Disentangle First, Then Distill: A Unified Framework for Missing Modality Imputation and Alzheimer's Disease Diagnosis

Yuanyuan Chen, Yongsheng Pan, *Member, IEEE*, Yong Xia, *Member, IEEE*, and Yixuan Yuan, *Member, IEEE*

***Abstract*— Multi-modality medical data provide complementary information, and hence have been widely explored for computer-aided AD diagnosis. However, the research is hindered by the unavoidable missing-data problem, *i.e.*, one data modality was not acquired on some subjects due to various reasons. Although the missing data can be imputed using generative models, the imputation process may introduce unrealistic information to the classification process, leading to poor performance. In this paper, we propose the *Disentangle First, Then Distill (DFTD)* framework for AD diagnosis using incomplete multi-modality medical images. First, we design a region-aware disentanglement module to disentangle each image into inter-modality relevant representation and intra-modality specific representation with emphasis on disease-related regions. To progressively integrate multi-modality knowledge, we then construct an imputation-induced distillation module, in which a lateral inter-modality transition unit is created to impute representation of the missing modality. The proposed DFTD framework has been evaluated against six existing methods on an ADNI dataset with 1248 subjects. The results show that our method has superior performance in both AD-CN classification and MCI-to-AD prediction tasks, substantially over-performing all competing methods.**

***Index Terms*— Alzheimer's disease, mild cognitive impairment, multi-modality diagnosis, modality imputation.**

Manuscript received 3 October 2022; revised 6 January 2023, 2 April 2023, and 26 May 2023; accepted 7 July 2023. Date of publication 14 July 2023; date of current version 30 November 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFC2009903/2022YFC2009900, in part by the China Postdoctoral Science Foundation under Grant BX2021333 and Grant 2021M703340, in part by the Innovation and Technology Commission-Innovation and Technology Fund under Grant ITS/100/20. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health) under Grant U01 AG024904. *(Corresponding authors: Yong Xia; Yixuan Yuan.)*

Yuanyuan Chen and Yong Xia are with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: yychen@mail.nwpu.edu.cn; yxia@nwpu.edu.cn).

Yongsheng Pan is with the School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China (e-mail: yspan@mail.nwpu.edu.cn).

Yixuan Yuan is with Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, SAR, China (e-mail: yxyuan@ee.cuhk.edu.hk).

Digital Object Identifier 10.1109/TMI.2023.3295489

I. Introduction

ALZHEIMER'S disease (AD) and mild cognitive impairment (MCI) are neurodegenerative disorders and pose a great threat to the health of elderly people [1], [2]. As deep learning has achieved excellent performance in medical image analysis [3], [4], numerous deep learning based techniques have been applied to the diagnosis of AD and the conversion from MCI to AD. Recently, the multi-modality PET and MRI imaging has attracted ever-growing research interests by providing complementary functional and structural information. However, due to a radioactive concern and high cost, patient dropout is a common and serious challenge in multi-modality imaging practice, leading to the modality missing issue [5], [6], [7]. For instance, all subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [8] have baseline structural magnetic resonance imaging (sMRI) data, while only approximately half subjects have baseline fluorodeoxyglucose positron emission tomography (PET) scans. Modality missing brings massive difficulties to fully explore the relevance among different modalities. Therefore, finding a way to utilize the incomplete multi-modality data is of great significance for improving the performance of multi-modality AD diagnosis.

An intuitive way to tackle this problem is to discard the subjects with incomplete scans [9], [10]. Such a solution, however, reduces the number of subjects available for model training, leading to performance degradation. Hence, extensive research efforts have been devoted to novel solutions based on subspace learning [6], [11], [12], knowledge distillation [13], [14], and missing data imputation [15], [16], [17], [18]. Subspace learning-based methods attempt to learn a common latent feature space using modality-complete data and an independent latent modality-specific feature space using modality-incomplete data, and then projected the latent representations into the label space for AD diagnosis [11]. Knowledge distillation-based methods distill the knowledge learned by a teacher network, which take multi-modality information as input, to a student network designed for mono-modality diagnosis [13]. Despite fully utilizing all available data, these methods reply only on modality-complete data to depict the relevance among different modalities. Moreover, they cannot learn from both modality-complete

and -incomplete data using a unified model, and hence bring extra difficulties to model design and training. Missing data imputation-based methods have been increasingly studied. These methods first impute the missing modality in the feature, kernel or image space [15], [16], [17], and then perform diagnosis using both real and synthetic modality-complete data. They achieve improved performance in disease diagnosis due to using an enlarged training dataset and the increased diversity of feature expression caused by imputation. Although not increasing input information, missing data imputation provides a different way to depict the disease-related features and thus helps the model to discovery useful information that is hard to be captured in the existing modality.

However, missing data imputation-based methods have two major limitations. First, they utilize the whole image to perform imputation, which may include redundant and biased information. Recent studies [19], [20] demonstrate that the information within a modality can be disentangled into two parts: the inter-modality relevant information and intra-modality specific information. The former characterizes the disease patterns in related but different ways, like the different impressions would be formed when viewing a mountain from different sides. Hence, translating this part of information from the existing modality to the missing modality is meaningful and feasible. However, the latter represents the exclusive patterns of the disease captured by each modality. Forcing the model to learn this part of information may introduce biases to the synthetic image, leading subsequently to less-accurate diagnosis. Therefore, we suggest separating the inter-modality relevant information from the existing modality and using it alone to impute the corresponding inter-modality relevant information of the missing modality. Second, most imputation-based methods first explicitly generate the missing-modality image and then fuse features of multi-modality images for diagnosis. Obviously, the missing-modality imputation and multi-modality disease diagnosis are performed sequentially as two sub-tasks with different objectives, which may result in sub-optimal solutions. Inspired by knowledge distillation methods [21], [22], in which the multi-modality knowledge is integrated into the mono-modality model that can be directly used for image analysis, we advocate to combine the merits of missing-modality imputation and knowledge distillation. Specifically, we perform modality imputation and multi-modality knowledge fusion in the training phase, and then distill the multi-modality knowledge to a mono-modality branch. Thus, $\mathcal{M}_{IID}$ can conduct multi-modality diagnosis using only single-modality data.

In this paper, we propose a ***Disentangle First, Then Distill (DFTD)*** framework for AD diagnosis using incomplete multi-modality PET and sMRI imaging. Under this framework, we first disentangle the existing-modality image into inter-modality relevant representations and intra-modality specific representations, and then perform multi-modality knowledge distillation to impute the missing-modality image based on the inter-modality relevant representations. Specifically, considering the variable contributions of brain regions to AD diagnosis, we design a region-aware disentanglement module to highlight the discriminative regions in the disentanglement process, resulting in the disentangled features with a strong discriminatory power. Meanwhile, we construct an imputation-induced distillation module to progressively integrate and distill multi-modality knowledge. In this module, a novel lateral inter-modality transition unit is proposed to impute the representations of those modality-missing subjects. The major contributions of this work are as follows:

- Different to conventional imputation-based methods which use the whole existing-modality image to impute the missing modality, we propose region-aware disentanglement to use only the inter-modality relevant information for imputation and to integrate diagnosis-related image regions into the disentanglement process, avoiding generating redundant and biased information.
- Combining the merits of imputation-based and knowledge distillation-based methods, we propose an imputation-induced distillation module, which can impute the representations of modality-missing subjects and diagnose a subject using only single-modality data.
- Experimental results on the ADNI dataset suggest that our DFTD framework achieves superior performance over five state-of-the-art approaches on both AD classification and MCI conversion prediction.

II. Related Work

A. Learning With Missing Modalities

Many research efforts have been devoted to addressing the missing modality issue for medical image analysis, resulting in many methods based mainly on subspace learning [23], [24], knowledge distillation [21], [22], and missing data imputation [25], [26]. In subspace learning based methods [23], [24], a common latent feature space is learned using modality-complete data and an independent modality-specific feature space is learned using modality-incomplete data. Zhou et al. [23] devised a correlation module to capture multi-modality relations, and then fused all available features in a latent representations with an attention module. Considering the distinct role of modality information in the segmentation of different tumor regions, Yang et al. [24] proposed to first decouple the modality-specific information from MRI data and then explicitly model the correlation of different modalities for brain tumor segmentation. Knowledge distillation-based methods distill multi-modality knowledge into mono-modality knowledge through a 'teacher-student' network. Chen et al. [21] designed a novel privileged knowledge learning framework, under which they used both a pixel-level and an image-level distillation schemes to distill the privileged multi-modality information for single-modality image segmentation. Wang et al. [22] adopted a novel adversarial co-training mechanism for both full-modality and missing-modality branches, which can supplement each other's feature representations and encourage the alignment of latent representations. Missing data imputation aims to impute missing modalities to construct complete multi-modality data. Gao et al. [25] proposed a generative model, which characterizes the joint distribution of image and non-image data with a class regularization loss on imputed data to recover discriminative information. Hamghalam et al. [26] designed a Multi-modal Gaussian process prior variational autoencode to impute one or more missing modalities by exploring sub-modality correlations.

B. Missing Data Imputation-Based AD Diagnosis

Due to the missing data issue, data imputation methods have been proposed for multi-modality AD diagnosis [11], [16], [17], [27], [28], [29], [30], [31]. Zhou et al. [11] recovered the missing modality by maximizing the dependency among different modalities and integrated missing data recovery, latent space learning, and image label prediction into a unified framework. Generative adversarial networks (GANs) have been used to synthesize the missing modality data based on the existing modality data [17], [27], [28], [30], [31]. Pan et al. [30] employed Cycle-GAN to synthesize the missing PET images based on the corresponding MRI scans and achieved good performance using both real and synthetic multi-modality PET and MRI images. Considering the variation of the discriminatory capability over modalities, they further proposed a disease-image-specific deep learning (DSDL) framework to impute the missing scans which are more consistent with real ones from a diagnostic perspective [31]. Cao et al. [17] proposed Auto-GAN to generate missing modalities and introduced a self-supervised learning scheme for better synthesis. AutoGAN can estimate any missing modality by imposing a modality mask vector to input images. Despite improved performance, these methods directly use the existing modality to impute the missing modality and may introduce redundant and biased information to synthetic images. Hence, we design a region-aware disentanglement module to split the information of existing modality into two parts, and then perform missing modality imputation based only on the inter-modality relevant representations.

