Deep Incomplete Multi-view Clustering via Mining Cluster Complementarity

Jie Xu, Chao Li, Yazhou Ren*, Liang Peng, Yujie Mo, Xiaoshuang Shi*, Xiaofeng Zhu

School of Computer Science and Engineering,
University of Electronic Science and Technology of China, Chengdu 611731, China
jiexuwork@outlook.com, lichao.cfm@gmail.com, yazhou.ren@uestc.edu.cn, larrypengliang@gmail.com,
moyujie2017@gmail.com, xsshi2013@gmail.com, seanzhuxf@gmail.com

Abstract

Incomplete multi-view clustering (IMVC) is an important unsupervised approach to group the multi-view data containing missing data in some views. Previous IMVC methods suffer from the following issues: (1) the inaccurate imputation or padding for missing data negatively affects the clustering performance, (2) the quality of features after fusion might be interfered by the low-quality views, especially the inaccurate imputed views. To avoid the above issues, this work presents an imputation-free and fusion-free deep IMVC framework. First, the proposed method builds a deep embedding feature learning and clustering model for each view individually. Our method then nonlinearly maps the embedding features of complete data into a high-dimensional space to discover linear separability. Concretely, this paper provides an implementation of the high-dimensional mapping as well as shows the mechanism to mine the multi-view cluster complementarity. This complementary information is then transformed to the supervised information with high confidence, aiming to achieve the multi-view clustering consistency for the complete data and incomplete data. Furthermore, we design an EM-like optimization strategy to alternately promote feature learning and clustering. Extensive experiments on real-world multi-view datasets demonstrate that our method achieves superior clustering performance over state-of-the-art methods.

1 Introduction

Data often has multiple views collected from diverse sources in practical applications, such as classification (Liu, Li, and Zhang 2016), community detection (Cao et al. 2019), dimensionality reduction (Liu et al. 2017), and cross-modal retrieval (Xu et al. 2020; Wei et al. 2021). Multi-view clustering is an important unsupervised approach, aiming to improve the model effectiveness by mining the complementary information hidden in multi-view data. Recently, many multi-view clustering methods have been proposed (Tao et al. 2017; Li et al. 2019; Peng et al. 2019; Tang et al. 2020; Chen et al. 2020; Huang et al. 2021; Xu et al. 2021b). These methods mainly deal with the complete multi-view data, where the information of all views is observed.

However, the information of multi-view data might be incomplete in some views, as known as incomplete multi-view data (Li, Jiang, and Zhou 2014; Xu, Tao, and Xu 2015). For example, different medical tests of a patient can be treated as different views, where cheap medical tests are easily available and expensive ones are often missed due to price. Images and texts are two views to describe scenes, but only some images have textual caption. The incomplete multi-view data inevitably makes existing multi-view clustering methods limited and inapplicable. To this end, an increasing attention is paid to partial multi-view clustering or incomplete multi-view clustering (IMVC) problems (Rai et al. 2016; Wang et al. 2018; Ye et al. 2018; Huang et al. 2020).

In the literature, existing IMVC methods can be categorized into two groups, *i.e.*, traditional methods and deep methods. Traditional IMVC methods usually adopt zero or mean values to pad missing data (Wen et al. 2021) firstly, and then design specific machine learning techniques to conduct multi-view clustering, such as non-negative matrix factorization methods (Li, Jiang, and Zhou 2014; Hu and Chen 2019), subspace learning methods (Kang et al. 2020; Liu et al. 2021a), kernel methods (Guo and Ye 2019; Liu et al. 2021b), and graph methods (Rai et al. 2016; Li, Wan, and He 2021). Nevertheless, traditional IMVC methods are limited in their representation capability and high complexity (Guo and Ye 2019). Recently, deep IMVC methods have gradually been attracting attentions due to their powerful generalization capability and scalability. Deep IMVC methods often utilize the imputation strategies to infer the possible values for missing data before conducting multi-view clustering. For instance, (Xu et al. 2019a; Wang et al. 2021) proposed to take advantage of generative adversarial networks to generate desired data for the missing data. (Lin et al. 2021) designed to recover the missing data with contrastive learning.

Although existing IMVC methods achieve important progress by padding missing data with imputation strategies, they have at least two issues. On the one hand, the effectiveness of imputation strategies depends on the quality of imputed data. It is difficult to correctly estimate the missing data based on the complete data, especially when the number of missing data is large. Moreover, it is also knotty to measure the quality of the imputation as the ground truth of the missing data is unknown. On the other hand, existing IMVC methods usually explore the complementary information among multi-view data by the fusion process. For instance, (Guo and Ye 2019) excavated the complementary in-

*Corresponding Authors.

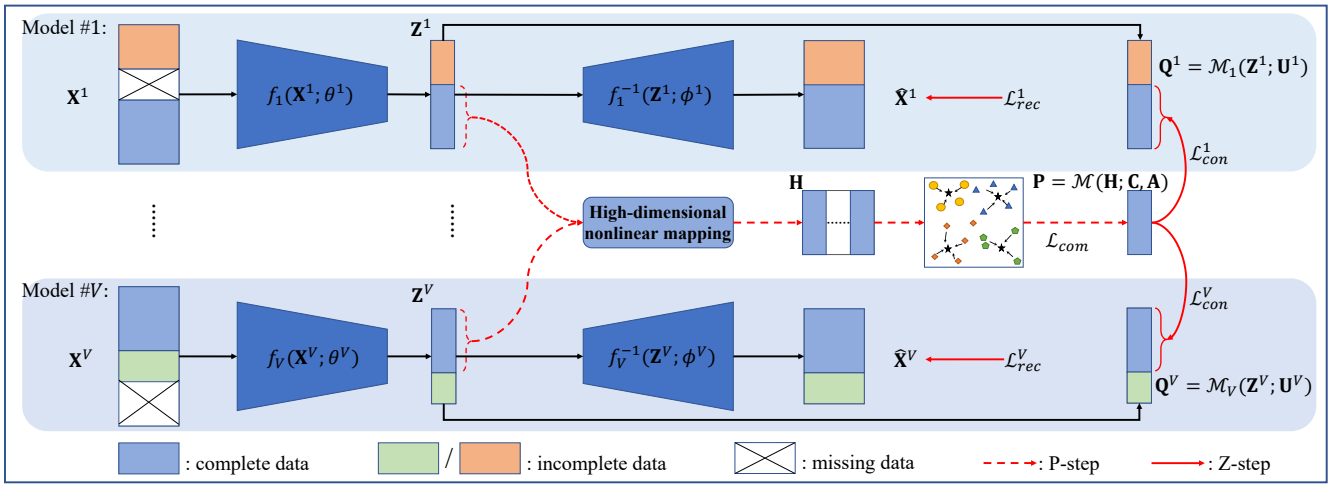


Figure 1: The framework of DIMVC. For the v-th view, the encoder f_v and decoder f_v^{-1} learn the embedding features $\mathbf{Z}^v$, from which the mapping $\mathcal{M}_v$ predicts the cluster assignments $\mathbf{Q}^v$ (Section 2.1). The high-dimensional nonlinear mapping is proposed to obtain the linearly separable features $\mathbf{H}$ (Section 2.2). The mapping $\mathcal{M}$ aims to generate supervised information $\mathbf{P}$ with high confidence (Section 2.3). For all views, P-step mines their complementary information by optimizing $\mathcal{L}_{com}$. Z-step achieves feature learning and clustering consistency across multiple views by optimizing $\{\mathcal{L}_{rec}^v + \mathcal{L}_{con}^v\}_{v=1}^V$ (Section 2.4).

formation by fusing multiple similarity matrices. (Wen et al. 2020; Wang et al. 2021) pointed out to utilize fusion layers to mine the complementary information. As some views might be inherently of low quality or inaccurately imputed, they will negatively affect the fusion process.

