Medical Image Analysis 69 (2021) 101953

Contents lists available at ScienceDirect

Medical Image Analysis

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

Challenge Report

Incomplete multi-modal representation learning for Alzheimer's disease diagnosis

Yanbei Liu[a,b], Lianxi Fan[c], Changqing Zhang[d,*], Tao Zhou[e], Zhitao Xiao[a], Lei Geng[a], Dinggang Shen[f,g,h,*]

[a] School of Life Sciences, Tiangong University, Tianjin 300387, China
[b] Tianjin Key Laboratory of Optoelectronic Detection Technology and Systems, Tianjin, China
[c] School of Electronics and Information Engineering, Tiangong University, Tianjin 300387, China
[d] College of Intelligence and Computing, Tianjin University, Tianjin, China
[e] Inception Institute of Artificial Intelligence, Abu Dhabi 51133, United Arab Emirates
[f] School of Biomedical Engineering, Shanghai Tech University, Shanghai, China
[g] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
[h] Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea

ARTICLE INFO

Article history:
Received 18 June 2020
Revised 25 December 2020
Accepted 28 December 2020
Available online 1 January 2021

Keywords:
Alzheimers disease diagnosis
multi-modal representation learning
incomplete multi-modality data
auto-encoder network
kernel completion

ABSTRACT

Alzheimers disease (AD) is a complex neurodegenerative disease. Its early diagnosis and treatment have been a major concern of researchers. Currently, the multi-modality data representation learning of this disease is gradually becoming an emerging research field, attracting widespread attention. However, in practice, data from multiple modalities are only partially available, and most of the existing multi-modal learning algorithms can not deal with the incomplete multi-modality data. In this paper, we propose an Auto-Encoder based Multi-View missing data Completion framework (AEMVC) to learn common representations for AD diagnosis. Specifically, we firstly map the original complete view to a latent space using an auto-encoder network framework. Then, the latent representations measuring statistical dependence learned from the complete view are used to complement the kernel matrix of the incomplete view in the kernel space. Meanwhile, the structural information of original data and the inherent association between views are maintained by graph regularization and Hilbert-Schmidt Independence Criterion (HSIC) constraints. Finally, a kernel based multi-view method is applied to the learned kernel matrix for the acquisition of common representations. Experimental results achieved on Alzheimers Disease Neuroimaging Initiative (ADNI) datasets validate the effectiveness of the proposed method.

1. Introduction

Alzheimer's disease (AD) and its early stage, Mild Cognitive Impairment (MCI), are progressive and irreversible neurodegenerative diseases causing many elderly deaths worldwide. Their early diagnosis and treatment are of great significance for improving the quality of patients' life. In the field of computer-aided research, there have been several studies (Weiner et al., 2017; Chen et al., 2016; Jia et al., 2012; Fan et al., 2007; Wu et al., 2006) exploring different aspects of the disease in the recent years. Hence, there are multiple modalities of data (i.e., Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET)) or multiple types of features available for this task (Zhang et al., 2011). Data from different modalities are complementary because they are the representations of each subject. Since each modal can be treated as a view of subjects, the multi-modal medical data processing problem is modeled as a multi-view machine learning framework.

Multi-view representation learning aims to learn new representations that can better fulfill the task than the original data. Earlier multi-view studies usually explore the minimum disagreement between views based on co-training (Kumar and Daumé, 2011). Canonical Correlation Analysis (CCA) based methods including CCA (Hotelling, 1992), Kernel Canonical Correlation Analysis (KCCA) (Akaho, 2006), deep neural networks based CCA (Andrew et al., 2013) are widely used in representation learning advocating learning a latent common subspace across different views. For AD diagnosis, the recent work (Zhu et al., 2014; Zhu et al., 2016) transform the original features from different modalities to a common space by using CCA. There are also some multi-kernel methods applied in multi-view learning (Zien and Ong, 2007)

* Corresponding authors.
E-mail addresses: zhangchangqing@tju.edu.cn (C. Zhang), Dinggang.Shen@gmail.com (D. Shen).

https://doi.org/10.1016/j.media.2020.101953
1361-8415/

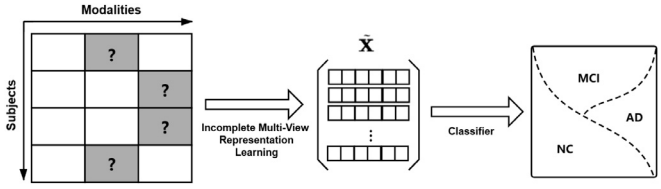


Fig. 1. The process of multi-view representation learning with incomplete views for Alzheimer's disease diagnosis. As shown in the figure, rows and columns represent subjects and modalities, respectively. White blocks represent existing subjects, and gray blocks represent missing subjects. We expect to complement the missing information and learn common representations $\tilde{\mathbf{X}}$. Then a classifier such as SVM is used to predict Alzheimer's disease.

(Liu et al., 2019), which calculate the kernel matrices for each view to obtain the optimal combination of kernels.

Recently, multi-view learning algorithms (Bickel and Scheffer, 2004; Perrin et al., 2009; Kumar and Daumé, 2011) are based on a hypothesis that all views are complete. However, we often face the dilemma that only partial data from multiple views can be obtained in practice, especially in the field of medical data analysis, which is featured by the incompleteness of multi-modality data, the scarcity of subjects and the complexity of data (Zhang et al., 2011; Zhou et al., 2019; Ghazi et al., 2019). Since the cost of PET examination is nearly ten times that of MRI in reality, some subjects may only take one examination for economic reasons. As a result, only one modal data can be acquired, which results in an incomplete view. General algorithms that require complete data are thus inapplicable, making it extremely difficult to accomplish the task, especially in the case of few subjects. Therefore, we have a motivation to develop a multi-view method that can deal with incomplete views and learn common representations of multi-view data as shown in Fig. 1. The representations obtained from existing data are crucial for exploring the relationship between subjects and subsequent analysis.

Under the above conditions, there are some multi-view methods (Li et al., 2014; Lei et al., 2016; Liu et al., 2016; Tran et al., 2017; Liu et al., 2018) having been developed. First, the missing modality imputation methods (Ngiam et al., 2011; Tran et al., 2017) work on a premise that a set of fully-paired training data can be obtained to learn the relationship between different views. In the test stage, it can predict the missing part from the observed one. Second, some low-rank based methods (Cai et al., 2010; Mazumder et al., 2010) are not applicable to this task since the missing views are usually blockwise that have been recognized (Tran et al., 2017; Cai et al., 2018). Finally, the most direct approach is to fill the missing values in feature space or kernel space with random or mean values. The advantage is that it can easily combine with other multi-view methods to gain the common latent representations. However, few methods are designed for AD data and they do not fully consider the correlation between views during completion.

In general, there are three main challenges in multi-view representation learning for AD diagnosis. Firstly, the complicated AD data and extracted features with noise pose a challenge to the learning of low dimensional representations while preserving the structural information of AD data and reducing the impact of noise. Secondly, completion is not sufficient to get a better performance, which also requires joint complementation of the missing part while exploring other information such as the inherent association between views. Thirdly, it is difficult to maintain the alignment of the completed matrix and the truth value in the absence of information and supervision.

To address these challenges, we propose an incomplete multi-view representation learning method for AD diagnosis. Specifically, we firstly map the original complete view to a latent space by using an auto-encoder network framework, which can reduce the noise of data and the dimensionality of features. Graph regularization is utilized to maintain the structural information of original data. Then, latent representations learned from the complete view are used to complement the kernel matrix of the incomplete view in the kernel space. Meanwhile, the correlation between different views is explored using Hilbert-Schmidt Independence Criterion (HSIC). Finally, a kernel-based multi-view method is applied to the learned kernel matrix to obtain the common representations. The proposed method complements information in the kernel space rather than in the original feature space, so it can better maintain the relationship of samples in the high dimensional space. In view of this, the proposed method is more applicable to the processing of complex AD data and can be easily docked with other kernel-based multi-view algorithms (e.g., KCCA and Spectral Clustering). The contributions of this study can be summarized as follows:

- We study an incomplete multi-modal representation learning problem for Alzheimers disease diagnosis in a new way of complementing the missing data.
- We propose an Auto-Encoder based Multi-View missing data Completion framework called AEMVC, that is able to complement the missing data in the kernel space while taking into account the structural information of data and the inherent association between multiple views.
- We conduct extensive experiments to evaluate the effectiveness of the proposed method. The proposed model is superior to the state-of-the-art models according to the experimental comparison results.

The rest of this paper is organized as follows. We overview the related work in Section 2. In Section 3, the problem is defined and the proposed method is introduced. We describe our optimization algorithm in Section 4. Experimental results and conclusion are shown in Sections 5 and 6.

2. Related work

Traditional machine learning methods are designed to identify groups of similar behavior in single view data (Von Luxburg, 2007; Steinwart et al., 2015). However, in many machine learning problems, data are often described by multiple distinct feature sets, each of which can be considered as a view of the original data sets. These feature sets can be divided into two parts, which are features of different types and features from different data sources (Zhang et al., 2018). For instance, an image can be described by morphological features and histogram features (i.e., different types). In content-based web-image retrieval, an object is simultaneously described by visual features from the image and the text surrounding the image (i.e., different data sources). Each feature describes different independent information of the same sample. In addition, a noteworthy feature of multi-view learning is that manufacturing splitting can improve the performance even when there is a lack of natural feature splitting (Sun, 2013). In this task, we combine the information of multiple views from different sources (i.e., MRI and PET) to learn better representations.