C. Feature Disentanglement

Deep learning models encode an image into a high-dimensional representation, which is highly entangled [32]. To obtain the specific information from the learned representation, many feature disentanglement methods have been proposed, such as InfoGAN [33], β-VAE [34], and JointVAE [35]. The effectiveness of feature disentanglement in image classification has been demonstrated extensively. After disentangling the modality-exclusive information from the learned representations, Guo et al. [19] reduced the impact of such information and thus improved the interpretability of extracted features. Lu et al. [20] proposed a cross-modality shared-specific feature transfer algorithm to improve the performance of person re-identification via exploring the potential of both modality-shared and modality-specific information. Sanchez et al. [36] proposed a method to disentangle the representations of images into shared representations and exclusive representations based on mutual information estimation. Li et al. [37] proposed a self-supervised learning algorithm to effectively disentangle modality-invariant features and patient-similarity features for retinal disease diagnosis. Chen et al. [38] utilized feature disentanglement to decompose the input into the modality-specific appearance code and the modality-invariant content code to perform brain tumor segmentation. In this work, we attempt to disentangle the image information into inter-modality relevant representations and inner-modality specific representations while considering the variable contributions of different image regions to AD diagnosis.

D. Multi-Modality Knowledge Distillation

Multi-modality knowledge distillation aims to transfer modality-correlated knowledge from a multi-modality network (teacher) to a mono-modality network (student) [39]. Hu et al. [40] proposed KD-Net, which uses multi-modality knowledge distillation to improve the performance of brain tumor segmentation. Sonsbeek et al. [41] proposed variational knowledge distillation to leverage the knowledge in both electronic health records and medical images for chest disease classification. Valverde et al. [42] presented the self-supervised MM-DistillNet framework, which consists of multiple teachers that leverage diverse modalities, including RGB, depth, and thermal images, to simultaneously exploit complementary cues and distill knowledge into a single audio student network. Recently, self-knowledge distillation has been proposed, which trains a student network progressively to distill its own knowledge without a pre-trained teacher network [43], [44]. Ji et al. [44] proposed the feature refinement via self-knowledge distillation (FRSKD) framework, which utilizes an auxiliary self-teacher network to transfer refined knowledge to the classification network. In this work, we formulate multi-modality image classification as a self-knowledge distillation task and then progressively discover the underlying multi-modality knowledge embedded in a specific modality. Different to other self-knowledge distillation methods, our method performs modality transition before fusing multi-modality features in the teacher network, aiming to impute representations for those modality-missing subjects. Thus, multi-modality feature integration and knowledge distillation can be performed on all subjects.

III. Method

A. Overview

Let a modality-complete case be denoted by $\{x_a, x_b, y\}$, where x_a and x_b are the images of modality a and b, respectively, and y is the disease label. Then, a modality-incomplete case (*e.g.*, with missing modality b) can be denoted by $\{x_a, y\}$. The proposed DFTD framework aims to predict the disease label y based on the images of either two modalities (*i.e.*, $\{x_a, x_b\}$) or only the existing modality (*i.e.*, $\{x_a\}$). To achieve this goal, the DFTD framework consists of a region-aware disentanglement module $\mathcal{M}_{RAD}$ and an imputation-induced distillation module $\mathcal{M}_{IID}$ (see Fig. 1). The inter-modality relevant representation and intra-modality specific representation produced by $\mathcal{M}_{RAD}$ are denoted by $\{c_a, c_b\}$ and $\{s_a, s_b\}$, respectively. Then, the global representations $f_a = [c_a, s_a]$ and $f_b = [c_b, s_b]$ are fed to $\mathcal{M}_{IID}$ for multi-modality feature distillation and disease classification. The classification process can be formally expressed as

$$\hat{y} = \mathcal{M}_{IID}(f_a, f_b). \tag{1}$$

The $\mathcal{M}_{IID}$ module contains three components: a student branch $\mathcal{S}$, an integrated teacher branch $\mathcal{T}$, and a lateral inter-modality transition unit $\mathcal{U}_{\mathcal{LIT}}$. The $\mathcal{U}_{\mathcal{LIT}}$ unit is trained using the inter-modality relevant representations (*i.e.*, c_a and c_b), aiming to learn a mapping from modality a to b in the latent feature space. For a modality-incomplete data, the trained

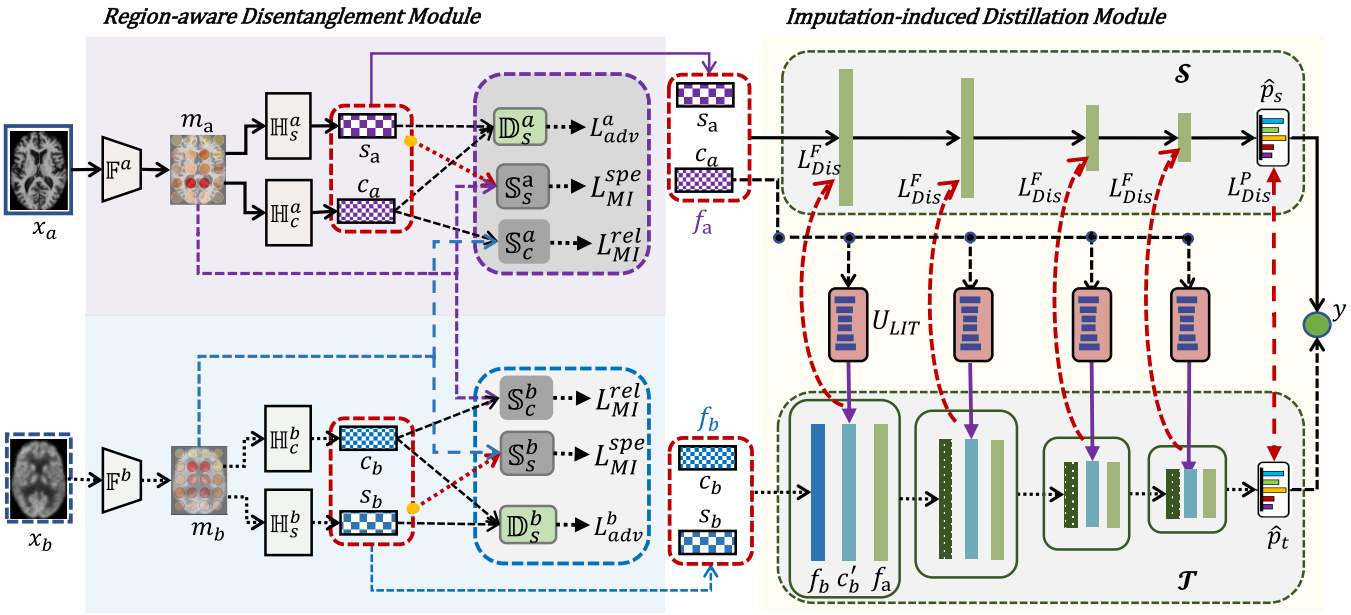

Fig. 1. Architecture of the proposed ***Disentangle First, Then Distill*** framework. There are two core modules: (1) a region-aware disentanglement module for disentangling each image into inter-modality relevant representation and intra-modality specific representation and (2) an imputation-induced distillation module for simultaneous missing modality imputation and multi-modality knowledge distillation. For modality-complete data (modality *a* and *b*), data flows on both solid lines and dashed lines. For modality-incomplete subjects (missing modality *b*), data flows on the solid lines only. Notations are explained in the main text.

TABLE I
SYMBOLS USED IN THIS PAPER

Symbol	Meaning
x_a/x_b	Input image from modality A/B
y	Disease label
f_a/f_b	Global representations of modality A/B
m_a/m_b	Intermediate feature maps of x_a/x_b
c_a/c_b	Inter-modality relevant representations of modality A/B
s_a/s_b	Intra-modality specific representations of modality A/B
r_{ai}/r_{bi}	The i_{th} region of the feature map from modality A/B
ω_{ai}/ω_{bi}	Learnable region weight of r_{ai}/r_{bi}
$\mathbb{F}^a/\mathbb{F}^b$	Base extractor to obtain intermediate feature maps m_a/m_b
$\mathbb{H}_c^a/\mathbb{H}_c^b$	Head extractor for inter-modality relevant representations c_a/c_b
$\mathbb{H}_s^a/\mathbb{H}_s^b$	Head extractor for intra-modality specific representations s_a/s_b
$\mathbb{S}_c^a/\mathbb{S}_c^b$	A discriminator to maximize the cross MI between c_a/c_b and r_{bi}/r_{ai}
$\mathbb{S}_s^a/\mathbb{S}_s^b$	A discriminator to maximize the MI between r_{ai}/r_{bi} and f_a/f_b
$\mathbb{D}_s^a/\mathbb{D}_s^b$	A discriminator to minimize the MI between s_a/s_b and c_a/c_b
$\mathbb{P}_{c_a s_a}$	The joint probability distribution of features c_a and s_a
$\mathbb{P}_{c_a}\mathbb{P}_{s_a}$	The product of the marginal distributions of features c_a and s_a
$\mathbb{E}/\mathbb{D}$	Encoder/Decoder of the lateral inter-modality transition unit $\mathcal{U}_{\mathcal{LIT}}$
o_i^t/o_i^s	Features of the teacher branch $\mathcal{T}$/student branch $\mathcal{S}$ at the i_{th} layer
$\hat{p}_t/\hat{p}_s$	Soft labels of $\mathcal{T}/\mathcal{S}$
$\hat{I}_{JSD}$	Cross mutual information (MI) between two modalities
L^{rel}	Loss for inter-modality relevant representation learning
L^{spe}	Loss for intra-modality specific representation learning
$\mathcal{L}_{CE}$	Cross-entropy loss
$\mathcal{L}_{Dis}^{P}$	Logit distillation loss
$\mathcal{L}_{Dis}^{F}$	Feature distillation loss

$\mathcal{U}_{\mathcal{LIT}}$ is able to impute the representation c_b' for modality b based on c_a. In this case, the classification process can be expressed as

$$\hat{y} \approx \mathcal{M}_{IID}(f_a, c_b'). \tag{2}$$

Table I describes all relevant symbols used in our DFTD framework. We now delve into the details of each module.