In this paper, we propose an imputation-free and fusion-free deep IMVC framework (named DIMVC) to address the aforementioned issues. That is, the missing data does not need to be imputed or padded and the cluster assignments do not depend on the fusion process of multiple views. To achieve this, there are two crucial challenges to be solved, *i.e.*, (i) new strategy needs to be developed to explore complementary information in the fusion-free model, and (ii) it is difficult to obtain consistent cluster assignments for both the complete data and incomplete data without imputation.

We illustrate the framework of our proposed DIMVC in Figure 1. More specifically, we build an individual model on all observed data for each view. Each model consists of an autoencoder and a clustering mapping. Based on our observation that the complementary information across multiple views can be described by nonlinear mappings, we tackle the challenge (i) by a high-dimensional mapping. Concretely, the embedding features of the complete multi-view data are nonlinearly projected into a concatenated weighted feature space, where the high-separability view is assigned with a high weight. Intuitively, the high-separability view means that there are well-separated cluster structures in the features. Moreover, we show that the linearly separable cluster information can be transferred to the high-dimensional features, called *multi-view cluster complementarity*. The complementary cluster information is then transformed to supervised information with high confidence, aiming to have consistent cluster assignments for all views, *i.e.*, solving the challenge (ii). In addition, we propose an EM-like optimization strategy, including P-step and Z-step, to alternately promote feature learning and clustering.

Different from existing traditional and deep IMVC methods, our contributions can be summarized as follows:

- We present a novel deep IMVC method with an imputation-free and fusion-free framework, which can avoid the noise caused by inaccurate imputation and alleviate the disturbance from the views with low quality.
- We propose to mine the complementary information in the high-dimensional feature space via a nonlinear mapping of multiple views. Moreover, we show the mechanism to achieve multi-view cluster complementarity and clustering consistency.
- We design an alternate (EM-like) optimization strategy to effectively optimize the proposed DIMVC framework. Extensive experiments demonstrate that our method achieves superior clustering performance, compared to state-of-the-art IMVC methods.

2 Method

Notations. Given an incomplete multi-view dataset of N samples $\{\mathbf{X}^v \in \mathbb{R}^{N_v \times D_v}\}_{v=1}^V$, V is the number of views. For the v-th view, D_v is the dimensionality of the data and N_v represents the number of samples, where $N_v \leq N$ due to missing data. K is the number of categories to be clustered. We denote the samples with complete data as a set $\mathcal{X}$.

2.1 Deep model of feature learning and clustering

Firstly, we introduce the feature learning and clustering model of each view, *i.e.*, Model #1, #2, . . . , #V in Figure 1.

Deep autoencoder can capture salient features of the data and has been applied in many unsupervised fields (Feng, Wang, and Li 2014; Song et al. 2018; Xu et al. 2019b; Zhang

et al. 2020; Cao et al. 2020; Lin et al. 2021). Therefore, we employ autoencoders to convert heterogeneous multi-view data into clustering-friendly embedding features. For the v-th view, the embedding features denoted as $\mathbf{Z}^v$ are learned by the encoder and decoder. The encoder is $f_v(\mathbf{X}^v;\theta^v) : \mathbf{X}^v \in \mathbb{R}^{N_v\times D_v} \longmapsto \mathbf{Z}^v \in \mathbb{R}^{N_v\times d_v}$ and the decoder is $f_v^{-1}(\mathbf{Z}^v;\phi^v) : \mathbf{Z}^v \in \mathbb{R}^{N_v\times d_v} \longmapsto \hat{\mathbf{X}}^v \in \mathbb{R}^{N_v\times D_v}$, where d_v is the dimensionality of embedding features, θ^v and ϕ^v are the learnable parameters of autoencoder network. The reconstruction loss between $\mathbf{X}^v$ and $\hat{\mathbf{X}}^v$ of all views is

$$\begin{aligned}\mathcal{L}_{rec} &= \sum_{v=1}^{V} \left\|\mathbf{X}^v - f_v^{-1}(\mathbf{Z}^v)\right\|_F^2 \\ &= \sum_{v=1}^{V}\sum_{i=1}^{N_v} \left\|\mathbf{x}_i^v - f_v^{-1}(f_v(\mathbf{x}_i^v))\right\|_2^2.\end{aligned} \tag{1}$$

In order to obtain clustering predictions, for each view, we utilize a parameterized mapping to obtain soft cluster assignments $\mathbf{Q}^v$, *i.e.*, $\mathcal{M}_v(\mathbf{Z}^v;\mathbf{U}^v) : \mathbf{Z}^v \in \mathbb{R}^{N_v\times d_v} \longmapsto \mathbf{Q}^v \in \mathbb{R}^{N_v\times K}$, where $\mathbf{U}^v = [\mathbf{u}_1^v;\mathbf{u}_2^v;\ldots;\mathbf{u}_K^v] \in \mathbb{R}^{K\times d_v}$ represent the learnable parameters. Concretely,

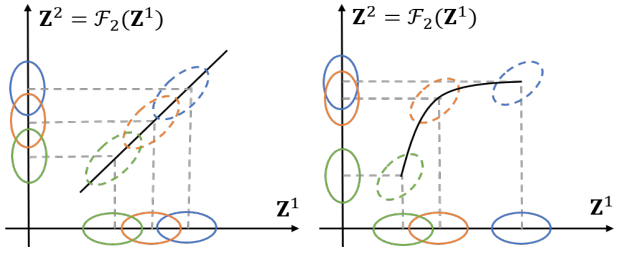
$$q_{ij}^v = \frac{(1+\|\mathbf{z}_i^v - \mathbf{u}_j^v\|_2^2)^{-1}}{\sum_{j=1}^{K}(1+\|\mathbf{z}_i^v - \mathbf{u}_j^v\|_2^2)^{-1}} \in \mathbf{Q}^v, \tag{2}$$

which is a commonly used manner to perform end-to-end clustering (Xie, Girshick, and Farhadi 2016; Guo et al. 2017; Xu et al. 2021a). In the v-th view, $\mathbf{u}_j^v$ is the j-th cluster centroid and q_{ij}^v is considered as the probability that the embedding feature $\mathbf{z}_i^v$ is assigned to the j-th cluster.

There is no connection among the different views so far, and the complete and incomplete data of each view can be learned without imputation for the missing data. Subsequently, we present our strategy to explore complementary information among all views for multi-view clustering.

