Incomplete Multi-view Spectral Clustering with Adaptive Graph Learning

Jie Wen, Yong Xu*, *Senior Member*, *IEEE*, Hong Liu

***Abstract*—In this paper, we propose a general framework for incomplete multi-view clustering. The proposed method is the first work that exploits the graph learning and spectral clustering techniques to learn the common representation for incomplete multi-view clustering. Firstly, owing to the good performance of low-rank representation in discovering the intrinsic subspace structure of data, we adopt it to adaptively construct the graph of each view. Secondly, a spectral constraint is used to achieve the low-dimensional representation of each view based on the spectral clustering. Thirdly, we further introduce a co-regularization term to learn the common representation of samples for all views, and then use the *k*-means to partition the data into their respective groups. An efficient iterative algorithm is provided to optimize the model. Experimental results conducted on seven incomplete multi-view datasets show that the proposed method achieves the best performance in comparison with some state-of-the-art methods, which proves the effectiveness of the proposed method in incomplete multi-view clustering.**

***Index Terms*—Incomplete multi-view clustering, low-rank representation, graph learning, co-regularization.**

I. Introduction

In the real-world, an image can be represented by various descriptors such as SIFT, LBP, and HOG, etc [1]. A web page can be described by the images, links, and texts, etc [2]. The values of blood test and magnetic resonance images can be viewed as two references for disease diagnosing [3]. The above phenomenons indicate that almost all things can be represented from different perspectives (views). These features acquired from different views are regarded as the multi-view features [4, 5]. Generally, multi-view features can represent data more comprehensive because features of different views provide many complementary information. Thus exploiting the multi-view features has the potential to improve the performance of different machine learning tasks [6-12].

Multi-view clustering is one of the hottest research directions in this filed, which focuses on adaptively partitioning data points into their respective groups without any label information [13-16]. In the past few years, many multi-view clustering methods have been proposed, such as the multi-view *k*-means clustering [17], canonical correlation analysis (CCA) based method [18], co-regularized multi-view spectral clustering [19], low-rank tensor based method [20], and the deep matrix factorization based method [4], etc. For multi-view clustering, ensuring the clustering agreement among all views is the key to achieving good performance [21]. To this end, Wang et al. introduced a views-agreement constraint to guarantee that the data-cluster graphs learned from all views are consistent to each other [21]. In [14], a structured low-rank matrix factorization based method is proposed to learn more consistent low-dimensional data-cluster representations for all views by exploiting the manifold structures and introducing the divergence constraint term jointly. Besides, in [22], based on the multi-view spectral clustering, a novel angular based regularizer is imposed on the sparse representation vectors of all views to learn the consensus similarity graph shared by all views for clustering. In [8], a more robust multi-view spectral clustering model is proposed, which can learn the optimal similarity graphs and data-cluster representations with views-agreement. For these multi-view clustering methods, they commonly require that all views of data are complete. However, the requirement is often impossible to satisfy because it is often the case that some views of samples are missing in the real-world applications, especially in the applications of disease diagnosing [3] and webpage clustering [23]. This incomplete problem of views leads to the failure of the conventional multi-view methods [24].

To address this issue, many efforts have been made in recent years, which can be generally categorized into two groups in terms of the exploited techniques, *i.e.*, matrix factorization based incomplete multi-view clustering (MFIMC) and graph based incomplete multi-view clustering (GIMC). MFIMC focuses on learning a consensus representation with low dimensionality for all views directly via the matrix factorization technique. For example, partial multi-view clustering (PMVC) seeks to learn a common latent subspace for all views, in which the instances of different views are enforced to have the same representation [3]. Different from PMVC, multi-incomplete-view clustering (MIC) first fills in the missing views with the average of all instances in the corresponding views, and then uses the weighted nonnegative matrix factorization technique to jointly learn the latent representations of different views and the consensus representation for all views [25]. For these matrix based methods, the common problem is that they only focus on learning the consensus representation while ignoring the intrinsic structure of data, which cannot guarantee the compactness and discriminability of the learned representation.

This work was supported in part by the Guangdong Province high-level personnel of special support program under Grant no. 2016TX03X164, and in part by the National Natural Science Foundation of Guangdong Province under Grant no. 2017A030313384. (Corresponding author: Yong Xu (Email: yongxu@ymail.com).)

Jie Wen and Yong Xu are with the Bio-Computing Research Center, Harbin Institute of Technology, Shenzhen, Shenzhen, 518055, Guangdong, China, are also with the Shenzhen Medical Biometrics Perception and Analysis Engineering Laboratory, Harbin Institute of Technology, Shenzhen, Shenzhen, 518055, Guangdong, China. (Email: jiewen_pr@126.com; yongxu@ymail.com.)

Hong Liu is with the Engineering Lab on Intelligent Perception for Internet of Things, Shenzhen Graduate School, Peking University, Shenzhen, 518055, Guangdong, China. (Email: hongliu@pku.edu.cn)

GIMC focuses on learning the low-dimensional representation from different graphs that reveal the relationships of all samples. Compared with the matrix factorization based methods, GIMC can effectively exploit the geometric structures of data. For GIMC, the graph construction is very crucial because it directly influences the clustering performance. However, since some views are missing, it is impossible to construct such graphs to connect all samples completely. To address the issue, Trivedi et al. proposed to complete the incomplete graph of the view with missing instances referring to the Laplacian matrix of the complete view, and then learned the low-dimensional representations for different views via the kernel CCA [23]. The biggest shortcoming of this method is that it requires at least one complete view. Gao et al. proposed to fill in the missing view with the average of instances in the corresponding view for graph construction and subspace learning [26]. However, when the multi-view data has large number of the missing instances in all views, this approach will fail because the filled missing views will dominate the subspace learning [25]. Thus completing the graph by filling in the missing views is not a good choice for incomplete multi-view clustering. In [27], Zhao et al. proposed a more novel graph learning method for incomplete multi-view clustering, in which a robust consensus graph is adaptively learned from the low-dimensional consensus representation. Different from the above graph based methods, the method proposed by Zhao et al. does not need to complete the graph or fill in the missing instances of any view.