B. Region-Aware Disentanglement

The region-aware disentanglement module $\mathcal{M}_{RAD}$ is designed to disentangle each brain image into the inter-modality relevant representation and intra-modality specific representation. Since there is little relevance between the intra-modality specific representations of two modalities, it is almost impossible for our model to translate this representation from the existing modality a to the missing modality b. By contrast, the inter-modality relevant representation is the part of information that can be translated between two modalities. Therefore, we only utilize this representation to perform missing modality imputation. Although such imputation does not bring any extra information, but it may make it possible to use the available information more effectively, as the imputed inter-modality relevant representation of the missing modality provides a different view of the disease pattern. Therefore, similar to other missing data imputation-based methods, jointly using the inter-modality relevant representation of the existing modality and the imputed inter-modality relevant representation of the missing modality for disease classification may produce better performance than using only a single modality. Besides, considering brain regions have different contributions to AD diagnosis, we introduce the region-aware disentanglement mechanism to highlight the discriminative regions in the disentanglement process, resulting in the disentangled features with a strong discriminatory ability.

1) Inter-Modality Relevant Representation Learning: Given a modality-complete data $\{x_a, x_b, y\}$, two networks which extract the inter-modality relevant representations c_a and c_b, respectively, are denoted as $\mathbb{F}^a \circ \mathbb{H}_c^a : x_a \rightarrow c_a$ and $\mathbb{F}^b \circ \mathbb{H}_c^b : x_b \rightarrow c_b$. To make the representations c_a have discriminatory power for AD diagnosis, encoder $\mathbb{F}^a$ and $\mathbb{H}_c^a$ are optimized

via maximizing the cross MI between its output c_a and the features of most disease-related local region in image x_b.

To highlight the brain regions with discriminatory power, we design a region saliency re-weighting strategy and incorporate it into the disentanglement process. For each modality (*e.g.*, modality a for simplicity), let the intermediate feature maps generated by $\mathbb{F}^a$ be denoted by m_a, which is split into $N = 4 \times 4 \times 4$ non-overlapping regions. Each region is represented by a feature vectors r_{ai} and is assigned a learnable weight $\{\omega_{ai}\}$ to measure its contribution to the diagnosis. The weighted average MI can be calculated as follows

$$L_{MI}^{rel} = \frac{1}{N}\sum_{i=1}^{N}(\omega_{ai}\hat{I}_{JSD}(c_b; r_{ai}) + \omega_{bi}\hat{I}_{JSD}(c_a; r_{bi})), \tag{3}$$

where $\hat{I}_{JSD}$ means the cross MI is maximized based on the Jensen-Shannon divergence (JSD). The weights $\{\omega_{ai}\}_{i=1}^{N}$ and $\{\omega_{bi}\}_{i=1}^{N}$ are updated with other model parameters.

When maximizing the cross MI between c_a and r_{bi}, we employ a negative sampling strategy similar to *Deep Infomax* [45]. For the j_{th} sample in a mini-batch, we regard its inter-modality relevant representation $c_{a(j)}$ of x_a and the feature vector of the i^{th} region $r_{bi(j)}$ of x_b as a real pair. Then, the mini-batch is shuffled and we obtained a new feature vector $r_{bi(k)}$ that belongs to the k_{th} sample. We now regard $c_{a(j)}$ and $r_{bi(k)}$ as a fake pair. We design a discriminator $\mathbb{S}_c^a$ to distinguish these two types of pairs, and the loss function of $\mathbb{S}_c^a$ is defined as follows:

$$L_{\mathbb{S}_c^a}^{i} = \log(1+e^z)(\mathbb{S}_c^a(c_{a(j)}, r_{bi(k)}) - \log(1+e^z)(\mathbb{S}_c^a(c_{a(j)}, r_{bi(j)}). \tag{4}$$

Similarly, the loss function of $\mathbb{S}_c^b$ is defined to maximize the MI between c_b and x_a.

$$L_{\mathbb{S}_c^b}^{i} = log(1+e^z)(\mathbb{S}_c^b(c_{b(j)}, r_{ai(k)}) - log(1+e^z)(\mathbb{S}_c^b(c_{b(j)}, r_{ai(j)}) \tag{5}$$

Therefore, the loss function for learning the inter-modality relevant representation is as follows:

$$L^{rel} = \frac{1}{N}\sum_{i=1}^{N}(\omega_{ai}L_{\mathbb{S}_c^b}^{i} + \omega_{bi}L_{\mathbb{S}_c^a}^{i}). \tag{6}$$

2) Intra-Modality Specific Representation Learning: Similarly, the intra-modality specific representations s_a and s_b, which capture the remaining information of each modality, are extracted by two networks, denoted as $\mathbb{F}^a \circ \mathbb{H}_s^a : x_a \rightarrow s_a$ and $\mathbb{F}^b \circ \mathbb{H}_s^b : x_b \rightarrow s_b$. For x_a, we first estimate and maximize the weighted average MI between the region features r_{ai} and global representation $f_a = [c_a, s_a]$. For x_b, we perform the same operation. Then, the loss function is defined as follows

$$L_{MI}^{spe} = \frac{1}{N}\sum_{i=1}^{N}(\omega_{ai}\hat{I}_{JSD}(f_a; r_{ai}) + \omega_{bi}\hat{I}_{JSD}(f_b; r_{bi})). \tag{7}$$

Similarly, we employ the negative sampling strategy to maximize the weighted average MI between r_{ai} and f_a. For the j_{th} sample in a mini-batch, we regard its global representation $f_{a(j)}$ and the feature vector of the i^{th} region $r_{ai(j)}$ as a real pair. Then, the mini-batch is shuffled and we obtained a new feature vector $r_{ai(k)}$ that belongs to the k_{th} sample. We now regard $f_{a(j)}$ and $r_{ai(k)}$ as a fake pair. We design a discriminator $\mathbb{S}_s^a$ to distinguish these two pairs, and the loss function of $\mathbb{S}_s^a$ is defined as follows:

$$L_{\mathbb{S}_s^a}^{i} = \log(1+e^z)(\mathbb{S}_s^a(f_{a(j)}, r_{ai(k)}) - \log(1+e^z)(\mathbb{S}_s^a(f_{a(j)}, r_{ai(j)}). \tag{8}$$

The loss function of $\mathbb{S}_2^b$ is defined to maximize the average MI between the region features r_{bi} and global representation $f_b = [c_b, s_b]$.

$$L_{\mathbb{S}_s^b}^{i} = log(1+e^z)(\mathbb{S}_2^b(f_{b(j)}, r_{bi(k)}) - log(1+e^z)(\mathbb{S}_2^b(f_{b(j)}, r_{bi(j)}) \tag{9}$$

Since the inter-modality relevant representation c_a has already been calculated, we enforce the intra-modality specific representation s_a to include the information which has not been captured by c_a. Therefore, the mutual information between s_a and c_a must be minimized. According to the information theory, the mutual information between features c_a and s_a can be expressed as follows

$$I(c_a, s_a) = \int_{c_a}\int_{s_a} p(c_a, s_a) log\left(\frac{p(c_a, s_a)}{p(c_a)p(s_a)}\right) d_{c_a} d_{s_a}. \tag{10}$$

The mutual information $I(c_a, s_a)$ can be rewritten as the Kullback-Leibler divergence between the joint probability distribution $\mathbb{P}_{c_a s_a}$ and the product of the marginal distributions $\mathbb{P}_{c_a}\mathbb{P}_{s_a}$, shown as follows

$$I(c_a, s_a) = D_{KL}(\mathbb{P}_{c_a s_a} || \mathbb{P}_{c_a}\mathbb{P}_{s_a}). \tag{11}$$

Following the optimization method in [36], we relax the minimization of $I(c_a, s_a)$ to the minimization of $D_{JS}(\mathbb{P}_{c_a s_a} || \mathbb{P}_{c_a}\mathbb{P}_{s_a})$ in an adversarial manner (see Eq. 12). A discriminator $\mathbb{D}_s^a$ is trained to classify representations drawn from $\mathbb{P}_{c_a s_a}$ as fake samples and representations drawn from $\mathbb{P}_{c_a}\mathbb{P}_{s_a}$ as real samples.

$$L_{adv}^{a} = \mathbb{E}_{p(c_a)p(s_a)}[log\mathbb{D}_s^a(c_a, s_a)] + \mathbb{E}_{p(c_a, s_a)}[log(1 - \mathbb{D}_s^a(c_a, s_a))] \tag{12}$$

Similarly, the mutual information between s_b and c_b can be minimized by Eq. 13.

$$L_{adv}^{b} = \mathbb{E}_{p(c_b)p(s_b)}[log\mathbb{D}_s^b(c_b, s_b)] + \mathbb{E}_{p(c_b, s_b)}[log(1 - \mathbb{D}_s^b(c_b, s_b))] \tag{13}$$

Therefore, the loss for intra-modality specific representation learning is as follows

$$max\ min\ L^{spe} = L_{MI}^{spe} - L_{adv}^{a} - L_{adv}^{b}. \tag{14}$$

C. Imputation-Induced Distillation

After obtaining the inter-modality relevant representation of the existing-modality image that is suitable for modality imputation, a conventional operation is to first impute the features of the missing-modality image and then fuse the features of two modalities for classification. In these methods, the missing-modality imputation and multi-modality

diagnosis are performed sequentially as two sub-tasks with different objectives, which may result in sub-optimal solutions. To address this issue, we propose the imputation-induced distillation module $\mathcal{M}_{IID}$, which can integrate the missing-modality imputation and multi-modality knowledge fusion into a unified framework, and can conduct multi-modality diagnosis using only single-modality data. The $\mathcal{M}_{IID}$ module consists of three components, including a student branch $\mathcal{S}$, an integrated teacher branch $\mathcal{T}$, and a lateral inter-modality transition unit $\mathcal{U}_{\mathcal{LIT}}$. The student branch $\mathcal{S}$ takes the global features f_a of modality a as its input and performs mono-modality classification. The teacher branch $\mathcal{T}$ takes the global features of both modalities (*i.e.*, f_a and f_b) as its input and performs multi-modality feature integration as well as classification. For modality-incomplete subjects, the lateral inter-modality transition unit $\mathcal{U}_{\mathcal{LIT}}$ imputes the missing modality representations at each scale (see Fig. 1). With the existing and imputed representations, $\mathcal{T}$ performs multi-modality feature integration and the integrated multi-modality features are then transferred to $\mathcal{S}$ to improve the classification performance.

