataset show that our framework can synthesize more realistic images, outperform the state-of-the-art AD diagnosis methods, and have better generalization on external AIBL and NACC datasets.

1. Introduction

Alzheimer's disease (AD) is one of the most common neurodegenerative disorders presenting irreversible progression of cognitive impairment, for which there is currently no cure and limited treatment (Winblad et al., 2016; Wang et al., 2022). Therefore, distinguishing AD from normal cognition (NC) as early as possible is critical for intervention with potential preventive measures, which can also reduce

* Corresponding author at: Department of Radiology, Huashan Hospital, Fudan University, Shanghai 200040, China.
** Corresponding author at: Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai 200433, China.
E-mail addresses: liyuxin@fudan.edu.cn (Y. Li), hmshan@fudan.edu.cn (H. Shan).

[1] Data used in preparation of this article was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

[2] Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (https://adni.loni.usc.edu). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au.

https://doi.org/10.1016/j.media.2023.103032
Received 15 December 2022; Received in revised form 31 August 2023; Accepted 13 November 2023
Available online 18 November 2023

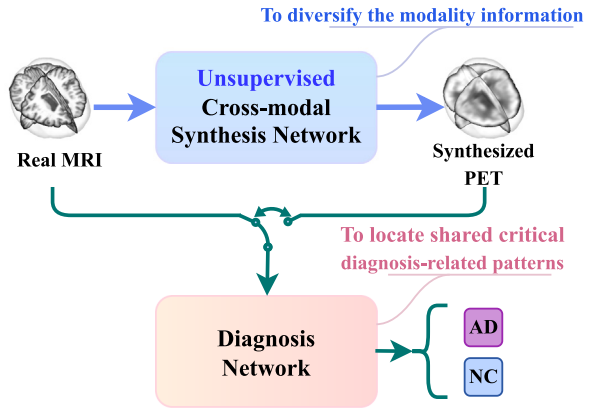

Fig. 1. The conceptual illustration of the proposed joint learning framework. The proposed framework jointly learns unsupervised cross-modal synthesis and diagnosis tasks by mining shared modality information, and uses either modality for diagnosis.

the high healthcare costs and heavy personal burdens associated with the treatment and care of AD patients (Wong, 2020).

Among various neuroimaging modalities used to diagnose AD, functional positron emission tomography (PET) has higher sensitivity than structural magnetic resonance imaging (MRI) (Pichler et al., 2010). However, PET is also costlier than MRI and often not available in many hospitals due to the high cost associated with multiple examinations, poorly equipped hospitals, and difficulties in data collection (Pan et al., 2020). How to leverage the more sensitive but less available PET to improve the diagnosis performance from the MRI becomes rather important. To address this challenge, traditional methods perform interpolation for the missing modality in radiomics analysis (Campos et al., 2015; Gillies et al., 2016), but they typically perform not well due to inaccurate interpolation (Pan et al., 2020).

The broad success of deep generative models (Zhu et al., 2017; Ho et al., 2020) has prompted the development of cross-modal medical image synthesis (Yi et al., 2019; Zhao and Zhao, 2021) that can directly synthesize the missing modality from available one for downstream tasks (Shin et al., 2020; Liu et al., 2020; Yang et al., 2020; Pan et al., 2021; Rahimpour et al., 2021; Zhang et al., 2022). However, most cross-modal synthesis tasks highly rely on supervised paired data (Dar et al., 2019; Sun et al., 2019; Hu et al., 2021; Liu et al., 2022; Luo et al., 2022). Collecting large-scale paired data for training, however, is time-consuming and cumbersome in practical clinical scenarios. Consequently, the synthesis network trained with limited supervised paired data may suffer from potential overfitting. To be more practical, cross-modal synthesis tasks should focus more on massive unpaired data rather than limited paired data (Yang et al., 2021b,a). On the other hand, although they can be used for downstream tasks, cross-modal synthesized images have little diagnosis information for AD, even with supervised paired data. Although looking realistic, the cross-modal synthesis task is independent of the diagnosis task. This situation worsens when the synthesis task is trained with unpaired data, where the diagnosis information is difficult to be learned.

In this work, we explore how to *leverage massive unpaired unlabeled PET data to improve AD diagnosis from MRI* while synthesizing more discriminative PET images. To this end, we propose a novel joint learning framework of unsupervised cross-modal synthesis and diagnosis for AD by mining underlying shared modality information. Fig. 1 presents the concept of the proposed joint learning framework, in which the first step is to synthesize pseudo-paired PET modality images through the unsupervised cross-modal synthesis network to diversify the modality information, and the second is to use either synthesized PET or real MRI as the input to the same diagnosis network to mine the shared information between them. The rationale behind the proposed workflow is that our framework considers the synthesized PET images to be task-specific data augmentation of the input MRI images, which can assist the downstream diagnosis network in filtering out diagnostically irrelevant features and mining underlying shared modality information. In addition, we jointly train the cross-modal synthesis and AD diagnosis networks to supervise each other, providing more discriminative diagnosis information in the synthesized images and improving the performance of AD diagnosis.

The contributions of this work are summarized as follows.

(1) We propose a novel joint learning framework of unsupervised cross-modal synthesis and diagnosis for AD by mining underlying shared modality information, improving the performance of both tasks. To the best of our knowledge, this is the *first* study to utilize the unsupervised cross-modal synthesis method to synthesize PET for AD diagnosis.
(2) To diversify the modality information, we propose a novel unsupervised cross-modal synthesis network that implements the inter-conversion between 3D PET and MRI in a single model modulated by the AdaIN module.
(3) To locate shared critical diagnosis-related patterns, we propose an interpretable diagnosis network based on fully 2D convolutions, which takes either 3D synthesized PET or real MRI as input. The regions of interest located by our network are consistent with those associated with AD.
(4) We extensively evaluate our framework on the ADNI dataset, demonstrating that our framework can synthesize more realistic images, outperform the state-of-the-art AD diagnosis networks, and have better generalization on external AIBL and NACC datasets.

The remainder of this paper is organized as follows. In Section 2, we detail the proposed framework along with the novel synthesis and diagnosis networks. We elaborate on the setup of the experiments and the experimental results in Section 3. Finally, Section 4 presents the discussion of our framework, followed by a concluding summary in Section 5.

2. Methodology

Fig. 2 presents the detailed training and inference phases of the proposed joint learning framework for AD diagnosis, which involves unsupervised cross-modal synthesis and AD diagnosis tasks. We first overview the overall framework in Section 2.1 and then describe the two novel networks for cross-modal synthesis and AD diagnosis in Sections 2.2 and 2.3, respectively. Finally, we present the objective function for optimizing the proposed framework in Section 2.4.

2.1. Overall framework

As shown in Fig. 2, the proposed framework is a joint learning framework of unsupervised cross-modal synthesis and diagnosis for AD, which can leverage massive unpaired unlabeled PET data to improve AD diagnosis from MRI by mining underlying shared modality information. We believe that the core of cross-modal synthesis for diagnosis is to mine the underlying shared information between different modalities. To this end, we mine the underlying shared modality information in two aspects: diversifying modality information through the cross-modal synthesis network and locating shared critical diagnosis-related patterns through the AD diagnosis network.

During the training phase, our framework requires a set of unpaired 3D MRI and PET images, within which the MRI images are labeled with AD or NC, while the PET images are unlabeled. To diversify modality information, we feed unpaired 3D MRI and PET images into the cross-modal synthesis network to synthesize pseudo-paired PET images. In our cross-modal synthesis network, termed ShareGAN, we employ a single shared synthesis model modulated by the AdaIN module to accomplish the bidirectional mapping between two modalities. At the

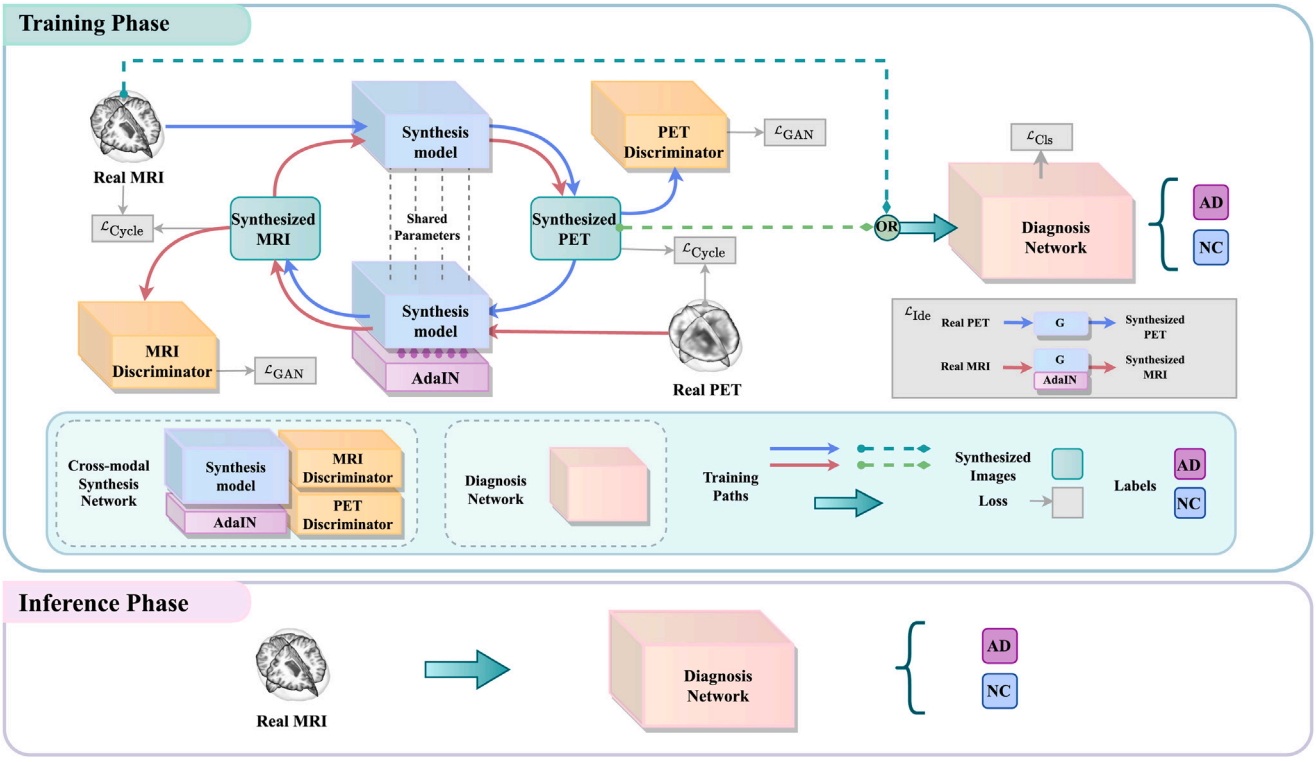

Fig. 2. Illustration of the training and inference phases of the proposed joint learning framework.

same time, inspired by Generative adversarial networks (GAN) (Zhu et al., 2017), two different discriminators are also used to discriminate the authenticity of the synthesized MRI and PET images. After the cross-modal synthesis process, our framework considers the synthesized PET images as the task-specific data augmentation for the corresponding MRI images. To locate shared critical diagnosis-related patterns, we feed *either* synthesized PET images *or* original MRI images into the same diagnosis network using MRI image labels.