Although many methods have been proposed to address the incomplete problem of multi-view clustering, they still have many problems to address. For example, a limitation of these methods is that they require few samples to have features of all views. In other words, for data with more than three views, these methods fail to deal with the case that no sample contains all of the three views. The second limitation is that these graph based methods cannot learn the global optimal consensus representation for clustering because the subspace learning and graph construction are treated independently in two separated steps. To solve the above issues, we in this paper propose a more general framework for multi-view scenarios, which is suitable to all kinds of the multi-view data including arbitrary incomplete cases and the complete case. The proposed method aims to jointly learn the low-dimensional consensus representation and similarity graphs for all views, which enables our method to obtain the global optimal consensus representation and thus has the potential to perform better. To this end, we exploit the low-rank representation technique to adaptively learn the similarity graph of each view owing to its good performance in discovering the intrinsic relationships of data [28, 29]. Meanwhile, considering that graphs constructed from different incomplete views generally have great differences in magnitude, structures, and dimensions owing to the diversity of missing views, it is impossible to directly learn a unified graph or representation for all views [30]. To tackle this problem, we provide an ingenious approach, which indirectly learns the consensus representation from the low-dimensional representations of all views rather than the graphs by introducing a co-regularization term. Meanwhile, the proposed method is robust to noise by introducing the sparse error term to compensate the noise. Overall, this work has the following contributions:

(1) In this paper, we provide a more general multi-view clustering framework, which can deal with both complete and incomplete multi-view cases.

(2) The proposed method is a pioneering work that integrates the graph construction and consensus representation learning into a joint optimization framework for incomplete multi-view clustering. Compared with the other methods, the proposed method is able to learn the optimal similarity graph of each view and the consensus cluster representation for all views, and thus has the potential to obtain a better performance.

(3) The proposed method is robust to noise to some extent. Specially, by introducing the sparse error term to compensate the noise, the proposed method can greatly reduce the negative influence of noise, which is beneficial to discover the intrinsic structure of the noisy data.

The rest of the paper is organized as follows. Section II briefly introduces some related works to the proposed method. In Section III, we present the proposed method and its optimization algorithm in detail. Section IV gives a deep analysis to the proposed method. In Section V, several experiments are conducted to prove the effectiveness of the proposed method. Section VI offers the conclusions of the paper.

II. Related works

A. Single-view spectral clustering

For a dataset $X = [x_1, x_2, \ldots, x_n] \in R^{m\times n}$ with n samples, spectral clustering tries to solve the following problem to learn the low-dimensional representation $F \in R^{n\times c}$ for clustering [31, 32]:

$$\min_{F^TF=I} Tr\left(F^TLF\right) \tag{1}$$

where $Tr\left(\cdot\right)$ denotes the trace operation. $L \in R^{n\times n}$ is the Laplacian graph, which is generally calculated as $L = D - W$ in the ratio cut method [33] and $L = I - D^{-1/2}WD^{-1/2}$ in the normalized cut method [31], where $D \in R^{n\times n}$ is a diagonal matrix and its ith diagonal element is calculated as $D_{i,i} = \sum_j \frac{(W_{i,j}+W_{j,i})}{2}$, $W\ (W \geq 0)$ is the symmetric similarity graph with non-negative elements constructed from the data, I is the identity matrix.

(1) is the typical eigenvalue decomposition problem and its solution is the eigenvector set corresponding to the first c minimum eigenvalues of L. For F, its each row can be viewed as the new representation of the corresponding original sample. Then spectral clustering utilizes the *k*-means algorithm to partition the new representation F into several clusters.

B. Multi-view subspace clustering (MVSC)

Generally, due to the different feature distributions of different views, graphs constructed from different views will have large differences. In this case, it is difficult to find the intrinsic consensus graphs of multiple views for clustering. Fortunately, it is possible to find the optimal and unique representation for

all views. Inspired by this motivation, Gao et al. proposed the multi-view subspace clustering (MVSC), which focuses on learning a consensus cluster indicator matrix for clustering [30]. For a dataset with k views whose vth view is represented as $X^{(v)} = [x_1^{(v)}, x_2^{(v)}, \ldots, x_n^{(v)}] \in R^{m\times n}$, the learning model of MVSC is designed as follows:

$$\begin{aligned}\min_{Z^{(v)},E^{(v)},F} &\sum_v \left\|X^{(v)} - X^{(v)}Z^{(v)} - E^{(v)}\right\|_F^2 \\ &+\sum_v \left(\lambda_1 Tr\left(F^T L^{(v)} F\right) + \lambda_2 \left\|E^{(v)}\right\|_1\right) \\ &s.t.\ Z^{(v)T}1 = 1, Z_{i.i}^{(v)} = 0, F^T F = I\end{aligned} \tag{2}$$

where 1 is a column vector with all ones, $Z_{i.i}^{(v)} = 0$ means that all diagonal elements of matrix $Z^{(v)}$ are 0, λ_1 and λ_2 are penalty parameters. $L^{(v)}$ is the Laplacian graph of the vth view and is calculated as $L^{(v)} = D^{(v)} - W^{(v)}$, where $W^{(v)} = \frac{\left|Z^{(v)}\right| + \left|Z^{(v)}\right|^T}{2}$ and $D_{i.i}^{(v)} = \sum_j W_{i.j}^{(v)}$. $E^{(v)}$ is the error matrix used to model different noises. In (2), $||\cdot||_F$ and $||\cdot||_1$ are the 'Frobenius' norm and 'l_1' norm, respectively [34, 35]. $Tr(\cdot)$ denotes the trace operator.