2.2 Multi-view cluster complementarity

Since multiple views share common semantic information, every view can be regarded as the mappings of the other views, *e.g.*, $\mathbf{Z}^2 = \mathcal{F}_2(\mathbf{Z}^1)$ is a mapping of $\mathbf{Z}^1$. As shown



(a) $\mathcal{F}_2$ is linear mapping (b) $\mathcal{F}_2$ is nonlinear mapping

Figure 2: Illustration of the mapping between two views.

in Figure 2(a), if one view is a linear mapping of the other views, there is no complementary information among them. If there exists complementary relationship among multiple views, *e.g.*, the inseparable clusters in one view are separable in other views, this complementarity can be described by nonlinear mappings as shown in Figure 2(b).

Based on the above observations, we treat the clustering problem as a classification problem by considering the cluster assignments as the pseudo labels of samples.

Assumption 1. *Cover's theorem (Cover 1965): a complex classification problem is more likely to be linearly separable when it is nonlinearly projected to a high dimensional space.*

Considering Assumption 1, we propose to map multi-view embeddings into a high-dimensional space by a nonlinear mapping $\mathcal{H}$. Many functions can lead to such $\mathcal{H}$. In this paper, we provide a simple method, *i.e.*, mapping the embeddings into a concatenated weighted feature space (CWFS):

$$\mathbf{H} = \mathcal{H}(\{\mathbf{Z}^v\}_{v=1}^V) = (w_1\mathbf{Z}^1,\ w_2\mathbf{Z}^2,\ \ldots,\ w_V\mathbf{Z}^V), \tag{3}$$

where $\mathbf{H} \in \mathbb{R}^{|\mathcal{X}|\times\sum_{v=1}^V d_v}$ denotes the obtained high-dimensional features. w_v is the weight calculated by

$$w_v = 1 + \log\left(1 + \frac{\sigma(\mathbf{U}^v)}{\sum_v \sigma(\mathbf{U}^v)}\right), \tag{4}$$

where $\sigma(\mathbf{U}^v)$ is the variance on cluster centroids $\mathbf{U}^v$ of the v-th view. Intuitively, the high-separability view means that the features have well-separated clusters, whose centroids have large variance. Therefore, in CWFS, the weights $\{w_v\}_{v=1}^V$ are proposed to increase the influence of the high-separability view as well as to reduce the influence of other views with unclear cluster structures. In this way, the mapping $\mathcal{H}$ can push the cluster assignment of a sample in CWFS in agreement with that in the high-separability view.

Additionally, Theorem 1 indicates that the linearly separable probability of $\mathbf{H}$ is improved compared with the embedding features $\{\mathbf{Z}^v\}_{v=1}^V$ of any single view.

Theorem 1. *$\mathcal{H}$ is a high-dimensional nonlinear mapping of $\{\mathbf{Z}^v\}_{v=1}^V$. The high-dimensional features $\mathbf{H}$ are more likely to be linearly separable than any $\mathbf{Z}^v$ for $v \in \{1,2,\ldots,V\}$.*

Proof. Considering the embedding feature $\mathbf{z}_i^v$ of any i-th sample in any v-th view, we assume that there are unknown mappings that make $\mathbf{z}_i^t = \mathcal{F}_t(\mathbf{z}_i^v) \in \mathbb{R}^{d_t}$ for $t \in \{1,2,\ldots V\}$. If some of the mappings $\{\mathcal{F}_t : \mathbf{Z}^v \longmapsto \mathbf{Z}^t\}_{t\neq v}^V$ are nonlinear, the concatenation of all embedding features, *i.e.*, $(\mathbf{z}_i^1,\mathbf{z}_i^2,\ldots,\mathbf{z}_i^V) = (\mathcal{F}_1(\mathbf{z}_i^v),\mathcal{F}_2(\mathbf{z}_i^v),\ldots,\mathcal{F}_V(\mathbf{z}_i^v)) \in \mathbb{R}^{\sum_{v=1}^V d_v}$ where $\sum_{v=1}^V d_v > d_v$, is a high-dimensional nonlinear mapping of $\mathbf{z}_i^v$. Additionally, w_v is a nonlinear weighting function, and thus the proposed $\mathcal{H} : \{\mathbf{Z}^v\}_{v=1}^V \longmapsto \mathbf{H}$ is also a high-dimensional nonlinear mapping for the embedding features of each view. Therefore, given Assumption 1, Theorem 1 holds. □

Based on Theorem 1, the proposed mapping $\mathcal{H}$ ensures that the obtained $\mathbf{H}$ contains more separable cluster patterns than that of single-view embedding features, which is achieved by utilizing the information of the clusters in the high-separability views to avoid the incorrect information from the views with unclear cluster structures (verified in Table 3). This is called *multi-view cluster complementarity*

in this paper. Consequently, in CWFS, we can obtain the new cluster centroids $\mathbf{C}$ by the following objective:

$$\begin{aligned}\mathcal{L}_{com} &= \min_{\{\mathbf{C}^v\}_{v=1}^V} \sum_{i\in\mathcal{X}}\sum_{j=1}^{K}\sum_{v=1}^{V} \left\|w_v\mathbf{z}_i^v - \mathbf{c}_j^v\right\|_2^2 \\ &= \min_{\mathbf{C}} \sum_{i\in\mathcal{X}}\sum_{j=1}^{K} \|\mathbf{h}_i - \mathbf{c}_j\|_2^2,\end{aligned} \tag{5}$$

where $\mathbf{C} \in \mathbb{R}^{K\times\sum_{v=1}^V d_v}$ and $\mathbf{c}_j = (\mathbf{c}_j^1, \mathbf{c}_j^2, \ldots, \mathbf{c}_j^V) \in \mathbb{R}^{\sum_{v=1}^V d_v}$. The multi-view cluster complementarity is learned from the complete data $\mathcal{X}$. We further discuss the details to obtain consistent cluster assignments for all data, *i.e.*, clustering consistency for incomplete multi-view data.

2.3 Multi-view clustering consistency

In CWFS, the linearly separable cluster information of all views is transferred to the high-dimensional features $\mathbf{H}$. The centroids $\mathbf{C}$ calculated on $\mathbf{H}$ reflect more accurate cluster structures due to the multi-view cluster complementarity. Motivated by (Xu et al. 2021c), we can generate supervised information for all views by a mapping $\mathcal{M}(\mathbf{H};\mathbf{C},\mathbf{A}) : \mathbf{H} \in \mathbb{R}^{|\mathcal{X}|\times\sum_{v=1}^V d_v} \longmapsto \mathbf{P} \in \mathbb{R}^{|\mathcal{X}|\times K}$, which is formulated by

$$\mathbf{P} = \mathcal{M}(\mathbf{H};\mathbf{C},\mathbf{A}) = \mathcal{E}(\mathcal{S}(\mathbf{H},\mathbf{C}))\mathbf{A}, \tag{6}$$

where the function $\mathcal{S}$ is leveraged to measure the confidence s_{ij} of the i-th sample being assigned to the j-th centroid:

$$s_{ij} = \mathcal{S}(\mathbf{h}_i, \mathbf{c}_j) = \frac{1}{1 + \|\mathbf{h}_i - \mathbf{c}_j\|_2^2} \in \mathbf{S}. \tag{7}$$