Recently, a number of unsupervised multi-view learning methods (De Sa, 2005; Cai et al., 2013; Xu et al., 2016) have been proposed to improve experimental performance. For example, the method in (Wang et al., 2014) co-regularize the clustering hypotheses to exploit the complementary information within the spectral clustering framework. A co-training method (Kumar and Daumé, 2011) is used to search for the clustering that agrees among different views. The work in (Zhang et al., 2016a) using the Hilbert-Schmidt Independence Criterion (HSIC) performs the kernel matching to regularize the dependence across multiple views and

obtains the low-dimensional projection for each view. The subspace clustering methods (Elhamifar and Vidal, 2013; Liu et al., 2012; Hu et al., 2014) have been proposed to explore the relationships between samples by self-representation. Inspired by the Laplace regularization, a clustering method with incomplete view (Trivedi et al., 2010) aligns the kernel matrices of multiple views to obtain closed solutions for incomplete views. Besides, a multi-view weak-label learning method (Tan et al., 2018) simultaneously learns a shared subspace, local label correlations and a predictor.

In the application of Alzheimer's disease diagnosis, although some methods use a single modal data (Zhang et al., 2016b; Lian et al., 2018), there are also a lot of methods that consider multi-modal data to improve the diagnostic performance. The work in (Zhang et al., 2011) adopts a kernel combination method to combine three modalities of biomarkers (i.e., MRI, FDG-PET and CSF biomarkers) to discriminate between AD (or MCI) and healthy controls. The method in (Zhang et al., 2018) captures the high-order complementarity among different views, by exploiting the underlying information with a low-rank tensor regularization. Considering that the extracted features from different brain regions are related to each other to some extent, the method in (Shi et al., 2019) leverages the coupled interactions in the feature level and modality level for diagnosis. More recently, the method in (Zhou et al., 2019) learns a common latent representation and uses subjects with incomplete data to learn independent modality specific latent representations for AD diagnosis. Yet it does not constrain the relationship between modalities, and the latent representations learned by subjects with missing modalities are modality specific and not shareable. The proposed method complements the missing data in the kernel space, which helps to explore the relationship between modalities and obtain common representations that can be shared.

The main purpose of multi-view learning is to make use of information from different views as fully as possible. In addition to the methods mentioned above, CCA-based methods (including CCA (Hotelling, 1992), KCCA (Akaho, 2006), deep neural networks based CCA (Andrew et al., 2013), *mathrmS²GCA* (Chen et al., 2012)) are extensively applied in multi-view representation learning (Hardoon and Shawe-Taylor, 2003; Hardoon et al., 2004). CCA aims to model the relationship between different sets of variables (views). CCA linearly computes low dimensional common representations of two different views so that the correlation between different views is maximized in this common space. Given a pair of datasets $\mathbf{X}$ and $\mathbf{Y}$ which can be treated as samples of two different views, two projection vectors $\mathbf{a}$ and $\mathbf{b}$ can be obtained by maximizing the following formula:

$$arg\max_{\mathbf{a},\mathbf{b}} \frac{cov\left(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}\right)}{\sqrt{D\left(\mathbf{a}^T\mathbf{X}\right)}\sqrt{D\left(\mathbf{b}^T\mathbf{Y}\right)}}. \tag{1}$$

However, the relationship between variables in a real dataset is generally non-linear, which cannot be modeled by CCA. Fortunately, kernel methods (Shawe-Taylor et al., 2004) are applicable to it. Kernel methods adopt the linear method (e.g., Support Vector Machines (Burges, 1998)) after using the kernel function (e.g., Polynomial kernel and Gaussian kernel) to find the inner product of the data in a high dimensional space without the need of an exact mapping from a low dimensional space to a high dimensional one. Inspired by the kernel methods, KCCA is proposed.

With the advantage of the kernel method, we expect to develop an incomplete multi-view representation learning method, which complements the missing information in the kernel space, and then uses KCCA to map the multi-view data to a common feature space.

3. Proposed model

In this section, we explicate the overall network architecture of the proposed AEMVC as shown in Fig. 2 and the optimization algorithm.

3.1. Problem definition

Consider that the feature vectors extracted from two modal data (i.e., MRI and PET) are two views, which are represented by $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ respectively. We can assume that $\mathbf{X}^{(1)} \in \mathbb{R}^{N\times d_1}$ having N subjects with d_1 dimension is complete whereas $\mathbf{X}^{(2)} \in \mathbb{R}^{C\times d_2}$ having C subjects with d_2 dimension is incomplete. Note that $\mathcal{D} = \left\{\mathbf{X}_n^{(1)}, \mathbf{X}_n^{(2)}\right\}_{n=1}^{C}$ is the set of subjects which are fully paired in two views, and $\mathcal{M} = \left\{\mathbf{X}_n^{(1)}\right\}_{n=C}^{N}$ is a set of subjects only in view 1 with $C + M = N$. Taking $\mathbf{X}^{(1)}$ as the input of the Auto-Encoder, we expect to get latent representations $\mathbf{H}$. $\mathbf{K}_1$ and $\mathbf{K}_2$, which should be two $N \times N$ symmetric matrices, denote kernel matrices computed by the original features. Yet, only $\mathbf{K}_2 \in \mathbb{R}^{C\times C}$ can be obtained due to incompletion of this view. $\tilde{\mathbf{K}}_1$ is calculated by $\mathbf{H}$. Afterward, the missing part of $\mathbf{K}_2$ is complemented using $\tilde{\mathbf{K}}_1$, and KCCA is adopted to obtain the common representation for the prediction task. More detailed notations used in this paper are listed in Table 1.

3.2. Auto-encoder based multi-view missing data completion

Kernel completion. $\mathbf{K}_1$ and $\mathbf{K}_2$ are both $N \times N$ kernel matrices. Only a $C \times C$ sub-block of $\mathbf{K}_2$ can be obtained since features for view $\mathbf{X}^{(2)}$ are merely available for a subset of all samples. Motivated by the work in (Trivedi et al., 2010), we reconstruct the full kernel matrix $\mathbf{K}_2$ by solving the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{K}_2\succeq 0} \ tr(\mathcal{L}_1\mathbf{K}_2) \\ &s.t.\ \mathbf{K}_2(i,j) = k\left(\mathbf{x}_i^{(2)}, \mathbf{x}_j^{(2)}\right),\ \forall 1 \le (i,j) \le C, \end{aligned} \tag{2}$$

where tr denotes the matrix trace and $\mathbf{x}_i^{(2)}$ stands for the i-th sample of view 2. The corresponding graph Laplacian is defined as $\mathcal{L}_1 = \mathbf{D}_1 - \mathbf{K}_1$, where $\mathbf{D}_1$ is the diagonal matrix consisting of the row sums of $\mathbf{K}_1$ along its diagonals. Then, a couple of matrices multiplication and inverses give a closed-form solution of $\mathbf{K}_2$. Note that $\mathbf{K}_2(i,j) = k\left(\mathbf{x}_i^{(2)}, \mathbf{x}_j^{(2)}\right),\ \forall 1 \le (i,j) \le C$ in Eq. (2) is $\mathbf{K}_2^{cc}$, which is a $C \times C$ sub kernel matrix of view $\mathbf{X}^{(2)}$. $\mathbf{K}_2$, a positive-definite matrix, can be rewritten as $\mathbf{A}\mathbf{A}^T$, where $\mathbf{A}$ is just a constant matrix of reals.

In order to explicitly state the problem, we split $\mathbf{A}$ into $(\mathbf{A}_c, \mathbf{A}_m)^T$, where $\mathbf{A}_c$ is a constant satisfying $\mathbf{A}_c\mathbf{A}_c^T = \mathbf{K}_2^{cc}$, and $\mathbf{A}_m$ is the missing part. Then $\mathcal{L}_1$ is divided into:

$$\mathcal{L}_1 = \begin{bmatrix} \mathcal{L}_1^{cc} & \mathcal{L}_1^{cm} \\ \left(\mathcal{L}_1^{cm}\right)^T & \mathcal{L}_1^{mm} \end{bmatrix}.$$

Therefore, based on $tr(\mathbf{X}) = tr(\mathbf{X}^T)$, Eq. (2) can be expressed as:

$$\begin{aligned} &\min_{\mathbf{A}} \ tr\left(\mathcal{L}_1\mathbf{A}\mathbf{A}^T\right) \\ &= \min_{\mathbf{A}} \ tr\left(\mathbf{A}^T\mathcal{L}_1\mathbf{A}\right) \\ &= \min_{\mathbf{A}_m} \ tr\left(\begin{pmatrix}\mathbf{A}_c \\ \mathbf{A}_m\end{pmatrix}^T \begin{bmatrix} \mathcal{L}_1^{cc} & \mathcal{L}_1^{cm} \\ \left(\mathcal{L}_1^{cm}\right)^T & \mathcal{L}_1^{mm} \end{bmatrix} \begin{pmatrix}\mathbf{A}_c \\ \mathbf{A}_m\end{pmatrix}\right) \\ &= \min_{\mathbf{A}_m} 2 \times tr\left(\mathbf{A}_c^T\mathcal{L}_1^{cm}\mathbf{A}_m\right) + tr\left(\mathbf{A}_m^T\mathcal{L}_1^{mm}\mathbf{A}_m\right). \end{aligned} \tag{3}$$