We note that the proposed framework does not involve the cross-modal synthesis network during the inference phase; that is, our network has the same inference speed as conventional AD diagnosis network, and is memory-friendly and computationally efficient.

2.2. Cross-modal synthesis network

The proposed unsupervised cross-modal synthesis network, ShareGAN, implements the inter-conversion between 3D PET and MRI in a *single* model modulated by the AdaIN module. This design aims to ensure that the synthesized images preserve underlying structures well and achieve better image quality. At the same time, two different modality discriminators, D_{m} and D_{p}, are used for MRI and PET, respectively, which aim to distinguish real modality from the synthesized one; see Supp. A.1 for detailed network architecture.

Fig. 3 illustrates our synthesis model and the AdaIN module. Our synthesis model is built upon the 3D UNet (Ronneberger et al., 2015). Differently, we introduce a fully convolution-implemented self-attention block (Dosovitskiy et al., 2020; Pan et al., 2022) at the bottleneck of the synthesis model to learn the global information of the input 3D images (see Supp. A.2 for details). In the following, we elaborate on how the AdaIN module modulates the synthesis model to achieve the inter-conversion between the two modalities.

Suppose that a multi-channel feature tensor $\boldsymbol{U}$ at a specific layer is represented as follows:

$$\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_C] \in \mathbb{R}^{D\times H\times W\times C}, \tag{1}$$

where $\boldsymbol{u}_i$ represents the feature of size $D \times H \times W$ at the i-th channel. Furthermore, the corresponding feature map $\boldsymbol{V}$ for the transformed modality image is given by:

$$\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_C] \in \mathbb{R}^{D\times H\times W\times C}. \tag{2}$$

Then, instance normalization (Ulyanov et al., 2016) and Adaptive Instance Normalization (AdaIN) (Huang and Belongie, 2017) convert the feature per channel using the following transform:

$$\mathcal{T}(\boldsymbol{u}_i, \boldsymbol{v}_i) = \mu(\boldsymbol{v}_i) + \sigma(\boldsymbol{v}_i)\left(\frac{\boldsymbol{u}_i - \mu(\boldsymbol{u}_i)}{\sigma(\boldsymbol{u}_i)}\right), \; i = 1, \ldots, C, \tag{3}$$

where $\mu(\boldsymbol{u}_i)$ and $\sigma(\boldsymbol{u}_i)$ represent the mean and standard deviation of the input modality feature $\boldsymbol{u}_i$. In Eq. (3), we scale the normalized input modality feature per channel with $\sigma(\boldsymbol{v}_i)$ and shift it with $\mu(\boldsymbol{v}_i)$. In terms of instance normalization, $\mu(\boldsymbol{v}_i)$ and $\sigma(\boldsymbol{v}_i)$ equal to 0 and 1, respectively. However, in the AdaIN module, $\mu(\boldsymbol{v}_i)$ and $\sigma(\boldsymbol{v}_i)$ are estimated using the transformed modality images.

As mentioned above, the proposed unsupervised cross-modal network uses the synthesis model to synthesize pseudo-paired PET images from MRI images ($\mathcal{M} \rightarrow \mathcal{P}$), and here uses the AdaIN module to modulate the conversion direction of the synthesis model ($\mathcal{P} \rightarrow \mathcal{M}$). In other words, for each of the two modalities, PET $\mathcal{P}$ or MRI $\mathcal{M}$, another one serves as the transformed modality. The AdaIN module modulates the inter-conversion between them as follows:

$$(\mu(\boldsymbol{V}), \sigma(\boldsymbol{V})) = \begin{cases} (0, 1), & \mathcal{M} \rightarrow \mathcal{P} \\ (\boldsymbol{\mu}_M, \boldsymbol{\sigma}_M), & \mathcal{P} \rightarrow \mathcal{M}. \end{cases} \tag{4}$$

Therefore, AdaIN module is defined as follows:

$$\mathrm{AdaIN}(\beta) := \begin{bmatrix} \mu(\beta) \\ \sigma(\beta) \end{bmatrix} = (1-\beta)\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \beta \begin{bmatrix} \mu_M \\ \sigma_M \end{bmatrix}, \tag{5}$$

where (μ_M, σ_M) are learnable parameters during the training, and $\beta \in \{0, 1\}$, where $\beta = 0$ corresponds to $\mathcal{M} \rightarrow \mathcal{P}$ while $\beta = 1$ corresponds to $\mathcal{P} \rightarrow \mathcal{M}$.

Given the modality (MRI) $\mathcal{X}_{\mathrm{M}}$ and the unpaired modality (PET) $\mathcal{X}_{\mathrm{P}}$, the AdaIN module acts as a "switch" in the single synthesis model to

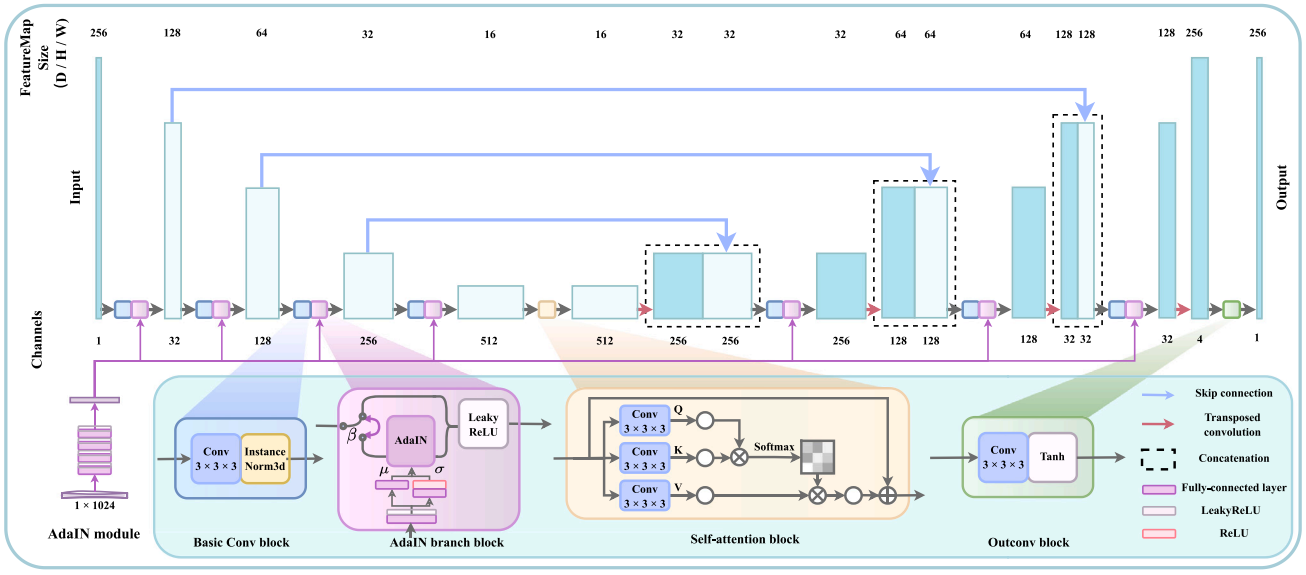

Fig. 3. Illustration of our unsupervised cross-modal synthesis ShareGAN network including synthesis model and AdaIN module.

determine whether specific modality information is introduced or not. Here, we denote the model involved in synthesized PET images as G_{p} and the model involved in synthesized MRI images as G_{m}. During the training phase, we only need to specify the conversion direction of the synthesis model G via the AdaIN module as follows:

$$G_{\mathrm{p}}(\mathcal{X}_{\mathrm{M}}) := G(\mathcal{X}_{\mathrm{M}}; \mathrm{AdaIN}(0)), \tag{6}$$

$$G_{\mathrm{m}}(\mathcal{X}_{\mathrm{P}}) := G(\mathcal{X}_{\mathrm{P}}; \mathrm{AdaIN}(1)). \tag{7}$$

In the concrete implementation, the AdaIN module takes a vector of size 1×1024 as input and outputs seven pairs of mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ vectors for all feature maps in the AdaIN branch blocks of our synthesis model, as shown in Fig. 3. Compared to the AdaIN module of Switchable CycleGAN (Yang et al., 2021a), we enhance its local learning capability by introducing an independent fully-connected layer, activated by the leaky ReLU function, at the start of each AdaIN branch block.

2.3. AD diagnosis network

3D medical imaging data can be viewed as a series of 2D slice images. During clinical examination, radiologists also tend to examine various 2D slice images. Inspired by this, we propose an efficient and interpretable AD diagnosis network implemented with fully 2D convolution and joint inter- and intra-slice modeling using a zero-parameter *slice-shift* module. In addition, it is also interpretable with the *slice-aware* module, as shown in Fig. 4.

We start by dividing 3D medical image data along a specific axis to obtain a series of 2D slice images (axial, coronal, or sagittal planes). We use a shared base feature extractor to extract the feature maps from each slice. Then, we use the *slice-aware* module to derive the distribution of slice attention scores and thus enhance the important slice information accordingly. At following different feature map scales, we design respective *slice-shift* modules to exchange feature information among neighboring slices, thus enabling joint inter- and intra-slice modeling. Moreover, each slice feature predicts the corresponding label via a fully-connected layer shared across all the slices. Finally, the overall result is obtained by averaging all predicted results.