MVSC integrates the graph construction and low-dimensional representation learning into a joint learning framework, which enables it to find the global optimal indicator matrix for clustering. By alternatively solving all variables, MVSC can find the local optimum of the consensus representation F. Then MVSC performs *k*-means on the consensus representation to obtain the final clustering results.

III. The proposed method

Fig.1 shows two cases of the incomplete multi-view data. In the first case, only few samples contain features of all views and the other samples contain only one view. In the second case, all views are arbitrarily missing. It includes the case that no samples contain the features of all views. For the first case, many methods, such as PMVC [3] and incomplete multi-model grouping (IMG) [27], have been proposed based on the matrix factorization. However, these methods fail to deal with the second incomplete case. To address this issue, in this section, we will propose a more general incomplete multi-view clustering framework that can handle all kinds of incomplete cases.

A. Incomplete multi-view spectral clustering with adaptive graph learning (IMSC_AGL)

From the previous presentation, we can find that all conventional graph based multi-view clustering methods including MVSC, learn the subspace based on the similarity graph of each view. These methods require that graphs constructed from all views are complete, which not only can reveal the similarity relationships of all samples, but also have the same size of $n\times n$ for the multi-view data with n samples. However, as shown in Fig.1, since all views are incomplete, graphs constructed from these incomplete views will have different sizes and also cannot reveal the relationships of all samples, which leads to the failure of the existing GIMC methods.

Some researchers propose to fill in the missing views with the average of instances of the corresponding view, and then construct the similarity graph of each view independently [26]. Although this approach can make the graphs of different views have the same size, it cannot produce the correct representation for each view and thus is harmful to the multi-view clustering. This is mainly because that these missing instances will be regarded as the same class and be connected with the same weight (weight of 1 in the binary nearest neighbor graph), which in turn pulls these missing instances together in the low-dimensional subspace whether they are from the same cluster or not. Therefore, this graph completion approach is unreasonable for incomplete multi-view clustering, especially for the case with large number of missing views. A more reasonable way to avoid this issue is to set the connected weights corresponding to these missing instances as 0 in the similarity graph of the corresponding view. In this way, the uncertain similarity information corresponding to the missing views will not play a negative role in learning the data-cluster representation. In contrast, only the authentic similarity information of the available instances are exploited to guide the representation learning, which is beneficial to achieve a more reliable data-cluster representation and reduces the negative influence of the missing views. Based on the above analysis, we rewrite MVSC as follows for incomplete multi-view learning:

$$\begin{aligned}\min_{Z^{(v)},E^{(v)},F} &\sum_v \left\|Y^{(v)} - Y^{(v)}Z^{(v)} - E^{(v)}\right\|_F^2 \\ &+\sum_v \left(\lambda_1 Tr\left(F^T \bar{L}^{(v)} F\right) + \lambda_2 \left\|E^{(v)}\right\|_1\right) \\ &s.t.\ Z^{(v)T}1 = 1, Z_{i.i}^{(v)} = 0, F^T F = I\end{aligned} \tag{3}$$

where $Y^{(v)} = \left[y_1^{(v)}, y_2^{(v)}, \ldots, y_{n_v}^{(v)}\right] \in R^{m_v\times n_v}$ denotes the set of the un-missing instances in the vth view, m_v and n_v are the number of the features and un-missing instances of the vth view, respectively. The original instance sets including the missing and un-missing instances in the vth view are still represented as $X^{(v)} = \left[x_1^{(v)}, x_2^{(v)}, \ldots, x_n^{(v)}\right] \in R^{m_v\times n}$ ($n_v < n$). $\bar{L}^{(v)} = \bar{D}^{(v)} - \bar{W}^{(v)}$, $\bar{W}^{(v)} = \frac{\left(\left|\bar{Z}^{(v)}\right| + \left|\bar{Z}^{(v)}\right|^T\right)}{2}$, $\bar{D}_{i.i}^{(v)} = \sum_j \bar{W}_{i.j}^{(v)}$. $\bar{Z}^{(v)} \in R^{n\times n}$ is the complete graph that connects all instances including the missing and un-missing instances of the vth view. In our method, we can exploit the following formula to obtain the completed graph $\bar{Z}^{(v)}$ based on the graph $Z^{(v)} \in R^{n_v\times n_v}$ learned from the un-missing instances:

$$\bar{Z}^{(v)} = G^{(v)T} Z^{(v)} G^{(v)} \tag{4}$$

where $G^{(v)} \in R^{n_v\times n}$ is an index matrix used to complete the graph, whose elements related to the missing instances are enforced to zero. Specially, matrix $G^{(v)}$ is defined as follows:

$$G_{i.j}^{(v)} = \begin{cases} 1, & if\ y_i^{(v)}\ is\ the\ original\ instance\ x_j^{(v)} \\ 0, & otherwise \end{cases} \tag{5}$$

It is simple to prove that the following equation is satisfied:

$$\bar{L}^{(v)} = G^{(v)T} L^{(v)} G^{(v)} \tag{6}$$

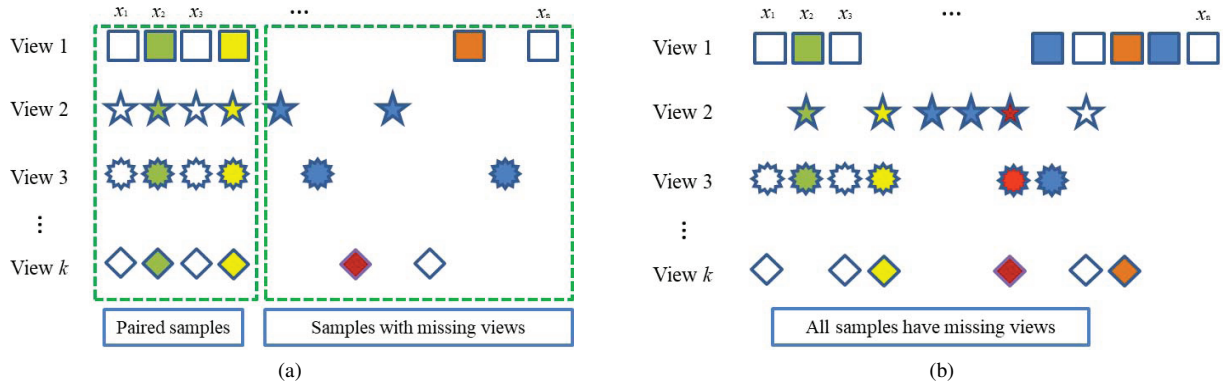


Fig. 1. Two types of the incomplete multi-view data. (a) Some samples contain the features of all views and the remaining samples contain only one view, (b) views are arbitrarily missing, including the case that none of samples contain the features of all views.

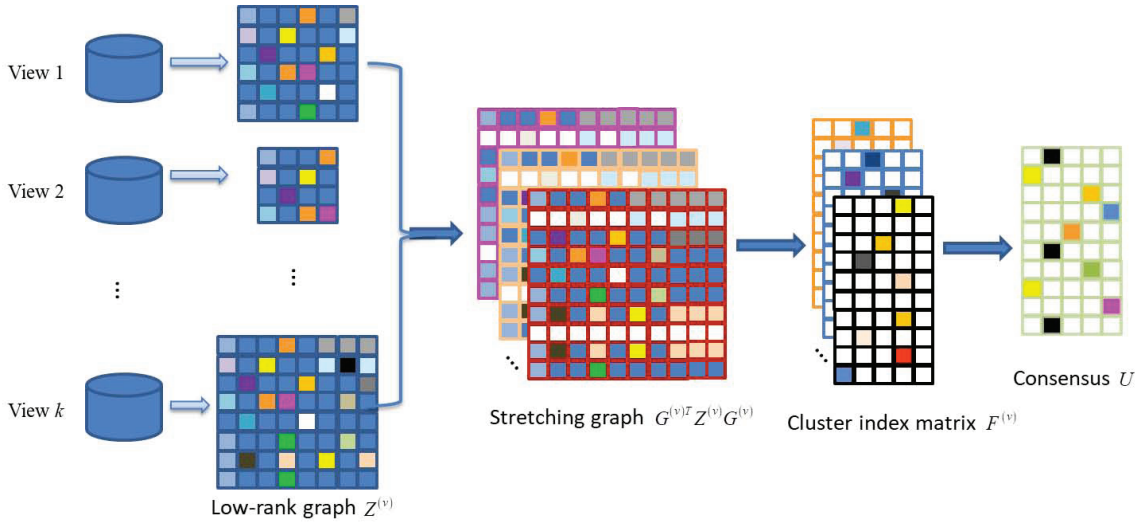


Fig. 2. The framework of the proposed incomplete multi-view clustering method. The proposed method focuses on learning a consensus representation U from multiple views for clustering.

where $L^{(v)}$ is the Laplacian matrix of graph $Z^{(v)}$.

Considering that the data is generally drawn from several low-rank subspaces, thus the learned graph should better discover the low-rank structure of data [28]. Moreover, inspired by the motivation that learning a non-negative graph is beneficial to improve the clustering performance and make the learned graph more interpretable [36], we rewrite model (3) as follows to learn multiple non-negative graphs for multi-view subspace learning:

$$\begin{aligned}\min_{Z^{(v)},E^{(v)},F} &\sum_v \left(\left\|Z^{(v)}\right\|_* + \lambda_2\left\|E^{(v)}\right\|_1\right) \\ &+\lambda_1 \sum_v Tr\left(F^T G^{(v)T} L^{(v)} G^{(v)} F\right) \\ s.t.\ &Y^{(v)} = Y^{(v)}Z^{(v)} + E^{(v)}, Z^{(v)}1 = 1, \\ &0 \le Z^{(v)} \le 1, Z^{(v)}_{i.i} = 0, F^T F = I\end{aligned} \tag{7}$$

where $||Z^{(v)}||_*$ is the nuclear norm of matrix $Z^{(v)}$, which is calculated as the summation of all singular values of matrix $Z^{(v)}$ [37-39]. Constraint $Z^{(v)}1 = 1$ is used to avoid that any sample has no contribution in the joint representation, 1 is a column vector with all elements as 1 [40]. In model (7), F is an $n \times c$ matrix whose each row denotes the representation of the corresponding sample.

In model (7), the third term is equivalent to $\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\|F_{i,:} - F_{j,:}\|_2^2 \sum_v \bar{W}_{i,j}^{(v)}\right)$. This demonstrates that the regularization weights to the target cluster indicator matrix F are the summation of all similarity weights of multiple graphs. For complete data, this approach may be effective to learn the optimal cluster indicator for each sample. However, for incomplete multi-view data, this approach will fail because the regularization weights to the missing instances and un-missing instances are too unfair. It may occur the case that through the summation of the weights of multiple graphs, the regularization weights of samples from different classes may larger than those of the samples from the same class, which will lead to the incorrect cluster indicators of samples. To solve this problem, we propose to learn the consensus representation from those

cluster indicator matrices of all views as follows:

$$\begin{aligned}\min_{Z^{(v)},E^{(v)},F}\sum_v\left(\left\|Z^{(v)}\right\|_*+\lambda_1 Tr\left(F^{(v)T}G^{(v)T}L^{(v)}G^{(v)}F^{(v)}\right)\right)\\+\sum_v\left(\lambda_2\left\|E^{(v)}\right\|_1+\frac{\lambda_3}{2}\Gamma\left(F^{(v)},U\right)\right)\\ s.t.\ Y^{(v)}=Y^{(v)}Z^{(v)}+E^{(v)},Z^{(v)}1=1,\\0\le Z^{(v)}\le 1,Z^{(v)}_{i.i}=0,F^{(v)T}F^{(v)}=I,U^TU=I\end{aligned}\tag{8}$$