In this way, the confidence s_{ij} is high when $\mathbf{h}_i$ is closer to $\mathbf{c}_j$. The function $\mathcal{E}(\mathbf{S})$ scales the confidence of each sample to $[0, 1]$ and meanwhile, enhances the confidence (such as s_{ij}) when it is the largest in $\{s_{i1}, s_{i2}, \ldots, s_{iK}\}$. Specifically,

$$s_{ij} = \mathcal{E}(\mathbf{s}_i) = \frac{(s_{ij}/\sum_j s_{ij})^2}{\sum_j (s_{ij}/\sum_j s_{ij})^2}, \tag{8}$$

by which the mined complementary cluster information is transformed to the supervised information with high confidence. $\mathbf{A}$ satisfies $\mathbf{A}\mathbf{A}^T = \mathbf{I}_K$, which is a boolean matrix to adjust the arrangement of $\mathbf{S}$. Furthermore, the cross-entropy loss between $\mathbf{P}$ and $\mathbf{Q}^v$ of all views is optimized:

$$\mathcal{L}_{con} = \sum_{v=1}^{V} H(\mathbf{P}, \mathbf{Q}^v) = -\sum_{v=1}^{V}\sum_{i\in\mathcal{X}} \mathbf{p}_i \log \mathbf{q}_i^v. \tag{9}$$

As the same $\mathbf{P}$ is shared by all views, the optimization of Eq. (9) can achieve the consistency of multi-view clustering, *i.e.*, the consistent $\{\mathbf{Q}^v\}_{v=1}^V$. Moreover, the consistency from the complete data can be generalized to the incomplete data through deep models (verified in Figure 3). After training the models, we can average the cluster assignments of the observed data to obtain robust results. Concretely, the clustering prediction of the i-th sample is inferred by

$$y_i = \arg\max_j \sum_v q_{ij}^v. \tag{10}$$

Significantly, the models trained by our method is fusion-free, because each view has its own clustering predictions and does not depend on the feature fusion of multiple views.

2.4 Optimization

In conclusion, the loss function of our proposed framework consists of three parts:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{rec} + \mathcal{L}_{com} + \mathcal{L}_{con} \\ &= \min_{\mathbf{C},\mathbf{A},\{\mathbf{Z}^v,\mathbf{U}^v\}_{v=1}^V} \sum_{v=1}^{V} \left\|\mathbf{X}^v - f_v^{-1}(\mathbf{Z}^v)\right\|_F^2 \\ &+ \sum_{i\in\mathcal{X}}\sum_{j=1}^{K} \|\mathbf{h}_i - \mathbf{c}_j\|_2^2 + \sum_{v=1}^{V} H(\mathbf{P}, \mathbf{Q}^v), \\ s.t.\ &\mathbf{P} = \mathcal{M}(\mathbf{H};\mathbf{C},\mathbf{A}), \mathbf{A}\mathbf{A}^T = \mathbf{I}_K, \mathbf{Q}^v = \mathcal{M}_v(\mathbf{Z}^v;\mathbf{U}^v),\end{aligned} \tag{11}$$

where $\mathbf{Z}^v = f_v(\mathbf{X}^v), \mathbf{H} = \mathcal{H}(\{\mathbf{Z}^v\}_{v=1}^V)$. $\mathcal{L}_{rec}$ is the reconstruction loss of autoencoders. $\mathcal{L}_{com}$ and $\mathcal{L}_{con}$ achieve the multi-view complementarity and consistency, respectively.

To optimize the above non-differentiable objective function, we present an alternate optimization strategy which is similar to Expectation Maximization algorithm, as follows:

Initialization: Firstly, the deep autoencoders are initialized by Eq. (1) to obtain meaningful embedding features. As thus, the cluster centroids $\{\mathbf{U}^1, \mathbf{U}^2, \ldots, \mathbf{U}^V\}$ can be initialized by K-means (MacQueen 1967). $\mathbf{A}$ is initialized as $\mathbf{I}_K$.

P-step: Update $\{\mathbf{P}, \mathbf{C}, \mathbf{A}\}$ with fixed $\{\mathbf{Z}^v, \mathbf{U}^v\}_{v=1}^V$.

In the first place, $\mathbf{C}$ is obtained by optimizing Eq. (5), which can be efficiently calculated with K-means.

Letting $l_i^{(t)} = \text{argmax}_j\, s_{ij}^{(t)}$ denote the cluster label of $\mathbf{h}_i$ in the t-th iteration, in unsupervised context, the clusters represented by $l_i^{(t+1)}$ and $l_i^{(t)}$ might be not consistent. Letting $\tilde{m}_{ij} = \sum_{n\in\mathcal{X}} \mathbb{1}[l_n^{(t+1)} = i]\mathbb{1}[l_n^{(t)} = j]$, we define a cost matrix $\mathbf{M} \in \mathbb{R}^{K\times K}$, where $m_{ij} = \max_{i,j} \tilde{m}_{ij} - \tilde{m}_{ij}$, and solve a maximum matching problem as follows:

$$\begin{aligned}\min_{\mathbf{A}} &\sum_{i=1}^{K}\sum_{j=1}^{K} m_{ij}a_{ij} \\ s.t.\ &\mathbf{A}\mathbf{A}^T = \mathbf{I}_K,\end{aligned} \tag{12}$$

where $\mathbf{A} \in \mathbb{R}^{K\times K}$ is a boolean matrix. Eq. (12) is optimized with Hungarian algorithm (Jonker and Volgenant 1986).

Subsequently, $\mathbf{P}$ can be computed directly by the mapping $\mathcal{M}(\mathbf{H};\mathbf{C},\mathbf{A})$ with the inputs of $\mathbf{H}$, $\mathbf{C}$, and $\mathbf{A}$.

Z-step: Update $\{\mathbf{Z}^v, \mathbf{U}^v\}_{v=1}^V$ with fixed $\{\mathbf{P}, \mathbf{C}, \mathbf{A}\}$.

Given fixed $\mathbf{C}$ and $\mathbf{A}$, $\mathbf{P}$ is treated as constant pseudo labels for all views. Eq. (11) can be divided into $\{\mathcal{L}^1, \mathcal{L}^2, \ldots, \mathcal{L}^V\}$, where $\mathcal{L}^v = \mathcal{L}_{rec}^v + \mathcal{L}_{con}^v = \left\|\mathbf{X}^v - f_v^{-1}(\mathbf{Z}^v)\right\|_F^2 + H(\mathbf{P}, \mathbf{Q}^v)$. In this way, the model of each view can be learned independently. Letting λ denote the learning rate and n be the batch size, we train the deep model via the mini-batch gradient descent algorithm:

$$\begin{aligned}\mathbf{U}^v &= \mathbf{U}^v - \frac{\lambda}{n}\sum_{i=1}^{n}\frac{\partial\mathcal{L}^v}{\partial\mathbf{U}^v}, \\ \mathbf{Z}^v &= \mathbf{Z}^v - \frac{\lambda}{n}\sum_{i=1}^{n}\frac{\partial\mathcal{L}^v}{\partial\mathbf{Z}^v},\end{aligned} \tag{13}$$

Algorithm 1: Optimization of the proposed DIMVC.