Since our diagnosis network is built with shared weights' 2D backbone model, it can be trained and tested efficiently compared to other 3D models, preventing potential overfitting. Meanwhile, our diagnosis network can be designed based on any 2D backbone. In our experiments, taking 2D ResNet18 as an example, ResNet18 consists of a down-sampling layer, four residual blocks, and a fully-connected layer. We can use the down-sampling layer, the first two residual blocks, as the base feature extractor. In order to uncover more information from adjacent slices, we use the slice-shift module before each of the last two residual blocks. The final fully-connected layer gives the corresponding predictions. Next, we detail the proposed slice-aware and slice-shift modules.

2.3.1. Slice-aware module

The slice-aware module is designed to interpret significant slices and regions in our diagnosis network. First, we feed the 3D images into a base feature extractor shared across all the slices, generating independent features for each slice. The resulting features have a size of $B \times S \times C \times H \times W$, where B, S, C, H, and W represent the size of mini-batch, the number of slices, the number of channels, the height, and the width of the features, respectively. In the next step, as illustrated in Fig. 5, we employ a 1×1 convolution layer with shared weights to fuse the whole information per slice into one channel. Then, one branch employs a 1×1 convolution layer to fuse all slices' information as the "global query". Another branch adopts a lighting depthwise convolution (Howard et al., 2017) to add different slices' local learning capability as the "local keys", where each slice has an individual convolution kernel. We then use a softmax function to translate the scaled dot product of "global query" and "local keys" into a relative measurement of slice attention as follows:

$$\mathrm{Softmax}(\boldsymbol{Q}\boldsymbol{K}^{\top} \times \alpha), \tag{8}$$

where $\boldsymbol{Q} \in \mathbb{R}^{B\times 1\times H\times W}$ and $\boldsymbol{K} \in \mathbb{R}^{B\times S\times H\times W}$ denote the normalized "global query" and "local keys", respectively. Moreover, a variable α is used to adjust the sharpness of slice attention distribution.

Furthermore, we expand the slice attention scores to the original input size and multiply it by the original input, enhancing the important slice information (Hu et al., 2018). Note that the slice attention distribution can reveal which slices are more important, and the "local keys" part can also reveal which parts of the slices are more critical.

2.3.2. Slice-shift module

As previously described, a 2D backbone model with shared weights extracts independent features from different slices. However, the features from 2D images divided along a specific direction cannot provide slice-to-slice 3D stereo-spatial information. Therefore, we propose a slice-shift module shown in Fig. 6(b), which allows for joint inter- and

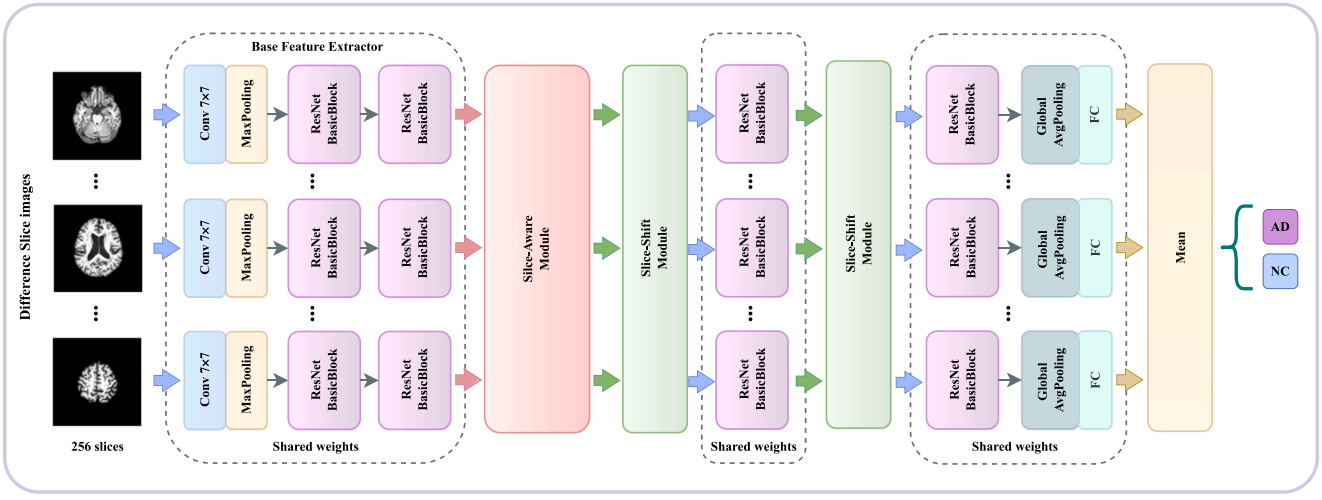

Fig. 4. The proposed AD diagnosis network with slice-aware module and slice-shift module.

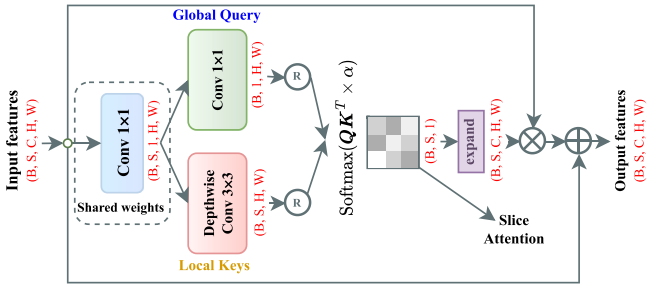

Fig. 5. Illustration of the proposed slice-aware module.

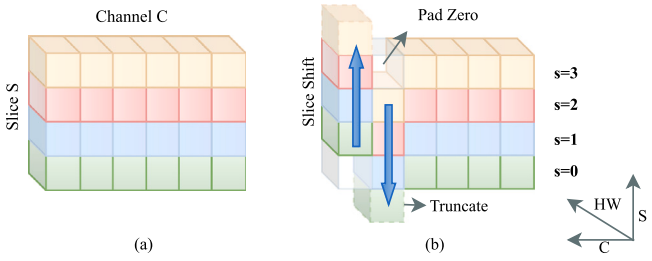

Fig. 6. Illustration of (**a**) the original feature map without shifting and (**b**) the changed feature map by the proposed slice-shift module.

intra-slice modeling. Unlike the original feature map without shifting in Fig. 6(a), to exchange information with neighboring slices, the slice-shift module shifts some channels forward and backward along the slice-dividing direction by ±1, truncates the surplus features, and pads the missing features with 0. The information from neighboring slices is mingled with the current slice by applying the slice-shift module. Note that the introduction of this module causes no additional computational burden on the network, allowing it to train and infer efficiently (Lin et al., 2019).

2.4. Objective function

In the following, we demonstrate the objective function of the proposed framework. During the training phase, we random sample unpaired 3D MRI $\boldsymbol{x}_m \in \mathcal{X}_{\mathrm{M}}$ $(m = 1, 2, \ldots)$ and PET $\boldsymbol{x}_p \in \mathcal{X}_{\mathrm{P}}$ $(p = 1, 2, \ldots)$ data for input. The ultimate goal of the cross-modal synthesis network is to learn a mapping $\mathcal{X}_{\mathrm{M}} \rightarrow \mathcal{X}_{\mathrm{P}}$. After completing the synthesis process, we feed either the synthesized PET image $\boldsymbol{x}_p^*$ or real MRI image $\boldsymbol{x}_m$ into the same diagnosis network for joint training, improving the performance of both tasks. The objective function consists of four different losses, including adversarial loss, cycle-consistency loss, identity loss, and diagnosis loss.

Two modality discriminators, D_{m} and D_{p}, use adversarial loss in the form of LSGAN loss (Mao et al., 2017) to distinguish the authenticity of synthesized and real images, defined as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}} = -\Big(& \mathbb{E}_{\boldsymbol{x}_p \in \mathcal{X}_{\mathrm{P}}}[(D_{\mathrm{p}}(\boldsymbol{x}_p)-1)^2] + \mathbb{E}_{\boldsymbol{x}_m \in \mathcal{X}_{\mathrm{M}}}[D_{\mathrm{p}}(G_{\mathrm{p}}(\boldsymbol{x}_m))^2] + \\ & \mathbb{E}_{\boldsymbol{x}_m \in \mathcal{X}_{\mathrm{M}}}[(D_{\mathrm{m}}(\boldsymbol{x}_m)-1)^2] + \mathbb{E}_{\boldsymbol{x}_p \in \mathcal{X}_{\mathrm{P}}}[D_{\mathrm{m}}(G_{\mathrm{m}}(\boldsymbol{x}_p))^2] \Big). \end{aligned} \tag{9}$$

Cycle-consistency loss constrains the conversion between original and transformed modalities, which is implemented by L_1 loss as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{Cycle}} = & \mathbb{E}_{\boldsymbol{x}_p \in \mathcal{X}_{\mathrm{P}}} \| G_{\mathrm{p}}(G_{\mathrm{m}}(\boldsymbol{x}_p)) - \boldsymbol{x}_p \|_1 + \\ & \mathbb{E}_{\boldsymbol{x}_m \in \mathcal{X}_{\mathrm{M}}} \| G_{\mathrm{m}}(G_{\mathrm{p}}(\boldsymbol{x}_m)) - \boldsymbol{x}_m \|_1. \end{aligned} \tag{10}$$

Identity loss forces the synthesis model (*e.g.* G_{p}) to achieve an identity mapping from the input (*e.g.* $\boldsymbol{x}_p \in \mathcal{X}_{\mathrm{P}}$) to the output, such that $G_{\mathrm{p}}(\boldsymbol{x}_p) \backsimeq \boldsymbol{x}_p$, which is defined as follows:

$$\mathcal{L}_{\mathrm{Ide}} = \mathbb{E}_{\boldsymbol{x}_p \in \mathcal{X}_{\mathrm{P}}} \| G_{\mathrm{p}}(\boldsymbol{x}_p) - \boldsymbol{x}_p \|_1 + \mathbb{E}_{\boldsymbol{x}_m \in \mathcal{X}_{\mathrm{M}}} \| G_{\mathrm{m}}(\boldsymbol{x}_m) - \boldsymbol{x}_m \|_1. \tag{11}$$

In this study, diagnosis loss is to distinguish between subjects with AD and NC, which is the cross-entropy loss defined as follows:

$$\mathcal{L}_{\mathrm{Cls}} = -\mathbb{E}_{(y_{\mathrm{m}}, y_{\mathrm{m}}^*)} y_{\mathrm{m}} \log(y_{\mathrm{m}}^*), \tag{12}$$

where $y_{\mathrm{m}}, y_{\mathrm{m}}^*$ are the ground-truth labels and the labels predicted by our AD diagnosis network, respectively.