1) Lateral Inter-Modality Transition: The lateral inter-modality transition unit $\mathcal{U}_{\mathcal{LIT}}$ consists of an encoder $\mathbb{E}$ and a decoder $\mathbb{D}$. During training, the encoder $\mathbb{E}$ converts the inter-modality relevant representation c_a into a latent representation, which is then sent to the decoder D to produce the translated feature c'_b. The transition error is defined as follows

$$L_{LIT} = \left\| c'_b - c_b \right\|_2 = \|\mathbb{D}(\mathbb{E}(c_a) - c_b\|_2, \tag{15}$$

where the operator $\|a - b\|_2$ represents the Euclidean distance between a and b. After training $\mathcal{U}_{\mathcal{LIT}}$ with modality-complete data, we can use it to impute the representation of missing modality b for modality-incomplete subjects.

2) Multi-Modality Feature Integration and Transfer: We concatenate the global representation f_a, global representation f_b and imputed inter-modality relevant representation c'_b together, and then employ a 1×1 convolution kernel to fuse them into a unified feature vector j_0^t, shown as follows

$$j_0^t = Conv(f_a, f_b, c'_b). \tag{16}$$

Note that, for modality-incomplete subjects, the global representation f_b is not available and zero-padding is used to keep the dimension of concatenated features unchanged.

Let o_i^t and o_i^s denote the features of the teacher branch $\mathcal{T}$ and student branch $\mathcal{S}$ at the i_{th} layer, respectively. We employed a 1×1 convolutional kernel to fuse o_i^t, o_i^s and $\mathcal{U}_{\mathcal{LIT}}(c_a)$ into a unified feature vector. To make the dimension of $\mathcal{U}_{\mathcal{LIT}}(c_a)$ same with o_i^t and o_i^s, we used the max-pooling operation to resize $\mathcal{U}_{\mathcal{LIT}}(c_a)$ before feature fusion. Thus, the feature of $\mathcal{T}$ and $\mathcal{S}$ can be fused as

$$j_i^t = Conv(o_i^s, resize(\mathcal{U}_{\mathcal{LIT}}(c_a)), o_{i-1}^t). \tag{17}$$

Following the strategy of generalized knowledge distillation, we transfer useful multi-modality knowledge from the teacher branch $\mathcal{T}$ to the student branch $\mathcal{S}$ using their soft labels. The soft labels of $\mathcal{T}$ and $\mathcal{S}$ are defined as follows

$$\hat{p}_t = softmax(\mathcal{T}(j_0^t)/K) \tag{18}$$

$$\hat{p}_s = softmax(\mathcal{S}(f_a)/K), \tag{19}$$

where K is the temperature scaling parameter that controls the softness of $\hat{p}_t$ and $\hat{p}_s$. The soft labels uncover the relations between classes that is harder to detect with hard labels. The knowledge distillation loss for soft labels is defined as

$$\mathcal{L}_{Dis}^P = D_{KL}(\hat{p}_s || \hat{p}_t). \tag{20}$$

Moreover, the knowledge distillation is also performed through the feature consistency constraint between two branches. The feature distillation loss $\mathcal{L}_{Dis}^F$ is defined as

$$\mathcal{L}_{Dis}^F = \sum_{i=1}^{z} ||\phi(o_i^t) - \phi(o_i^s)||_2, \tag{21}$$

where ϕ is a L_2 normalization operation.

The student branch $\mathcal{S}$ and teacher branch $\mathcal{T}$ are supervised by their common ground-truth label using the cross-entropy loss. Therefore, the loss for multi-modality feature integration and transfer is as follows

$$\begin{aligned}\mathcal{L}_{\mathcal{M}_{IID}} = \mathcal{L}_{CE}(f_a, y) + \mathcal{L}_{CE}(j_0^t, y) + \alpha \cdot \mathcal{L}_{Dis}^P(\hat{p}_s, \hat{p}_t; K) \\ + \beta \cdot \mathcal{L}_{Dis}^F(o^t, o^s) + L_{LIT}(c_a, c_b),\end{aligned} \tag{22}$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss, and α and β are two weighting parameters.

D. Optimization

Training the DFTD framework consists of optimizing the region-aware disentanglement module $\mathcal{M}_{RAD}$ and imputation-induced distillation module $\mathcal{M}_{IID}$ via minimizing the following total loss

$$L_{total} = L^{rel} + L^{spe} + \mathcal{L}_{\mathcal{M}_{IID}}. \tag{23}$$

In the training stage, we first use modality-complete subjects to train $\mathcal{M}_{RAD}$ and $\mathcal{M}_{IID}$ under the supervision of their disease labels. Next, for each modality-incomplete subject (i.e., only modality a available), we obtain two disentangled representations c_a and s_a using the trained $\mathcal{M}_{RAD}$, and impute the representation of the missing modality b using the trained $\mathcal{U}_{\mathcal{LIT}}$. The imputed representations are then used to fine-tune $\mathcal{M}_{IID}$ under the supervision of the corresponding disease label, while fixing $\mathcal{M}_{RAD}$. The workflow of the training stage is summarized in Algorithm 1.

In the inference stage, we assume that each subject only has the scan of modality a. We first obtain the global representation f_a using the trained $\mathcal{M}_{RAD}$, and then perform the diagnosis using the student branch $\mathcal{S}$ of $\mathcal{M}_{IID}$. Due to the knowledge distillation from the teacher branch $\mathcal{T}$ to $\mathcal{S}$, the student branch $\mathcal{S}$ is able to implicitly discover the multi-modality knowledge (*i.e.*, the global representation of modality a and the imputed inter-modality relevant representation of modality b), and hence produce accurate diagnosis.

IV. Experiments and Results

A. Dataset

T1-weighted sMRI and FDG PET scans from the ADNI database (adni.loni.usc.edu), including ADNI-1, ADNI-2, and ADNI-GO [8], were used for this study. All scans were acquired from 1248 subjects, including 347 AD patients, 417 cognitively normal (CN) subjects, and 484 mild cognitive

Algorithm 1 Optimization for the Proposed ***DFTD*** Framework

Input:

Modlity-complete training samples $\boldsymbol{x_c}$

Modlity-incomplete training samples $\boldsymbol{x_{in}}$

Diagnosis labels of all training samples $\boldsymbol{y}$

Pre-defined Learning rate l_r, Maximum epoch E, K, α and β

Output:

Updated parameters of $\mathcal{M}_{RAD}$ and $\mathcal{M}_{IID}$

Randomly Initialize the weights of $\mathcal{M}_{RAD}$ and $\mathcal{M}_{IID}$
repeat
  Choose a batch of modality-complete data from $\boldsymbol{x_c}$
  Calculate c_a and c_b through $\mathbb{F}^a \circ \mathbb{H}^a_c$ and $\mathbb{F}^b \circ \mathbb{H}^b_c$
  Update $\mathbb{F}^a$, $\mathbb{H}^a_c$, $\mathbb{H}^b_c$, $\mathbb{S}^a_c$ and $\mathbb{S}^b_c$ by formula 6
  Calculate s_a and s_b through $\mathbb{F}^a \circ \mathbb{H}^a_s$ and $\mathbb{F}^b \circ \mathbb{H}^b_s$
  Update $\mathbb{F}^a$, $\mathbb{H}^a_s$, $\mathbb{H}^b_s$, $\mathbb{S}^b_s$, $\mathbb{D}^a_s$ and $\mathbb{D}^b_s$ by formula 14
  Calculate c'_b based on c_a through $\mathcal{U}_{\mathcal{LIT}}$
  Update $\mathbb{E}$ and $\mathbb{D}$ by formula 15
  Calculate j^t_i through formula 17
  Update the $\mathcal{M}_{IID}$ by formula 22
until iter reaches its desired maximum
repeat
  Choose a batch of modality-incomplete samples from $\boldsymbol{x_{in}}$
  Calculate c_a through $\mathbb{F}^a \circ \mathbb{H}^a_c$
  Calculate c'_b based on c_a through $\mathcal{U}_{\mathcal{LIT}}$
  Using zero-padding for f_b and calculate j^t_i through formula 17
  Update the $\mathcal{M}_{IID}$ by formula 22
until iter reaches its desired maximum

impairment (MCI) individuals. These subjects were selected solely based on their diagnostic labels, and other detailed criteria like sex, age, slice thickness, and manufacturer were not specifically considered. For the AD-CN classification task, subjects that have label 'AD' or 'CN' at any time point were selected. For each subject, scans at all available time points, including the baseline and the following time points were utilized to train the proposed DFTD. As for the test set, only the baseline image of each subject (*i.e.*, 1 image per subject) was included to avoid misleading the model's prediction. For the MCI-to-AD prediction task, the selected subjects should have disease label 'MCI' at baseline and label 'AD' or 'MCI' after 18 months. Those MCI subjects who would progress to AD within 18 months were regarded as progressive MCI (pMCI) cases and those who would not progress to AD within 18 months were considered as static MCI (sMCI) cases. Notably, only the baseline data of each subject (*i.e.*, one image per subject) was used for both training and test. The demography of this dataset is shown in Table II.

B. Preprocessing

For MRI data, we downloaded original scans from the ADNI database and performed the following pre-processing procedures. First, we applied the Anterior Commissure (AC)-Posterior Commissure (PC) reorientation to all scans and corrected their small motions. Then, we applied the non-parametric non-uniform intensity normalization (N3) to each scan. Next, we employed the MINC program mritotal to transform each original scan to the MNI305 atlas, followed by skull stripping and cerebellum removal. All these procedures were conducted by using the FreeSurfer software [46].

