Deep Variational Incomplete Multi-View Clustering: Exploring Shared Clustering Structures

Gehui Xu[1], Jie Wen[1*], Chengliang Liu[1], Bing Hu[1], Yicheng Liu[1], Lunke Fei[2], Wei Wang[1]

[1]Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen, China
[2]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
tkkxgh@foxmail.com, jiewen_pr@126.com, liucl1996@163.com, {200111412,190110406}@stu.hit.edu.cn, flksxm@126.com, wangwei2019@hit.edu.cn

Abstract

Incomplete multi-view clustering (IMVC) aims to reveal shared clustering structures within multi-view data, where only partial views of the samples are available. Existing IMVC methods primarily suffer from two issues: 1) Imputation-based methods inevitably introduce inaccurate imputations, which in turn degrade clustering performance; 2) Imputation-free methods are susceptible to unbalanced information among views and fail to fully exploit shared information. To address these issues, we propose a novel method based on variational autoencoders. Specifically, we adopt multiple view-specific encoders to extract information from each view and utilize the Product-of-Experts approach to efficiently aggregate information to obtain the common representation. To enhance the shared information in the common representation, we introduce a coherence objective to mitigate the influence of information imbalance. By incorporating the Mixture-of-Gaussians prior information into the latent representation, our proposed method is able to learn the common representation with clustering-friendly structures. Extensive experiments on four datasets show that our method achieves competitive clustering performance compared with state-of-the-art methods.

Introduction

Multi-view data widely exist in real-world application scenarios, where data are often collected from different sources or by different sensors (Jiang et al. 2022b; Yan et al. 2021). For instance, a 3D object can be described from different angles; an image can be characterized by heterogeneous feature descriptions, e.g., HOG, SIFT, and LBP. Multi-view learning focuses on developing methods to effectively utilize multi-view data for a range of tasks (Jiang et al. 2022a, 2023). A fundamental challenge within this domain is multi-view clustering, which aims to partition multi-view data into distinct groups in an unsupervised manner by exploiting the consistent and complementary characteristics inherent across different views.

Existing multi-view clustering methods heavily rely on the assumption that samples have all views (Li and He 2020; Wang et al. 2021b; Xu et al. 2022b). However, such an assumption may not always hold in practical applications where some samples may only contain partial views due to unstable sensors or different acquisition costs. To this end, many incomplete multi-view clustering (IMVC) methods have been proposed and can be roughly categorized into two groups: imputation-based methods and imputation-free methods.

To address the incomplete learning issue, most IMVC methods typically impute missing views before exploring clustering information, *i.e.*, imputation-based IMVC methods (Wen et al. 2019; Liu 2021; Yin and Sun 2021; Wen et al. 2021c; Liu et al. 2020). For traditional IMVC methods, such imputation strategies are generally built on matrix factorization, multiple kernel learning, and graph learning (Wen et al. 2022). For example, Wen et al. proposed a unified embedding alignment framework based on matrix factorization to jointly reconstruct the missing views and learn the consensus representation (Wen et al. 2019). Although imputation-based traditional IMVC methods are known for their interpretability and physical significance, their effectiveness is constrained by the limited representation extraction capacity of shallow models. Furthermore, these methods commonly involve high computational complexity and struggle with mixed data types like text, audio, and video. In recent years, with the advancements of deep learning, many deep incomplete multi-view clustering (DIMVC) methods have also been developed. By leveraging the superior representation extraction and generalization capability of deep neural networks, many imputation-based DIMVC methods have shown significant performance improvements over traditional ones (Tang and Liu 2022; Liu et al. 2023; Lin et al. 2021; Xu et al. 2022a, 2023). Intuitively, the effectiveness of imputation-based IMVC methods largely hinges on the quality of the imputed missing views. While some works visualize recovered missing views to demonstrate their effectiveness in missing view recovery, regrettably, none can rigorously prove their ability to accurately restore missing views from a theoretical perspective. Accurately estimating missing data is challenging without ground-truth, especially when the proportion of missing views is large. Therefore, the imputation process carries a high risk of degrading clustering performance.

Different from the aforementioned imputation-based IMVC methods, imputation-free IMVC methods directly

*Corresponding author: Jie Wen.

learn clustering information from available views, thereby avoiding noise that could be introduced by potentially inaccurate imputation (Wen, Xu, and Liu 2018; Zhuge et al. 2019; Zhao et al. 2018; Wen et al. 2021b). Most of these methods aim to transform the original incomplete multi-view data into a common space and obtain a shared latent representation by fusing the available views (Hu and Chen 2019; Wen et al. 2020; Li, Wan, and He 2021). In these methods, concatenation and weighting summation are two popular fusion strategies used to derive the common representation from the view-specific representations associated with each view. However, these fusion strategies exhibit limitations: On the one hand, such transformations are challenging to capture the complex nonlinear correlations across views. On the other hand, since different views may have different numbers of missing instances, these fusion strategies could introduce bias and result in low-quality common representations.

To address the above limitations, we propose an imputation-free DIMVC method based on variational autoencoders (VAEs). Compared to existing DIMVC methods, which are generally based on conventional autoencoders, the proposed method offers more freedom in capturing the nonlinear correlations across views at the latent space. Extensive experiments demonstrate the superiority of the proposed method over other advanced methods. The major contributions of this paper are summarized as follows:

- We propose a flexible DIMVC method based on VAEs. To the best of our knowledge, this is the first imputation-free work based on VAE in the field of incomplete multi-view clustering. Our method is applicable to all kinds of incomplete multi-view data clustering tasks with arbitrary missing views.
- We employ the Product-of-Experts (Hinton 2002) approach coupled with a coherence constraint to effectively address the incomplete learning issue and mitigate the clustering bias caused by the information imbalance across multiple views. By incorporating the shared clustering assignment learning with coherence loss, our method can simultaneously obtain common representations with clustering-friendly structure and optimal clustering results for incomplete multi-view data.

Preliminaries and Related Works

Problem Definition of IMVC

Let $\{\{\boldsymbol{x}_i^v\}_{v=1}^V\}_{i=1}^N$ represent an incomplete multi-view dataset consisting of N samples with V views. Here, $\boldsymbol{x}_i^v \in \mathbb{R}^{D_v}$ denotes the v-th view of the i-th sample. The available views of samples is described by an indicator matrix $\mathbf{M} \in \{0,1\}^{N\times V}$, where $\mathbf{M}_{iv} = 1$ indicates that the v-th view of the i-th sample is available; otherwise $\mathbf{M}_{iv} = 0$. To improve readability, we will omit the subscript i and use $\{\boldsymbol{x}^v\}_{v=1}^V$ to represent a general sample with multiple views in the remaining part. The target of IMVC is to discover the shared clustering structure across views, *i.e.*, to group the N samples, with locations of missing views indicated by $\mathbf{M}$, into K disjoint clusters.