Ultimately, the overall objective function as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{GAN}} + \lambda_{\mathrm{Cycle}} \mathcal{L}_{\mathrm{Cycle}} + \lambda_{\mathrm{Ide}} \mathcal{L}_{\mathrm{Ide}} + \lambda_{\mathrm{Cls}} \mathcal{L}_{\mathrm{Cls}}, \tag{13}$$

where λ_{Cycle}, λ_{Ide}, and λ_{Cls} are hyperparameters that adjust the weights between these four objective functions, and set to 10, 3, and 1, respectively, according to the experimental results of hyperparameter selection on the validation set (see Supp. C.1 for details).

3. Experiment results

In this section, we first elaborate on the setup of the experiments in Section 3.1. Then, we present evaluation results on the cross-modal synthesis and AD diagnosis tasks in Sections 3.2 and 3.3, respectively. Finally, the overall joint learning framework results against other state-of-the-art Alzheimer's diagnosis methods are presented in Section 3.4.

Table 1
Demographic information of the datasets used in the experiment.

Modality	Source♯	Type	Subject	Gender *		Age *		MMSE †		APOE4 †	
			no.	Male	(%)	Mean	(std)	Mean	(std)	Positive	(%)
MRI	ADNI	NC	317	148	(46.69)	75.604	(6.231)	28.588	(2.573)	85	(26.81)
		AD	247	129	(52.23)	74.847	(7.934)	22.495	(3.321)	157	(63.56)
	AIBL	NC	105	47	(44.76)	70.514	(6.075)	28.790	(1.174)	5	(4.76)
		AD	28	11	(37.93)	74.233	(7.684)	20.433	(4.599)	21	(75.00)
	NACC	NC	43	18	(41.86)	72.353	(8.438)	29.137	(1.167)	1	(2.32)
		AD	49	21	(42.86)	71.314	(9.092)	21.824	(5.384)	9	(18.37)
PET	ADNI	NC	316	163	(51.58)	75.961	(6.394)	28.967	(1.449)	84	(26.58)
		AD	233	136	(58.37)	75.386	(7.898)	22.149	(3.715)	154	(66.09)

♯ In the ADNI, AIBL, and NACC study cohort, Mini-Mental State Examination (MMSE) and Apolipoprotein E4 (APOE4) allele genetic information are unavailable for some subjects. All the scans considered for this study are performed on individuals within ±6 months from the date of clinical diagnosis.
∗ No significant differences are found between the different subject groups of the corresponding modality for gender and age ($p > 0.05$).
† Significant differences are found between the different subject groups of the corresponding modality for MMSE and APOE4 ($p < 0.05$), where the chi-square test is used for gender and APOE4, while age and MMSE are all tested using t-tests.

Table 2
Quantitative results of synthesized images across different subjects by different cross-modal synthesis networks on the ADNI testing set. The best and second best quantitative results in different supervised types are highlighted in **bold** and underlined fonts, respectively.

Type	Methods	MAE (%) ↓	PSNR (dB) ↑	SSIM ↑
Supervised	Pix2Pix	1.589 ± 0.133	42.551 ± 0.328	0.936 ± 0.003
	Pix2Pix (**after joint learning**)	**1.581 ± 0.101**	**42.884 ± 0.275**	**0.937 ± 0.001**
Unsupervised	CycleGAN	2.947 ± 0.231	37.563 ± 0.454	0.896 ± 0.007
	CycleAttGAN	2.603 ± 0.187	38.158 ± 0.358	0.904 ± 0.006
	ShareGAN (**ours**)	2.329 ± 0.132	38.078 ± 0.301	0.908 ± 0.003
	ShareGAN (**ours after joint learning**)	**2.144 ± 0.117**	**38.373 ± 0.281**	**0.916 ± 0.003**

3.1. Experiment setup

3.1.1. Dataset and preprocessing

The data used in this experiment are obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI), Australian Imaging, Biomarkers and Lifestyle (AIBL), and National Alzheimer's Coordinating Center (NACC) cohorts. Resources and data in ADNI are from North America. ADNI researchers collect, validate, and utilize data, including MRI and PET images, genetics, cognitive tests, CSF, and blood biomarkers, as predictors of AD (Petersen et al., 2010). AIBL is a long-term longitudinal investigation aiming to advance understanding of the causes of AD, collected in Australia (Ellis et al., 2009). Over the past two decades, NACC has been built in collaboration with more than 42 Alzheimer's Disease Research Centers (ADRCs) throughout the US (Beekly et al., 2004). All study participation protocols are reviewed by each subject's local review committee and signed with the subject's consent (Petersen et al., 2010). The subjects' demographic information in each dataset is summarized in Table 1.

In our experiment, only one MRI image from the baseline visit was selected for each subject. The ADNI dataset is randomly split according to the ratio of 6:2:2, where 60% is used for model training, 20% of the data is used for internal validation, and the rest is used for internal testing. We select the best-performing model on the validation set for making predictions on the internal ADNI testing data and the external AIBL and NACC testing data. The ADNI dataset used in our experiments has 564 T1 MRI images and 549 FDG PET images. We highlight that the data in the training set is unpaired for the ADNI dataset. However, the MRI and PET data in the ADNI validation and testing sets are paired for better quantitative comparison. The criterion for selection included individuals aged $\geqslant 55$ years, within ±6 months from the clinically confirmed diagnosis. We excluded cases including AD with mixed dementia, non-Alzheimer's disease dementias, and incident major systemic illnesses (Qiu et al., 2020).

For each modality, MRI and PET, we use different standard preprocessing pipelines to preprocess them, with details given in Supp. B. The data are preprocessed and aligned to the MNI152 template space using Freesurfer (Fischl, 2012), ANTs (Avants et al., 2009), and FSL software. In order to ensure the quality of the synthesized images, we do not change the image size, resulting in the final image size of $256 \times 256 \times 256$.

3.1.2. Implementation details

We implement all the methods in this paper in the PyTorch library (Paszke et al., 2019) and train them on NVIDIA V100 32G Tensor Core GPUs. All the networks are initialized by the kaiming method (He et al., 2015) and trained using the Adam optimization algorithm with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We use the early stopping method to select better network weights during the training phase. For a fair and reliable performance evaluation (Song and Chai, 2018; Qian et al., 2021), we repeat the experiments with different random seeds three times and report their mean results in this paper. The overall framework training process is divided into three stages, and the concrete training stages are described below.

In the first stage, we pre-train the cross-modal synthesis network by randomly cropping the original 3D medical images to the size of $128 \times 128 \times 128$. A learning rate of 2×10^{-4} is used first to train 15 epochs, followed by an exponential learning rate with a decay rate of 0.95. Secondly, the whole 3D brain image with the size of $256 \times 256 \times 256$ is input into the cross-modal synthesis network obtained in the previous stage and the diagnosis network to get a better version of both network weights. An exponential learning rate of $1{\times}10^{-4}$ with a decay rate of 0.95 is used in this stage for the cross-modal synthesis and diagnosis networks. In the last stage, we perform end-to-end joint learning to fine-tune all the parameters. To ensure the stability of the synthesis quality and the capacity of the diagnosis network, the learning rate of the diagnosis network is set to 1×10^{-5}, which is ten times that of the cross-modal synthesis network at this stage, *i.e.* 1×10^{-6}. Source code is available at https://github.com/thibault-wch/Joint-Learning-for-Alzheimer-disease.

3.1.3. Evaluation metrics

For the cross-modal synthesis task, three evaluation metrics are used to measure the quality of synthesized images, including the mean absolute error (MAE) (Willmott and Matsuura, 2005), the peak signal-to-noise ratio (PSNR), and the structural similarity index measure (SSIM) (Hore and Ziou, 2010). For the diagnosis task, five metrics

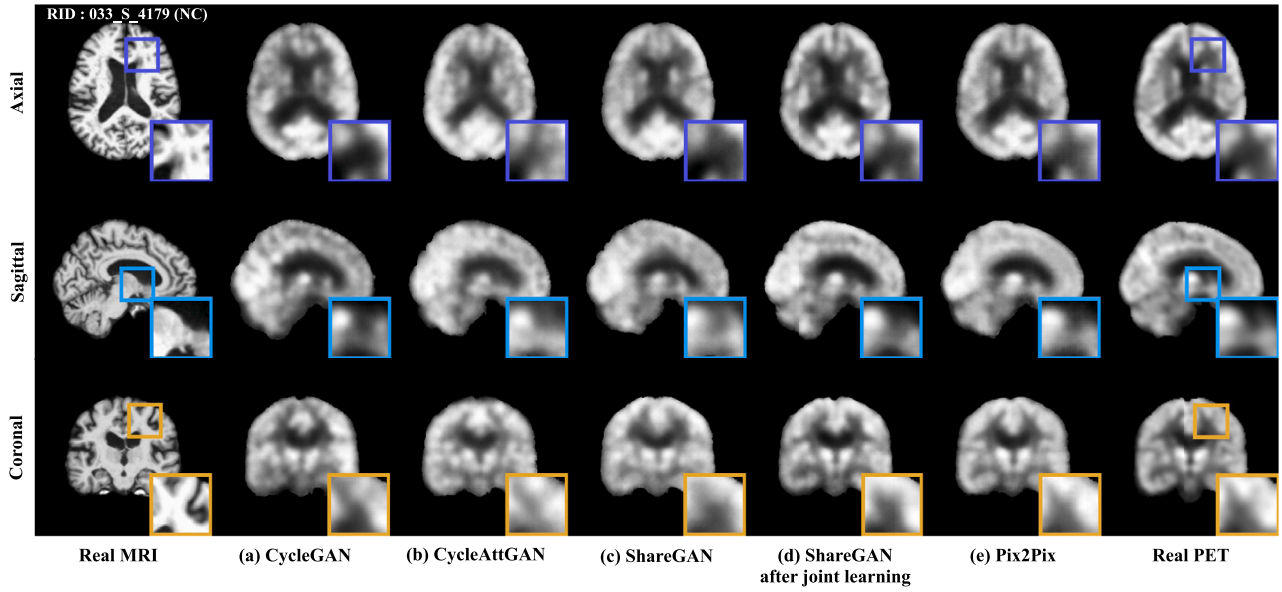

Fig. 7. Qualitative results of different cross-modal synthesis networks.