For PET data, we downloaded the preprocessed scans with the third preprocessing type of 'CO-REG, AVG, STANDARDIZED IMAGE AND VOXEL SIZE' (http://adni.loni.usc.edu/methods/pet-analysis-method/pet-analysis). Specifically, the preprocessing procedures provided by the ADNI database consist of (1) smoothing, (2) calculating and applying coregistration, (3) averaging frames, (4) computing AC-PC orient baseline, (5) standardizing to the baseline, and (6) intensity normalization. Besides, we also aligned each downloaded PET data to the space of the corresponding MRI data using the SPM12 toolbox [47]. Finally, both MRI and PET data were resized to 256 × 256 × 256 voxels and the voxel values are normalized to the range between 0 and 1.

C. Experimental Settings

We evaluated the proposed DFTD framework on two tasks: AD diagnosis (AD-CN classification) and MCI conversion prediction (pMCI-sMCI classification). We utilized the five-fold cross-validation for evaluation. (1) We divided all data into five folds at the subject level. Specifically, we first divided the data of each class (*i.e.*, AD, CN, sMCI, or pMCI) into a modality-complete group and a modality-incomplete group, and then randomly sampled the scans from each group to form five folds, in which the number of scans and the distribution of categories and modality (in)completeness are nearly balanced. (2) We performed five complete training sessions in total. In each training session, one fold was used for testing and the remaining four folds were used for training. In each training session, the validation set was constructed by selecting randomly 20% modality-complete subjects from the training set, and hence has non-overlap with the test set. The validation set was used to tune hyper-parameters and monitor the training process to prevent over-fitting. Evaluation metrics we used include the area under receiver operating characteristic (AUC), average precision (AP), sensitivity (SEN), specificity (SPE), and Matthews correlation coefficient (MCC).

In the region-aware disentanglement module $\mathcal{M}_{RAD}$, each of the four encoders $\mathbb{F}^a \circ \mathbb{H}^a_c$, $\mathbb{F}^a \circ \mathbb{H}^a_s$, $\mathbb{F}^b \circ \mathbb{H}^b_c$ and $\mathbb{F}^b \circ \mathbb{H}^b_s$ consists of five convolutional layers with 16, 32, 64, 64 and 64 channels. All layers are followed by 3× 3×3 max-pooling layers with a stride of 2, as well as the instance normalization and $ReLU$ activation. The $\mathbb{F}^a$ and $\mathbb{F}^b$ are composed of the first three convolutional layers and are used to extract the intermediate feature maps m_a and m_b. The discriminators $\mathbb{S}^a_c$, $\mathbb{S}^b_c$, $\mathbb{S}^a_s$, $\mathbb{S}^b_s$, $\mathbb{D}^a_s$, $\mathbb{D}^b_s$ are all composed of three fully-connected layers with $ReLU$ activation, and the output dimension of each layer is 256, 64 and 1, respectively. In the imputation-induced distillation module $\mathcal{M}_{IID}$, both the student branch $\mathcal{S}$ and the teacher branch $\mathcal{T}$ consist of four fully-connected layers, with output dimension of 512, 256, 64 and 1, respectively. There are three fully connected layers followed by $ReLU$ activation in the encoder $\mathbb{E}$ and the decoder $\mathbb{D}$ of the lateral inter-modality transition unit $\mathcal{U}_{\mathcal{LIT}}$, respectively. Each layer performs down-sampling or up-sampling with ratio of 2 to

TABLE II
DEMOGRAPHIC INFORMATION OF 1248 SUBJECTS. THE RIGHT UPPER MARK $(\cdot)$ REPRESENTS THE NUMBER OF SCANS AVAILABLE

Class	Number of Subjects	Modality (sMRI/PET)	Gender (F/M)	Age (Mean ± Std)
AD	$347^{(1136)}$	$347^{(871)}/146^{(265)}$	195/152	74.8 ± 9.1
CN	$417^{(1235)}$	$417^{(988)}/185^{(247)}$	199/218	73.8 ± 8.2
pMCI	$179^{(568)}$	$179^{(438)}/87^{(130)}$	74/105	75.2 ± 6.1
sMCI	$305^{(952)}$	$305^{(753)}/135^{(199)}$	127/178	75.7 ± 5.8

its input representations. All sub-models were implemented under Pytorch and optimized using the Adam optimizer with an initial learning rate of $1e^{-3}$. The batch size is set to 8 and the training stops when the model fails to improve performance on the validation set over 50 epochs.

D. Comparing to Existing Methods

We compared our DFTD framework against six methods, including (1) a conventional deep learning framework for AD diagnosis using only sMRI scans [48]; (2) a modality fusion network for disease diagnosis using subjects with complete two-modality data [49]; (3) a subspace learning method which can utilize all avalible data to conduct diagnosis [11]; (4) two GAN-based image imputation methods which impute missing PET scans first and then conduct multi-modality diagnosis [30], [31]; and (5) a knowledge distillation-based method [13]. We implemented these methods using their default hyper-parameters and tested them on our dataset.

The performance of each method is recorded in Table III. For AD-CN classification, our DFTD framework achieves an AUC of 96.85%, an AP of 90.23%, a SEN of 91.73%, a SPE of 93.69% and a MCC of 84.21%, ranking the first in all four metrics. Specifically, the method using only sMRI data [48] achieves the worst performance among these methods, suggesting that using multi-modality data [11], [30], [31], [49] could explore more information to help improve the performance of AD diagnosis. Comparing to the subspace learning-based method [11], our DFTD achieves an AUC improvement of 4.68 percent points. It implies that imputing missing representations could increase the multiformity information of the features and is effective in boosting the classification performance. Besides, comparing to two imputation-based methods, the AUC of the proposed framework achieves an improvement of more than 1.77 percent points, suggesting that disentangling the inter-modality relevant representations before conducting synthesis and performing imputation-induced distillation can further improve diagnosis performance. For MCI conversion prediction, the proposed DFTD framework achieves an AUC of 83.81%, an AP of 77.26%, a SEN of 72.97%, a SPE of 79.02% and a MCC of 51.27%, all ranking the best among competing methods. Table III shows that the performance of multi-modality based methods [11], [30], [31], [49] remains better than that of single-modality based method [48] on this task. Meanwhile, the performance of methods which involve modality-missing images is better than that of methods using only modality-complete data, and image imputation based methods further improve subspace learning-based methods. What's more, the proposed DFTD framework outperforms the other methods with a substantial improvement on all metrics, demonstrating the ideas of region-aware disentanglement and imputation-induced distillation is especially effective on this challenging task.

E. Effectiveness of Disentanglement

To evaluate the effectiveness of the proposed region-aware disentanglement module $\mathcal{M}_{RAD}$, we chose the AD-CN classification task as a case study and compared the performance of each DFTD variant. In Model I, Model II, and Model III, we conducted missing modality imputation using the lateral inter-modality transition unit $\mathcal{U}_{\mathcal{LIT}}$ based on the global representations, intra-modality specific representations, and inter-modality relevant representations of the existing modality, respectively. Specifically, in Model I, the global representations (*i.e.*, the intra-modality specific representations and inter-modality relevant representations) of the existing modality are needed for missing modality imputation. In Model II, inter-modality relevant representations should be calculated, before obtaining the intra-modality specific representations which are needed for missing modality imputation. Therefore, all loss functions, including L_{MI}^{rel}, L_{MI}^{spe}, $L_{\mathbb{S}_c^a}^i$, $L_{\mathbb{S}_c^b}^i$, $L_{\mathbb{S}_s^a}^i$, $L_{\mathbb{S}_s^b}^i$, L_{adv}^a and L_{adv}^b, are used in both models. In Model III, only the inter-modality relevant representations need to be calculated. Thus, only L_{MI}^{rel}, $L_{\mathbb{S}_c^a}^i$, $L_{\mathbb{S}_c^b}^i$, $L_{\mathbb{S}_s^a}^i$, and $L_{\mathbb{S}_s^b}^i$ are utilized.

The performance of these models is shown in Table IV. The performance of Model II is obviously worse than others, demonstrating that using the intra-modality specific representations of the existing modality to impute the representations of missing modality is not effective. This is not surprising, since there is little modality relevance in these intra-modality specific representations. Model I performs better than Model II, due to the modality relevance that exists in the global representation. However, the performance of Model I is worse than Model III, probably since the modality-specific features of the global representations introduce biased information to the imputation process. Besides, the proposed DFTD further improves Model III with an AUC increase of 0.87 percent points, verifying the effectiveness of taking disease-related regions into account when conducting disentanglement.

F. Ablation Study for $\mathcal{M}_{IID}$

In the imputation-induced distillation module $\mathcal{M}_{IID}$, a lateral inter-modality transition unit $\mathcal{U}_{\mathcal{LIT}}$ is inserted between each layer of the student branch $\mathcal{S}$ and integrated teacher branch $\mathcal{T}$ to impute the representations of the missing modality. To validate its effectiveness, we discard the $\mathcal{U}_{\mathcal{LIT}}$ unit and perform knowledge distillation only for those modality-complete sMRI-PET pairs based on their global feature representations f_a and f_b, as implemented by Model 1. During this process, the loss function L_{LIT} is not utilized for missing modality imputation. Only $\mathcal{L}_{CE}(f_a, y)$, $\mathcal{L}_{Dis}^F(o^t, o^s)$, $\mathcal{L}_{CE}(j_0^t, y)$, and $\mathcal{L}_{Dis}^P(\hat{p}_s, \hat{p}_t; K)$ are utilized for multi-modality knowledge distillation and disease diagnosis. Meanwhile, we conducted an ablation study to evaluate the

TABLE III

RESULTS ((MEAN±STD)/%) OF SEVEN METHODS IN AD-CN CLASSIFICATION AND MCI-TO-AD PREDICTION ON OUR DATASET. NOTE THAT '±STD' REPRESENTS THE EMPIRICAL STANDARD DEVIATION ACROSS THE 5 FOLDS