Deep Incomplete Multi-view Clustering

Inspired by the expressive power of deep neural networks, many DIMVC methods have been developed, which can be categorized into two groups: imputation-based methods and imputation-free methods. (1) Imputation-based methods employ various strategies to impute missing views and then conduct clustering on the imputed complete multi-view dataset. Using deep generative models, Xu et al. (Xu et al. 2021a) and Wang et al. (Wang et al. 2021a) trained Generative Adversarial Networks to recover missing views and then learned the common representations for all views. Based on information theory, Lin et al. proposed a framework to maximize mutual information between different views of samples and used the additional prediction networks to predict missing views (Lin et al. 2022). Tang et al. utilized contrastive learning objectives to ensure that samples within the same cluster have similar latent representations, subsequently filling missing views based on similarity (Tang and Liu 2022). (2) Imputation-free methods commonly focus on designing distinct strategies to aggregate available view-specific representations and directly obtain the clustering result from the aggregated representation (Wen et al. 2021b; Xu et al. 2022a). For instance, Xu et al. designed a weighting summation approach that adaptively weights view-specific representations based on the representation separability of different views (Xu et al. 2023).

Variational Autoencoders

Variational Autoencoders (VAEs) (Kingma and Welling 2013) offer an unsupervised approach to estimate latent distributions of data. By incorporating different prior distributions of latent representations and various generative assumptions, VAEs can encode flexible latent representations. For example, some researchers assumed a Mixture-of-Gaussians (MoG) prior, enabling simultaneous representation learning and clustering (Jiang et al. 2017; Dilokthanakul et al. 2016). Yang et al. introduced graph constraints on the latent representation, enhancing the clustering structure within representations (Yang et al. 2019). Dupont designed a disentangled latent space, learning continuous and discrete representations together (Dupont 2018). Numerous studies have extended VAEs to multi-view representation learning, focusing on employing reasonable principles to aggregate information from single views and derive the joint representation (Suzuki, Nakayama, and Matsuo 2016; Wu and Goodman 2018; Shi et al. 2019; Sutter, Daunhawer, and Vogt 2020a). Sutter et al. (Sutter, Daunhawer, and Vogt 2020b) and Hwang et al. (Hwang et al. 2021) enhanced the expressiveness of the shared representation by adopting the adaptive prior and introducing an auxiliary objective function, respectively. Some researchers designed partitioned latent spaces, effectively disentangling the shared and private information in multi-view data (Xu et al. 2021b; Hsu and Glass 2018; Tsai et al. 2018). Yin et al. employed an adaptive weighting strategy to aggregate information from each view and assume a MoG prior for latent representation, thereby acquiring clustering-friendly shared representations (Yin, Huang, and Gao 2020).

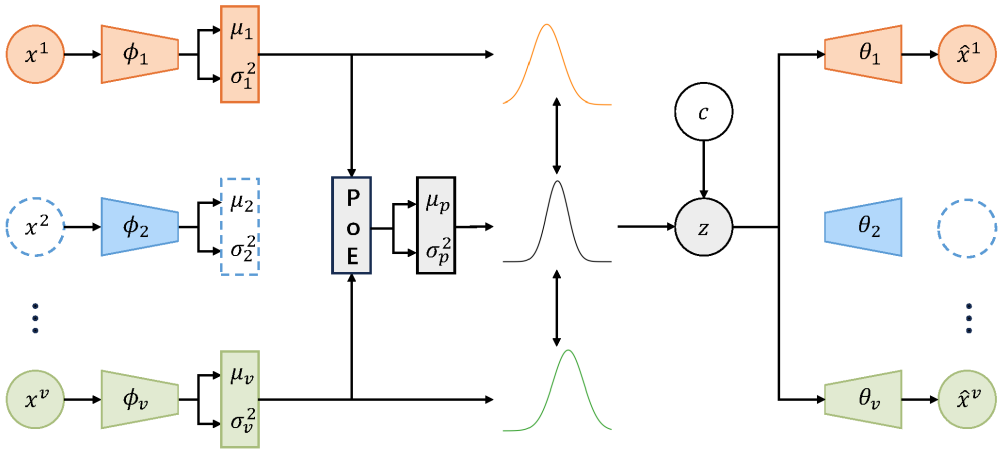


Figure 1: The framework of the proposed method. In the inference process, the Product-of-Experts approach derives a common representation by aggregating the representations of each available view. The coherence objective function maintains information consistency by reducing Kullback-Leibler divergence between these representations. In the generative process, each observed view is generated from the latent representation, which incorporates a clustering structure.

Proposed Method

In this section, we first discuss the deep multi-view Gaussian mixture model. Then, we present our inference method for the incomplete multi-view data. Similar to VAEs, we transform the issue of approximate inference into an optimization problem and provide a detailed introduction to the overall objective function. The architecture of the proposed method, called Deep Variational Incomplete Multi-View Clustering (DVIMC), is shown in Figure 1.

Deep Multi-view Gaussian Mixture Model

In our method, we assume that the multi-view data $\{\boldsymbol{x}^v\}_{v=1}^V$ is generated from a random process consisting of three steps (Jiang et al. 2017). First, the shared clustering assignment $\boldsymbol{c}$, represented as the binary vector $\boldsymbol{c} \in \{0,1\}^K$, is sampled from a categorical distribution parameterized by $\boldsymbol{\pi} \in \mathbb{R}_+^K$. Here, π_k represents the prior probability of the k-th class and satisfies $\sum_{k=1}^K \pi_k = 1$. Second, the common continuous variable $\boldsymbol{z} \in \mathbb{R}^D$ is sampled from a Gaussian distribution conditioned on $\boldsymbol{c}$. Finally, each view of the sample is generated from its respective view-specific distribution conditioned on $\boldsymbol{z}$. The generative process can be summarized as follows:

$$p(c_k = 1 \mid \boldsymbol{\pi}) = \pi_k \tag{1}$$

$$p(\boldsymbol{z} \mid c_k = 1) = \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2 \mathbf{I}) \tag{2}$$