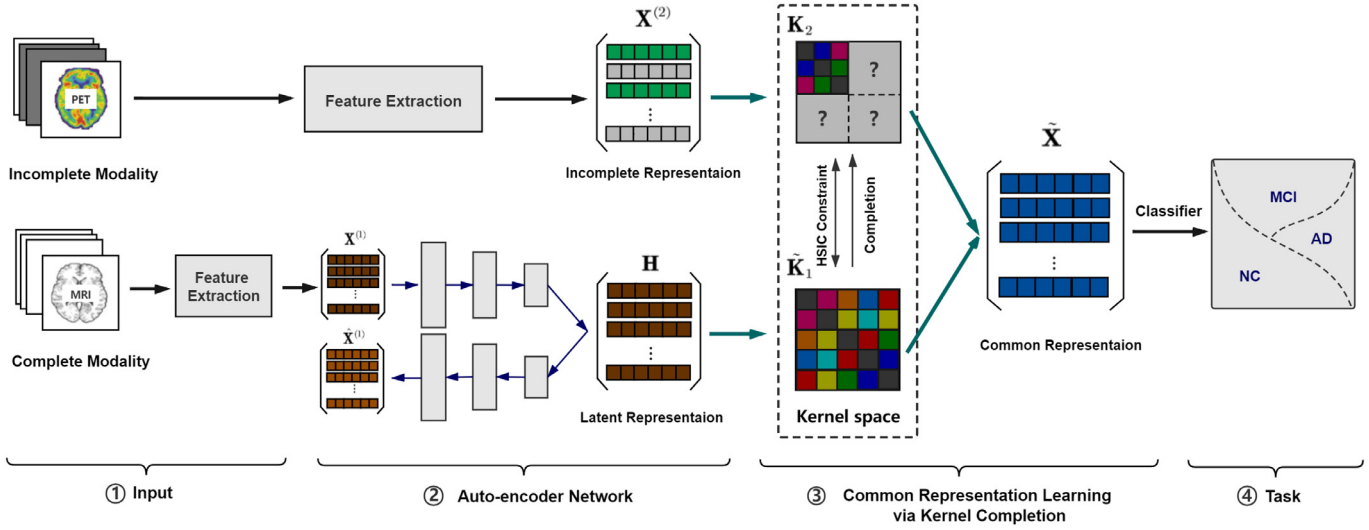


Fig. 2. Illustration of the proposed framework. It mainly consists of four parts. First of all, features of the original data are extracted to obtain the feature vectors. Then, we obtain the latent representations $\mathbf{H}$ of the complete view $\mathbf{X}^{(1)}$ by using an auto-encoder network. Meanwhile, the graph regularization is utilized to maintain the structural information. The latent representations $\mathbf{H}$ are used to complement the missing part of view $\mathbf{X}^{(2)}$ in the kernel space. Farther, HSIC constraint explores the correlation between different views. Finally, Kernel CCA is adopted to obtain the common representations $\tilde{\mathbf{X}}$, which can be used for the next task.

Table 1
Table of main notations used in this paper .

Model Specification	
Notation	Meaning
$\mathbf{X}^{(1)} \in \mathbb{R}^{N\times d_1}$	feature matrix of the view 1 (complete)
$\mathbf{X}^{(2)} \in \mathbb{R}^{C\times d_2}$	feature matrix of the view 2 (incomplete)
$\mathbf{K}_n \in \mathbb{R}^{N\times N}$	kernel matrix of view n
$\mathbf{H} \in \mathbb{R}^{N\times h}$	latent representations of view 1 encoded by the auto-encoder network
$\tilde{\mathbf{K}}_1 \in \mathbb{R}^{N\times N}$	kernel matrix of view 1 calculated by the latent representations
$\tilde{\mathbf{X}} \in \mathbb{R}^{N\times d}$	common representations of two views to accomplish the specific task
$\mathbf{\Theta}$	parameters of the auto-encoder network
$\lambda_n > 0$	hyperparameters balancing the effects of individual loss functions

Taking the derivative $\mathbf{A}_m$ and setting it to zero, the closed-form solution of $\mathbf{K}_2$ can be achieved according to $\mathbf{K}_2 = \mathbf{A}\mathbf{A}^T$.

Auto-encoder network. Unlike the original kernel completion method, we firstly calculate latent representations $\mathbf{H}$ relying on an Auto-Encoder network framework that can reduce noise and maintain self information of samples simultaneously. Then a new kernel matrix $\tilde{\mathbf{K}}_1$ can be computed by $\mathbf{H}$ to complement the missing part of $\mathbf{K}_2$. The preliminary objective function is defined as follows:

$$\mathscr{L} = \mathscr{L}_{rec} + \mathscr{L}_{com}, \tag{4}$$

where $\mathscr{L}_{reg}$ represents the reconstruction loss of the auto-encoder and $\mathscr{L}_{com}$ stands for the completion loss defined by Eq. (2).

The auto-encoder is a commonly used unsupervised learning method that maps the original features into latent low-dimensional representations, capable of reducing dimensions and noise. In this paper, we use a neural network to build an auto-encoder. The auto-encoder network with M layers is denoted as $f\left(\mathbf{X}^{(1)};\mathbf{\Theta}\right) = \hat{\mathbf{X}}$, where $\mathbf{X}^{(1)}$ is the input data (i.e., the original features) and $\mathbf{\Theta}$ is the parameter set consisting of wights and bias. Generally, the first $M/2$ layers are defined as an encoder whose output is latent representations containing the main information denoted by $\mathbf{H} \in \mathbb{R}^{N\times h}$ with h dimension. The last $M/2$ layers reconstruct the input by decoding $\mathbf{H}$ to output $\hat{\mathbf{X}}^{(1)}$. We have $\mathbf{H} = f\left(\mathbf{X}^{(1)};\mathbf{\Theta}^{\frac{M}{2}}\right)$ by minimizing the following equation:

$$\mathscr{L}_{rec} = \|\mathbf{X}^{(1)} - \hat{\mathbf{X}}^{(1)}\|_F^2. \tag{5}$$

As shown by the formula, input and output of the auto-encoder are expected to be as consistent as possible so that the latent representations can well preserve their own information. Since the latent representations $\mathbf{H}$ are learnable, the addition of some constraints may help achieve a better performance.

Structural constraint. Furthermore, $\mathbf{H}$ should have more structural information to reflect the differences among samples of varied categories. Spurred by the graph regularization, we add a structural constraint on $\mathbf{H}$ as follows:

$$\mathscr{L}_{graph} = \frac{1}{2}\sum_{i,j=1}^{N} \|\mathbf{h}_i - \mathbf{h}_j\|^2 \mathbf{W}_{ij}, \tag{6}$$

where $\mathbf{h}_i$ represents the i-th sample and $\mathbf{W}_{ij}$ is defined as the similarity of samples (Yang et al., 2017). Here $\mathbf{W}$ can be replaced by the kernel matrix $\mathbf{K}_1$, which reflects the similarity between samples in the kernel space. For simplicity, we can further derive the Eq. (6) as follows:

$$\begin{aligned}\mathscr{L}_{graph} &= \frac{1}{2}\sum_{i,j=1}^{N} \|\mathbf{h}_i - \mathbf{h}_j\|^2 \mathbf{K}_{ij} \\ &= \frac{1}{2}\sum_{i=1}^{N} \mathbf{h}_i^T\mathbf{h}_i\mathbf{D}_{ii} - \sum_{i,j=1}^{N} \mathbf{h}_i^T\mathbf{h}_j\mathbf{K}_{ij} \\ &= \frac{1}{2} tr\left(\mathbf{H}^T\mathcal{L}_1\mathbf{H}\right),\end{aligned} \tag{7}$$

where $\mathbf{D}_1$ is a diagonal matrix consisting of the row sums of $\mathbf{K}_1$ and $\mathcal{L}_1 = \mathbf{D}_1 - \mathbf{K}_1$ is called the graph Laplacian matrix. Since $\mathbf{H}$ is

just latent representations of the auto-encoder, after optimizing $\mathbf{\Theta}$, the above formula can be represented:

$$\min_{\mathbf{\Theta}} tr\left(f\left(\mathbf{X}^{(1)};\mathbf{\Theta}^{\frac{M}{2}}\right)^{T}\mathcal{L}_1 f\left(\mathbf{X}^{(1)};\mathbf{\Theta}^{\frac{M}{2}}\right)\right). \tag{8}$$

By minimizing the Eqs. (5) and (7), we obtain a new latent representations $\mathbf{H}$ in which $\mathbf{h}_i$ and $\mathbf{h}_j$ are mapped close to each other in this $\mathbb{R}^{N\times h}$ space if data points $\mathbf{x}_i^{(1)}$ and $\mathbf{x}_j^{(1)}$ are close in the original space.

Correlation constraint. Since different views are deemed as different representations of the same sample in the multi-view learning, these representations should be as relevant as possible. With the view that representations of view $\mathbf{X}^{(1)}$ and the missing part of view $\mathbf{X}^{(2)}$ are learnable, a correlation constraint is added in the training stage. However, there is a gap between the problem of incomplete data and the task of calculating the correlation. Fortunately, the complete kernel matrix of two views can be acquired by the above formula (4), (8). It has been proven both theoretically (Gretton et al., 2005) and empirically (Xiao and Guo, 2014; Song et al., 2007) that HSIC (Gretton et al., 2005; Liu et al., 2020) is an appropriate measure of (in)dependence between different views when it is associated with a generic kernel.