where λ_3 is also a penalty parameter, matrix $U\in R^{n\times c}$ is the target cluster indicator matrix (or consensus representation) to learn. Function $\Gamma\left(F^{(v)},U\right)$ measures the disagreement of the consensus representation U and representation $F^{(v)}$. $\Gamma\left(F^{(v)},U\right)$ is defined as follows [19]:

$$\Gamma\left(F^{(v)},U\right)=\left\|\frac{K_U}{\|K_U\|_F^2}-\frac{K_{F^{(v)}}}{\|K_{F^{(v)}}\|_F^2}\right\|_F^2\tag{9}$$

where K_U and $K_{F^{(v)}}$ are the similarity matrixes of U and $F^{(v)}$, respectively. For simplicity, we also choose the linear kernel, *i.e.*, K_U=UU^T as the similarity measure metric [19]. Based on the facts that $\|K_U\|_F^2=c$ and $\|K_{F^{(v)}}\|_F^2=c$, (9) can be rewritten as follows:

$$\Gamma\left(F^{(v)},U\right)=\frac{2\left(c-Tr\left(F^{(v)}F^{(v)T}UU^T\right)\right)}{c^2}\tag{10}$$

As a result, our final model is expressed as follows:

$$\begin{aligned}\min_{Z^{(v)},E^{(v)},F^{(v)},U}\sum_v\left(\left\|Z^{(v)}\right\|_*+\lambda_3\left(c-Tr\left(F^{(v)}F^{(v)T}UU^T\right)\right)\right)\\+\sum_v\lambda_1 Tr\left(F^{(v)T}G^{(v)T}L^{(v)}G^{(v)}F^{(v)}\right)+\sum_v\lambda_2\left\|E^{(v)}\right\|_1\\ s.t.\ Y^{(v)}=Y^{(v)}Z^{(v)}+E^{(v)},Z^{(v)}1=1,\\0\le Z^{(v)}\le 1,Z^{(v)}_{i.i}=0,F^{(v)T}F^{(v)}=I,U^TU=I\end{aligned}\tag{11}$$

Since the proposed IMC framework is based on the adaptive graph learning and spectral clustering, we refer to the proposed method as incomplete multi-view spectral clustering with adaptive graph learning (IMSC_AGL). The framework of our method is briefly outlined in Fig.2.

B. Solution to IMSC_AGL

For the optimization problem (11), we choose the alternating direction method of multipliers (ADMM) to calculate its local optimal solution [41]. At the beginning, we introduce several variables to make problem (11) separable as follows:

$$\begin{aligned}\min_{Z^{(v)},E^{(v)},F^{(v)},U,S^{(v)},W^{(v)}}\sum_v\left(\left\|S^{(v)}\right\|_*-\lambda_3 Tr\left(F^{(v)}F^{(v)T}UU^T\right)\right)\\+\sum_v\lambda_1 Tr\left(F^{(v)T}G^{(v)T}L_w^{(v)}G^{(v)}F^{(v)}\right)+\sum_v\lambda_2\left\|E^{(v)}\right\|_1\\ s.t.\ Y^{(v)}=Y^{(v)}Z^{(v)}+E^{(v)},Z^{(v)}=S^{(v)},Z^{(v)}=W^{(v)},\\W^{(v)}1=1,0\le W^{(v)}\le 1,W^{(v)}_{i.i}=0,F^{(v)T}F^{(v)}=I,U^TU=I\end{aligned}\tag{12}$$

where $L_w^{(v)}$ denotes the Laplacian graph of matrix $W^{(v)}$. Note: since c is a constant, we can ignore it in (11). The augmented Lagrangian function of (12) is formulated as follows:

$$\begin{aligned}L=\sum_v\left(\left\|S^{(v)}\right\|_*+\lambda_1 Tr\left(F^{(v)T}G^{(v)T}L_w^{(v)}G^{(v)}F^{(v)}\right)\right)\\-\sum_v\lambda_3 Tr\left(F^{(v)}F^{(v)T}UU^T\right)+\sum_v\lambda_2\left\|E^{(v)}\right\|_1\\+\frac{\mu}{2}\sum_v\left(\left\|Z^{(v)}-S^{(v)}+\frac{C_2^{(v)}}{\mu}\right\|_F^2+\left\|Z^{(v)}-W^{(v)}+\frac{C_3^{(v)}}{\mu}\right\|_F^2\right)\\+\frac{\mu}{2}\sum_v\left\|Y^{(v)}-Y^{(v)}Z^{(v)}-E^{(v)}+\frac{C_1^{(v)}}{\mu}\right\|_F^2\\-\sum_v\psi^{(v)T}\left(W^{(v)}1-1\right)\end{aligned}\tag{13}$$

where matrixes $C_1^{(v)}$, $C_2^{(v)}$, $C_3^{(v)}$, and vector $\psi^{(v)}$ are Lagrange multipliers, μ is a penalty parameter.