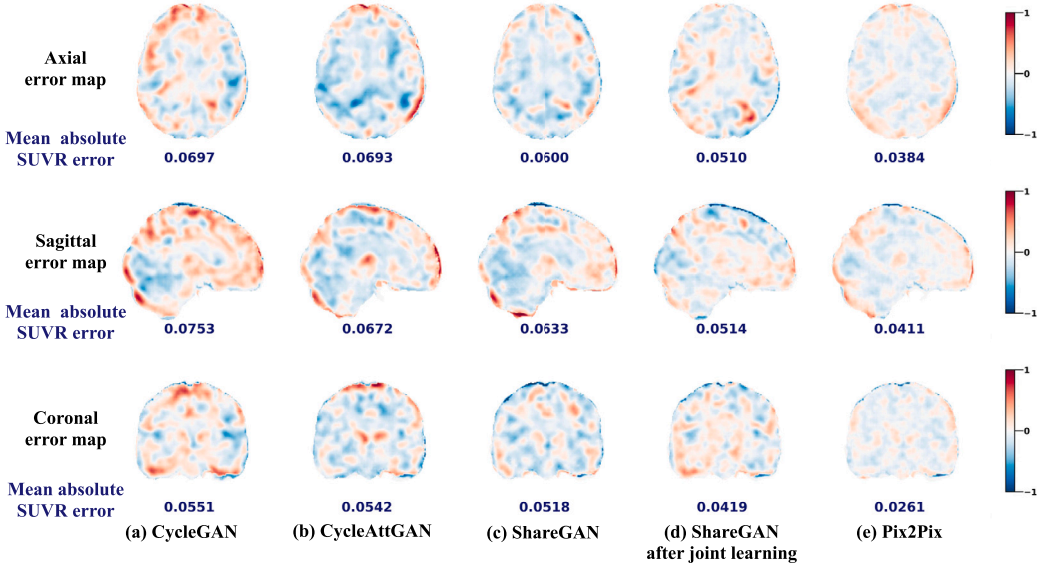

Fig. 8. The SUVR error map ($x_p - x_p^*$) results between the real PET images x_p and synthesized PET images x_p^*.

are used for performance evaluation, including accuracy (ACC), the area under the receiver operating characteristic curve (AUC), F1-Score (F1S), specificity (SPE), and sensitivity (SEN).

3.2. Evaluation on cross-modal synthesis task

For the unsupervised cross-modal synthesis task, we compare the proposed cross-modal synthesis network (before and after joint learning), ShareGAN, to the competitive baselines such as 3D UNet-based (Ronneberger et al., 2015) Cycle-Consistent Generative Adversarial Networks (CycleGAN) (Zhu et al., 2017) and CycleAttGAN networks. CycleAttGAN only introduces a self-attention block at the bottleneck of the 3D UNet synthesis model. Moreover, we also compare our network with the supervised Pix2Pix (Isola et al., 2017) network (the synthesis model is the same as CycleAttGAN). For all comparison methods, the discriminators have the same structure.

Table 3
Comparison of model size in synthesis models.

CycleAttGAN		ShareGAN (ours)	
Network	Parameters	Network	Parameters
G_m	27.637M	G	27.637M
G_p	27.637M	AdaIN	5.318M
Total	55.274M	Total	32.955M

Quantitative comparison Table 2 presents the quantitative results across different subjects on the ADNI testing set of all compared methods. For the unsupervised cross-modal synthesis, our ShareGAN outperforms others in all quantitative metrics after joint learning. Meanwhile, Table 3 compares the number of parameters in the synthesis model between CycleAttGAN and our network. Note that the synthesis model structure of CycleAttGAN is the same as that of our network. However,

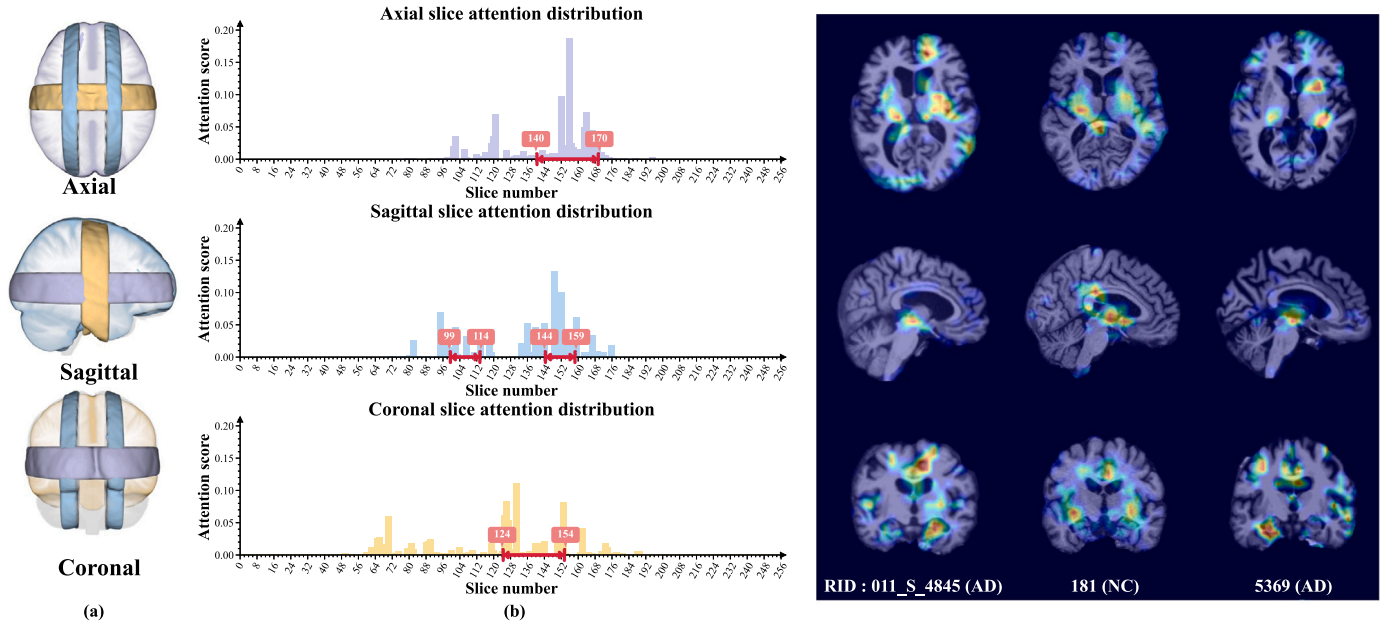

Fig. 9. Interpretable results of diagnosis network. **(a)** 3D visualization of the region of interest. **(b)** Slice attention distribution of the different slice-dividing directions. **(c)** Grad-CAM visualization of the slice-aware module's "local keys" part in randomly selected brains from different datasets.

CycleAttGAN uses two synthesis models, and we only need one plus some additional parameters from AdaIN. Although the number of parameters needed for synthesis models has been reduced by half, the performance of the network has been improved thanks to more stable training.

Table 2 shows a large gap in synthesized image quality between unsupervised and supervised training. Nonetheless, both show performance improvement after joint learning with the downstream diagnosis tasks. As for the Pix2pix network after joint learning, the improvement in quantitative performance metrics is not as large as our unsupervised network, possibly due to the relatively high degree of pixel-level supervision in the original supervised task.

Qualitative comparison Fig. 7 presents the qualitative results of all the methods. In addition, we select the pons (Thibeau-Sutre et al., 2022) as the reference region and present the Standardized Uptake Value Ratio (SUVR) (Thie, 2004) error map to validate the quality of synthesized images in Fig. 8. The PET images synthesized by CycleGAN contain many artifacts that do not exist in the original PET images and lose some of the original structural information. The CycleAttGAN network augments the CycleGAN with a self-attention module, resulting in a smoother image with global information. The synthesized images of the proposed ShareGAN have a more coherent and realistic pattern than the previous images but have poor local detail synthesis. However, after joint learning, our ShareGAN extracts more discriminative diagnosis information from the input MRI image based on the downstream diagnosis task, and the quality of the synthesized images is greatly improved; see Supp. C.2 for more qualitative results.

From the qualitative results, it is evident that there are substantial differences between PET images synthesized by unsupervised and supervised training. However, even with the supervised Pix2Pix network, we can see that there is still a large gap between the synthesized and real images. In other words, some information about the molecular functionality of PET modality cannot be derived from structural MRI modality. In cross-modal image synthesis, we should focus on more than the authenticity of synthesized images; that is, the synthesis task can be more meaningful to make the synthesized image have more discriminative information related to the downstream diagnosis task, as described in the following.

Table 4
Ablation studies of the diagnosis network on the ADNI testing set. The best quantitative results are highlighted in **bold** font.

Slice aware	Slice shift	Shift folds	ACC	AUC	F1S	SPE	SEN
✗	✗	–	0.817	0.874	0.811	0.831	0.810
✔	✗	–	0.833	0.890	0.827	0.864	0.837
✗	✔	**16**	0.839	0.913	0.836	0.865	0.837
✗	✔	**8**	0.858	0.921	0.854	0.871	0.853
✗	✔	**4**	0.837	0.909	0.833	0.870	0.831
✔	✔	**8**	**0.867**	**0.924**	**0.864**	**0.873**	**0.865**

3.3. Evaluation on AD diagnosis task

This subsection concentrates on our diagnosis network's quantitative and interpretable visualization results. Moreover, all the experimental results in this subsection are obtained on the ADNI testing set without the joint learning. Unless noted otherwise, axial planes are used in our experiment design because they are the most commonly used in clinical applications.