$$p_{\boldsymbol{\theta}_v}(\boldsymbol{x}^v \mid \boldsymbol{z}) = \begin{cases} \textbf{Bernoulli}\left(\boldsymbol{\mu}_{\boldsymbol{x}^v}\right) & \text{binary} \\ \mathcal{N}\left(\boldsymbol{x}^v \mid \boldsymbol{\mu}_{\boldsymbol{x}^v}, \boldsymbol{\sigma}_{\boldsymbol{x}^v}^2 \mathbf{I}\right) & \text{real-valued} \end{cases} \tag{3}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k^2$ represent the mean and variance of the Gaussian distribution corresponding to cluster k in the latent space, $\mathbf{I}$ denotes the identity matrix. If $\boldsymbol{x}^v$ is real-valued, we define $p_{\boldsymbol{\theta}_v}(\boldsymbol{x}^v \mid \boldsymbol{z})$ as the multivariate Gaussian distribution with diagonal covariance, $\left[\boldsymbol{\mu}_{\boldsymbol{x}^v}; \boldsymbol{\sigma}_{\boldsymbol{x}^v}^2\right] = g(\boldsymbol{z}; \boldsymbol{\theta}_v)$; if $\boldsymbol{x}^v$ is binary, we define $p_{\boldsymbol{\theta}_v}(\boldsymbol{x}^v \mid \boldsymbol{z})$ as the multivariate Bernoulli distribution, $\boldsymbol{\mu}_{\boldsymbol{x}^v} = g(\boldsymbol{z}; \boldsymbol{\theta}_v)$. The function $g(\boldsymbol{z}; \boldsymbol{\theta}_v)$ denotes a neural network with trainable parameters $\boldsymbol{\theta}_v$, called view-specific decoder.

According to the generative process described above, the joint probability for a multi-view sample is formulated as:

$$p(\{\boldsymbol{x}^v\}_{v=1}^V, \boldsymbol{z}, \boldsymbol{c}) = p_{\boldsymbol{\theta}}(\{\boldsymbol{x}^v\}_{v=1}^V \mid \boldsymbol{z}) p(\boldsymbol{z} \mid \boldsymbol{c}) p(\boldsymbol{c}). \tag{4}$$

In the multi-view setting, we assume that each view of a sample is conditionally independent given the common latent variable $\boldsymbol{z}$. That is, $p_{\boldsymbol{\theta}}(\{\boldsymbol{x}^v\}_{v=1}^V \mid \boldsymbol{z}) = \prod_{v=1}^V p_{\boldsymbol{\theta}_v}(\boldsymbol{x}^v \mid \boldsymbol{z})$. This factorization allows us to evaluate the likelihood only for the available views indicated by $V_a = \{v \mid \mathbf{M}_{iv} = 1\}$. Consequently, the joint probability for incomplete data can be written as:

$$p(\{\boldsymbol{x}^v\}_{v=1}^V, \boldsymbol{z}, \boldsymbol{c}) = p(\boldsymbol{z} \mid \boldsymbol{c}) p(\boldsymbol{c}) \prod_{v \in V_a} p_{\boldsymbol{\theta}_v}(\boldsymbol{x}^v \mid \boldsymbol{z}). \tag{5}$$

Inference Method

Our goal is to learn the common representation as well as the corresponding clustering assignment, *i.e.*, to obtain the posterior $p(\boldsymbol{z}, \boldsymbol{c} \mid \{\boldsymbol{x}^v\}_{v=1}^V)$. According to Bayes rule, the posterior can be written as:

$$p(\boldsymbol{z}, \boldsymbol{c} \mid \{\boldsymbol{x}^v\}_{v=1}^V) = \frac{p(\{\boldsymbol{x}^v\}_{v=1}^V, \boldsymbol{z}, \boldsymbol{c})}{\int_{\boldsymbol{z}} \sum_{\boldsymbol{c}} p(\{\boldsymbol{x}^v\}_{v=1}^V, \boldsymbol{z}, \boldsymbol{c}) \mathrm{d}\boldsymbol{z}}. \tag{6}$$

Considering that it is intractable to calculate the true posterior in Eq.(6), we introduce a mean-field variational posterior $q(\boldsymbol{z}, \boldsymbol{c} \mid \{\boldsymbol{x}^v\}_{v=1}^V)$ to approximate it. This surrogate posterior can be formulated as:

$$q(\boldsymbol{z}, \boldsymbol{c} \mid \{\boldsymbol{x}^v\}_{v=1}^V) = q(\boldsymbol{z} \mid \{\boldsymbol{x}^v\}_{v=1}^V) q(\boldsymbol{c} \mid \{\boldsymbol{x}^v\}_{v=1}^V). \tag{7}$$

Then, we can perform separate approximate inferences for each component.

Common Representation. Following the general approach in VAEs, we parametrize the approximate posterior of each individual view by the Gaussian distribution:

$$q_{\boldsymbol{\phi}_v}\left(\boldsymbol{z} \mid \boldsymbol{x}^v\right) = \boldsymbol{\mathcal{N}}\left(\boldsymbol{z} \mid \boldsymbol{\mu}_v, \boldsymbol{\sigma}_v^2 \mathbf{I}\right), \tag{8}$$

where $[\boldsymbol{\mu}_v, \boldsymbol{\sigma}_v] = f\left(\boldsymbol{x}^v; \boldsymbol{\phi}_v\right)$. The function $f\left(\boldsymbol{x}^v; \boldsymbol{\phi}_v\right)$ denotes a neural network with trainable parameters $\boldsymbol{\phi}_v$, called view-specific encoder.

We employ the Product-of-Experts (PoE) approach to integrate information from each view and derive the common representation. Since the product of the Gaussian distributions remains a Gaussian distribution, the PoE aggregated posterior can be formulated as:

$$q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \{\boldsymbol{x}^v\}_{v=1}^V) = \boldsymbol{\mathcal{N}}\left(\boldsymbol{z} \mid \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}^2 \mathbf{I}\right), \tag{9}$$

where $\tilde{\boldsymbol{\mu}} = \frac{\sum_{v \in V} \boldsymbol{\mu}_v / \boldsymbol{\sigma}_v^2}{\sum_{v \in V} 1/\boldsymbol{\sigma}_v^2}$ and $\tilde{\boldsymbol{\sigma}}^2 = \frac{1}{\sum_{v \in V} 1/\boldsymbol{\sigma}_v^2}$.

By utilizing PoE, we can flexibly perform inference on samples with arbitrarily available views, while maintaining a valid and analytic form of the aggregated posterior. The mean and variance of the aggregated posterior can be calculated as:

$$\tilde{\boldsymbol{\mu}} = \frac{\sum_{v \in V_a} \boldsymbol{\mu}_v / \boldsymbol{\sigma}_v^2}{\sum_{v \in V_a} 1/\boldsymbol{\sigma}_v^2} \text{ and } \tilde{\boldsymbol{\sigma}}^2 = \frac{1}{\sum_{v \in V_a} 1/\boldsymbol{\sigma}_v^2}. \tag{10}$$

Shared Clustering Assignment. Although approximating $q(\boldsymbol{c} \mid \{\boldsymbol{x}^v\}_{v=1}^V)$ directly from the sample using a neural network is common with complete multi-view data, this approach has limitations in terms of flexibility and validity in scenarios with incomplete data.