Then we can iteratively solve all unknown variables one by one as follows:

Step 1. Update variable $Z^{(v)}$. Fixing all of the other variables, the problem to solve variable $Z^{(v)}$ is degraded to minimize the following problem:

$$\begin{aligned}L\left(Z^{(v)}\right)=\left\|Z^{(v)}-S^{(v)}+\frac{C_2^{(v)}}{\mu}\right\|_F^2+\left\|Z^{(v)}-W^{(v)}+\frac{C_3^{(v)}}{\mu}\right\|_F^2\\+\left\|Y^{(v)}-Y^{(v)}Z^{(v)}-E^{(v)}+\frac{C_1^{(v)}}{\mu}\right\|_F^2\end{aligned}\tag{14}$$

Then we can obtain variable $Z^{(v)}$ by setting the derivative of $L\left(Z^{(v)}\right)$ with respect to $Z^{(v)}$ to zero as follows:

$$\begin{aligned}\frac{\partial L\left(Z^{(v)}\right)}{\partial Z^{(v)}}=2Z^{(v)}-2M_1^{(v)}+2Z^{(v)}-2M_2^{(v)}\\+2Y^{(v)T}\left(Y^{(v)}Z^{(v)}-M_3^{(v)}\right)=0\\\Leftrightarrow Z^{(v)}=\left(Y^{(v)T}Y^{(v)}+2I\right)^{-1}\left(Y^{(v)T}M_3^{(v)}+M_1^{(v)}+M_2^{(v)}\right)\end{aligned}\tag{15}$$

where $M_1^{(v)}=S^{(v)}-\frac{C_2^{(v)}}{\mu}$, $M_2^{(v)}=W^{(v)}-\frac{C_3^{(v)}}{\mu}$, and $M_3^{(v)}=Y^{(v)}-E^{(v)}+\frac{C_1^{(v)}}{\mu}$.

Step 2. Update variable $S^{(v)}$. Fixing the other variables, the sub-problem to calculate variable $S^{(v)}$ is degraded to the following formula:

$$\min_{S^{(v)}}\left\|S^{(v)}\right\|_*+\frac{\mu}{2}\left\|Z^{(v)}-S^{(v)}+\frac{C_2^{(v)}}{\mu}\right\|_F^2\tag{16}$$

(16) can be computed by the singular value thresholding (SVT) shrinkage operator as follows [42-44]:

$$S^{(v)}=\Theta_{1/\mu}\left(Z^{(v)}+\frac{C_2^{(v)}}{\mu}\right)\tag{17}$$

where Θ denotes the SVT shrinkage operator.

Step 3. Update variables $W^{(v)}$ and $\psi^{(v)}$. Fixing the other variables, the optimization problem to calculate variable $W^{(v)}$

is formulated as follows:

$$\begin{aligned}\min_{W^{(v)}\geq 0, W^{(v)}_{i.i}=0} & \lambda_1 Tr\left(F^{(v)T}G^{(v)T}L_w^{(v)}G^{(v)}F^{(v)}\right)\\ +\frac{\mu}{2}&\left\|Z^{(v)}-W^{(v)}+\frac{C_3^{(v)}}{\mu}\right\|_F^2-\psi^{(v)T}\left(W^{(v)}1-1\right)\end{aligned} \tag{18}$$

Define $P^{(v)}=G^{(v)}F^{(v)}$, $R^{(v)}=Z^{(v)}+\frac{C_3^{(v)}}{\mu}$, (18) is equivalent to the following optimization problem:

$$\begin{aligned}\min_{W^{(v)}\geq 0, W^{(v)}_{i.i}=0} & \frac{\lambda_1}{2}\sum_{i,j}^{n_v}\left\|P_{i,:}^{(v)}-P_{j,:}^{(v)}\right\|_2^2 W_{i,j}^{(v)}\\ +\frac{\mu}{2}&\left\|W^{(v)}-R^{(v)}\right\|_F^2-\psi^{(v)T}\left(W^{(v)}1-1\right)\end{aligned} \tag{19}$$

where $P_{i,:}^{(v)}$ and $P_{j,:}^{(v)}$ represent the ith and jth row vectors of matrix $P^{(v)}$, respectively. It is obvious to see that problem (19) is independent to each row. Define $H_{i,j}^{(v)}=\left\|P_{i,:}^{(v)}-P_{j,:}^{(v)}\right\|_2^2$, then we can simplify problem (19) into the following problem:

$$\begin{aligned}&\min_{W_{i,:}^{(v)}\geq 0, W_{i.i}^{(v)}=0}\frac{\lambda_1}{2}W_{i,:}^{(v)}H_{i,:}^{(v)T}+\frac{\mu}{2}\left\|W_{i,:}^{(v)}-R_{i,:}^{(v)}\right\|_2^2\\ &\qquad\qquad -\psi_i^{(v)}\left(W_{i,:}^{(v)}1-1\right)\\ \Leftrightarrow &\min_{W_{i,:}^{(v)}\geq 0, W_{i.i}^{(v)}=0}\frac{\mu}{2}\left\|W_{i,:}^{(v)}-\left(R_{i,:}^{(v)}-\frac{\lambda_1}{2\mu}H_{i,:}^{(v)}\right)\right\|_2^2\\ &\qquad\qquad -\psi_i^{(v)}\left(W_{i,:}^{(v)}1-1\right)\end{aligned} \tag{20}$$

where $\psi_i^{(v)}$ is the ith element of vector $\psi^{(v)}$. The optimal solution to problem (20) is as follows [45]:

$$W_{i,j}^{(v)}=\begin{cases}0, & j=i\\ R_{i,j}^{(v)}-\frac{\lambda_1}{2\mu}H_{i,j}^{(v)}+\frac{\psi_i^{(v)}}{\mu}, & otherwise\end{cases} \tag{21}$$

Then we further enforce all elements of matrix $W^{(v)}$ to be not less than 0 by $W^{(v)}=\max\left(W^{(v)},0\right)$, in which all elements less than 0 in $W^{(v)}$ are enforced to 0, and the remaining elements are preserved. According to constraint $W_{i,:}^{(v)}1=1$, we can obtain that the Lagrange multiplier $\psi_i^{(v)}$ is updated as follows:

$$\psi_i^{(v)}=\mu\left(1-\sum_{j=1,j\neq i}^{n_v}\left(R_{i,j}^{(v)}-\frac{\lambda_1}{2\mu}H_{i,j}^{(v)}\right)\right)\Big/(n_v-1) \tag{22}$$

Step 4: Update variable $E^{(v)}$. Fixing the other variables, the sub-problem to solve variable $E^{(v)}$ is as follows:

$$\min_{E^{(v)}}\lambda_2\left\|E^{(v)}\right\|_1+\frac{\mu}{2}\left\|Y^{(v)}-Y^{(v)}Z^{(v)}-E^{(v)}+\frac{C_1^{(v)}}{\mu}\right\|_F^2 \tag{23}$$

Problem (23) is a typical sparsity constraint optimization problem and has the following closed form solution [46-48]:

$$E^{(v)}=\vartheta_{\lambda_2/\mu}\left(Y^{(v)}-Y^{(v)}Z^{(v)}+\frac{C_1^{(v)}}{\mu}\right) \tag{24}$$

where ϑ denotes the shrinkage operator.