Assume two views $\mathbf{Z}^{(v)}$ and $\mathbf{Z}^{(w)}$ containing n samples $\left\{\left(\mathbf{z}_i^{(v)},\mathbf{z}_i^{(w)}\right)\in\mathcal{Y}^{(v)}\times\mathcal{Y}^{(w)}\right\}_{i=1}^{n}$ that are jointly drawn from a probability distribution $P_{\mathbf{z}^{(v)}\mathbf{z}^{(w)}}$. The correlation measured by HSIC is calculated based on the norm of the cross-covariance operator in the domain $\mathcal{Y}^{(v)}\times\mathcal{Y}^{(w)}$ of the Hilbert space. A larger HSIC value indicates strong dependence on kernel selection. Note that $\phi\left(\mathbf{z}^{(v)}\right)$ and $\psi\left(\mathbf{z}^{(w)}\right)$ are functions obtained by mapping $\mathbf{z}^{(v)}\in\mathcal{Y}^{(v)}$ and $\mathbf{z}^{(w)}\in\mathcal{Y}^{(w)}$ to a higher dimensional space respectively. The respective kernel spaces $\mathcal{F}$ and Ω for the kernel functions $k_v\left(\mathbf{z}_i^{(v)},\mathbf{z}_j^{(v)}\right)=\left\langle\phi\left(\mathbf{z}_i^{(v)}\right),\phi\left(\mathbf{z}_j^{(v)}\right)\right\rangle$ and $k_w\left(\mathbf{z}_i^{(w)},\mathbf{z}_j^{(w)}\right)=\left\langle\psi\left(\mathbf{z}_i^{(w)}\right),\psi\left(\mathbf{z}_j^{(w)}\right)\right\rangle$ are then achieved. The cross-covariance function that gives the covariance of two random variables and is defined as follows:

$$\mathcal{C}_{\mathbf{z}^{(v)},\mathbf{z}^{(w)}}=E_{\mathbf{z}^{(v)}\mathbf{z}^{(w)}}\left[\left(\phi\left(\mathbf{z}^{(v)}\right)-\mu_{\mathbf{z}^{(v)}}\right)\otimes\psi\left(\mathbf{z}^{(w)}\right)-\mu_{\mathbf{z}^{(w)}}\right], \tag{9}$$

where $\otimes$ denotes the tensor product, $\mathcal{F}$ and Ω are Reproducing Kernel Hilbert Space (RKHS) on $\mathcal{Y}^{(v)}$ and $\mathcal{Y}^{(w)}$. Then the HSIC is defined as:

$$\mathrm{HSIC}(P_{\mathbf{z}^{(v)}\mathbf{z}^{(w)}},\mathcal{F},\Omega):=\|\mathcal{C}_{\mathbf{z}^{(v)},\mathbf{z}^{(w)}}\|_{HS}^2, \tag{10}$$

where $\|A\|_{HS}=\sqrt{\sum_{i,j}a_{ij}^2}$. Accordingly, the empirical version of HSIC is given as:

$$\mathrm{HSIC}\left(\mathbf{z}^{(v)},\mathbf{z}^{(w)}\right)=(n-1)^{-2}tr(\mathbf{K}_v\mathbf{E}\mathbf{K}_w\mathbf{E}), \tag{11}$$

where $\mathbf{K}_v,\mathbf{K}_w,\mathbf{E}\in\mathbb{R}^{n\times n}$, $k_{v,ij}=k_v\left(\mathbf{z}_i^{(v)},\mathbf{z}_j^{(v)}\right)$, $k_{w,ij}=k_w\left(\mathbf{z}_i^{(w)},\mathbf{z}_j^{(w)}\right)$ and $e_{ij}=\delta_{ij}-1/n$ which centers the matrix to gain a zero mean in the feature space.

In our implementation, HSIC is used to investigate the correlation between the $\tilde{\mathbf{K}}_1$ computed by $\mathbf{H}$ and the complemented kernel matrix $\mathbf{K}_2$. Minimizing the HSIC loss helps render the representation of the two views more relevant. Therefore, the $e_{ij}=\delta_{ij}-1/N$ and Eq. (11) can be rewritten as:

$$\mathscr{L}_{hsic}=-(N-1)^{-2}tr\left(\tilde{\mathbf{K}}_1\mathbf{E}\mathbf{K}_2\mathbf{E}\right). \tag{12}$$

In general, the learned latent representations $\mathbf{H}$ are better than the original features $\mathbf{X}^{(1)}$. The kernel matrix calculated by $\mathbf{H}$ is more accurate than the one calculated by the original data. A better kernel matrix of view $\mathbf{X}^{(2)}$ can be obtained by replacing $\mathcal{L}_1$ in the Eq. (2) with a new Laplace matrix $\tilde{\mathcal{L}}_1=\tilde{\mathbf{D}}_1-\tilde{\mathbf{K}}_1$ computed by $\mathbf{H}$. Eq. (2) can be rewritten as follows:

$$\begin{aligned}&\mathscr{L}_{com}=tr\left(\tilde{\mathcal{L}}_1\mathbf{K}_2\right)\\&s.t.\ \mathbf{K}_2(i,j)=k\left(\mathbf{x}_i^{(2)},\mathbf{x}_j^{(2)}\right),\ \forall 1\le(i,j)\le C.\end{aligned} \tag{13}$$

In summary, the total objective function that should be minimized is as follows:

$$\begin{aligned}\mathscr{L}&=\mathscr{L}_{rec}+\lambda_1\mathscr{L}_{graph}+\lambda_2\mathcal{L}_{hsic}+\lambda_3\mathscr{L}_{com}\\&=\|\mathbf{X}^{(1)}-f\left(\mathbf{X}^{(1)};\mathbf{\Theta}^{M}\right)\|_F^2\\&\quad+\lambda_1 tr\left(f\left(\mathbf{X}^{(1)};\mathbf{\Theta}^{\frac{M}{2}}\right)^{T}\mathcal{L}_1 f\left(\mathbf{X}^{(1)};\mathbf{\Theta}^{\frac{M}{2}}\right)\right)\\&\quad-\lambda_2 tr\left(\tilde{\mathbf{K}}_1\mathbf{E}\mathbf{K}_2\mathbf{E}\right)\\&\quad+\lambda_3 tr\left(\tilde{\mathcal{L}}_1\mathbf{K}_2\right),\end{aligned} \tag{14}$$

where non-negative $\lambda_1,\lambda_2,\lambda_3$ are hyperparameters to balancing the effects of individual loss functions. The proposed model is able to complement the missing data in the kernel space and take into consideration the structural information of data and the inherent association between multiple views in an unsupervised manner.

4. Optimization

Two variables (i.e., the auto-encoder network parameters $\mathbf{\Theta}$ and the kernel matrix $\mathbf{K}_2$) in the above objective function need to be optimized jointly and iteratively. The general optimization steps are expounded next.

4.1. Update the auto-encoder network

According to the total objective function Eq. (14), the following function should be minimized to update the auto-encoder network:

$$\begin{aligned}\mathscr{L}_{ae}(\mathbf{\Theta})&=\|\mathbf{X}^{(1)}-f\left(\mathbf{X}^{(1)};\mathbf{\Theta}^{M}\right)\|_F^2\\&\quad+\lambda_1 tr\left(f\left(\mathbf{X}^{(1)};\mathbf{\Theta}^{\frac{M}{2}}\right)^{T}\mathcal{L}_1 f\left(\mathbf{X}^{(1)};\mathbf{\Theta}^{\frac{M}{2}}\right)\right)\\&\quad-\lambda_2 tr\left(\tilde{\mathbf{K}}_1\mathbf{E}\mathbf{K}_2\mathbf{E}\right)\\&\quad+\lambda_3 tr\left(\tilde{\mathcal{L}}_1\mathbf{K}_2\right).\end{aligned} \tag{15}$$

In particular, it is not straightforward to solve $tr\left(\tilde{\mathcal{L}}_1\mathbf{K}_2\right)$ in a matrix form. Instead, inspired by the work (Tao et al., 2017), we introduce an auxiliary matrix $\mathbf{P}\in\mathbb{R}^{N\times N}$ as:

$$\mathbf{P}=\left[\mathbf{P}_1\ldots\mathbf{P}_j\ldots\mathbf{P}_N\right],\mathbf{P}_j=\begin{bmatrix}\|\mathbf{A}_1-\mathbf{A}_j\|\\\vdots\\\|\mathbf{A}_N-\mathbf{A}_j\|\end{bmatrix}. \tag{16}$$

Consequently, $tr\left(\tilde{\mathcal{L}}_1\mathbf{K}_2\right)$ is rewritten as $tr\left(\mathbf{P}^T\tilde{\mathbf{K}}_1\right)$ in Eq. (15). Finally, parameters are updated with gradient descent as: $\mathbf{\Theta}^{(t+1)}=\mathbf{\Theta}^{(t)}-\alpha\frac{\partial\mathscr{L}_{ae}(\mathbf{\Theta})}{\partial\mathbf{\Theta}^{(t)}}$ where α is defined as the learning rate.