To avoid the requirement for complex alternative networks while maintaining efficiency, we adopt the VaDE trick proposed in (Jiang et al. 2017; Falck et al. 2021). This method indirectly approximates the clustering assignment from the latent representation. Specifically, the VaDE trick involves sampling by using *reparameterization* trick from $q(\boldsymbol{z} \mid \{\boldsymbol{x}^v\}_{v=1}^V)$ to estimate the clustering assignment as follows:

$$q(c_k = 1 \mid \{\boldsymbol{x}^v\}_{v=1}^V) = \frac{p(\boldsymbol{z}^{(1)} \mid c_k = 1) p(c_k = 1)}{\sum_{\boldsymbol{c}} p(\boldsymbol{z}^{(1)} \mid \boldsymbol{c}) p(\boldsymbol{c})}, \tag{11}$$

where $\boldsymbol{z}^{(1)} = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\sigma}} \circ \boldsymbol{\epsilon}^{(1)}$ and $\boldsymbol{\epsilon}^{(1)} \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{I})$. $\circ$ denotes element-wise multiplication.

Given $\boldsymbol{z}^{(1)}$, we can analytically calculate the clustering assignment of this sample by Eq.(11).

Training Objective Loss

Variational Lower Bound. The Kullback-Leibler (KL) divergence between the variational posterior and true posteior can be written as:

$$\begin{aligned} &D_{KL}(q(\boldsymbol{z}, \boldsymbol{c} \mid \{\boldsymbol{x}^v\}_{v=1}^V) \parallel p(\boldsymbol{z}, \boldsymbol{c} \mid \{\boldsymbol{x}^v\}_{v=1}^V)) \\ &= \log p(\{\boldsymbol{x}^v\}_{v=1}^V) - \boldsymbol{\mathcal{L}}_{\mathrm{ELBO}}(\{\boldsymbol{x}^v\}_{v=1}^V), \end{aligned} \tag{12}$$

where $\boldsymbol{\mathcal{L}}_{\mathrm{ELBO}}(\{\boldsymbol{x}^v\}_{v=1}^V)$ is the evidence lower bound (ELBO) on the marginal log-likelihood of multi-view data $\{\boldsymbol{x}^v\}_{v=1}^V$.

Intuitively, to closely approximate the true posterior, we can minimize the KL divergence between the approximate and true posterior distributions by maximizing the ELBO. Integrating the proposed deep multi-view Gaussian mixture model and inference approach, the ELBO is formulated as:

$$\begin{aligned} &\boldsymbol{\mathcal{L}}_{\mathrm{ELBO}}(\{\boldsymbol{x}^v\}_{v=1}^V) \\ &= \mathbb{E}_{q(\boldsymbol{z}, \boldsymbol{c} \mid \{\boldsymbol{x}^v\}_{v=1}^V)}[\log \frac{p(\{\boldsymbol{x}^v\}_{v=1}^V, \boldsymbol{z}, \boldsymbol{c})}{q(\boldsymbol{z}, \boldsymbol{c} \mid \{\boldsymbol{x}^v\}_{v=1}^V)}] \\ &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \{\boldsymbol{x}^v\}_{v=1}^V)}[\sum_{v \in V_a} \log p_{\boldsymbol{\theta}_v}(\boldsymbol{x}^v \mid \boldsymbol{z})] \\ &- \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{c} \mid \{\boldsymbol{x}\}_{v=1}^V)}[D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \{\boldsymbol{x}^v\}_{v=1}^V)) \parallel p(\boldsymbol{z} \mid \boldsymbol{c}))] \\ &- D_{KL}(q(\boldsymbol{c} \mid \{\boldsymbol{x}^v\}_{v=1}^V)) \parallel p(\boldsymbol{c})). \end{aligned} \tag{13}$$

where the first term of the ELBO, known as the reconstruction term, encourages the common representation to encode shared information across views. The KL divergence terms enforce the clustering structure within the latent representation.

Coherence Objective Loss. Checking the formulation in Eqs.(9)-(11), the mean and variance of the PoE aggregated posterior are often dominated by component posteriors with significantly lower variance. Incomplete multi-view data amplifies this characteristic of PoE in the inference process due to two factors: the inherent information imbalance across different views and the presence of arbitrary missing views in samples. This leads to the neglect of views with higher uncertainty in the latent space, resulting in suboptimal training of corresponding view-specific encoders. To mitigate this issue, we introduce a coherence objective loss:

$$\begin{aligned} &\boldsymbol{\mathcal{L}}_{\mathrm{CH}}(\{\boldsymbol{x}^v\}_{v=1}^V) \\ &= \sum_{v \in V_a} -\frac{1}{|V_a|} D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \{\boldsymbol{x}^v\}_{v=1}^V) \| q_{\boldsymbol{\phi}_v}(\boldsymbol{z} \mid \boldsymbol{x}^v)). \end{aligned} \tag{14}$$

Maximizing $\boldsymbol{\mathcal{L}}_{\mathrm{CH}}$ is equivalent to minimizing the average KL divergence between the aggregated posterior and each available single-view posterior. This term aims to enforce the consistency of information encoded by the aggregated representation.

Overall Objective Loss Combining the above two objective functions with a regularization parameter α, the overall objective loss (to maximize) is formulated as:

$$\boldsymbol{\mathcal{L}} = \boldsymbol{\mathcal{L}}_{\mathrm{ELBO}} + \alpha \boldsymbol{\mathcal{L}}_{\mathrm{CH}}. \tag{15}$$

Following the general approach in VAEs, we utilize the Stochastic Gradient Variational Bayes estimator, along with the *reparameterization* trick and the VaDE trick, to estimate the ELBO. Additionally, given that both the aggregated posterior and single-view posterior have the Gaussian form, the KL divergence in the coherence objective loss can be calculated analytically. Thus, stochastic gradient descent can be employed to train and update parameters efficiently. The optimization process is summarized in Algorithm 1.

Algorithm 1: Optimization of the proposed method

Input: Incomplete multi-view dataset $\{\{\boldsymbol{x}_i^v\}_{v=1}^V\}_{i=1}^N$ with indicator matrix $\mathbf{M}$; Number of clusters K; Regularization parameter α.
Initialize $\{\boldsymbol{\theta}_v, \boldsymbol{\phi}_v\}_{v=1}^V, \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2\}_{k=1}^K$.
while not reaching the maximal epochs **do**
1. Calculate $\{\boldsymbol{\mu}_v, \boldsymbol{\sigma}_v^2\}_{v\in V_a}$ by view-specific encoders.
2. Calculate $\{\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}^2\}$ by Eq.(10).
3. Calculate $q(\boldsymbol{c} \mid \{\boldsymbol{x}^v\}_{v=1}^V)$ by Eq.(11).
4. Generate $\{p_{\boldsymbol{\theta}_v}(\boldsymbol{x}^v \mid \boldsymbol{z})\}_{v\in V_a}$ by view-specific decoders with *reparameterization* trick.
5. Update $\{\boldsymbol{\theta}_v, \boldsymbol{\phi}_v\}_{v=1}^V, \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2\}_{k=1}^K$ by maximizing Eq.(15).
end while
Calculate the clustering assignment of each sample by Eq.(11) and predict the cluster by the max probability

Experiments

Experimental Settings

Datasets. Four real-world datasets are used in our experiments, namely Caltech7-5V (Fei-Fei, Fergus, and Perona 2007; Li et al. 2015), Scene-15 (Fei-Fei and Perona 2005), Multi-Fashion (Xiao, Rasul, and Vollgraf 2017), and NoisyMNIST (Wang et al. 2015; Basu et al. 2017). Statistics of these datasets are shown in Table 1.