Ablation study We start by conducting ablation experiments to demonstrate the effectiveness of the slice-aware and slice-shift modules in Table 4. In the slice-shift module, we divide the channels' number of the corresponding feature map's layer into 16, 8, and 4 folds, respectively. Then, we shift one fold unit forward and one fold unit backward along the slice-dividing direction to learn the information between adjacent slices; the other folds retain the information from the original slices.

We have the following observations. (1) Both modules can improve diagnosis performance independently. (2) Dividing the total number of channels in the feature map into 16 and 4 is too little or too much exchanged information between neighborhood slices, respectively. In contrast, the number of channels divided is eight is better than others. (3) When using the slice-aware and slice-shift modules together, performance can be further improved.

Interpretability As illustrated in Fig. 9, the slice-aware module offers the unique interpretability of our diagnosis network. We train the diagnosis network separately in three planes: axial, sagittal, and coronal. The slice attention distribution of the different slice-dividing directions is shown in Fig. 9(b). Meanwhile, we select the most critical

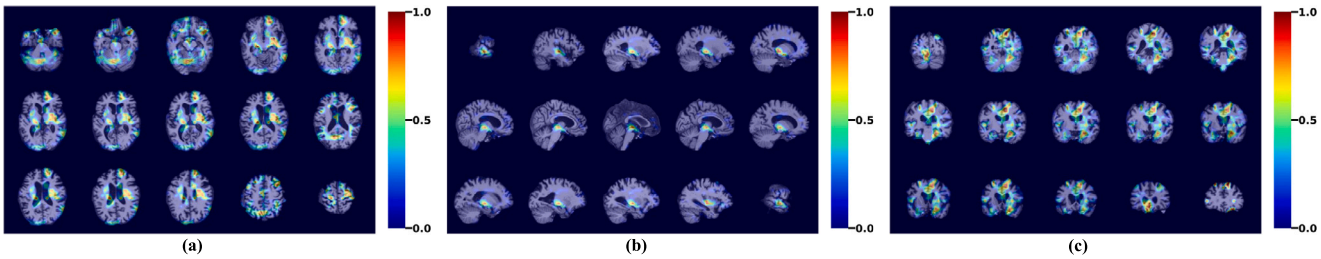

Fig. 10. Grad-CAM interpretable results for three different slice-dividing directions of one brain randomly selected from the ADNI testing set; RID: 011_S_4845 (AD). **(a)** Axial planes, **(b)** Sagittal planes, and **(c)** Coronal planes.

Table 5
Diagnosis results of different slice-dividing directions on the ADNI testing set. The best quantitative results are highlighted in **bold** font.

Slice-dividing directions	ACC	AUC	F1S	SPE	SEN
Sagittal planes	0.850	0.917	0.845	0.880	0.843
Axial planes	0.867	0.924	0.864	0.873	**0.865**
Coronal planes	**0.872**	**0.927**	**0.868**	**0.919**	0.864

Table 6
Diagnosis results of different backbones on the ADNI testing set. The best quantitative results are highlighted in **bold** font.

Backbones	Param.	ACC	AUC	F1S	SPE	SEN
ResNet18	11.24M	0.867	0.924	0.864	0.873	0.865
ResNet34	21.35M	**0.872**	**0.928**	**0.869**	**0.891**	**0.869**

30 slices from three different directions and plot them in Fig. 9(a) with 3D visualization. We find that the crossover region of interest drawn in three directions roughly coincided with the location of the hippocampus, whose changes have been validated to link with the worsening of AD. In Fig. 9(c), we also employ gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017) to visualize the slice-aware module's "local keys" part in randomly selected brains from different datasets. We observe that our diagnosis network locates in different regions based on individual differences. However, the common located regions are consistent with the crossover region in Fig. 9(a) and are majorly around the hippocampus and ventricles.

Slice direction In Table 5, we show the quantitative results of our diagnosis network in three directions: axial, sagittal, and coronal. The coronal planes performed the best, followed by the axial and sagittal planes. Although the axial planes are the most commonly used in clinical practice, coronal planes are more easily used to identify AD-related regions. The radiologist also often assesses AD diagnosis based on a reliable biomarker visual rating of medial temporal lobe atrophy (MTA) (Mårtensson et al., 2020; Custodio et al., 2022; Wan et al., 2022; Ma et al., 2022) on coronal planes. Fig. 10 presents the interpretable results from different directions of one brain randomly selected from the ADNI dataset. We discover similarities in the ROIs from slice to slice due to the use of a shared 2D backbone and the proposed slice-shift module. More Grad-CAM interpretable results are illustrated in Supp. C.3.

2D backbone As described in Section 2.3, our slice-aware and slice-shift modules are not dependent on any specific backbone network, and the diagnosis experiments described above are all based on ResNet18. Here, we also introduce these two modules into ResNet34, and the results are shown in Table 6. The improvement in the backbones contributes to the corresponding enhancement in performance. In addition, the result of ResNet18 also demonstrates that our method can achieve satisfied diagnosis performance with a remarkably low number of network parameters.

3.4. Overall framework results

This subsection focuses on the effectiveness of joint learning and the overall framework results against other state-of-the-art AD diagnosis methods.

Effectiveness of joint learning For the cross-modal synthesis task, whether training in a supervised or unsupervised way, Table 2 shows an improvement in the quality of the synthesized images after joint learning with the downstream diagnosis task. For the diagnosis task, we trained four variants of our diagnosis network in Table 7. V1 is a single-modality variant of our diagnosis network that solely takes labeled MRI images as input; V2 is a modality-agnostic variant of our diagnosis network that takes either labeled MRI or labeled PET images as input; V3 adopts the supervised Pix2Pix synthesis network to achieve joint learning, which takes both paired, labeled MRI and PET images; V4 adopts our unsupervised synthesis network to achieve joint learning, which takes labeled MRI images and unpaired, unlabeled PET images. Note that all the variants only take MRI images as input in the inference phase.

Based on the results in Table 7, we have the following observations. (**i**) By directly comparing V1 to V2, we find that the classification network trained to be agnostic to the modality performs worse than the one trained with only MRI images. Although V2 has access to additional labeled PET images, the performance drop may be explained by: (1) real PET images are more complicated than real MRI images due to large inter-individual *functional* variances, which is more challenging in extracting diagnostic-related features; (2) the modality difference between real PET and MRI images is large, which is challenging to align these two real modalities in the feature space. (**ii**) By comparing V3 to V1, we find that leveraging paired PET images to train a cross-modal synthesis network can help improve the diagnosis performance. Unlike V2, V3 can synthesize PET images whose information is derived *solely* from input MRI without introducing additional information for the diagnosis network compared to real PET images, facilitating the diagnosis network in extracting AD-relevant features more effectively. Moreover, the joint learning of cross-modal and diagnosis networks allows for mutual supervision, leading to feature alignment across different modalities and mining shared modality information more easily. (**iii**) By comparing V4 to V3, we find that our unsupervised synthesis network can better aid the diagnosis network than the supervised one when joint learning. One reason may be that the *pixel-level* supervision in supervised training could force the synthesis model to synthesize each pixel in PET images, which is challenging for the diagnosis network to extract discriminative *feature-level* diagnostic information since many pixels are not discriminative. In contrast, our unsupervised training can better focus on diagnosis-related features through mining shared modality information.

Comparison to state-of-the-art methods We performed a comprehensive comparison of various classical classification methods, including 2D ResNet18 (channel-wise), 3D ResNet18 (Korolev et al., 2017), I3D (Carreira and Zisserman, 2017), and 3D ViT (Dosovitskiy et al., 2020), as well as other state-of-the-art AD diagnosis methods. These

Table 7
Effectiveness of joint learning for diagnosis on the ADNI testing set. The best quantitative results are highlighted in **bold** font.

Variants[♯]		Input modality	Joint learning	ACC	AUC	F1S	SPE	SEN
V1	Single-modality network	Only MRI	✗	0.867	0.924	0.864	0.873	0.865
V2	Modality-agnostic network	MRI/PET	✗	0.837	0.910	0.834	0.864	0.832
V3	Joint learning (Pix2Pix)	Paired MRI + PET	✔	0.872	0.931	0.870	**0.874**	0.871
V4	Joint learning (Ours)	Unpaired MRI + PET	✔	**0.875**	**0.933**	**0.874**	**0.874**	**0.875**

♯ All the methods adopt the proposed diagnosis network and are evaluated using MRI images from the testing set. Note that Pix2Pix is supervised cross-modal synthesis method, while ours is unsupervised.

Table 8
Diagnosis results of the proposed framework for AD diagnosis with other state-of-the-art methods. The best and second best quantitative results are highlighted in **bold** and underlined fonts.