4.2. Update $\mathbf{K}_2$

To update the kernel matrix $\mathbf{K}_2$, the following loss function should be minimized:

$$\mathscr{L}_K(\mathbf{K}_2)=tr\left(\tilde{\mathcal{L}}_1\mathbf{K}_2\right)-\lambda tr\left(\tilde{\mathbf{K}}_1\mathbf{E}\mathbf{K}_2\mathbf{E}\right). \tag{17}$$

Satisfactorily, HSIC can be transformed into the form of trace as shown in Section 3.2. Hence, the optimal solution is obtained when the first-order derivative of $\mathbf{K}_2$ is set to zero. As suggested in Section 3.2, $\mathbf{K}_2$ can still be rewritten as $\mathbf{A}\mathbf{A}^T$ where $\mathbf{A}=(\mathbf{A}_c,\mathbf{A}_m)^T$. Then, taking the derivative $\mathbf{A}_m$ and setting it to zero, we can solve:

$$\mathbf{A}_m=-\left(\tilde{\mathcal{L}}_1^{mm}+\mathbf{L}^{mm}\right)^{-1}\left(\tilde{\mathcal{L}}_1^{cm}+\mathbf{L}^{cm}\right)^T\mathbf{A}_c, \tag{18}$$

where $\mathbf{L} = \mathbf{E}\tilde{\mathbf{K}}_1\mathbf{E}$ is a constant. For convenience, $\mathbf{A}_m$ is rewritten as $-\mathbf{B}^{-1}\mathbf{C}^T\mathbf{A}_c$. Finally, the closed-form expression for $\mathbf{K}_2$ is given as:

$$\mathbf{K}_2 = \begin{pmatrix} \mathbf{K}_2^{cc} & -\mathbf{K}_2^{cc}\mathbf{C}\mathbf{B}^{-1} \\ -\mathbf{B}^{-1}\mathbf{C}^T\mathbf{K}_2^{cc} & \mathbf{B}^{-1}\mathbf{C}^T\mathbf{K}_2^{cc}\mathbf{C}\mathbf{B}^{-1} \end{pmatrix}. \tag{19}$$

In conclusion, two variables are updated alternately using Eqs. (15) and (19) in order to seek the optimal solution. To clarify, we summarize the optimization process in Algorithm 1 .

Algorithm 1: Algorithm of AEMVC.

Input: data set: $\mathcal{X} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\}$, where $\mathbf{X}^{(2)}$ is an incomplete view.
Initialization: initialize network parameter $\mathbf{\Theta}$.**while** *not converged* **do**
  for *each sample of view* $\mathbf{X}^{(1)}$ **do**
    update the parameter of the auto-encoder Network with Eq (15)
  end
  update the missing part of $\mathbf{K}_2$ with Eq (19)
end
Calculate $\tilde{\mathbf{K}}_1$, $\mathbf{K}_2$ andcommon representations $\tilde{\mathbf{X}}$.
Adopt the K-means clustering and SVM classification algorithm for $\tilde{\mathbf{X}}$.
Output: Category list.

5. Experiments

5.1. Materials

In this paper, the multi-modality dataset with two modalities (i.e., 1.5T MR and PET images) obtained from 85 Alzheimer's disease (AD), 185 mild cognitive impairment (MCI), and 90 normal control (NC) subjects are used to evaluate the proposed algorithm. All data are downloaded from the Alzheimers Disease Neuroimaging Initiative (ADNI) from the ADNI website[2].

We follow the work in (Zhou et al., 2019) for data preprocessing steps. In order to guarantee the quality of MR images collected by using various scanners following their respective protocols, the spatial distortions of MR images in homogeneities and gradient nonlinearities caused by B1 field are resolved. These images are subjected to the following steps: anterior commissure-posterior commissure (AC-PC) correction, intensity inhomogeneity correction, brain extraction, cerebellum removal, tissues segmentation, registration to a template with 93 ROIs (Kabani et al., 1998) and ROI labels projection. For each ROI, we use the gray matter volume normalized with the intracranial volume in the labeled image as a feature representation. Moreover, the PET images are collected by 30–60 min post Fluoro-Deoxy Glucose (FDG) injection and aligned to their corresponding T1 MR images using affine registration. Then, in the same template, the average PET intensity value of each ROI is computed as a feature representation in the labeled image. Therefore, for each subject, we extract a 93-dimensional ROI-based feature vector from a specific modality (i.e., MRI or PET). Furthermore, in each modality, we use Chi-Square test to rank these 93 ROI-based features according to their influence on the task and select 40 most representative ROI-based features as experimental data, so as to obtain more effective and stable. Since the top ROIs selected for each task is similar, we visualize top 15 representative ROIs in each modality as shown in Fig. 3 (including thalamus, putamen, fornix, etc.). To verify the effectiveness of the extracted features on the diagnosis task, we report the classification accuracy in the AD/NC task at the missing rate of 50% as shown in Fig. 4. Note that the results of MRI+PET is obtained by the multi-modal method MDcR(Zhang et al., 2016a).

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

5.2. Comparison method

We compare the proposed AEMVC framework with the following methods:

FeaCon: FeaCon method can simply concatenate the features from multiple view.

CCA: CCA (Hotelling, 1992) (Canonical Correlation Analysis) method mentioned in Section 2 maps features from multiple view into a common space, capable of keeping the maximum correlation between views.

KCCA: KCCA (Kernel Canonical Correlation Analysis) method mentioned in Section 2 also maps features from multiple view into a common space, and employs the kernel method to keep the maximum correlation between views.

DCCA: DCCA (Andrew et al., 2013) (Deep Canonical Correlation Analysis) explores the common space and keeps the maximum correlation between views with the deep neural network, just like the CCA.

DCCAE: DCCAE (Wang et al., 2015b) (Deep Canonical Correlated AutoEncoders) adopts auto-encoders for common representation, then combines these projected *low* dimensional features together.

MDcR: MDcR (Zhang et al., 2016a) (Multi-view Dimensionality co-Reduction) applies the kernel matching to regularize the dependence across multiple views and projects each view onto a low dimensional space.

MCIV: MCIV (Trivedi et al., 2010) (Multiview Clustering with Incomplete Views) complements missing views in the kernel space via kernel alignment, and then obtains common representations between views through KCCA.

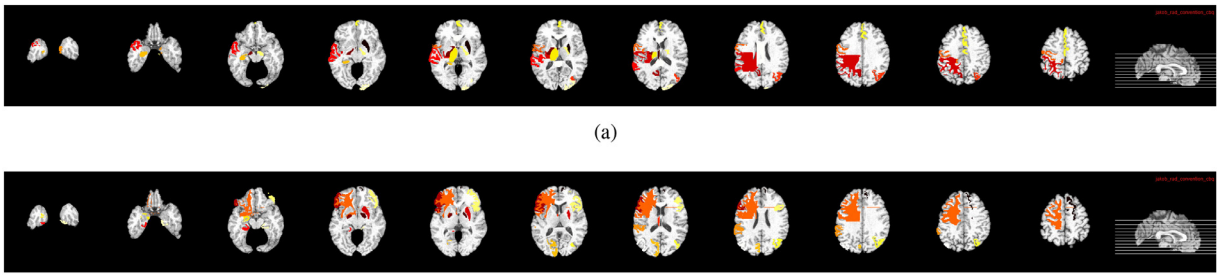
K-Com: K-Com(Zhang et al., 2011)(Multi-modal classification of Alzheimer's disease and mild cognitive impairment) is a multi-modal data fusion and classification method based on kernel combination for AD and MCI.

iMVWL: iMVWL(Tan et al., 2018)(Incomplete Multi-View Weak-Label Learning) jointly addresses incomplete views and missing labels. It learns a shared subspace from incomplete views with weak labels, label correlations, and a predictor in this subspace simultaneously.

5.3. Experiment setup

We conduct experiments on commonly used ADNI dataset. Specifically, the Gaussian kernel function (i.e., $k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left\langle \phi(\mathbf{x}_i), \phi\left(\mathbf{x}_j\right)\right\rangle = exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$) is used for steps that requires the kernel method. For comparison methods that cannot directly handle the incomplete data, we complement the missing part with random values in experiments. The missing rate is defined as $\eta = \frac{M}{N}$, where M indicates the number of missing samples in the incomplete view. Due to the high cost, some subjects do not perform PET examinations, resulting in the missing PET view. Therefore, we set the MRI data as the complete view and the PET data as the incomplete view by default. At the same time, we carry out different settings to conduct more experiments to verify the effectiveness of our method.

To verify the diagnosis performance of the proposed method, we perform multiple classification experiments on the learned representations (AD/MCI/NC, AD/NC, MCI/NC, AD/MCI). Note that a



(a)

(b)

Fig. 3. Visualization of ROIs selected in two modalities, where figure (a) and figure (b) denote ROIs for MRI and PET respectively.

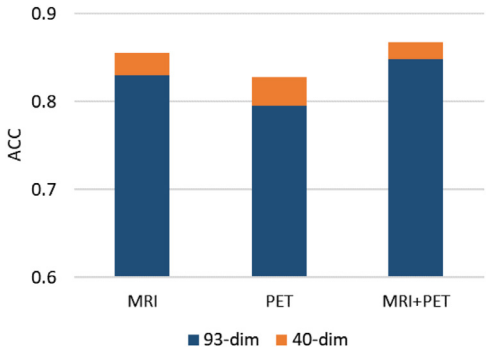


Fig. 4. Classification results with different feature dimensions in AD / NC task.

support vector classification model from the LIBSVM toolbox[3] publicly available with the margin parameter $C = 1$ is employed as the basic classifier for the compared methods. A ten-fold cross-validation strategy is applied to evaluating all compared methods. First, all 360 subjects are randomly divided into 10 subsets on average, 9 of which are selected to train the classifier and the remaining 1 subset is taken as the test set to verify classification performance. This step is repeated 30 times to avoid possible biases. The accuracy and standard deviation of the classification experiment are reported. Moreover, the common representations of our model are achieved in an unsupervised manner, so we further conduct clustering experiments to evaluate the performance of the representations. Finally, a kernel alignment experiment is performed in order to verify the superiority of the proposed method over other comparison methods using kernel method.

In the operation, subjects are randomly chosen as the missing parts in the incomplete modality. The clustering accuracy, classification accuracy and kernel alignment of all methods are mainly examined. For the hyperparameters in the objective function, we tune the algorithm to the best performance and set λ_1, λ_2, and λ_3 to 0.01, 1 and 1e-5, respectively.