Input: dataset $\{\mathbf{X}^v \in \mathbb{R}^{N_v \times D_v}\}_{v=1}^V$, number of clusters K
Output: clustering predictions

Initialization: initialize $\{\mathbf{Z}^v\}_{v=1}^V$ by Eq. (1), initialize $\{\mathbf{U}^v\}_{v=1}^V$ by K-means, $\mathbf{A} = \mathbf{I}_K$
while not reaching the maximal iterations **do**
  P-step: fix $\{\mathbf{Z}^v, \mathbf{U}^v\}_{v=1}^V$
    update $\mathbf{C}$ by Eq. (5)
    update $\mathbf{A}$ by Eq. (12)
    update $\mathbf{P}$ by Eq. (6)
  Z-step: fix $\{\mathbf{P}, \mathbf{C}, \mathbf{A}\}$
    update $\{\mathbf{Z}^v, \mathbf{U}^v\}_{v=1}^V$ by Eq. (13)
end while
Calculate the clustering predictions by Eqs. (2 and 10)

where $\mathbf{Z}^v$ is optimized by updating the neural network parameters (*i.e.*, θ^v and ϕ^v) of the autoencoder.

Furthermore, the alternate (EM-like) strategy can make the feature learning and the clustering promote each other (verified in Table 3). Concretely, the P-step produces more precise supervised information by mining cluster complementarity from the embedding features of all views. The Z-step makes the model of each view learn better clustered embedding features with the supervised information.

Complexity analysis. Letting V, K, and N, respectively, denote the number of views, clusters and samples, D represent the maximum number of neurons in deep autoencoders' hidden layers, and $M = \sum_{v=1}^V d_v$ denote the dimensionality of high-dimensional features, $N \gg V, K, M$ generally holds. In Algorithm 1, the complexity to optimize Eq. (5) and Eq. (12) in the P-step is $O(NMK)$ and $O(K^3 + NK)$, respectively, while the complexity to optimize Eq. (13) in the Z-step is $O(NVD^2)$. In conclusion, the total complexity of our algorithm is $O(K^3 + NK + NMK + NVD^2)$ in each iteration, which is linear to data size N.

3 Experiments

We evaluate the effectiveness of our proposed DIMVC by comparing it with seven state-of-the-art IMVC methods on real-world multi-view datasets, in terms of three clustering metrics including accuracy (ACC), normalized mutual information (NMI), and adjusted rand index (ARI).

3.1 Settings

Comparison methods. The comparison methods include four traditional methods (*i.e.*, SRLC (Zhuge et al. 2019), APMC (Guo and Ye 2019), TMBSD (Li et al. 2021), and IMVTSC-MVI (Wen et al. 2021)) and three deep methods (*i.e.*, DiMVMC (Wei et al. 2020), CDIMC-net (Wen et al. 2020), and COMPLETER (Lin et al. 2021)).

Datasets. We use four datasets in our experiments, *i.e.*, BDGP (Cai et al. 2012), Caltech (Fei-Fei, Fergus, and Perona 2004), RGB-D (Kong et al. 2014), and Scene (Fei-Fei and Perona 2005). Table 1 presents the description of the used datasets. We construct incomplete multi-view datasets with varying missing rates (0.1, 0.3, 0.5, 0.7). When the missing rate is 0.5, for example, we randomly select 50% samples and randomly drop partial views of these samples.

Dataset	Type	# Sample	# Class
BDGP	Visual and textual views	2,500	5
Caltech	HOG and GIST	2,386	20
RGB-D	Visual and textual views	1,449	13
Scene	GIST, PHOG, and LBP	4,485	15

Table 1: The description of datasets.

Implementation details. For our DIMVC, the following settings are adopted for all datasets. Concretely, the autoencoders of all views are implemented by fully connected neural networks with the same structure. For the v-th view, the network structure can be denoted as $\mathbf{X}^v - \text{Fc}_{500} - \text{Fc}_{500} - \text{Fc}_{2000} - \mathbf{Z}^v - \text{Fc}_{2000} - \text{Fc}_{500} - \text{Fc}_{500} - \hat{\mathbf{X}}^v$, where Fc_{500} represents the fully connected neural network with 500 neurons. The dimensionality of embeddings $\mathbf{Z}^v$ is reduced to 10. The activation function is ReLU (Glorot, Bordes, and Bengio 2011). We adopt Adam (Kingma and Ba 2014) to optimize the deep models with a learning rate of 0.001. In the initialization phase, the autoencoders are pre-trained for 500 epochs. The batch size is set to 256. In every iteration of the proposed alternate (EM-like) optimization strategy, the Z-step will train the deep models for 1000 batches after the P-step updates the learning targets. The number of iterations is set to 10. The code is provided in the website[1].

3.2 Experimental results and analysis

The clustering performance of all methods on four datasets is listed in Table 2, from which we have the following observations: (1) Our DIMVC obtains the best performance on all datasets. Compared with the second-best methods, DIMVC has considerable improvements especially on BDGP, Caltech, and Scene. (2) It is obvious that the clustering performance of all methods is reduced when the missing rate varies from 0.1 to 0.7. Nevertheless, our DIMVC still achieves superior clustering performance in most cases.

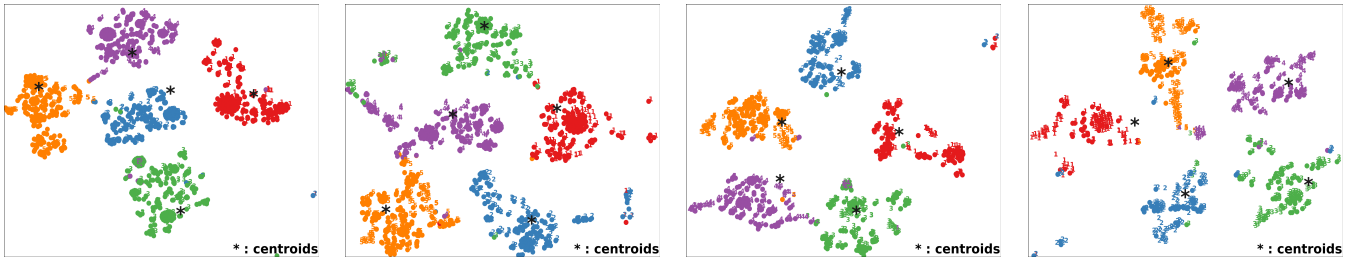
The reasons for the above observations can be explained as follows: (1) If the missing rate of multi-view data becomes high, the complementary information among multiple views becomes rare. This results in the reduction of clustering quality of all methods. (2) Imputation methods depend on the estimation of data distribution. Hence, the cumulative error increases when the missing rate is high, *e.g.*, the performance of certain methods (such as CDIMC-net and TMBSD) is poor when the missing rate is 0.5 or 0.7. (3) Fusion methods might be influenced by the views with low quality, especially for incomplete multi-view learning, *e.g.*, even on BDGP with low missing rates, the performance of certain methods (like DiMVMC and COMPLETER) is bad.

Different from existing traditional and deep IMVC methods, our DIMVC is imputation-free and fusion-free. It explores the cluster complementarity by the proposed high-dimensional mapping, and obtains robust results through the EM-like optimization strategy. In addition to clustering

[1] https://github.com/SubmissionsIn/DIMVC.