Task	Metric	Method						
		Feng *et al.* [48]	Shi *et al.* [49]	Zhou *et al.* [11]	Pan *et al.* [30]	Pan *et al.* [31]	Guan *et al.* [13]	**Proposed**
AD vs. CN	AUC	90.22±0.64	92.25±0.76	91.98±1.17	94.22±0.57	95.16±0.38	95.08±0.36	**96.85±0.21**
	AP	85.77±0.56	85.34±0.69	85.96±0.75	88.62±0.70	89.19±0.65	89.71±0.41	**90.23±0.25**
	SEN	85.21±0.41	85.83±0.92	87.08±0.87	87.92±0.56	89.25±0.33	90.39±0.43	**91.73±0.38**
	SPE	86.52±0.51	88.86±0.97	89.91±1.02	91.26±0.68	92.86±0.29	92.18±0.34	**93.69±0.30**
	MCC	75.06±1.05	77.22±0.98	76.43±0.89	79.82±0.81	81.09±0.92	82.28±0.93	**84.21±0.62**
	Param.	1.8M	0.7M	0.1M	4.6M	4.8M	1.2M	1.5M
	T_{train}	6.3h	4.2h	2.6h	10.2h	12.5h	6.1h	7.2h
	$T_{inf.}$	8.2ms	6.1ms	4.9ms	22.6ms	25.7ms	9.3ms	7.8ms
pMCI vs. sMCI	AUC	77.16±0.61	77.46±0.82	78.21±0.95	80.18±0.69	80.85±0.49	79.92±0.51	**83.81±0.25**
	AP	72.17±0.48	73.04±0.96	73.75±0.89	74.28±0.43	75.18±0.46	75.42±0.41	**77.26±0.21**
	SEN	68.50±0.58	68.61±0.74	69.33±0.79	71.72±0.45	71.50±0.58	70.86±0.55	**72.97±0.21**
	SPE	72.39±0.79	72.11±0.92	73.46±0.92	74.37±0.42	75.89±0.62	77.81±0.53	**79.02±0.41**
	MCC	41.89±2.75	43.77±2.96	44.38±2.91	47.15±2.59	47.62±2.67	49.75±2.42	**51.27±2.03**
	Param.	1.8M	0.7M	0.1M	4.6M	4.8M	1.2M	1.5M
	T_{train}	6.4h	4.6h	2.1h	10.4h	12.4h	5.8h	7.3h
	$T_{inf.}$	9.1ms	7.3ms	6.4ms	28.3ms	30.6ms	10.3ms	8.5ms

TABLE IV

PERFORMANCES OF EACH VARIANT OF THE PROPOSED M_{RAD}. ISR: INTRA-MODALITY SPECIFIC REPRESENTATIONS. IRR: INTER-MODALITY RELEVANT REPRESENTATIONS. RA: REGION-AWARE DISENTANGLEMENT. NOTE THAT '±STD' REPRESENTS THE EMPIRICAL STANDARD DEVIATION ACROSS THE 5 FOLDS

M	$\mathcal{M}_{RAD}$			AUC(%)	AP(%)	SEN(%)	SPE(%)
	ISR	IRR	RA				
I	√	√		93.97±0.26	87.75±0.39	88.64±0.40	90.59±0.36
II	√			82.81±0.35	80.17±0.47	76.83±0.32	80.90±0.38
III		√		95.98±0.31	89.16±0.38	90.35±0.43	92.73±0.29
Ours		√	√	**96.85±0.21**	**90.23±0.25**	**91.73±0.38**	**93.69±0.30**

TABLE V

PERFORMANCES OF EACH COMPONENT OF THE PROPOSED M_{IID}. FD: FEATURE DISTILLATION. SD: SOFT LABEL DISTILLATION. NOTE THAT '±STD' REPRESENTS THE EMPIRICAL STANDARD DEVIATION ACROSS THE 5 FOLDS

M	$\mathcal{M}_{IID}$			AUC(%)	AP(%)	SEN(%)	SPE(%)
	$\mathcal{U}_{LIT}$	FD	SD				
1		√	√	94.04±0.31	87.47±0.56	88.38±0.35	91.72±0.34
2	√	√		96.13±0.22	89.59±0.47	90.81±0.32	92.91±0.26
3	√		√	95.83±0.25	88.91±0.39	90.53±0.36	92.15±0.27
Ours	√	√	√	**96.85±0.21**	**90.23±0.25**	**91.73±0.38**	**93.69±0.30**

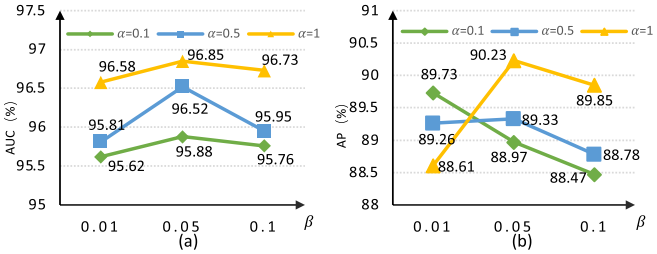

Fig. 2. Plots of model performance (AUC and AP) on validation set versus settings of hyper-parameters α and β.

TABLE VI

PERFORMANCES OF THE PROPOSED FRAMEWORK ON AD-CN CLASSIFICATION WITH DIFFERENT HYPER-PARAMETER VALUES OF K. NOTE THAT '±STD' REPRESENTS THE EMPIRICAL STANDARD DEVIATION ACROSS THE 5 FOLDS

Values		AUC(%)	AP(%)	SEN(%)	SPE(%)
K	0.1	96.01±0.37	89.43±0.35	90.67±0.29	92.53±0.34
	0.5	96.58±0.21	89.83±0.40	91.02±0.26	93.28±0.23
	1	**96.85±0.21**	**90.23±0.25**	**91.73±0.38**	**93.69±0.30**

effectiveness of feature distillation and soft label distillation, implemented by Model 2 and Model 3, respectively. In Model 2, only L_{LIT}, $\mathcal{L}_{CE}(f_a, y)$, $\mathcal{L}_{CE}(j_0^t, y)$, and $\mathcal{L}_{Dis}^F(o^t, o^s)$ are utilized for representation imputation, feature distillation and disease diagnosis. In Model 3, only the soft label distillation is performed, where L_{LIT}, $\mathcal{L}_{CE}(f_a, y)$, $\mathcal{L}_{CE}(j_0^t, y)$, and $\mathcal{L}_{Dis}^P(\hat{p}_s, \hat{p}_t; K)$ are utilized. The performance of these variants is shown in Table V. It reveals that the proposed DFTD framework outperforms Model 1 with a significant AUC improvement of 2.81 percent points, demonstrating that imputing missing features for those modality-incomplete samples is helpful for disease diagnosis. Moreover, comparing to Model 2 and Model 3, our DFTD framework achieves the best performance when utilizing both feature distillation and soft label distillation to perform multi-modality knowledge transfer.

G. Hyper-Parameter Settings

When optimizing the proposed imputation-induced distillation module $\mathcal{M}_{IID}$, α and β are two hyper-parameters which jointly balance the importance of the feature distillation loss and the soft label distillation loss (see Eq. 22). We empirically adjust their values simultaneously and performed the AD-CN classification task with $\alpha \in \{0.1, 0.5, 1\}$ and $\beta \in \{0.01, 0.05, 0.1\}$. Fig. 2 shows the changing curve of AUC and

TABLE VII
PERFORMANCES OF OUR DFTD FRAMEWORK WITH DIFFERENT BACKBONES ON AD-CN CLASSIFICATION TASK. NOTE THAT '±STD' REPRESENTS THE EMPIRICAL STANDARD DEVIATION ACROSS THE 5 FOLDS

Architecture	AUC(%)	AP(%)	SEN(%)	SPE(%)
w/o $\mathcal{M}_{RAD}$	93.97±0.26	87.75±0.39	88.64±0.40	90.59±0.36
w $\mathcal{M}_{RAD}$ +3D ResNet-10	96.46±0.39	89.71±0.34	90.37±0.43	93.05±0.37
w $\mathcal{M}_{RAD}$ +3D ShuffleNet-V1	96.26±0.28	88.95±0.38	89.62±0.35	92.73±0.32
w $\mathcal{M}_{RAD}$ +5Conv(proposed)	**96.85±0.21**	**90.23±0.25**	**91.73±0.38**	**93.69±0.30**

TABLE VIII
THE P-VALUES OF THE STUDENT T-TEST PERFORMED ON AD-CN CLASSIFICATION TASK. THE SIGNIFICANCE LEVEL IS SET TO $\alpha' = 0.00167$ AFTER BONFERRONI CORRECTION

Methods	p-value (AUC)	p-value (AP)	p-value (SEN)	p-value (SPE)	p-value (MCC)
DFTD vs. Feng *et al.* [48]	$4.76E-6$	$6.97E-6$	$5.45E-9$	$6.60E-8$	$1.38E-6$
DFTD vs. Shi *et al.* [49]	$8.18E-5$	$2.35E-5$	$2.79E-5$	$1.69E-4$	$3.91E-6$
DFTD vs. Zhou *et al.* [11]	$5.85E-4$	$8.07E-5$	$6.28E-5$	$6.82E-4$	$7.30E-7$
DFTD vs. Pan *et al.* [30]	$1.86E-4$	$4.69E-4$	$4.40E-6$	$4.95E-4$	$1.77E-5$
DFTD vs. Pan *et al.* [31]	$1.03E-4$	$1.96E-3$	$4.77E-6$	$2.15E-3$	$4.06E-4$
DFTD vs. Guan *et al.* [13]	$5.07E-5$	$4.80E-4$	$8.41E-4$	$7.89E-5$	$6.26E-3$

AP with varying values of α and β, keeping other settings fixed. We find that DFTD is robust on different values of α and β, but achieves the best performance on both metrics when setting α to 1 and β to 0.05, respectively. What's more, with the determined values of α and β, we further discuss different values $\{0.1, 0.5, 1\}$ of the hyperparameters K, which is a temperature scaling parameter in the knowledge distillation process (see Eq. 18 and Eq. 19). Table VI shows that when setting K to the value of 1, the proposed DFTD framework obtains an AUC of 96.85%, an AP of 90.23%, a SEN of 91.73% and a SPE of 93.69%, achieving the best performance on all four metrics.

H. Backbone Architecture of $\mathcal{M}_{RAD}$

The backbone we used consists of five convolutional layers. Besides that, we attempted to use 3D ResNet-10 [50] and 3D ShuffleNet-V1 [51], respectively, as the backbone of $\mathcal{M}_{RAD}$. Experimental results were listed in Table VII. It shows that, no matter which backbone is used, the model with $\mathcal{M}_{RAD}$ substantially outperforms the one without $\mathcal{M}_{RAD}$, indicating that the proposed $\mathcal{M}_{RAD}$ is effective with different backbones. Besides, our $\mathcal{M}_{RAD}$ achieves slightly better performance when using the backbone with five convolutional layers.