Dateset	Instance No.	View No.	Cluster No.
Caltech7-5V	1400	5	7
Scene-15	4485	3	15
Multi-Fashion	10000	3	10
NoisyMNIST	70000	2	10

Table 1. Description of the four multi-view datasets.

Incomplete Multi-view Data Construction. Following (Tang and Liu 2022), we randomly removed some views from the dataset while ensuring that at least one view remained in each sample. For instance, in a dataset with V views and N samples, we randomly selected M samples to be incomplete, with 1 to $V-1$ views randomly removed for each sample. The missing rate p is then calculated as $p = M/N$.

Compared Methods. We compared our proposed method with the following methods. **MVAE** extended the variational autoencoder to the multi-view learning task (Wu and Goodman 2018). **DMVCVAE** extended the probabilistic clustering framework to the multi-view scene (Yin, Huang, and Gao 2020). **Multi-VAE** applied the Gumbel-Softmax trick (Jang, Gu, and Poole 2016; Maddison, Mnih, and Teh 2017) for the approximate inference of the clustering variable (Xu et al. 2021b). **DIMVC** proposed to project representations into a more easily separable high-dimension space to enhance clustering accuracy (Xu et al. 2022a). **DCP** utilized the principle of mutual information maximization to obtain consistent representations and designed additional networks to impute missing views (Lin et al. 2022). **DSIMVC** proposed an imputation-based DIMVC framework incorporating a meta-learning objective to mitigate the impact of inaccurate imputation (Tang and Liu 2022). **APADC** introduced a distribution alignment objective based on the Maximum Mean Discrepancy distance to obtain consistent common representations (Xu et al. 2023). Among these, **MVAE** and **DMVCVAE** are most closely related to our method. **MVAE** utilized the PoE approach to obtain the common representation and **DMVCVAE** employed the weight sum approach to obtain the common representation. For **Multi-VAE**, which cannot directly handle incomplete data, we imputed the missing view using the average value of the corresponding views before evaluation.

Evaluation Metrics. The clustering performance is evaluated by four widely-used metrics, including Accuracy (ACC), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Purity (PUR) (Zhang et al. 2015; Wen et al. 2021a).

Implementation Details. In our experiments, we use the same network structure for all datasets. Specifically, for each view, we adopt a fully connected network with the layer dimensions of D^v-500-500-2000-10 (10-2000-500-500–D^v) as the encoder (decoder), where D^v denotes the feature dimension of the original data. 'ReLU' is used as the activation function for these layers. We implement the experiments on Linux with an NVIDIA 4090 GPU. We first train each view-specific encoder-decoder network using MSE loss for 200 epochs, updating the network parameters with default Adam optimizer. Then, we apply K-means clustering on the latent embedding to initialize the means of the latent Mixture-of-Gaussian (MoG) prior. After this initialization, the learning rate for Adam optimizer is set to 0.0005 for the encoder-decoders parameters and 0.05 for the latent MoG prior parameters, both with a decay rate of 0.9 every 10 epochs. The training batch size is set as 512 for NoisyMNIST and set as 256 for the other three datasets. The regularization parameter is set to 5 for Caltech7-5V, 10 for NoisyMNIST and Multi-Fashion, and 20 for Scene-15. The source codes of our method based on Pytorch and Mindspore are released at https://sites.google.com/view/jerry-wen-hit/publications. The mindspore source code is developed on OpenI Community.

Experimental Results and Analysis

Table 2 reports the average clustering results obtained by repeating each method 10 times on the randomly constructed incomplete datasets with different missing view rates. From Table 2, we have the following observations:

- Our method achieves competitive performance in comparison with both imputation-free and imputation-based methods. For example, on the Caltech7-5V dataset with a missing rate of 0.5, the clustering results in terms of ACC and NMI of our method are 10% and 15% higher than the second-best method, DIMVC, an imputation-free method. On the Multi-Fashion dataset, the cluster-