	Missing rates	0.1			0.3			0.5			0.7		
	Evaluation metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
BDGP	SRLC (Zhuge et al. 2019)	0.691	0.514	0.470	0.697	0.458	0.430	0.626	0.366	0.320	0.566	0.379	0.333
	APMC (Guo and Ye 2019)	0.859	0.662	0.683	0.814	0.589	0.594	0.770	0.672	0.626	0.749	0.580	0.556
	TMBSD (Li et al. 2021)	0.739	0.620	0.582	0.714	0.597	0.546	0.605	0.273	0.275	0.510	0.236	0.218
	IMVTSC-MVI (Wen et al. 2021)	0.962	0.880	0.908	0.934	0.816	0.844	0.931	0.813	0.836	0.592	0.411	0.216
	DiMVMC (Wei et al. 2020)	0.730	0.673	0.560	0.730	0.677	0.565	0.595	0.511	0.282	0.479	0.310	0.181
	CDIMC-net (Wen et al. 2020)	0.875	0.755	0.640	0.757	0.692	0.467	0.657	0.530	0.271	0.549	0.426	0.225
	COMPLETER (Lin et al. 2021)	0.596	0.528	0.254	0.552	0.511	0.255	0.541	0.504	0.228	0.529	0.426	0.203
	DIMVC (**ours**)	**0.964**	**0.892**	**0.912**	**0.954**	**0.866**	**0.889**	**0.947**	**0.845**	**0.873**	**0.929**	**0.802**	**0.831**
Caltech	SRLC (Zhuge et al. 2019)	0.478	0.588	0.349	0.453	0.566	0.311	0.433	0.570	0.304	0.416	0.468	0.269
	APMC (Guo and Ye 2019)	0.523	0.625	0.421	0.437	0.600	0.302	0.429	0.597	0.295	0.427	0.585	0.293
	TMBSD (Li et al. 2021)	0.404	0.608	0.274	0.415	0.615	0.281	0.407	0.620	0.286	0.418	0.606	0.284
	IMVTSC-MVI (Wen et al. 2021)	0.625	0.642	0.507	0.590	0.668	0.445	0.549	0.520	0.395	0.461	0.517	0.333
	DiMVMC (Wei et al. 2020)	0.373	0.551	0.315	0.358	0.521	0.301	0.322	0.454	0.285	0.304	0.342	0.242
	CDIMC-net (Wen et al. 2020)	0.452	0.589	0.325	0.443	0.439	0.265	0.362	0.400	0.184	0.327	0.353	0.106
	COMPLETER (Lin et al. 2021)	0.742	0.712	0.843	0.741	0.690	0.835	0.716	0.681	0.784	0.697	**0.655**	0.763
	DIMVC (**ours**)	**0.772**	**0.726**	**0.870**	**0.761**	**0.697**	**0.842**	**0.758**	**0.685**	**0.835**	**0.710**	0.638	**0.802**
RGB-D	SRLC (Zhuge et al. 2019)	0.383	0.237	0.152	0.352	0.214	0.126	0.322	0.184	0.104	0.317	0.170	0.099
	APMC (Guo and Ye 2019)	0.412	0.319	0.216	0.373	0.271	0.166	0.345	0.254	0.149	0.308	0.226	0.123
	TMBSD (Li et al. 2021)	0.377	0.323	**0.263**	0.317	0.210	0.132	0.299	0.182	0.107	0.274	0.167	0.096
	IMVTSC-MVI (Wen et al. 2021)	0.422	0.322	0.229	0.401	0.311	0.193	0.362	0.267	0.151	0.318	0.228	0.124
	DiMVMC (Wei et al. 2020)	0.371	0.323	0.209	0.355	0.269	0.178	0.251	0.222	0.116	0.248	0.173	0.079
	CDIMC-net (Wen et al. 2020)	0.358	0.334	0.249	0.393	**0.368**	0.161	0.302	0.270	0.149	0.260	0.187	0.106
	COMPLETER (Lin et al. 2021)	0.418	0.265	0.237	0.398	0.236	0.213	0.370	0.236	0.191	0.348	0.203	0.102
	DIMVC (**ours**)	**0.436**	**0.353**	0.258	**0.405**	0.316	**0.215**	**0.391**	**0.288**	**0.193**	**0.380**	**0.289**	**0.159**
Scene	SRLC (Zhuge et al. 2019)	0.366	0.351	0.189	0.332	0.307	0.163	0.333	0.292	0.148	0.299	0.264	0.137
	APMC (Guo and Ye 2019)	0.433	0.434	0.269	0.414	0.401	0.248	0.408	0.383	0.236	0.388	0.346	0.213
	TMBSD (Li et al. 2021)	0.437	0.398	0.271	0.364	0.325	0.185	0.344	0.293	0.166	0.300	0.253	0.131
	IMVTSC-MVI (Wen et al. 2021)	0.330	0.302	0.161	0.277	0.245	0.117	0.265	0.211	0.101	0.226	0.184	0.082
	DiMVMC (Wei et al. 2020)	0.315	0.291	0.156	0.241	0.206	0.083	0.183	0.136	0.045	0.173	0.137	0.042
	CDIMC-net (Wen et al. 2020)	0.346	0.374	0.143	0.246	0.219	0.112	0.309	0.288	0.136	0.264	0.228	0.121
	COMPLETER (Lin et al. 2021)	–	–	–	–	–	–	–	–	–	–	–	–
	DIMVC (**ours**)	**0.474**	**0.465**	**0.306**	**0.440**	**0.403**	**0.254**	**0.428**	**0.394**	**0.252**	**0.401**	**0.355**	**0.221**

Table 2: Clustering results of all methods on four datasets. The best and second-best results are highlighted with bold and underline, respectively. The symbol '–' denotes unknown results as COMPLETER mainly focuses on two-view clustering.



(a) Missing rate = 0.1 (b) Missing rate = 0.3 (c) Missing rate = 0.5 (d) Missing rate = 0.7

Figure 3: Visualization of the embedding features and centroids on BDGP (textual view) with 4 missing rates via t-SNE (Maaten and Hinton 2008). Dots denote the samples with complete data and digits represent the samples with incomplete data.

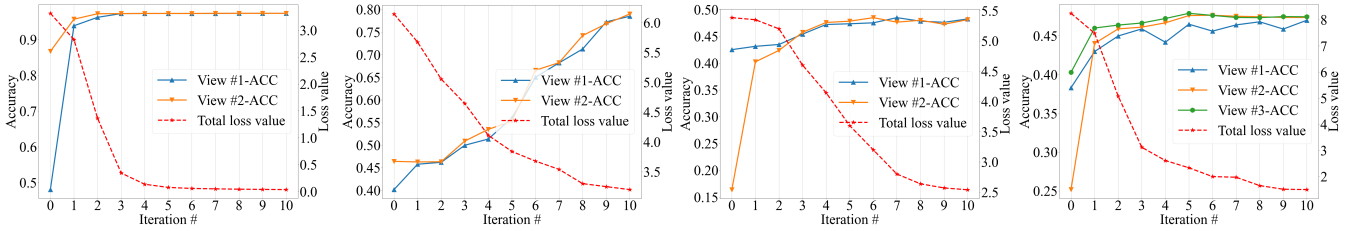
performance, the learned embedding features and centroids are visualized in Figure 3. We find that the incomplete data points have the similar cluster structures in accord with that of the complete data points. This indicates that our method achieves the consistency for the complete data and incomplete data, and has good feature generalization capability.