Evaluation metrics. On one hand, clustering accuracy reflects the quality of the representation learned by the proposed method. The accuracy used in our clustering experiments is defined as:

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} \Gamma(\mathbf{s}_i, map(\mathbf{r}_i))}{n}, \tag{20}$$

where $\mathbf{r}_i$ and $\mathbf{s}_i$ are cluster label and ground-truth label of sample $\mathbf{x}_i$ respectively, and $\Gamma(x, y) = 1$ if $x = y$, or otherwise $\Gamma(x, y) = 0$. $map(\cdot)$ is a permutation map function mapping the cluster label into the class labels. Then the best map can be obtained by Kuhn-Munkres algorithm.

On the other hand, the kernel alignment index is used to evaluate the degree of alignment of different kernel matrices after completion. Referring to the work (Wang et al., 2015a), the kernel alignment index in experiments is defined as:

$$\mathrm{S}(\mathbf{K}, \mathbf{K}^*) = \frac{\langle \mathbf{K}, \mathbf{K}^* \rangle_{\mathrm{F}}}{\sqrt{\langle \mathbf{K}, \mathbf{K} \rangle_{\mathrm{F}} \langle \mathbf{K}^*, \mathbf{K}^* \rangle_{\mathrm{F}}}}, \tag{21}$$

where $\mathbf{K}$ and $\mathbf{K}^*$ denote the kernel matrix. $\langle \mathbf{K}, \mathbf{K}^* \rangle_{\mathrm{F}}$ is the Frobenius inner product between two matrices. This alignment index $\mathrm{S}(\mathbf{K}, \mathbf{K}^*)$ can be treated as a similarity index based on the cosine of the angle, ranging between -1 and 1. A greater value indicates a higher similarity between the two kernel matrices. In practice, $\mathbf{K}^*$ is replaced with $\mathbf{Y}\mathbf{Y}^T$, where $\mathbf{Y}$ is a one-hot type label matrix. Therefore, $\mathbf{K}^*$ can be regarded as the target kernel matrix, which reflects the ideal similarity between samples (that is the similarity between samples with the same label is 1, or otherwise 0). It is worth noting that a good kernel matrix can exactly describe relationships between the samples and contributes to the later task.

5.4. Results and analysis

The clustering and classification results of various methods at different missing rates are reported in Table 2 and Fig. 5, respectively.

The vertical analysis of the data in Table 2 suggests that the clustering ACC of the proposed method is higher than that of other methods at all missing rates. Meanwhile, the standard deviation of the proposed method is lower than that of other methods. From the horizontal perspective, the clustering ACC of most compared methods fluctuates greatly as the missing rate increases. While, the proposed method can maintain stable. Hence, the proposed method can learn better common representations with the incomplete view, and thus is of importance for AD diagnosis. It should be noted that the K-Com and iMVWL methods are not considered in clustering experiments since they are not a representation learning approach, and K-means is not applicable to them.

We report the classification results of comparison methods at different missing rates as shown in Fig. 5. The proposed method is represented with red bars, and other methods are shown in blue. It is observed that the proposed method has the best and stable classification performance in all experiments at different missing rates, even at 50% missing rate, indicating that the proposed method has good robustness. The classification results obtained by two-category experiments are shown in Fig. 5 from row 2 to row 4. From the figure, the proposed method exhibits the best performance. Especially in the classification of AD/NC, our algorithm achieves the accuracy of 84.85% at the missing rate of 50%. Specifically, we can see that the DCCAE method performs better than other CCA-based methods in most experiments. One of the main possible reasons is that DCCAE maintains self information of views and correlations between the views. The K-Com method has poor performance due to its simple strategy of exploring the relationship between views (i.e., directly linear fusion). The MDcR method

[3] https://www.csie.ntu.edu.tw/cjlin/libsvm/

Table 2
Clustering results with different methods at the missing rates of 10% to 50%.

Method	10%	20%	30%	40%	50%
FeaCon	0.454±0.012	0.457±0.027	0.462±0.022	0.450±0.024	0.449±0.021
CCA	0.460±0.015	0.455±0.016	0.457±0.016	0.458±0.017	0.455±0.016
KCCA	0.457±0.020	0.468±0.010	0.443±0.022	0.424±0.033	0.424±0.011
DCCA	0.475±0.051	0.453±0.047	0.472±0.025	0.449±0.040	0.451±0.008
DCCAE	0.462±0.031	0.463±0.039	0.455±0.32	0.473±0.035	0.466±0.038
MDcR	0.445±0.027	0.435±0.027	0.433±0.021	0.425±0.026	0.422±0.032
MCIV	0.481±0.013	0.455±0.021	0.438±0.011	0.422±0.015	0.427±0.033
AEMVC	**0.523**±0.012	**0.516**±0.005	**0.519**±0.002	**0.518**±0.010	**0.520**±0.012

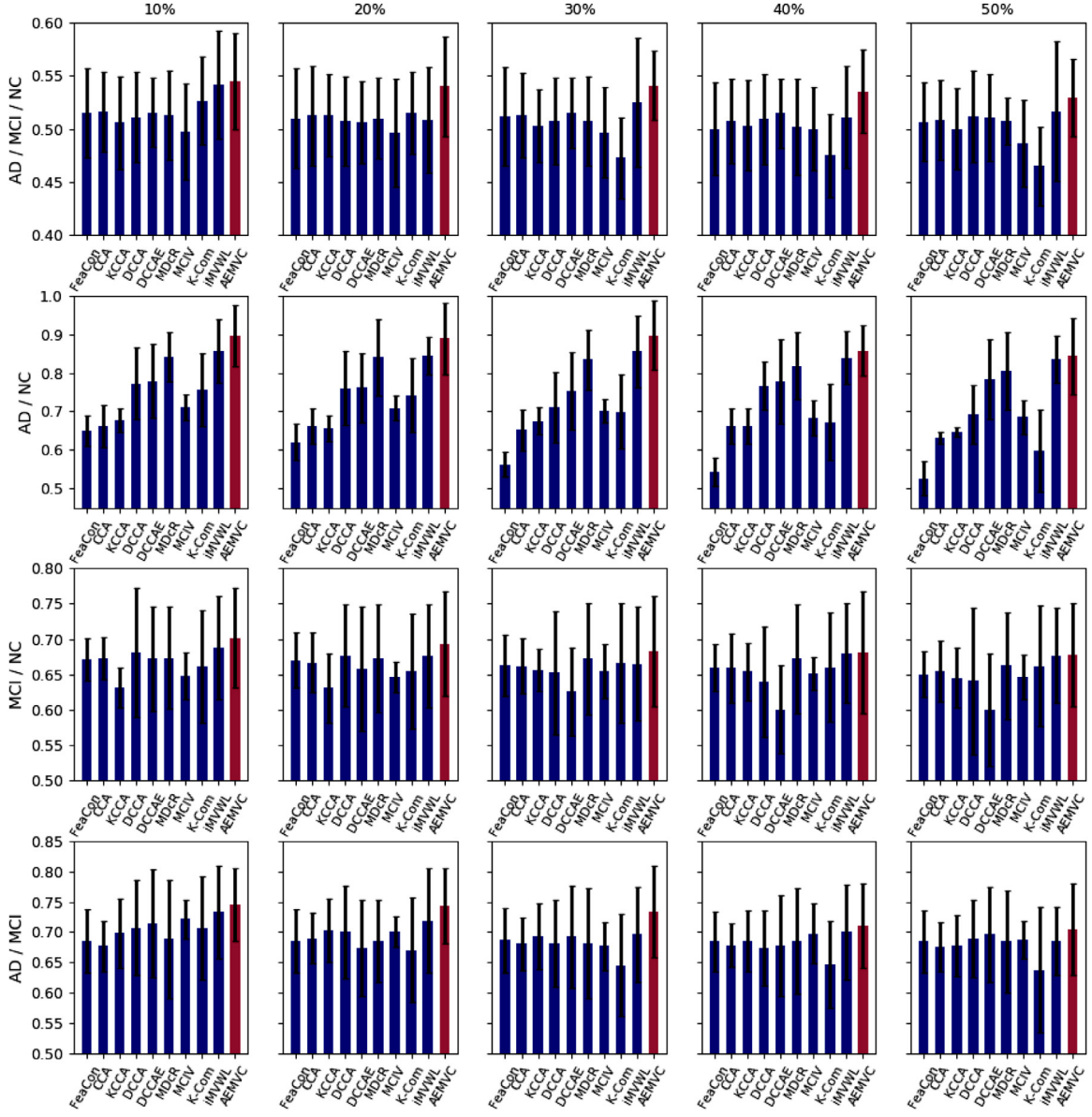


Fig. 5. Classification results achieved by different methods at different missing rates, where each row represents a classification task (i.e., AD/MCI/NC, AD/NC, MCI/NC, AD/MCI) and each column has the same missing rate. The red bar represents the proposed method, and blue bars represent other compared methods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

achieves a better performance than other multi-view methods with complete data, because it introduces the kernel method and the correlation constraint between views. In addition, as a multi-view method with incomplete view, the MCIV method complements the missing information in the kernel space, and has a better accuracy than other methods in MCI/NC classification. This verifies the effectiveness of kernel matrix completion. It is worthwhile noting that iMVWL achieves the second-best result, because it simultaneously learns a shared subspace, local label correlations and a predictor. Yet it does not consider the deeper correlation of samples in higher-dimensional space (e.g., kernel space). The proposed method integrates the advantages of DCCAE and MCIV methods, complements the missing data in the kernel space while taking into account the structural information of data and the inherent

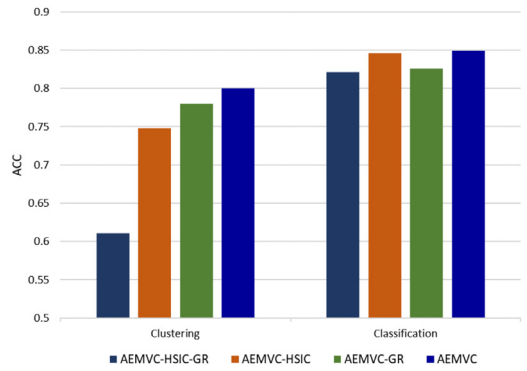
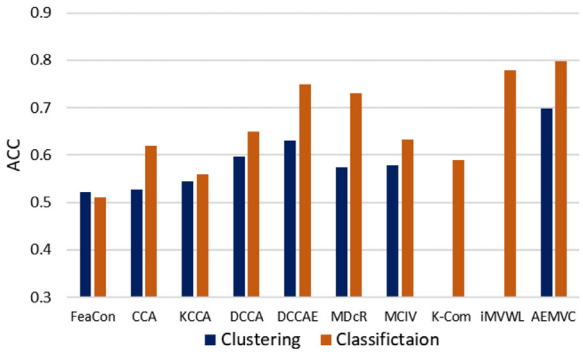


Fig. 6. Ablation experimental results (AD/NC task at the missing rate of 50%) .