	Missing rates	0.1				0.3				0.5				0.7			
	Methods	ACC	NMI	ARI	PUR	ACC	NMI	ARI	PUR	ACC	NMI	ARI	PUR	ACC	NMI	ARI	PUR
Caltech7-5V	MVAE	0.671	0.561	0.473	0.679	0.614	0.495	0.406	0.627	0.591	0.474	0.383	0.603	0.610	0.489	0.404	0.625
	DMVCVAE	0.745	0.597	0.547	0.751	0.657	0.514	0.414	0.669	0.595	0.457	0.319	0.602	0.527	0.407	0.210	0.527
	Multi-VAE	0.610	0.515	0.428	0.626	0.555	0.446	0.337	0.571	0.504	0.385	0.522	0.244	0.436	0.334	0.451	0.183
	DIMVC	0.834	0.735	0.701	0.834	0.798	0.677	0.643	0.803	0.760	0.605	0.576	0.760	0.715	0.550	0.516	0.719
	DCP	0.643	0.636	0.456	0.670	0.542	0.503	0.331	0.565	0.452	0.379	0.208	0.462	0.346	0.250	0.144	0.351
	DSIMVC	0.765	0.672	0.601	0.766	0.781	0.670	0.614	0.782	0.729	0.617	0.541	0.731	0.621	0.530	0.430	0.625
	APADC	0.612	0.614	0.464	0.623	0.609	0.610	0.457	0.622	0.550	0.540	0.358	0.558	0.539	0.501	0.328	0.547
	Ours	**0.895**	**0.812**	**0.786**	**0.895**	**0.886**	**0.798**	**0.774**	**0.886**	**0.868**	**0.764**	**0.746**	**0.868**	**0.844**	**0.730**	**0.707**	**0.844**
Scene-15	MVAE	0.421	0.369	0.238	0.453	0.372	0.323	0.200	0.401	0.363	0.305	0.188	0.389	0.352	0.295	0.177	0.379
	DMVCVAE	0.375	0.371	0.216	0.411	0.341	0.325	0.157	0.369	0.281	0.275	0.108	0.312	0.245	0.230	0.063	0.272
	Multi-VAE	0.305	0.294	0.144	0.347	0.256	0.234	0.096	0.291	0.245	0.222	0.081	0.274	0.234	0.208	0.068	0.260
	DIMVC	0.456	0.444	0.285	0.495	0.424	0.402	0.254	0.463	0.412	0.379	0.237	0.449	0.381	0.334	0.205	0.417
	DCP	0.405	0.448	0.247	0.442	0.399	0.422	0.234	0.427	0.3831	0.408	0.221	0.412	0.357	0.386	0.197	0.393
	DSIMVC	0.268	0.295	0.137	0.320	0.267	0.292	0.135	0.318	0.255	0.277	0.126	0.310	0.262	0.274	0.127	0.308
	APADC	0.409	9.420	0.246	0.449	0.399	0.405	0.230	0.440	0.390	0.389	0.222	0.428	0.386	0.379	0.218	0.421
	Ours	**0.479**	**0.467**	**0.308**	**0.509**	**0.454**	**0.442**	**0.288**	**0.480**	**0.442**	**0.422**	**0.263**	**0.461**	**0.411**	**0.394**	**0.245**	**0.428**
Multi-Fashion	MVAE	0.747	0.718	0.636	0.787	0.731	0.689	0.600	0.762	0.725	0.659	0.573	0.748	0.6545	0.567	0.446	0.660
	DMVCVAE	0.770	0.760	0.671	0.784	0.663	0.654	0.535	0.668	0.565	0.556	0.415	0.568	0.441	0.459	0.296	0.451
	Multi-VAE	0.729	0.826	0.543	0.812	0.677	0.759	0.463	0.708	0.6563	0.772	0.468	0.750	0.661	0.669	0.523	0.693
	DIMVC	0.760	0.836	0.704	0.801	0.707	0.784	0.637	0.758	0.691	0.737	0.601	0.736	0.679	0.688	0.563	0.719
	DCP	0.837	0.843	0.765	0.850	0.718	0.709	0.525	0.731	0.608	0.595	0.331	0.622	0.499	0.484	0.191	0.513
	DSIMVC	0.880	0.864	0.811	0.886	0.873	0.850	0.789	0.876	**0.835**	0.803	0.737	0.835	0.796	0.774	0.690	0.797
	APADC	0.814	0.865	0.733	0.808	0.809	0.850	0.731	0.813	0.754	0.815	0.676	0.758	0.699	0.759	0.614	0.717
	Ours	**0.883**	**0.876**	**0.822**	**0.898**	**0.879**	**0.859**	**0.805**	**0.886**	0.824	**0.832**	**0.755**	**0.847**	**0.815**	**0.811**	**0.727**	**0.826**
NoisyMNIST	MVAE	0.558	0.459	0.361	0.564	0.609	0.506	0.429	0.616	0.566	0.491	0.412	0.574	0.544	0.454	0.382	0.550
	DMVCVAE	0.479	0.491	0.336	0.509	0.447	0.427	0.274	0.4534	0.448	0.420	0.226	0.449	0.381	0.364	0.138	0.389
	Multi-VAE	0.625	0.710	0.3848	0.781	0.593	0.727	0.382	0.687	0.503	0.626	0.167	0.562	0.375	0.524	0.047	0.503
	DIMVC	0.691	0.723	0.597	0.709	0.682	0.689	0.570	0.701	0.6327	0.623	0.510	0.657	0.628	0.591	0.491	0.642
	DCP	0.870	0.889	0.817	0.890	0.912	0.892	0.865	0.922	0.894	0.855	0.822	0.904	0.875	0.812	0.793	0.885
	DSIMVC	0.735	0.708	0.633	0.772	0.637	0.607	0.514	0.677	0.589	0.565	0.461	0.624	0.581	0.555	0.443	0.634
	APADC	0.891	0.895	0.855	0.910	0.864	0.839	0.775	0.878	0.868	0.820	0.777	0.881	0.733	0.727	0.634	0.764
	Ours	**0.965**	**0.946**	**0.941**	**0.965**	**0.943**	**0.919**	**0.912**	**0.943**	**0.923**	**0.883**	**0.876**	**0.923**	**0.911**	**0.833**	**0.835**	**0.912**

Table 2. Clustering results of all methods on four datasets. The best results are highlighted in bold.

ing performance of our method surpasses the state-of-the-art imputation-based methods, DSIMVC and DCP. This suggests that our method is particularly effective in accurately capturing the shared clustering structure from incomplete multi-view data.

- Our method significantly outperforms other VAE-based methods on four datasets in terms of all metrics. Especially, when the multi-view data have a large number of missing views, VAE-based methods exhibit poor clustering performance. For instance, on the NoisyMNIST dataset with a missing rate of 0.7, the clustering ACCs of our method and the best VAE-based method, MVAE, are 0.9118 and 0.5443, respectively. An approximately 37% improvement in ACC demonstrates the superiority and effectiveness of our method for the incomplete multi-view clustering task.

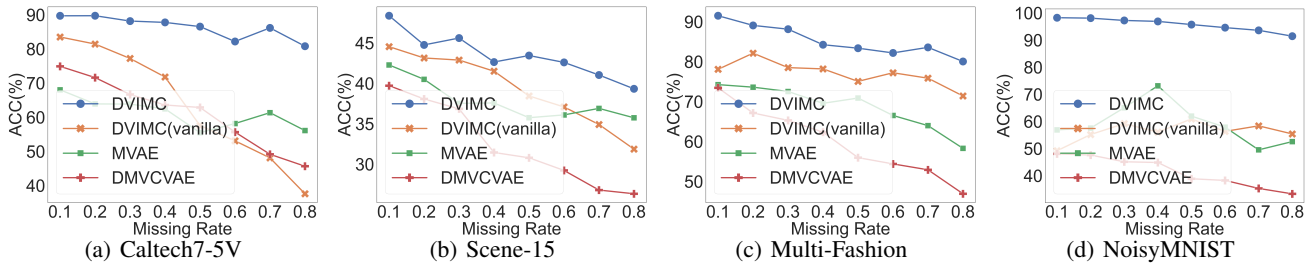
In Figure 2, we visualize the common representation obtained by our proposed method and the fused representation obtained by summing multiple single-view representations of the autoencoder method on the NoisyMNIST and Caltech7-5V datasets with a missing rate of 0.5. The T-SNE visualization of these two types of representations reveals that the representation obtained by our method exhibits more distinct clustering structures, where data points in different clusters are clearly separated. This illustrates the effectiveness of our method in learning the shared clustering structure.