3.3 Ablation study

We conduct the ablation study to demonstrate the importance of each component of our method. As shown in Table 3, View-v denotes the K-means clustering performance on the v-th view's embedding features obtained by the autoencoder (AE). Item-1 obtains better performance than any single view, which experimentally validates our Theorem 1, *i.e.*, the multi-view complementary information mined by the mapping $\mathcal{H}$ improves clustering performance. Item-2 does not apply the proposed EM-like strategy to optimize the framework. The results indicate that Item-2 cannot effectively leverage the mined complementary information. The improvement of Item-3 is limited as it does not apply the function $\mathcal{E}(\mathbf{S})$ to enhance the confidence of samples. The

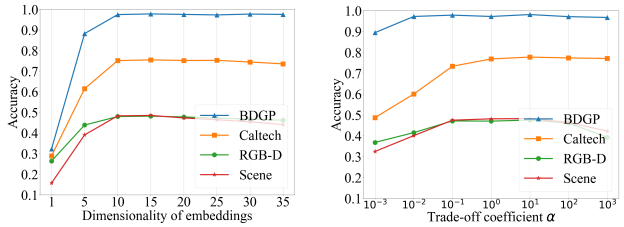
	Components				BDGP			Caltech			RGB-D			Scene		
	AE	$\mathcal{H}$	$\mathcal{E}(\mathbf{S})$	EM	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
View-1	√				0.473	0.298	0.260	0.454	0.611	0.402	0.411	0.344	0.234	0.357	0.370	0.192
View-2	√				0.876	0.750	0.708	0.413	0.546	0.348	0.162	0.056	0.017	0.274	0.198	0.106
View-3	√													0.370	0.401	0.224
Item-1	√	√			0.935	0.833	0.844	0.455	0.636	0.419	0.433	0.392	0.268	0.437	0.429	0.271
Item-2	√	√	√		0.506	0.393	0.270	0.416	0.652	0.398	0.237	0.181	0.131	0.235	0.256	0.099
Item-3	√	√		√	0.962	0.912	0.909	0.476	0.659	0.436	0.366	0.224	0.158	0.424	0.421	0.256
Item-4	√	√	√	√	0.982	0.932	0.945	0.792	0.731	0.879	0.475	0.419	0.302	0.490	0.482	0.319

Table 3: Ablation experiments on four datasets with missing rate $= 0$.



(a) BDGP (b) Caltech (c) RGB-D (d) Scene

Figure 4: Accuracy (ACC) and loss values on four datasets.



(a) ACC $v.s.$ Dimensionality (b) ACC $v.s.$ α

Figure 5: Dimensionality and parameter analysis.

best performance is achieved by Item-4, which indicates that the different parts of our framework are important.

3.4 Model analysis

Convergence and training process. In our proposed alternate (EM-like) optimization strategy, both the P-step and the Z-step are performed at each iteration. We report the convergence of our optimization strategy in Figure 4, where we record the total loss at the end of every iteration. In the training process, the clustering accuracy of all views is gradually improved. It indicates that the complementary information mined by the P-step improves the separability of embedding features learned by the Z-step and vice versa. In this way, the P-step and the Z-step can promote each other alternately.
Dimensionality. We investigate the influence of the dimensionality of embedding features for our proposed high-dimensional mapping and show the results in Figure 5(a). Obviously, the accuracy is low when the dimensionality is set to 1. That is because 1-dimensional embedding features cannot capture the salient information well. The performance is stable for the high-dimensional features, which demonstrates that our method is still feasible in a very high-dimensional feature space. In our experiments, the dimensionality of embedding features is set to 10 for all datasets.

Parameter analysis. In the Z-step of our proposed optimization strategy, the loss of each view is $\mathcal{L}^v = \mathcal{L}^v_{rec} + \mathcal{L}^v_{con}$. We investigate whether a coefficient is needed to balance $\mathcal{L}^v_{rec}$ and $\mathcal{L}^v_{con}$, *i.e.*, $\mathcal{L}^v = \mathcal{L}^v_{rec} + \alpha\mathcal{L}^v_{con}$. As shown in Figure 5(b), the performance reduces to some extend when the value of α is too small and too large (*e.g.*, 10^{-3} and 10^3). This indicates that both $\mathcal{L}^v_{rec}$ and $\mathcal{L}^v_{con}$ are important. They contribute on maintaining the salient information contained in embedding features of autoencoders as well as achieving the consistency of multi-view clustering, respectively. In conclusion, the hyper-parameter α is insensitive in the range of $[10^{-1}, 10^1]$ and we let $\alpha = 1.0$ for all datasets.

4 Conclusion

In this paper, we presented an imputation-free and fusion-free deep incomplete multi-view clustering framework. Firstly, we developed a novel strategy to mine the complementary information among multiple views, *i.e.*, mapping embedding features of all views into the concatenated weighted feature space (CWFS). As a result, the new features in CWFS are more likely linearly separable due to the multi-view cluster complementarity. Furthermore, the feature learning and clustering were conducted with an alternate (EM-like) optimization strategy, where the complementary information was transformed to the supervised information to achieve the consistency of multiple views. In conclusion, our method can effectively explore the multi-view complementary information from the complete data, and has good generalization capability to handle the incomplete data. Extensive experiments demonstrated that our method achieves the state-of-the-art clustering performance.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grants No: 61836016, 61876046, and 61806043) and the Guangxi "Bagui" Teams for Innovation and Research, China.

References

Cai, X.; Wang, H.; Huang, H.; and Ding, C. 2012. Joint stage recognition and anatomical annotation of Drosophila gene expression patterns. *Bioinformatics*, 28(12): i16–i24.

Cao, J.; Bu, Z.; Wang, Y.; Yang, H.; Jiang, J.; and Li, H.-J. 2019. Detecting prosumer-community groups in smart grids from the multiagent perspective. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(8): 1652–1664.

Cao, J.; Wang, Y.; He, J.; Liang, W.; Tao, H.; and Zhu, G. 2020. Predicting Grain Losses and Waste Rate Along the Entire Chain: A Multitask Multigated Recurrent Unit Autoencoder Based Method. *IEEE Transactions on Industrial Informatics*, 17(6): 4390–4400.

Chen, M.; Huang, L.; Wang, C.-D.; and Huang, D. 2020. Multi-View Clustering in Latent Embedding Space. In *AAAI*, 3513–3520.

Cover, T. M. 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 326–334.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR*, 178–178.

Fei-Fei, L.; and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, volume 2, 524–531.

Feng, F.; Wang, X.; and Li, R. 2014. Cross-modal retrieval with correspondence autoencoder. In *ACM MM*, 7–16.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Deep sparse rectifier neural networks. In *Journal of Machine Learning Research*, 315–323.

Guo, J.; and Ye, J. 2019. Anchors bring ease: An embarrassingly simple approach to partial multi-view clustering. In *AAAI*, 118–125.

Guo, X.; Gao, L.; Liu, X.; and Yin, J. 2017. Improved deep embedded clustering with local structure preservation. In *IJCAI*, 1753–1759.

Hu, M.; and Chen, S. 2019. One-pass incomplete multi-view clustering. In *AAAI*, 3838–3845.