I. Statistical Analysis

We adopted the Student t-test to determine whether the performance gain achieved by the proposed DFTD framework over the competing methods is statistically significant. We assumed that the AUC/AP/SEN/SPE/MCC values of DFTD and each competing method are random variables X_1 and X_2, respectively, each following a Gaussian distribution, *i.e.*, $X_1 \sim N(\mu_1, {\sigma_1}^2)$, $X_2 \sim N(\mu_2, {\sigma_2}^2)$. The difference between X_1 and X_2 is defined as $\Delta = X_1 - X_2$. The hypotheses to be tested are H_0: $\mu_\Delta \leq 0$ versus H_1: $\mu_\Delta > 0$. To enhance the rigor of this statistical testing and control the overall false positive rate, we applied the Bonferroni correction to adjust the significance level. To achieve this, we divided the original level of significance ($\alpha = 0.05$) by the total number of tests performed ($m = 6 \times 5$), which yielded a new significance threshold of $\alpha' = \alpha/m = 0.00167$. Our analysis, as presented in Table VIII, indicates that for the vast majority of comparisons with competing methods, the calculated p-values were below the adjusted significance level of $\alpha' = 0.00167$. As a result, we were able to reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1), indicating that the DFTD framework performed significantly better than the other competing methods in terms of five evaluation metrics.

J. Embeddings Visualization

To illustrate the effectiveness of the region-aware disentanglement module $\mathcal{M}_{RAD}$, we visualized three groups of representations on the validation set in Fig. 3, including the inter-modality-relevant representations from a randomly initialized $\mathcal{M}_{RAD}$ (Fig. 3(a)), the inter-modality-relevant representations from a well-trained $\mathcal{M}_{RAD}$ (Fig. 3(b)) and the disentangled inter-modality-relevant representations and intra-modality specific representations (Fig. 3(c)). For each paradigm, the "embedding" is a 512-dimensional feature vector. The embedding representations of each sample are projected into a two-dimensional feature space using the principal component analysis (PCA) for the visualization purpose. As shown in Fig. 3(a), although the distributions of the inter-modality-relevant representations of sMRI and PET from a randomly initialized $\mathcal{M}_{RAD}$ overlap in some parts, the overall distance between two distributions are large and the two-modality embeddings are not relevant closely. By contrast, the distributions of the two inter-modality-relevant representations for sMRI and PET in Fig. 3(b) are very similar, which meets the expectation that the well-learned inter-modality representations c_a and c_b are highly relevant with each other. Therefore, due to the high relevance and close distribution distance, the feature translation becomes possible. Meanwhile, the classification boundaries in Fig. 3(b) remain separable while the relevance between inter-modalities features improves a lot, which demonstrates the effectiveness of our $\mathcal{M}_{RAD}$. In Fig. 3(c), the inter-modality-relevant representations and intra-modality-specific representations of sMRI or PET are separable while the inter-modality-relevant representations of two modalities are relevant with each other, which further demonstrates our $\mathcal{M}_{RAD}$ can successfully disentangle two kinds of representations from the original images.

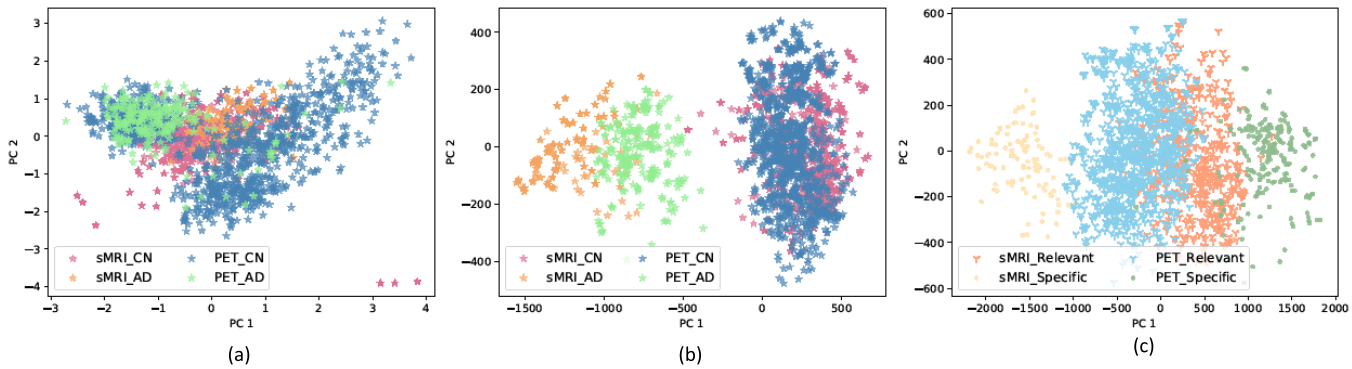

Fig. 3. Embeddings Visualization. (a) Inter-modality-relevant representations from a randomly initialized $\mathcal{M}_{\text{RAD}}$. (b) Inter-modality-relevant representations from a well-trained $\mathcal{M}_{\text{RAD}}$. (c) Disentangled inter-modality-relevant representations and intra-modality specific representations from a well-trained $\mathcal{M}_{\text{RAD}}$.

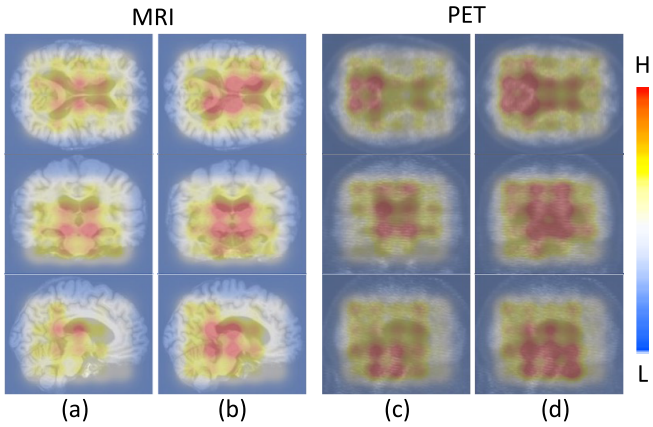

Fig. 4. Visualization of salience maps (overlaid on the original images *x*) for sMRI and PET, with each element denoting the salience of a specific brain region.

K. Visualization of Salient Regions

We visualized the salience of each brain region in Fig. 4. In Fig. 4(a), we first visualized the salience of each brain region for the MRI modality according to the region weight learned by our DFTD framework. Specifically, in the proposed region-aware disentanglement module $\mathcal{M}_{RAD}$, we split each feature map into $\mathrm{N} = 4 \times 4 \times 4$ non-overlapping regions. We assigned each region a learnable weight ω_{ai}, aiming to incorporate the brain region salience into the disentanglement process. The weight ω_{ai} of each region is a learnable parameter that participates in the optimization of the region-aware disentanglement module $\mathcal{M}_{RAD}$. It can be updated automatically by minimizing Eq.3 and Eq.7 when optimizing $\mathcal{M}_{RAD}$. Thus, after the model converges at the end of training stage, we can obtain the optimized weight of each region. Then, we take each region weight as a coefficient, and utilize a Gaussian kernel function to draw a Gaussian distribution based on this coefficient. We utilized the Gaussian distribution map of each coefficient as a mask and covered it on the corresponding brain region of a MRI image, aiming to show the salience of this brain region. The high (H) and low (L) salience was denoted by red and blue, respectively. To prove the reliability of the regional weights learned in the $\mathcal{M}_{RAD}$ module, we further visualized the discrimination ability of each region in Fig. 4(b) by considering the contribution of each region to the diagnosis separately. For the i_{th} region, instead of setting its weight ω_{ai} as a learnable parameter, we set it to a fixed value 1, and set the weights of other regions to 0. This ensures the disentangled inter-modality relevant representation c_a and intra-modality specific representation s_a are mainly related to this region. Then, we optimized the modules $\mathcal{M}_{RAD}$ and $\mathcal{M}_{IID}$ of our DFTD framework, and used the AUC of classification as the diagnosis contribution of the i_{th} region. We normalize the contribution of all regions and used them as coefficients to draw Gaussian distribution maps. It reveals that the most relative regions in Fig. 4(a) are also the most discriminative regions in Fig. 4(b). It suggests that the proposed DFTD framework can automatically locate the disease-related regions in the brain. For the PET modality, the same operations were performed for each region. The salient regions determined by the proposed $\mathcal{M}_{RAD}$ module and the diagnosis contribution of each region was visualized in Fig. 4(c) and Fig. 4(d), respectively. It shows that the locations of the most discriminative regions are partially overlapped in MRI and PET modalities. On one hand, the existence of non-overlapping regions suggests that both modalities can provide complementary information for brain disease diagnosis. On the other hand, the difference in discriminatory abilities suggests that PET provides a way to represent features that is different from MRI. Thus, conducting missing modality imputation based on the inter-modality relevance between two modalities can contribute to improving the diagnosis performance.

V. Discussions

A. Advantages

The proposed DFTD framework has three distinct advantages over existing methods. First, traditional multi-modality methods [9], [10] discard the subjects with incomplete scans, leading to a reduced number of training cases and consequent performance degradation. While the proposed DFTD utilizes all scans to perform AD and MCI diagnosis, and thus achieves better performance by making full use of multi-modality knowledge from available data. Second, subspace learning-based methods [11] use different learning schemes for modality-complete data and modality-incomplete data, and

cannot learn the modality correlation due to the missing modality. On the contrary, the proposed DFTD depicts the relevance among different modalities more explicitly, and can explore more abundant correlations between the existing and missing modalities. Third, existing modality imputation-based methods [30], [31] first generate the whole missing-modality image, and then extract features from both existing images and generated images for classification. Hence, these methods inevitably suffer from complex computation and biased information in image generation. Moreover, in the inference phase, these methods [30], [31] cannot perform disease diagnosis before the completion of modality imputation, leading to increased inference time and limited clinical value. By contrast, the proposed DFTD only needs to impute the modality-relevant representation for the missing modality, which reduces the computation complexity and avoids redundant information. In the inference phase, our DFTD can perform multi-modality diagnosis of a subject by using only single-modality data, resulting in faster inference speed and better clinical value.