Methods	Param.	*Internal* ADNI testing set					*External* AIBL testing set					*External* NACC testing set				
		ACC	AUC	F1S	SPE	SEN	ACC	AUC	F1S	SPE	SEN	ACC	AUC	F1S	SPE	SEN
2D ResNet18	11.97M	0.802	0.871	0.798	0.870	0.800	0.812	0.901	0.766	0.858	0.831	0.772	0.860	0.764	0.829	0.784
3D ResNet18	33.14M	0.835	0.897	0.829	0.882	0.826	0.835	0.908	0.787	0.857	0.843	0.812	0.909	0.808	0.863	0.817
I3D	12.25M	0.844	0.914	0.830	0.856	0.847	0.837	0.914	0.795	0.867	0.821	0.826	0.911	0.818	0.861	0.831
TripleMRNet	34.31M	0.841	0.911	0.833	0.880	0.827	0.855	0.919	0.809	0.889	0.832	0.837	0.915	0.817	0.854	0.840
MedicalNet	33.14M	0.848	0.915	0.841	0.878	0.836	0.853	0.923	0.799	0.888	0.847	0.831	0.915	0.827	0.851	0.841
FCNlinksCNN	16.30M	0.865	0.923	0.863	0.870	0.865	0.869	0.921	0.818	0.872	0.863	0.855	0.918	0.854	0.860	**0.863**
3D ViT	14.69M	0.811	0.873	0.799	0.889	0.794	0.829	0.868	0.757	0.869	0.774	0.750	0.826	0.748	0.767	0.748
3D ResAttNet18	35.25M	0.844	0.910	0.839	0.877	0.838	0.855	0.918	0.796	0.883	0.817	0.832	0.911	0.831	0.859	0.834
$RES_{late\text{-}dyn}$	11.17M	0.851	0.918	0.848	0.867	0.848	0.860	0.937	0.819	0.904	0.864	0.846	0.912	0.843	0.855	0.844
JSRL	28.70M	0.865	0.916	0.860	**0.896**	0.835	0.865	0.943	0.822	**0.913**	0.852	0.857	**0.921**	0.854	**0.877**	0.852
Our framework	11.24M	**0.875**	**0.933**	**0.874**	0.874	**0.875**	**0.879**	**0.947**	**0.827**	0.907	**0.867**	**0.859**	0.919	**0.856**	0.865	**0.863**

AD diagnosis methods included MedicalNet (Chen et al., 2019), FCNlinksCNN (Qiu et al., 2020), 3D ResAttNet18 (Zhang et al., 2021), TripleMRNet (Bien et al., 2018), $RES_{late\text{-}dyn}$ (Liang et al., 2021), and the joint image synthesis and representation learning (JSRL) framework (Liu et al., 2022). We assessed the performance of these methods on both internal and external datasets, and the comparative outcomes are presented in Table 8. Note that all the methods employ only MRI as input except for JSRL, which utilizes paired MRI and PET data as input. However, all the methods were tested using only MRI as the input. 2D ResNet18 (channel-wise) is a standard 2D ResNet18 network treating slices as channels. 3D ResNet18, I3D, and 3D ViT are the classical 3D classification networks that have also been used in the literature for various disease diagnoses very recently. We finally chose the relatively optimal parameters after extensive experiments in 3D ViT: patch size of $16 \times 16 \times 16$, depth of 4, number of heads equal to 8, hidden layer dimension of 512, and intermediate layer dimension of 2048. MedicalNet employs 3D ResNet18 weights trained on 23 different medical databases. The FCNlinksCNN network begins by extracting high-confidence regions using a 3D CNN and finishes the prediction with a fully connected layer. A self-attention module is used at the end of each residual block in the 3D ResAttNet18 network to extract more global features. Considering the spatio-temporal complexity due to the large size of the input images, the self-attention we implemented is based on the approach developed by Zamir et al. (2022). TripleMRNet and $RES_{late\text{-}dyn}$ are two recent well-performing 2D convolution-based algorithms for Alzheimer's disease diagnosis. To be fair, they are all built with ResNet18 as the backbone. The JSRL uses the underlying shared features between MRI and PET for the joint learning of cross-modal synthesis and AD diagnosis networks.

Our framework performs well in terms of diagnosis performance and generalization ability across different external datasets. And we draw three major conclusions from Table 8. First, slice-to-slice information is critical in 2D or 3D convolution-based methods. Even the simplest 2D ResNet18 (channel-wise) can achieve reasonable results. TripleMRNet combined three slice-dividing directions, ResAttNet18 introduced slice global information, the feature maps' late fusion of $RES_{late\text{-}dyn}$, and our network's slice-shift module will all help us learn more information between slices. Second, many 2D convolution-based algorithms outperform many 3D convolutional methods. 3D networks are prone to overfitting due to significant parameters, and the performance tested in the external dataset drops quickly unless the network is pre-trained with many datasets to maintain good generalization. Due to the lack of local induction bias like CNN (Naseer et al., 2021), 3D ViT has poor experimental results with limited data. Third, the joint learning framework can help cross-modal synthesis and diagnosis networks supervise each other, improving the performance of both tasks. The results of JSRL and our joint learning method both validate the effectiveness of joint learning. Unlike JSRL, our framework does not involve the cross-modal synthesis network during the inference phase, which is memory-friendly and computationally efficient. Supp. C.4 also shows the visualization of data using *t*-distributed stochastic neighbor embedding (*t*-SNE) (Van der Maaten and Hinton, 2008), which validates the network's generalization ability again, and it can be seen that our network generalizes well on external datasets.

4. Discussion

This section discusses the related works and the proposed framework's advantages and limitations.

Discussion on the related work Related works on leveraging PET to improve AD diagnosis from MRI (Li et al., 2014; Campos et al., 2015; Pan et al., 2020, 2021; Liu et al., 2022) usually contain two steps: the first is to synthesize the missing PET modality images, and the second is to use both synthesized PET and original MRI for so-called multi-modal diagnosis. However, during the entire workflow, the information for the synthesized PET images is all derived from the input MRI images, so no additional information is introduced for AD diagnosis; that is, the so-called AD multi-modal diagnosis is essentially single-modal. For example, Pan et al. (2021) propose a feature-consistency generative adversarial network (FGAN) using two synthesis models and four auxiliary components to implement the inter-conversion between MRI and PET, and then develops a disease-image-specific network (DSNet) based on 3D convolution to extract different modalities' features individually and integrate them together for AD diagnosis. Although (Liu et al., 2022) use the underlying shared features of MRI and PET for the joint learning of cross-modal synthesis and diagnosis networks, the quality of synthesized PET images is not being considered, and the cross-modal synthesis network still needs to be deployed during the diagnosis inference phase. Unlike existing methods that take the synthesized PET images as an independent modality, our framework considers the synthesized PET images to be task-specific data augmentation of the

input MRI images and feeds either synthesized PET or real MRI into the same diagnosis network to mine the underlying shared information between them. Please refer to Supp. D for detailed comparison.

Furthermore, our framework differs from existing methods in terms of practical application, training process, and task type, which are summarized as follows. (**i**) Regarding the practical application, the diagnosis network in existing workflows requires the simultaneous input of synthetic PET and real MRI to complete the diagnosis. It indicates that during deployment, they must first synthesize PET images before decision-making. In contrast, once the training of the proposed joint learning framework is completed, our framework does not require the synthesized PET images. (**ii**) Regarding the training process, the existing works train the synthesis and diagnosis networks *separately*. In contrast, we employ joint learning in our framework, where these two networks mutually supervise each other, resulting in high-quality image synthesis and better diagnostic performance. (**iii**) Regarding the task type, most existing workflows are designed for supervised cross-modal synthesis tasks. However, our model focuses more on unsupervised cross-modal synthesis, indicating that our framework is more practical.

In addition, existing works typically demand vast computational resources and have complicated training procedures. Unlike them, we have tried to simplify our framework, such as using *only* one synthesis model in the cross-modal synthesis network and proposing an *efficient 2D network* for the 3D diagnosis tasks.

Discussion on the advantages We highlight the advantages of the proposed framework as follows. (1) *More applicability*. Because collecting large-scale paired data for training is time-consuming and cumbersome in practical scenarios, suffering from potential overfitting. The proposed framework focuses on cross-modality synthesis using massive unpaired unlabeled data, which is more practical than the supervised one. In addition, our diagnosis network is interpretable. It can locate key slices and critical slice regions that are consistent with those associated with AD. (2) *More discriminative diagnosis information in the synthesized missing modality images*. Regardless of training in a supervised or unsupervised way, the synthesized images are challenging to be applied to task-specific diagnosis. The proposed joint learning framework of cross-modal synthesis and diagnosis tasks can connect them together, provide more discriminative diagnosis information in the synthesized images, and improve AD diagnosis performance. In addition, we propose a novel unsupervised cross-modal synthesis network that implements the inter-conversion between 3D PET and MRI in a single model modulated by the AdaIN module, which can help preserve the underlying structure well and achieve better image quality. (3) *Efficient and easy to be deployed*. The proposed AD diagnosis network is implemented based on fully 2D convolution, joint inter- and intra-slice modeling, and zero-params slice-shift modules, which is much more efficient than the conventional 3D one. Moreover, only the diagnosis network needs to be deployed during the diagnosis inference phase. In addition, we can still parallelly synthesize the corresponding PET images to aid radiologists in their diagnoses.

Discussion on the limitations Here, we acknowledge some limitations in this work. (1) *Costly and complicated training*. The proposed framework requires more computational resources in terms of floating point operations and memory usage compared to 2D synthesis tasks, and has complicated training procedures similar to existing works. However, our framework is efficient once trained during the inference phase. How to train our joint learning framework with fewer computational resources in one step needs to be further optimized. (2) *Potentially biased diagnosis for other disease.* Because our joint learning framework is mainly concentrated on AD, which is a *task-specific* framework, we acknowledge that the synthesis process guided by the downstream diagnosis network may introduce specific diagnostic information, potentially biasing the computer-aided diagnosis for other diseases. (3) *Subjects in other stages of AD progression are not considered*. In the setting of unsupervised training, it is more challenging to distinguish Mild Cognitive Impairment (MCI) from NC and AD subjects. How to leverage more information from subjects in other stages of AD development needs to be considered.

5. Conclusion

This paper proposes a novel joint learning framework of cross-modal synthesis and diagnosis for AD by mining underlying shared modality information. Our joint learning framework has the following advantages: (1) more applicability; (2) more discriminative diagnosis information in the synthesized missing modality images; and (3) efficient and easy to be deployed. Moreover, our joint learning framework outperforms other state-of-the-art methods regarding synthesized image quality, diagnosis capacity, and generalization ability across external datasets. In the future, we will integrate other downstream tasks (segmentation, registration, etc.) in our framework to synthesize more realistic images and achieve better performance, and also apply our framework to other brain diseases, such as Parkinson's disease, depression, and autistic disorder, for efficient and interpretable diagnosis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hongming Shan and Chenhui Wang have a patent pending to Hongming Shan and Chenhui Wang.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors would like to thank anonymous reviewers for their insightful comments that helped to improve the quality of the paper and also thank Tao Chen and Zhihao Chen for their helpful suggestions. This work was partially supported by Natural Science Foundation of Shanghai (No. 21ZR1403600), Key Projects of "Double First-Class" Initiative of Fudan University (No. XM03231648), Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01) and ZJLab, Shanghai Municipal of Science and Technology Project (No. 20JC1419500), Shanghai Center for Brain Science and Brain-inspired Technology, and CCF-Tencent Open Research Fund.