Ablation Study

In this section, we conduct experiments to compare the clustering performance of the proposed DVIMC with its degraded version DVIMC (vanilla) (*i.e.*, removing $\mathcal{L}_{\mathrm{CH}}$). In addition, MVAE and DMVCVAE are used as baseline methods to demonstrate the synergy of the PoE, VaDE trick, and coherence objective loss. The experimental results, as shown in Figure 3, lead to the following observations: 1) Methods not equipped with the coherence objective loss (*i.e.*, DMVCVAE, MVAE, and DVIMC (vanilla)) cannot achieve optimal clustering results, and their performance significantly decreases with increasing missing view rate. This demonstrates the challenge of learning from incomplete multi-view data, as mentioned earlier. 2) Comparing DMVCVAE and DVIMC (vanilla), which differ in their aggregation approaches, the superiority of DVIMC(vanilla) demonstrates

(a) AE on NoisyMNIST (b) DVIMC on NoisyMNIST (c) AE on Caltech7-5V (d) DVIMC on Caltech7-5V

Figure 2: T-SNE visualization. Comparing the fused representation from general AutoEncoder (AE) and the common representation by our DVIMC method on NoisyMNIST and Caltech7-5V datasets with a missing view rate of 0.5.



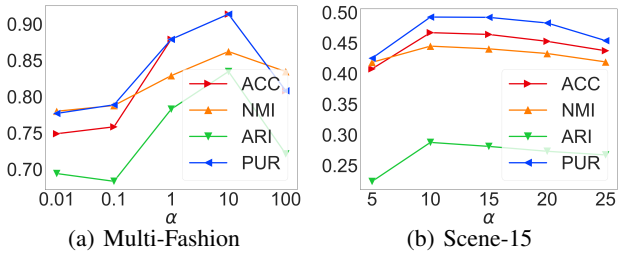
(a) Caltech7-5V (b) Scene-15 (c) Multi-Fashion (d) NoisyMNIST

Figure 3: Comparison of the proposed DVIMC and its degraded version DVIMC (vanilla) with MVAE and DMVCVAE on the four datasets with different missing view rates.

the PoE approach is a more reasonable and efficient method than the weight summarization approach under the VAEs framework, especially for incomplete multi-view data. This is attributed to the characteristic of PoE to form a valid and analytical distribution for Gaussian experts. 3) Comparing MVAE and DVIMC (vanilla), which differ in the prior of latent representations, the improved clustering performance of DVIMC (vanilla) over MVAE highlights the efficacy of explicitly guiding the clustering structure in the representation and utilizing the VaDE trick to infer this structure. 4) The best clustering performance of DVIMC suggests that using PoE to aggregating information from incomplete multi-view data, introducing the coherence objective loss, and employing the VaDE trick can effectively discover the hidden clustering structure.

Parameter Analysis

We conducted experiments on the Multi-Fashion and Scene-15 datasets with a missing view rate of 0.5 to analyze the influence of the regularization parameter α on the clustering performance of our proposed method. The results, depicted in Figure 4, show that neither excessively high nor low values of α are beneficial for clustering. With a small α, the method may suffer from the drawbacks of the PoE approach, while a large α might overly emphasize the coherence objective. Such overemphasis can reduce the capacity of shared information in the aggregated representation, leading to suboptimal clustering results. It is important to note that, in some cases, this level of information may still be adequate for distinguishing samples from different clusters. Empirically, we suggest setting α to 10 for our method.



(a) Multi-Fashion (b) Scene-15

Figure 4: The parameter analysis on the Multi-Fashion and Scene-15 datasets with a missing view rate of 0.5.

Conclusion

In this paper, we introduced DVIMC, a novel imputation-free deep incomplete multi-view clustering method. DVIMC is designed to be flexibly applied to multi-view datasets with arbitrary missing views. By leveraging classical techniques such as the Product-of-Experts and the VaDE trick, our method efficiently addresses the challenges inherent in learning from incomplete multi-view data and effectively explores shared clustering structures. The experimental results demonstrate that DVIMC significantly enhances performance in incomplete multi-view clustering tasks.

Acknowledgements

This work is supported by Shenzhen Higher Education Stability Support Program Project under Grant No. GXWD20220811173317002, National Natural Science Foundation of China under Grant No. 62372136, and Chinese Association for Artificial Intelligence (CAAI)-Huawei MindSpore Open Fund under Grant No. CAAIXSJLJJ-2022-011C. Sponsored by CAAI-MindSpore Open Fund, developed on OpenI Community.

References

Basu, S.; Karki, M.; Ganguly, S.; DiBiano, R.; Mukhopadhyay, S.; Gayaka, S.; Kannan, R.; and Nemani, R. 2017. Learning sparse feature representations using probabilistic quadtrees and deep belief nets. *Neural Processing Letters*, 45: 855–867.

Dilokthanakul, N.; Mediano, P. A.; Garnelo, M.; Lee, M. C.; Salimbeni, H.; Arulkumaran, K.; and Shanahan, M. 2016. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*.

Dupont, E. 2018. Learning disentangled joint continuous and discrete representations. In *Proceedings of the International Conference on Neural Information Processing Systems*, 708–718.

Falck, F.; Zhang, H.; Willetts, M.; Nicholson, G.; Yau, C.; and Holmes, C. C. 2021. Multi-Facet Clustering Variational Autoencoders. In *Proceedings of the International Conference on Neural Information Processing Systems*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2007. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 1(106): 59–70.

Fei-Fei, L.; and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 524–531. IEEE.

Hinton, G. E. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8): 1771–1800.

Hsu, W.-N.; and Glass, J. 2018. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv preprint arXiv:1805.11264*.

Hu, M.; and Chen, S. 2019. One-pass incomplete multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3838–3845.

Hwang, H.; Kim, G.-H.; Hong, S.; and Kim, K.-E. 2021. Multi-View Representation Learning via Total Correlation Objective. In *Proceedings of the International Conference on Neural Information Processing Systems*.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*.

Jiang, B.; Wu, X.; Zhou, X.; Liu, Y.; Cohn, A. G.; Sheng, W.; and Chen, H. 2022a. Semi-supervised multiview feature selection with adaptive graph learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Jiang, B.; Xiang, J.; Wu, X.; Wang, Y.; Chen, H.; Cao, W.; and Sheng, W. 2022b. Robust multi-view learning via adaptive regression. *Information Sciences*, 610: 916–937.

Jiang, B.; Zhang, C.; Zhong, Y.; Liu, Y.; Zhang, Y.; Wu, X.; and Sheng, W. 2023. Adaptive collaborative fusion for multi-view semi-supervised classification. *Information Fusion*, 96: 37–50.

Jiang, Z.; Zheng, Y.; Tan, H.; Tang, B.; and Zhou, H. 2017. Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Li, L.; and He, H. 2020. Bipartite graph based multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 34(7): 3111–3125.

Li, L.; Wan, Z.; and He, H. 2021. Incomplete multi-view clustering with joint partition and graph learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(1): 589–602.