B. Limitations

Meanwhile, the proposed DFTD framework also has three major limitations. First, the data used for this study are from the ADNI database and are preprocessed after data collection. However, the data collected in clinical practice may differ a lot in quality from the ADNI data we used, which may degrade the model performance to some extent. Thus, our future work is to explore data harmonization / domain adaptation techniques and capture unified features regardless of image quality. Second, comparing to the dataset used for computer vision research such as ImageNet, the dataset used for this study is relatively small. As a deep learning model requires a myriad number of data for training, our DFTD relies its performance heavily on the number of training data. In our future work, we will explore how to expand the data distribution with the inspiration from augmentation-based few-shot learning techniques. Third, the disease-related regions were determined by the region-aware disentanglement module in a data-driven manner, without considering the related findings reported in the literature. Another future work is to collaborate with clinical experts and incorporate the clinical prior knowledge into the decision process of the proposed framework.

C. Transfer to Other Tasks

Although the proposed DFTD framework is designed for the diagnosis of neural degenerative diseases, the principles behind it are generic and can be transferred to other multi-modality medical image analysis tasks. Taking brain tumor segmentation from multi-sequence MRI for example, the modality missing issue is usually encountered, and we can easily extend our DFTD framework to accomplish this task. Let the available sequence be denoted by a and the missing sequence be denoted by b. We first impute the missing representation of sequence b and then fuse the representations of both sequences for tumor segmentation. When generating the representation of sequence b, the idea of disentangling the inter-sequence relevant information from sequence a and using it alone for imputation to avoid redundant information is rational. Thus, the first region-aware disentanglement module $\mathcal{M}_{RAD}$ of the proposed DFTD can be directly adopted to extract the disentangled representation. When fusing multi-sequence representations for tumor segmentation, the idea of distilling the multi-sequence knowledge from both the existing representation and imputed representation to single-sequence network is also rational. Therefore, in the second imputation-induced distillation module $\mathcal{M}_{IID}$, we need (1) replace the student and teacher branches with two encoder-decoder based segmentation networks; and (2) replace the cross-entropy classification loss with a Dice loss to perform segmentation.

VI. Conclusion

To deal with the modality-missing problem encountered in multi-modality AD diagnosis, we propose the DFTD framework in this paper, which consists of a region-aware disentanglement module and an imputation-induced distillation module. The region-aware disentanglement module disentangles inter-modality relevant representations and intra-modality specific representations with emphasis in diagnosis-related image regions. The imputation-induced distillation module performs multi-modality knowledge transfer and incorporates a lateral inter-modality transition unit to impute the representation of missing modality. We evaluated this framework against five existing methods on an ADNI dataset with 1248 subjects. Our results show that the proposed DFTD framework achieves the best performance in both AD-NC classification and MCI-AD conversion prediction, suggesting DFTD is effective for tackling the modality-missing issue. In our future work, we will extend the proposed DFTD to an approach that could deal with more modalities.

References

[1] M. Baumgart, H. M. Snyder, M. C. Carrillo, S. Fazio, H. Kim, and H. Johns, "Summary of the evidence on modifiable risk factors for cognitive decline and dementia: A population-based perspective," *Alzheimer's Dementia*, vol. 11, no. 6, pp. 718–726, Jun. 2015.

[2] R. Petersen, J. Stevens, M. Ganguli, M. Tangalos, J. Cummings, and S. DeKosky, "Early detection of dementia: Mild cognitive impairment (an evidence-based review)," *Neurology*, vol. 56, no. 9, pp. 1133–1142, 2001.

[3] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[4] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.

[5] K. Ritter, J. Schumacher, M. Weygandt, R. Buchert, C. Allefeld, and J. Haynes, "Multimodal prediction of conversion to Alzheimer's disease based on incomplete biomarkers," *Alzheimer's Dementia, Diagnosis, Assessment Disease Monit.*, vol. 1, no. 2, pp. 206–215, Jun. 2015.

[6] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, and J. Ye, "Bi-level multi-source learning for heterogeneous block-wise missing data," *NeuroImage*, vol. 102, pp. 192–206, Nov. 2014.

[7] R. Ossenkoppele et al., "Associations between tau, aβ, and cortical thickness with cognition in Alzheimer disease," *Neurology*, vol. 92, no. 6, pp. e601–e612, Feb. 2019.

[8] C. R. Jack et al., "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J. Magn. Reson. Imag.*, vol. 27, no. 4, pp. 685–691, 2008.

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

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

[11] T. Zhou, M. Liu, K. Thung, and D. Shen, "Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2411–2422, Oct. 2019.

[12] T. Zhou, M. Liu, H. Fu, J. Wang, J. Shen, and L. Shao, "Deep multi-modal latent representation learning for automated dementia diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, vol. 11766, 2019, pp. 629–638.

[13] H. Guan, C. Wang, and D. Tao, "MRI-based Alzheimer's disease prediction via distilling the knowledge in multi-modal data," *NeuroImage*, vol. 244, Dec. 2021, Art. no. 118586.

[14] Q. Wang, L. Zhan, P. Thompson, and J. Zhou, "Multimodal learning with incomplete modalities by knowledge distillation," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 1828–1838.

[15] T. Zhou et al., "Inter-modality dependence induced data recovery for MCI conversion prediction," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, vol. 11767, 2019, pp. 186–195.

[16] Y. Liu et al., "Incomplete multi-modal representation learning for Alzheimer's disease diagnosis," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101953.

[17] B. Cao, H. Zhang, N. Wang, X. Gao, and D. Shen, "Auto-GAN: Self-supervised collaborative learning for medical image synthesis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10486–10493.

[18] Y. Pan, M. Liu, Y. Xia, and D. Shen, "Disease-image specific generative adversarial network for brain disease diagnosis with incomplete multi-modal neuroimages," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 137–145.

[19] W. Guo, H. Huang, X. Kong, and R. He, "Learning disentangled representation for cross-modal retrieval with deep mutual information estimation," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1712–1720.

[20] Y. Lu et al., "Cross-modality person re-identification with shared-specific feature transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13379–13389.

[21] C. Chen, Q. Dou, Y. Jin, Q. Liu, and P. A. Heng, "Learning with privileged multimodal knowledge for unimodal segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 3, pp. 621–632, Mar. 2022.

[22] Y. Wang et al., "ACN: Adversarial co-training network for brain tumor segmentation with missing modalities," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 410–420.

[23] T. Zhou, S. Canu, P. Vera, and S. Ruan, "Latent correlation representation learning for brain tumor segmentation with missing MRI modalities," *IEEE Trans. Image Process.*, vol. 30, pp. 4263–4274, 2021.

[24] Q. Yang, X. Guo, Z. Chen, P. Y. M. Woo, and Y. Yuan, "D2-Net: Dual disentanglement network for brain tumor segmentation with missing modalities," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2953–2964, Oct. 2022.

[25] R. Gao et al., "Lung cancer risk estimation with incomplete data: A joint missing imputation perspective," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 647–656.

[26] M. Hamghalam, A. F. Frangi, B. Lei, and A. L. Simpson, "Modality completion via Gaussian process prior variational autoencoders for multi-modal glioma segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 442–452.

[27] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[28] D. Lee, J. Kim, W.-J. Moon, and J. C. Ye, "CollaGAN: Collaborative GAN for missing image data imputation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2487–2496.

[29] Y. Pan, M. Liu, C. Lian, T. Zhou, Y. Xia, and D. Shen, "Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2018, pp. 455–463.

[30] Y. Pan, M. Liu, C. Lian, Y. Xia, and D. Shen, "Spatially-constrained Fisher representation for brain disease identification with incomplete multi-modal neuroimages," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2965–2975, Sep. 2020.

[31] Y. Pan, M. Liu, Y. Xia, and D. Shen, "Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6839–6853, Oct. 2022.

[32] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–51.

[33] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 2180–2188.

[34] I. Higgins et al., "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, vol. 2, 2017, pp. 1–22.

[35] E. Dupont, "Learning disentangled joint continuous and discrete representations," in *Proc. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 708–718.

[36] E. H. Sanchez, M. Serrurier, and M. Ortner, "Learning disentangled representations via mutual information estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 205–221.

[37] X. Li, M. Jia, M. T. Islam, L. Yu, and L. Xing, "Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4023–4033, Dec. 2020.

[38] C. Chen, Q. Dou, Y. Jin, H. Chen, J. Qin, and P.-A. Heng, "Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2019, pp. 447–456.

[39] Q. Dou, Q. Liu, P. A. Heng, and B. Glocker, "Unpaired multi-modal segmentation via knowledge distillation," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2415–2425, Jul. 2020.

[40] M. Hu et al., "Knowledge distillation from multi-modal to mono-modal segmentation networks," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2020, pp. 772–781.

[41] T. van Sonsbeek, X. Zhen, M. Worring, and L. Shao, "Variational knowledge distillation for disease classification in chest X-rays," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2021, pp. 334–345.

[42] F. R. Valverde, J. V. Hurtado, and A. Valada, "There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11612–11621.

[43] T. Xu and C. Liu, "Data-distortion guided self-distillation for deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 5565–5572.

[44] M. Ji, S. Shin, S. Hwang, G. Park, and I.-C. Moon, "Refine myself by teaching myself: Feature refinement via self-knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10664–10673.

[45] R. D. Hjelm et al., "Learning deep representations by mutual information estimation and maximization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–24.

[46] B. Fischl, "FreeSurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, Aug. 2012.

[47] F. Kurth, C. Gaser, and E. Luders, "A 12-step user guide for analyzing voxel-wise gray matter asymmetries in statistical parametric mapping (SPM)," *Nature Protocols*, vol. 10, no. 2, pp. 293–304, Feb. 2015.

[48] C. Feng et al., "Deep learning framework for Alzheimer's disease diagnosis via 3D-CNN and FSBi-LSTM," *IEEE Access*, vol. 7, pp. 63605–63618, 2019.

[49] J. Shi, X. Zheng, Y. Li, Q. Zhang, and S. Ying, "Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 173–183, Jan. 2018.

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

[51] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.