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

of Southern California. AIBL Data was collected by the AIBL study group. AIBL study methodology has been reported previously (Ellis et al., 2009). The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD),P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Eric Reiman, MD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, PhD), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P20 AG068024 (PI Erik Roberson, MD, PhD), P20 AG068053 (PI Justin Miller, PhD), P20 AG068077 (PI Gary Rosenberg, MD), P20 AG068082(PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), P30 AG072959 (PI James Leverenz, MD).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2023.103032.

References

Avants, B.B., Tustison, N., Song, G., et al., 2009. Advanced normalization tools (ANTs). Insight J. 2 (365), 1–35.

Beekly, D.L., Ramos, E.M., van Belle, G., Deitrich, W., Clark, A.D., Jacka, M.E., Kukull, W.A., et al., 2004. The National Alzheimer's Coordinating Center (NACC) database: an Alzheimer disease database. Alzheimer Dis. Assoc. Dis. 18 (4), 270–277.

Bien, N., Rajpurkar, P., Ball, R.L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B.N., Yeom, K.W., Shpanskaya, K., et al., 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLOS Med. 15 (11), e1002699.

Campos, S., Pizarro, L., Valle, C., Gray, K.R., Rueckert, D., Allende, H., 2015. Evaluating imputation techniques for missing data in ADNI: A patient classification study. In: Iberoamerican Congress on Pattern Recognition. Springer, pp. 3–10.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? A new model and the Kinetics dataset. In: CVPR. pp. 6299–6308.

Chen, S., Ma, K., Zheng, Y., 2019. Med3D: Transfer learning for 3D medical image analysis. arXiv:1904.00625.

Custodio, N., et al., 2022. Combining visual rating scales to identify prodromal Alzheimer's disease and Alzheimer's disease dementia in a population from a low and middle-income country. Front. Neurol. 13, 1891.

Dar, S.U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., Çukur, T., 2019. Image synthesis in multi-contrast MRI with conditional generative adversarial networks. Trans. Med. Imag. 38 (10), 2375–2388.

Dosovitskiy, A., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR.

Ellis, K.A., et al., 2009. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int. Psychogeriatr. 21 (4), 672–687.

Fischl, B., 2012. FreeSurfer. NeuroImage 62 (2), 774–781.

Gillies, R.J., Kinahan, P.E., Hricak, H., 2016. Radiomics: Images are more than pictures, they are data. Radiology 278 (2), 563–577.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: ICCV. pp. 1026–1034.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. In: NeurIPS. pp. 6840–6851.

Hore, A., Ziou, D., 2010. Image quality metrics: PSNR vs. SSIM. In: ICPR. IEEE, pp. 2366–2369.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861.

Hu, S., Lei, B., Wang, S., Wang, Y., Feng, Z., Shen, Y., 2021. Bidirectional mapping generative adversarial networks for brain MR to PET synthesis. Trans. Med. Imag. 41 (1), 145–157.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141.

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with Adaptive Instance Normalization. In: ICCV. pp. 1501–1510.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 1125–1134.

Korolev, S., Safiullin, A., Belyaev, M., Dodonova, Y., 2017. Residual and plain convolutional neural networks for 3D brain MRI classification. In: ISBI. IEEE, pp. 835–838.

Li, R., Zhang, W., Suk, H.-I., Wang, L., Li, J., Shen, D., Ji, S., 2014. Deep learning based imaging data completion for improved brain disease diagnosis. In: MICCAI. Springer, pp. 305–312.

Liang, G., Xing, X., Liu, L., Zhang, Y., Ying, Q., Lin, A.-L., Jacobs, N., 2021. Alzheimer's disease classification using 2D convolutional neural networks. In: EMBC. IEEE, pp. 3008–3012.

Lin, J., Gan, C., Han, S., 2019. TSM: Temporal shift module for efficient video understanding. In: ICCV. pp. 7083–7093.

Liu, Y., Yue, L., Xiao, S., Yang, W., Shen, D., Liu, M., 2022. Assessing clinical progression from subjective cognitive decline to mild cognitive impairment with incomplete multi-modal neuroimages. Med. Image Anal. 75, 102266.

Liu, F., et al., 2020. JSSR: A joint synthesis, segmentation, and registration system for 3D multi-modal image alignment of large-scale pathological CT scans. In: ECCV. Springer, pp. 257–274.

Luo, Y., Zhou, L., Zhan, B., Fei, Y., Zhou, J., Wang, Y., Shen, D., 2022. Adaptive rectification based adversarial network with spectrum constraint for high-quality PET image synthesis. Med. Image Anal. 77, 102335.

Ma, J., Ma, L.-Y., Man, F., Zhang, G., 2022. Association of homocysteine levels with medial temporal lobe atrophy among carriers and non-carriers of APOE $\varepsilon 4$ in MCI subjects. Front. Psychiatry 13, 823605.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks. In: ICCV. pp. 2794–2802.

Mårtensson, G., Håkansson, C., Pereira, J.B., Palmqvist, S., Hansson, O., van Westen, D., Westman, E., 2020. Medial temporal atrophy in preclinical dementia: visual and automated assessment during six year follow-up. NeuroImage: Clin. 27, 102310.

Naseer, M.M., Ranasinghe, K., Khan, S.H., Hayat, M., Shahbaz Khan, F., Yang, M.-H., 2021. Intriguing properties of vision transformers. In: NeurIPS. pp. 23296–23308.

Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z., Huang, G., 2022. On the integration of self-attention and convolution. In: CVPR. pp. 815–825.

Pan, Y., Liu, M., Lian, C., Xia, Y., Shen, D., 2020. Spatially-constrained Fisher representation for brain disease identification with incomplete multi-modal neuroimages. Trans. Med. Imag. 39 (9), 2965–2975.

Pan, Y., Liu, M., Xia, Y., Shen, D., 2021. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. Trans. Pattern Anal. Mach. Intell. 44 (10), 6839–6853.

Paszke, A., et al., 2019. PyTorch: An imperative style, high-performance deep learning library. In: NeurIPS.

Petersen, R.C., et al., 2010. Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. Neurology 74 (3), 201–209.

Pichler, B.J., Kolb, A., Nägele, T., Schlemmer, H.-P., 2010. PET/MRI: Paving the way for the next generation of clinical multimodality imaging applications. J. Nucl. Med. 51 (3), 333–336.

Qian, S., et al., 2021. Are my deep learning systems fair? An empirical study of fixed-seed training. In: NeurIPS. pp. 30211–30227.

Qiu, S., Joshi, P.S., Miller, M.I., Xue, C., Zhou, X., Karjadi, C., Chang, G.H., Joshi, A.S., Dwyer, B., Zhu, S., et al., 2020. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain 143 (6), 1920–1933.

Rahimpour, M., et al., 2021. Cross-modal distillation to improve MRI-Based brain tumor segmentation with missing MRI sequences. Trans. Biomed. Eng. 69 (7), 2153–2164.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. Springer, pp. 234–241.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626.

Shin, H.-C., et al., 2020. GANDALF: Generative adversarial networks with discriminator-adaptive loss fine-tuning for Alzheimer's disease diagnosis from MRI. In: MICCAI. Springer, pp. 688–697.

Song, G., Chai, W., 2018. Collaborative learning for deep neural networks. In: NeurIPS.

Sun, H., Mehta, R., Zhou, H.H., Huang, Z., Johnson, S.C., Prabhakaran, V., Singh, V., 2019. DUAL-GLOW: Conditional flow-based generative model for modality transfer. In: ICCV. pp. 10611–10620.

Thibeau-Sutre, E., Diaz, M., Hassanaly, R., Routier, A., Dormont, D., Colliot, O., Burgos, N., 2022. ClinicaDL: An open-source deep learning software for reproducible neuroimaging processing. Comput. Methods Programs Biomed. 220, 106818.

Thie, J.A., 2004. Understanding the standardized uptake value, its methods, and implications for usage. J. Nucl. Med. 45 (9), 1431–1434.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2016. Instance Normalization: The missing ingredient for fast stylization. arXiv:1607.08022.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).

Wan, M.-D., et al., 2022. Associations of multiple visual rating scales based on structural magnetic resonance imaging with disease severity and cerebrospinal fluid biomarkers in patients with Alzheimer's disease. Front. Aging Neurosci. 14, 906519.

Wang, R., et al., 2022. Human microRNA (miR-20b-5p) modulates Alzheimer's disease pathways and neuronal function, and a specific polymorphism close to the MIR20B gene influences Alzheimer's biomarkers. Mol. Psychiatry 27 (2), 1256–1273.

Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30 (1), 79–82.

Winblad, B., et al., 2016. Defeating Alzheimer's disease and other dementias: a priority for European science and society. Lancet Neurol. 15 (5), 455–532.

Wong, W., 2020. Economic burden of Alzheimer disease and managed care considerations. Am. J. Manag. Care 26 (8 Suppl), S177–S183.

Yang, S., Kim, E.Y., Ye, J.C., 2021a. Continuous conversion of CT kernel using switchable CycleGAN with AdaIN. Trans. Med. Imag. 40 (11), 3015–3029.

Yang, Q., Li, N., Zhao, Z., Fan, X., Chang, E.I., Xu, Y., et al., 2020. MRI cross-modality image-to-image translation. Sci. Rep. 10 (1), 1–18.

Yang, H., Sun, J., Yang, L., Xu, Z., 2021b. A unified Hyper-GAN for unpaired multi-contrast MR image translation. In: MICCAI. Springer, pp. 127–137.

Yi, X., Walia, E., Babyn, P., 2019. Generative adversarial network in medical imaging: A review. Med. Image Anal. 58, 101552.

Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H., 2022. Restormer: Efficient transformer for high-resolution image restoration. In: CVPR. pp. 5728–5739.

Zhang, X., Han, L., Zhu, W., Sun, L., Zhang, D., 2021. An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. J. Biomed. Health Inf. 26 (11), 5289–5297.

Zhang, J., He, X., Qing, L., Gao, F., Wang, B., 2022. BPGAN: Brain PET synthesis from MRI using generative adversarial network for multi-modal Alzheimer's disease diagnosis. Comput. Methods Programs Biomed. 217, 106676.

Zhao, X., Zhao, X.-M., 2021. Deep learning of brain magnetic resonance images: A brief review. Methods 192, 131–140.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232.