Table 3
Clustering and classification results achieved by different methods in AD/NC task at the missing rate of 50% (PET data is complete and MRI data is incomplete).

Method	Clustering ACC	Classification ACC
FeaCon	0.520±0.026	0.530±0.105
CCA	0.533±0.021	0.655±0.099
KCCA	0.549±0.021	0.616±0.085
DCCA	0.614±0.013	0.682±0.108
DCCAE	0.636±0.027	0.772±0.071
MDcR	0.603±0.024	0.782±0.085
MCIV	0.571±0.022	0.664±0.075
K-Com	-	0.548±0.106
iMVWL	-	0.828±0.079
AEMVC	**0.654**±0.021	**0.836**±0.077



(a) MRI data is complete and PET data is missing

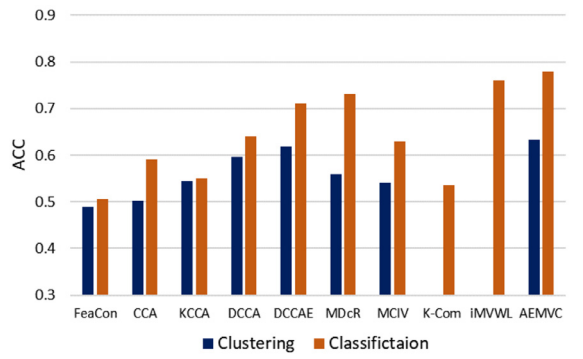


Fig. 7. Clustering and classification results achieved by different methods in AD/NC task when one view is completely missing.

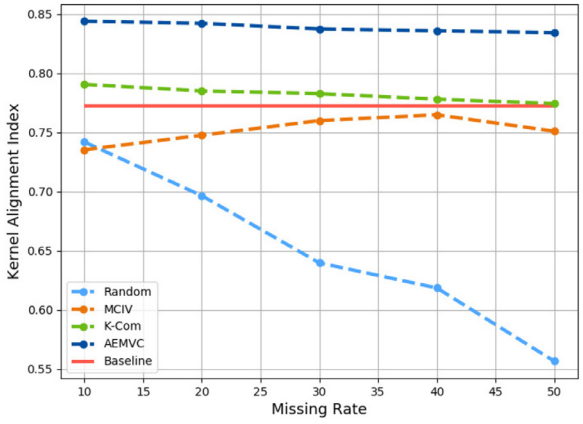


Fig. 8. Kernel alignment indices with different methods.

association between multiple views, and shows the best performance. As the missing rate changes from 50% to 10%, the accuracy of the proposed method improves by about 4%. Although the best performance (black lines) changes with some minor fluctuations, the mean values (red bars) consistently improve with less missing data and achieve the best performances.

To further verify the effectiveness of HSIC constraint and graph regularization, we report the clustering and classification accuracies of the variants of the proposed AEMVC method in AD/NC task at the missing rate of 50% in Fig. 6. The variants of AEMVC are as follows: **AEMVC-HSIC-GR** represents the AEMVC method without the HSIC and graph regularization constraints, **AEMVC-HSIC** represents the AEMVC method without HSIC constraint and **AEMVC-GR** denotes the AEMVC method without graph regularization. It can be seen that the AEMVC method achieves the best performance with two constraints.

Furthermore, we conduct more experiments in different settings to verify the effectiveness of the proposed method. When PET data is complete and MRI data is incomplete, the accuracies of clustering and classification with different methods in AD/NC task at the missing rate of 50% are shown in Table 3. It can be observed that AEMVC has the best performance compared with other methods in this scenario. Moreover, we compare different methods in AD/NC task when the missing rate is 100% as shown in Fig. 7. Fig. 7 (a) shows the performance when MRI data is complete and PET data is completely missing and Fig. 7 (b) shows the performance when PET data is complete and MRI data is completely missing. It can be observed that when a modality is completely missing, the classification accuracy of most methods including AEMVC drops by nearly 10%. However, the proposed method still has the best performance in this extreme case.

Fig. 8 shows the comparison of kernel alignment indices for different methods in the AD/NC classification task at different missing rates. The red line represents the baseline. The blue line denotes a random filling method used by KCCA, which indicates the alignment degree between the incomplete $\mathbf{K}_2$ obtained by filling the missing parts with random values. The K-Com method represented by the green line also uses random values to fill the missing information, but it calculates the kernel alignment index with the weighted kernel matrix of two views. The kernel alignment indices of the AEMVC and MCIV method both are calculated by the complemented $\mathbf{K}_2$. Compared with other methods, the proposed method has the highest degree of kernel alignment at any missing rate, implying that it can better mine the relationship between samples than other methods and is therefore more conducive to AD diagnosis.

To determine the dimensions of the learned common representation, as shown in Fig. 9, we report the classification accuracy on different dimensions of the common representation in AD/NC task

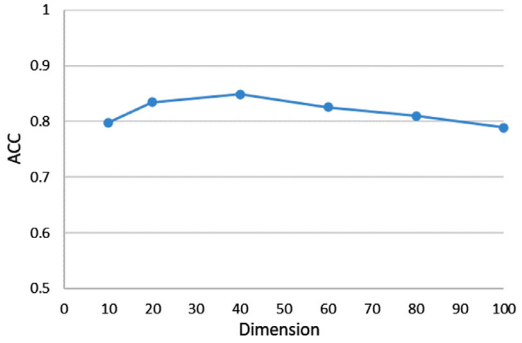


Fig. 9. Classification results on different dimensions of the common representation in AD/NC task at the missing rate of 50% .

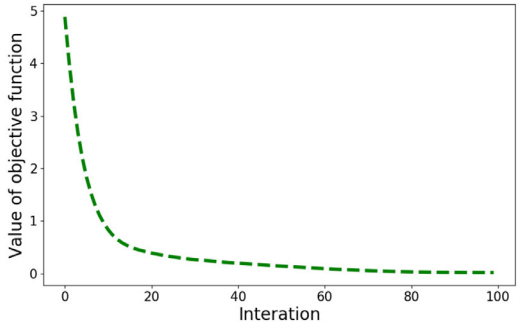


Fig. 10. Convergence curve of the objective function.

at the missing rate of 50%. According to observations, when the common dimension is 40, the proposed method achieves the best performance. Moreover, To verify the convergence of the proposed method, as shown in Fig. 10, we report the objective function value at each iteration step with 50% missing rate in AD/NC classification task. It is obvious to be observed that objective function value decreases in the first 30 iterations and then gradually converges within 100 iterations.

6. Conclusion

This paper studies an multi-modal representation learning problem for Alzheimers disease diagnosis with incomplete modalities and proposes an Auto-Encoder based Multi-View missing data Completion framework(AEMVC). The original complete view is mapped to a latent space through an auto-encoder network framework. Then, the latent representations learned from the complete view are used to complement the kernel matrix of the incomplete view while graph regularization and HSIC constraints are adopted to maintain the structural information of original data and the inherent association between views. Finally, Kernel CCA is applied to the learned kernel matrix to obtain the common representation. The experimental comparison of all methods verifies the best performance of the proposed method on ADNI datasets. There are many other directions to explore in the future, such as mining commonality of the same group of diseased subjects using unsupervised clustering, and improving the flexibility of the algorithm by taking arbitrary missing (i.e., each modality can be incomplete) into consideration.

Declaration of Competing Interest

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

CRediT authorship contribution statement

Yanbei Liu: Conceptualization, Methodology, Writing - review & editing. **Lianxi Fan:** Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing. **Changqing Zhang:** Conceptualization, Methodology, Data curation, Writing - review & editing. **Tao Zhou:** Methodology, Writing - review & editing, Data curation. **Zhitao Xiao:** Supervision, Writing - review & editing. **Lei Geng:** Investigation, Validation. **Dinggang Shen:** Supervision, Data curation, Writing - review & editing.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (No. 61901297), Tianjin Science and Technology Major Projects and Engineering (No. 17ZXSCSY00060, 17ZXSCSY00090), Natural Science Foundation of Tianjin of China (19JCYBJC15200), Program for Innovative Research Team in University of Tianjin (No. TD13-5034).