Li, Y.; Nie, F.; Huang, H.; and Huang, J. 2015. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Lin, Y.; Gou, Y.; Liu, X.; Bai, J.; Lv, J.; and Peng, X. 2022. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4447–4461.

Lin, Y.; Gou, Y.; Liu, Z.; Li, B.; Lv, J.; and Peng, X. 2021. COMPLETER: Incomplete Multi-View Clustering via Contrastive Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11174–11183.

Liu, C.; Wen, J.; Wu, Z.; Luo, X.; Huang, C.; and Xu, Y. 2023. Information Recovery-Driven Deep Incomplete Multiview Clustering Network. *IEEE Transactions on Neural Networks and Learning Systems*, 1–11.

Liu, X. 2021. Incomplete multiple kernel alignment maximization for clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, X.; Li, M.; Tang, C.; Xia, J.; Xiong, J.; Liu, L.; Kloft, M.; and Zhu, E. 2020. Efficient and effective regularized incomplete multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8): 2634–2646.

Maddison, C.; Mnih, A.; and Teh, Y. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.

Shi, Y.; Siddharth, N.; Paige, B.; and Torr, P. H. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Proceedings of the International Conference on Neural Information Processing Systems*, 15718–15729.

Sutter, T. M.; Daunhawer, I.; and Vogt, J. E. 2020a. Generalized Multimodal ELBO. In *International Conference on Learning Representations*.

Sutter, T. M.; Daunhawer, I.; and Vogt, J. E. 2020b. Multimodal generative learning utilizing jensen-shannon-divergence. In *Proceedings of the International Conference on Neural Information Processing Systems*, 6100–6110.

Suzuki, M.; Nakayama, K.; and Matsuo, Y. 2016. Joint Multimodal Learning with Deep Generative Models. arXiv:1611.01891.

Tang, H.; and Liu, Y. 2022. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *International Conference on Machine Learning*, 21090–21110.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2018. Learning Factorized Multimodal Representations. In *International Conference on Learning Representations*.

Wang, Q.; Ding, Z.; Tao, Z.; Gao, Q.; and Fu, Y. 2021a. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Transactions on Image Processing*, 30: 1771–1783.

Wang, S.; Liu, X.; Zhu, X.; Zhang, P.; Zhang, Y.; Gao, F.; and Zhu, E. 2021b. Fast parameter-free multi-view subspace clustering with consensus anchor guidance. *IEEE Transactions on Image Processing*, 31: 556–568.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015. On deep multi-view representation learning. In *International conference on Machine Learning*, 1083–1092.

Wen, J.; Xu, Y.; and Liu, H. 2018. Incomplete multi-view spectral clustering with adaptive graph learning. *IEEE Transactions on Cybernetics*, 50(4): 1418–1429.

Wen, J.; Yan, K.; Zhang, Z.; Xu, Y.; Wang, J.; Fei, L.; and Zhang, B. 2021a. Adaptive graph completion based incomplete multi-view clustering. *IEEE Transactions on Multimedia*, 23: 2493–2504.

Wen, J.; Zhang, Z.; Fei, L.; Zhang, B.; Xu, Y.; Zhang, Z.; and Li, J. 2022. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2): 1136–1149.

Wen, J.; Zhang, Z.; Xu, Y.; Zhang, B.; Fei, L.; and Liu, H. 2019. Unified embedding alignment with missing views inferring for incomplete multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 5393–5400.

Wen, J.; Zhang, Z.; Xu, Y.; Zhang, B.; Fei, L.; and Xie, G.-S. 2021b. CDIMC-net: Cognitive deep incomplete multi-view clustering network. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, 3230–3236.

Wen, J.; Zhang, Z.; Zhang, Z.; Fei, L.; and Wang, M. 2020. Generalized incomplete multiview clustering with flexible locality structure diffusion. *IEEE Transactions on Cybernetics*, 51(1): 101–114.

Wen, J.; Zhang, Z.; Zhang, Z.; Zhu, L.; Fei, L.; Zhang, B.; and Xu, Y. 2021c. Unified tensor framework for incomplete multi-view clustering and missing-view inferring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 10273–10281.

Wu, M.; and Goodman, N. 2018. Multimodal generative models for scalable weakly-supervised learning. In *Proceedings of the International Conference on Neural Information Processing Systems*, 5580–5590.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xu, C.; Liu, H.; Guan, Z.; Wu, X.; Tan, J.; and Ling, B. 2021a. Adversarial incomplete multiview subspace clustering networks. *IEEE Transactions on Cybernetics*, 52(10): 10490–10503.

Xu, J.; Li, C.; Peng, L.; Ren, Y.; Shi, X.; Shen, H. T.; and Zhu, X. 2023. Adaptive feature projection with distribution alignment for deep incomplete multi-view clustering. *IEEE Transactions on Image Processing*, 32: 1354–1366.

Xu, J.; Li, C.; Ren, Y.; Peng, L.; Mo, Y.; Shi, X.; and Zhu, X. 2022a. Deep incomplete multi-view clustering via mining cluster complementarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8761–8769.

Xu, J.; Ren, Y.; Tang, H.; Pu, X.; Zhu, X.; Zeng, M.; and He, L. 2021b. Multi-VAE: Learning Disentangled View-Common and View-Peculiar Visual Representations for Multi-View Clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9234–9243.

Xu, J.; Ren, Y.; Tang, H.; Yang, Z.; Pan, L.; Yang, Y.; Pu, X.; Philip, S. Y.; and He, L. 2022b. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*.

Yan, X.; Hu, S.; Mao, Y.; Ye, Y.; and Yu, H. 2021. Deep multi-view learning methods: A review. *Neurocomputing*, 448: 106–129.

Yang, L.; Cheung, N.-M.; Li, J.; and Fang, J. 2019. Deep Clustering by Gaussian Mixture Variational Autoencoders With Graph Embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6440–6449.

Yin, J.; and Sun, S. 2021. Incomplete multi-view clustering with reconstructed views. *IEEE Transactions on Knowledge and Data Engineering*.

Yin, M.; Huang, W.; and Gao, J. 2020. Shared generative latent representation learning for multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 6688–6695.

Zhang, C.; Fu, H.; Liu, S.; Liu, G.; and Cao, X. 2015. Low-Rank Tensor Constrained Multiview Subspace Clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1582–1590.

Zhao, L.; Chen, Z.; Yang, Y.; Wang, Z. J.; and Leung, V. C. 2018. Incomplete multi-view clustering via deep semantic mapping. *Neurocomputing*, 275: 1053–1062.

Zhuge, W.; Hou, C.; Liu, X.; Tao, H.; and Yi, D. 2019. Simultaneous representation learning and clustering for incomplete multi-view data. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 4482–4488.