Step 5: Update variable $F^{(v)}$. Fixing the other variables, the cluster indicator matrix $F^{(v)}$ of each view can be calculated by minimizing the following formula:

$$\begin{aligned}&\min_{F^{(v)T}F^{(v)}=I}\lambda_1 Tr\left(F^{(v)T}G^{(v)T}L_w^{(v)}G^{(v)}F^{(v)}\right)\\ &\qquad -\lambda_3 Tr\left(F^{(v)}F^{(v)T}UU^T\right)\\ \Leftrightarrow &\max_{F^{(v)T}F^{(v)}=I}Tr\left(F^{(v)T}\left(\lambda_3 UU^T-\lambda_1 G^{(v)T}L_w^{(v)}G^{(v)}\right)F^{(v)}\right)\end{aligned} \tag{25}$$

Problem (25) can be solved by the eigenvalue decomposition, where the first c eigenvectors corresponding to the first c largest eigenvalues of matrix $\left(\lambda_3 UU^T-\lambda_1 G^{(v)T}L_w^{(v)}G^{(v)}\right)$ are chosen as the optimal solution to variable $F^{(v)}$.

Step 6: Update variable U. Fixing the other variables, the problem to obtain the consensus cluster representation U is degraded to the following problem:

$$\begin{aligned}&\min_{U^TU=I}-\sum_v\lambda_3 Tr\left(F^{(v)}F^{(v)T}UU^T\right)\\ \Leftrightarrow &\max_{U^TU=I}Tr\left(U^T\left(\sum_v F^{(v)}F^{(v)T}\right)U\right)\end{aligned} \tag{26}$$

Problem (26) can also be simply computed by the eigenvalue decomposition. The optimal solution to variable U is the eigenvector set corresponding to the first c largest eigenvalues of matrix $\left(\sum_v F^{(v)}F^{(v)T}\right)$.

Step 7: Update variables $C_1^{(v)}$, $C_2^{(v)}$, $C_3^{(v)}$, and μ. These four variables are updated as follows, respectively:

$$C_1^{(v)}=C_1^{(v)}+\mu\left(Y^{(v)}-Y^{(v)}Z^{(v)}-E^{(v)}\right) \tag{27}$$

$$C_2^{(v)}=C_2^{(v)}+\mu\left(Z^{(v)}-S^{(v)}\right) \tag{28}$$

$$C_3^{(v)}=C_3^{(v)}+\mu\left(Z^{(v)}-W^{(v)}\right) \tag{29}$$

$$\mu=\min\left(\rho\mu,\mu_0\right) \tag{30}$$

where ρ and μ_0 are constants.

Algorithm 1 summarizes the computation procedures presented above. After obtaining the consensus representation U, we perform the *k*-means algorithm on it to obtain the final clustering results.

IV. Analysis of the Proposed Method

A. Computational complexity

For simplicity, we do not take into account the computational costs of matrix multiplication, elements based matrix division, matrix addition and subtraction, etc., since these operations are very simple in comparison with the other matrix operations. For the proposed algorithm listed in Algorithm 1, the major computational costs are the operations like matrix inverse, singular value decomposition (SVD), and eigenvalue decomposition. Generally, the computational complexities of the above three operations are $O(n^3)$ for an $n\times n$ matrix, $O(mn^2)$ for an $m\times n$ matrix, and $O(n^3)$ for an $n\times n$ matrix, respectively. Therefore, the computational complexities of

Algorithm 1 : IMSC_AGL (solving (12))

Input: Incomplete multi-view data $Y^{(v)}$, index matrix $G^{(v)}$, $v \in [1, k]$, parameters λ_1,λ_2, and λ_3.
Initialization: Initialize $Z^{(v)}$ with the k-nearest neighbor graph of each view; initialize $F^{(v)}$ via the eigenvalue decomposition on the Laplacian graph of each view; using (26) to initialize the cluster matrix U, $\mu_0 = 10^8$, $\rho = 1.1$, $\mu = 0.01$.
while not converged **do**
  for v from 1 to k
    1. Update variable $Z^{(v)}$ via (15);
    2. Update variable $S^{(v)}$ via (17);
    3. Update variables $\psi^{(v)}$ and $W^{(v)}$ via (22) and (21), respectively, and then implement $W^{(v)} = \max\left(W^{(v)}, 0\right)$;
    4. Update variable $E^{(v)}$ via (24);
    5. Update variable $F^{(v)}$ by solving (25);
    6. Update variables $C_1^{(v)}$, $C_2^{(v)}$, and $C_3^{(v)}$ via (27), (28), and (29), respectively.
  end
  7. Update U by solving (26);
  8. Update μ via (30).
end while
Output: U

Steps 2, 5, and 6 are about $O(n_v^3)$, $O(n^3)$, and $O(n^3)$, respectively. For Step 1, although it needs to calculate the inverse operation, we can still ignore its computational complexity because the inverse operation about $(Y^{(v)T}Y^{(v)}+2I)^{-1}$ can be pre-computed before the iteration. For the remaining steps, their computational complexities can also be ignored since they only contain the basic matrix operations. Therefore, the whole computational complexity of the proposed method is about $O(\tau(kn^3 + n^3 + \sum_v n_v^3))$, where τ denotes the iteration number, k denotes the number of views, n_v denotes the number of un-missing instances in the vth view.