References

Akaho, S., 2006. A kernel method for canonical correlation analysis. arXiv preprint cs/0609071.
Andrew, G., Arora, R., Bilmes, J., Livescu, K., 2013. Deep canonical correlation analysis. In: International Conference on Machine Learning, pp. 1247–1255.
Bickel, S., Scheffer, T., 2004. Multi-view clustering.. In: ICDM, 4, pp. 19–26.
Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2 (2), 121–167.
Cai, J.-F., Candès, E.J., Shen, Z., 2010. A singular value thresholding algorithm for matrix completion. SIAM J. Optim. 20 (4), 1956–1982.
Cai, L., Wang, Z., Gao, H., Shen, D., Ji, S., 2018. Deep adversarial learning for multi-modality missing data completion. In: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1158–1166.
Cai, X., Nie, F., Huang, H., 2013. Multi-view k-means clustering on big data. In: International Joint Conference on Artificial Intelligence.
Chen, X., Chen, S., Xue, H., Zhou, X., 2012. A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data. Pattern Recognit 45 (5), 2005–2018.
Chen, X., Zhang, H., Gao, Y., Wee, C.Y., Li, G., Shen, D., Initiative, A.D.N., 2016. High-order resting-state functional connectivity network for mci classification. Human Brain Mapping 37, 3282–3296.
De Sa, V.R., 2005. Spectral clustering with two views. In: ICML Workshop on Learning with Multiple Views, pp. 20–27.
Elhamifar, E., Vidal, R., 2013. Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans Pattern Anal Mach Intell 35 (11), 2765–2781.
Fan, Y., Rao, H., Hurt, H., Giannetta, J., Korczykowski, M., Shera, D., Avants, B.B., Gee, J.C., Wang, J., Shen, D., 2007. Multivariate examination of brain abnormality using both structural and functional MRI. NeuroImage 36, 1189–1199.
Ghazi, M.M., Nielsen, M., Pai, A., Cardoso, M.J., Modat, M., Ourselin, S., Sørensen, L., Initiative, A.D.N., et al., 2019. Training recurrent neural networks robust to incomplete data: application to Alzheimers disease progression modeling. Med Image Anal 53, 39–46.
Gretton, A., Bousquet, O., Smola, A., Schölkopf, B., 2005. Measuring statistical dependence with hilbert-schmidt norms. In: International Conference on Algorithmic Learning Theory, pp. 63–77.
Hardoon, D.R., Shawe-Taylor, J., 2003. Kcca for different level precision in content-based image retrieval. International Workshop on Content-Based Multimedia Indexing, IRISA, Rennes, France.
Hardoon, D.R., Szedmak, S., Shawe-Taylor, J., 2004. Canonical correlation analysis: an overview with application to learning methods. Neural Comput 16 (12), 2639–2664.
Hotelling, H., 1992. Relations between two sets of variates. In: Breakthroughs in Statistics, pp. 162–190.
Hu, H., Lin, Z., Feng, J., Zhou, J., 2014. Smooth representation clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3834–3841.
Jia, H., Yap, P.T., Shen, D., 2012. Iterative multi-atlas-based multi-image segmentation with tree-based registration. NeuroImage 59, 422–430.
Kabani, N.J., MacDonald, D.J., Holmes, C.J., Evans, A.C., 1998. 3D anatomical atlas of the human brain. Neuroimage 7 (4), S717.
Kumar, A., Daumé, H., 2011. A co-training approach for multi-view spectral clustering. In: International Conference on Machine Learning (ICML-11), pp. 393–400.
Lei, B., Chen, S., Ni, D., Wang, T., 2016. Discriminative learning for alzheimer's disease diagnosis via canonical correlation analysis and multimodal fusion. Frontiers in Aging Neuroscience 8, 77.
Li, S.-Y., Jiang, Y., Zhou, Z.-H., 2014. Partial multi-view clustering. In: AAAI Conference on Artificial Intelligence.
Lian, C., Liu, M., Zhang, J., Shen, D., 2018. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural mri. IEEE Trans Pattern Anal Mach Intell.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y., 2012. Robust recovery of subspace structures by low-rank representation. IEEE Trans Pattern Anal Mach Intell 35 (1), 171–184.
Liu, M., Zhang, J., Yap, P.-T., Shen, D., 2016. Diagnosis of Alzheimers disease using view-aligned hypergraph learning with incomplete multi-modality data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 308–316.
Liu, X., Zhu, X., Li, M., Wang, L., Tang, C., Yin, J., Shen, D., Wang, H., Gao, W., 2018. Late fusion incomplete multi-view clustering. IEEE Trans Pattern Anal Mach Intell.
Liu, X., Zhu, X., Li, M., Wang, L., Zhu, E., Liu, T., Kloft, M., Shen, D., Yin, J., Gao, W., 2019. Multiple kernel k-means with incomplete kernels. IEEE Trans Pattern Anal Mach Intell.
Liu, Y., Wang, X., Wu, S., Xiao, Z., 2020. Independence promoted graph disentangled networks.. In: AAAI, pp. 4916–4923.
Mazumder, R., Hastie, T., Tibshirani, R., 2010. Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research 11 (Aug), 2287–2322.
Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y., 2011. Multimodal deep learning. In: International Conference on Machine Learning (ICML-11), pp. 689–696.
Perrin, R.J., Fagan, A.M., Holtzman, D.M., 2009. Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. Nature 461 (7266), 916–922.
Shawe-Taylor, J., Cristianini, N., et al., 2004. Kernel methods for pattern analysis. Cambridge University Press.
Shi, Y., Suk, H.-I., Gao, Y., Lee, S.-W., Shen, D., 2019. Leveraging coupled interaction for multimodal Alzheimers disease diagnosis. IEEE Trans Neural Netw Learn Syst 31 (1), 186–200.
Song, L., Smola, A., Gretton, A., Borgwardt, K.M., Bedo, J., 2007. Supervised feature selection via dependence estimation. In: International Conference on Machine Learning, pp. 823–830.
Steinwart, I., et al., 2015. Fully adaptive density-based clustering. The Annals of Statistics 43 (5), 2132–2167.
Sun, S., 2013. A survey of multi-view machine learning. Neural Computing and Applications 23 (7–8), 2031–2038.
Tan, Q., Yu, G., Domeniconi, C., Wang, J., Zhang, Z., 2018. Incomplete multi-view weak-label learning.. In: IJCAI, pp. 2703–2709.
Tao, Z., Liu, H., Li, S., Ding, Z., Fu, Y., 2017. From ensemble clustering to multi-view clustering. In: International Joint Conference on Artificial Intelligence.
Tran, L., Liu, X., Zhou, J., Jin, R., 2017. Missing modalities imputation via cascaded residual autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1405–1414.
Trivedi, A., Rai, P., Daumé III, H., DuVall, S.L., 2010. Multiview clustering with incomplete views. NIPS Workshop, 224.
Von Luxburg, U., 2007. A tutorial on spectral clustering. Stat Comput 17 (4), 395–416.
Wang, H., Weng, C., Yuan, J., 2014. Multi-feature spectral clustering with minimax optimization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4106–4113.
Wang, T., Zhao, D., Tian, S., 2015. An overview of kernel alignment and its applications. Artif Intell Rev 43 (2), 179–192.
Wang, W., Arora, R., Livescu, K., Bilmes, J., 2015. On deep multi-view representation learning. In: International Conference on Machine Learning, pp. 1083–1092.
Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack Jr, C.R., Jagust, W., Morris, J.C., et al., 2017. Recent publications from the Alzheimer's disease neuroimaging initiative: reviewing progress toward improved ad clinical trials. Alzheimer's & Dementia 13 (4), e1–e85.
Wu, G., Qi, F., Shen, D., 2006. Learning-based deformable registration of MR brain images. IEEE Trans Med Imaging 25, 1145–1157.
Xiao, M., Guo, Y., 2014. Feature space independent semi-supervised domain adaptation via kernel matching. IEEE Trans Pattern Anal Mach Intell 37 (1), 54–66.
Xu, J., Han, J., Nie, F., 2016. Discriminatively embedded k-means for multi-view clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5356–5364.
Yang, S., Li, L., Wang, S., Zhang, W., Huang, Q., 2017. A graph regularized deep neural network for unsupervised image representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1203–1211.
Zhang, C., Adeli, E., Zhou, T., Chen, X., Shen, D., 2018. Multi-layer multi-view classification for Alzheimers disease diagnosis. In: AAAI Conference on Artificial Intelligence.
Zhang, C., Fu, H., Hu, Q., Zhu, P., Cao, X., 2016. Flexible multi-view dimensionality co-reduction. IEEE Trans. Image Process. 26 (2), 648–659.
Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., Initiative, A.D.N., et al., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55 (3), 856–867.
Zhang, J., Gao, Y., Gao, Y., Munsell, B.C., Shen, D., 2016. Detecting anatomical landmarks for fast Alzheimers disease diagnsosis. IEEE Trans Med Imaging 35 (12), 2524–2533.
Zhou, T., Liu, M., Thung, K.-H., Shen, D., 2019. Latent representation learning for Alzheimers disease diagnosis with incomplete multi-modality neuroimaging and genetic data. IEEE Trans Med Imaging 38 (10), 2411–2422.
Zhu, X., Suk, H.-I., Lee, S.-W., Shen, D., 2016. Canonical feature selection for joint regression and multi-class identification in Alzheimers disease diagnosis. Brain Imaging Behav 10 (3), 818–828.
Zhu, X., Suk, H.-I., Shen, D., 2014. Multi-modality canonical feature selection for Alzheimers disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 162–169.
Zien, A., Ong, C.S., 2007. Multiclass multiple kernel learning. In: International Conference on Machine Learning, pp. 1191–1198.