Huang, Z.; Hu, P.; Zhou, J. T.; Lv, J.; and Peng, X. 2020. Partially View-aligned Clustering. *NeurIPS*, 33.

Huang, Z.; Ren, Y.; Pu, X.; and He, L. 2021. Non-Linear Fusion for Self-Paced Multi-View Clustering. In *ACM MM*, 3211–3219.

Jonker, R.; and Volgenant, T. 1986. Improving the Hungarian assignment algorithm. *Operations Research Letters*, 5(4): 171–175.

Kang, Z.; Zhao, X.; Peng, C.; Zhu, H.; Zhou, J. T.; Peng, X.; Chen, W.; and Xu, Z. 2020. Partition level multiview subspace clustering. *Neural Networks*, 122: 279–288.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kong, C.; Lin, D.; Bansal, M.; Urtasun, R.; and Fidler, S. 2014. What are you talking about? text-to-image coreference. In *CVPR*, 3558–3565.

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*.

Li, R.; Zhang, C.; Fu, H.; Peng, X.; Zhou, T.; and Hu, Q. 2019. Reciprocal Multi-Layer Subspace Learning for Multi-View Clustering. In *ICCV*, 8172–8180.

Li, S.-Y.; Jiang, Y.; and Zhou, Z.-H. 2014. Partial multi-view clustering. In *AAAI*, 1968–1974.

Li, Z.; Tang, C.; Liu, X.; Zheng, X.; Zhang, W.; and Zhu, E. 2021. Tensor-Based Multi-View Block-Diagonal Structure Diffusion for Clustering Incomplete Multi-View Data. In *ICME*, 1–6.

Lin, Y.; Gou, Y.; Liu, Z.; Li, B.; Lv, J.; and Peng, X. 2021. COMPLETER: Incomplete multi-view clustering via contrastive prediction. In *CVPR*, 11174–11183.

Liu, H.; Li, X.; and Zhang, S. 2016. Learning instance correlation functions for multilabel classification. *IEEE Transactions on Cybernetics*, 47(2): 499–510.

Liu, H.; Liu, L.; Le, T. D.; Lee, I.; Sun, S.; and Li, J. 2017. Nonparametric sparse matrix decomposition for cross-view dimensionality reduction. *IEEE Transactions on Multimedia*, 19(8): 1848–1859.

Liu, J.; Teng, S.; Fei, L.; Zhang, W.; Fang, X.; Zhang, Z.; and Wu, N. 2021a. A novel consensus learning approach to incomplete multi-view clustering. *Pattern Recognition*, 115: 107890.

Liu, X.; Li, M.; Tang, C.; Xia, J.; Xiong, J.; Liu, L.; Kloft, M.; and Zhu, E. 2021b. Efficient and effective regularized incomplete multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8): 2634–2646.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9: 2579–2605.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *BSMSP*, 281–297.

Peng, X.; Huang, Z.; Lv, J.; Zhu, H.; and Zhou, J. T. 2019. COMIC: Multi-view Clustering Without Parameter Selection. In *ICML*, 5092–5101.

Rai, N.; Negi, S.; Chaudhury, S.; and Deshmukh, O. 2016. Partial multi-view clustering using graph regularized NMF. In *ICPR*, 2192–2197.

Song, J.; Zhang, H.; Li, X.; Gao, L.; Wang, M.; and Hong, R. 2018. Self-supervised video hashing with hierarchical binary auto-encoder. *IEEE Transactions on Image Processing*, 27(7): 3210–3221.

Tang, C.; Liu, X.; Zhu, X.; Zhu, E.; Luo, Z.; Wang, L.; and Gao, W. 2020. CGD: Multi-view clustering via cross-view graph diffusion. In *AAAI*, 5924–5931.

Tao, Z.; Liu, H.; Li, S.; Ding, Z.; and Fu, Y. 2017. From ensemble clustering to multi-view clustering. In *IJCAI*, 2843–2849.

Wang, Q.; Ding, Z.; Tao, Z.; Gao, Q.; and Fu, Y. 2018. Partial multi-view clustering via consistent GAN. In *ICDM*, 1290–1295.

Wang, Q.; Ding, Z.; Tao, Z.; Gao, Q.; and Fu, Y. 2021. Generative Partial Multi-View Clustering With Adaptive Fusion and Cycle Consistency. *IEEE Transactions on Image Processing*, 30: 1771–1783.

Wei, J.; Yang, Y.; Xu, X.; Zhu, X.; and Shen, H. T. 2021. Universal Weighting Metric Learning for Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wei, S.; Wang, J.; Yu, G.; Domeniconi, C.; and Zhang, X. 2020. Deep Incomplete Multi-View Multiple Clusterings. In *ICDM*, 651–660.

Wen, J.; Zhang, Z.; Xu, Y.; Zhang, B.; Fei, L.; and Xie, G.-S. 2020. CDIMC-net: Cognitive Deep Incomplete Multi-view Clustering Network. In *IJCAI*, 3230–3236.

Wen, J.; Zhang, Z.; Zhang, Z.; Zhu, L.; Fei, L.; Zhang, B.; and Xu, Y. 2021. Unified Tensor Framework for Incomplete Multi-view Clustering and Missing-view Inferring. In *AAAI*, 10273–10281.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *ICML*, 478–487.

Xu, C.; Guan, Z.; Zhao, W.; Wu, H.; Niu, Y.; and Ling, B. 2019a. Adversarial Incomplete Multi-view Clustering. In *IJCAI*, 3933–3939.

Xu, C.; Tao, D.; and Xu, C. 2015. Multi-view learning with incomplete views. *IEEE Transactions on Image Processing*, 24(12): 5812–5825.

Xu, J.; Ren, Y.; Li, G.; Pan, L.; Zhu, C.; and Xu, Z. 2021a. Deep embedded multi-view clustering with collaborative training. *Information Sciences*, 573: 279–290.

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 *ICCV*, 9234–9243.

Xu, J.; Ren, Y.; Tang, H.; Yang, Z.; Pan, L.; Yang, Y.; and Pu, X. 2021c. Self-Supervised Discriminative Feature Learning for Deep Multi-View Clustering. *arXiv preprint arXiv:2103.15069*.

Xu, X.; Lin, K.; Yang, Y.; Hanjalic, A.; and Shen, H. T. 2020. Joint Feature Synthesis and Embedding: Adversarial Cross-modal Retrieval Revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xu, X.; Lu, H.; Song, J.; Yang, Y.; Shen, H. T.; and Li, X. 2019b. Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. *IEEE Transactions on Cybernetics*, 50(6): 2400–2413.

Ye, Y.; Liu, X.; Liu, Q.; Guo, X.; and Yin, J. 2018. Incomplete multiview clustering via late fusion. *Computational intelligence and neuroscience*, 6148456:1–6148456:11.

Zhang, C.; Cui, Y.; Han, Z.; Zhou, J. T.; Fu, H.; and Hu, Q. 2020. Deep partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhuge, W.; Hou, C.; Liu, X.; Tao, H.; and Yi, D. 2019. Simultaneous Representation Learning and Clustering for Incomplete Multi-view Data. In *IJCAI*, volume 7, 4482–4488.