B. Convergence analysis

For the ADMM-style optimization approach, it is difficult to prove its strong convergence property with more than two unknown variables. Fortunately, we can prove a weak convergence property of the proposed method based on the following theorem [42, 49].

Theorem 1. Let the solution of the optimization problem (12) at the tth iteration step be $\Upsilon_t = (Z_t^{(v)}, S_t^{(v)}, W_t^{(v)}, E_t^{(v)}, F_t^{(v)}, U_t, (C_1^{(v)})_t, (C_2^{(v)})_t, (C_3^{(v)})_t, \psi_t^{(v)})$, $v = [1, k]$. If the sequence solutions $\{\Upsilon_t\}_{t=1}^{\infty}$ of problem (12) are bounded and satisfy condition $\lim_{t\to\infty}(\Upsilon_{t+1} - \Upsilon_t) = 0$, then we can conclude that the accumulated point of sequence $\{\Upsilon_t\}_{t=1}^{\infty}$ is a Karush-Kuhn-Tucker (KKT) point of problem (12). Whenever $\{\Upsilon_t\}_{t=1}^{\infty}$ converges, it converges to the KKT point.

Proof: The detailed proof to Theorem 1 is moved to the supplementary file.

Theorem 1 provides some assurances to the convergence property of the proposed algorithm. In the subsequent section, we will conduct several experiments to further prove it.

V. Experiments and Analyses

In this section, we conduct experiments to prove the effectiveness of the proposed method. The Matlab code of our IMSC_AGL is released at: https://sites.google.com/view/jerry-wen-hit/publications.

A. Experimental settings

1) Baseline algorithms and evaluation metrics: The following six methods are selected to compare with the proposed method:

(1) Best single view (BSV) [27]: For BSV, the missing instances in every view are first filled in the average of instances in the corresponding view. Then it performs *k*-means on all views independently and reports their best clustering results.

(2) Concat [27]: Concat adopts the same approach as BSV to fill in the missing instances. Their difference is that it concatenates all views into a single view with long dimensions, and then performs *k*-means to obtain the clustering results.

(3) Partial multi-view clustering (PMVC) [3]: Based on the non-negative matrix factorization, PMVC learns a common latent representation for all views, then performs *k*-means on the learned representation to obtain the clustering results.

(4) Incomplete multi-modality grouping (IMG) [27]: IMG also uses the matrix factorization technique to learn a common latent representation for all views. Different with PMVC, IMG simultaneously learns a graph from the common representation and finally exploits the spectral clustering algorithm to obtain the clustering results.

(5) Double constrained non-negative matrix factorization (DCNMF) [50]: DCNMF uses local geometric structure of each view to guide the representation learning.

(6) Graph regularized partial multi-view clustering (GPMVC) [51]: GPMVC can be viewed as a variant of DCNMF, which learns the common representation from the normalized individual representations of all views.

IMG, DCNMF, and GPMVC have many tunable parameters (more than two parameters). In our experiments, we implement these methods with a wide candidate parameter set, such as $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$ and report their best clustering results for fair comparison.

There are many criterions to evaluate the clustering performances of different methods [52]. In our experiments, we choose three most well-known evaluation criterions, *i.e.*, clustering accuracy (ACC), normalized mutual information (NMI), and purity to compare the above methods [17]. Generally, we expect the values of these evaluation criterions as big as possible.

2) Databases: Seven real-world databases listed in Table I are adopted for evaluation. Their detailed information are as follows:

(1) **Handwritten digit database** [53]: Following the experimental settings in [53], we conduct the experiments on the multi-view handwritten database[1] with 2000 samples in total. There are 10 digits, *i.e.*, 0-9, in the used handwritten

[1]https://www.dropbox.com/s/b00cdxc2s96o1to/handwritten.mat.

digit database. In our experiments, we only use two views, *i.e.*, pixel average features and Fourier coefficient features, to conduct experiments, in which the pixel average view has 240 features per sample and the Fourier coefficient view contains 76 features for one sample.

(2) **BUAA-visnir face database (BUAA)** [54]: Following the experimental settings in [27], we implement the experiments on the subset of the BUAA face database, which contains 90 visual images and 90 near infrared images of the first 10 classes. The used BUAA face database is available at https://github.com/hdzhao/IMG/tree/master/data. For the used BUAA face database, two types of images, *i.e.*, visual images and near infrared images are regarded as the two views of persons. All images were resized into an 10×10 matrix and vectorized in advance.

(3) **Cornell database** [2, 55]: Cornell database is one of the popular WebKB databases[2], which is composed of 195 documents over the 5 labels, *i.e.*, student, project, course, staff, and faculty. Each document is described by two views, *i.e.*, 1703 content features and 195 citation features.

(4) **Caltech101 database** [56]: The Caltech101 database is one of the popular object databases which contains 101 objects in total. Each object provides 40-800 images. In our experiments, a multi-view subset which contains 7 classes and 1474 images is adopted [53]. For convenience, we refer to the subset as Caltech7[3]. The original multi-view Caltech7 database contains 5 types of features. In our experiments, we only select two views to implement the experiments, in which the one is GIST features [57] with 512 dimensions per sample and the other one is the Local Binary Patterns (LBP) features [57] with 928 dimensions per sample.

(5) **ORL database**[4]: The ORL face database is composed of 400 faces taken from 40 individuals. In the experiments, we first pre-resized all images into the size of 32×32, and then extract the features of LBP, GIST, and Pyramid of Histogram of Oriented Gradients (PHOG) [58]. Combining the original pixel features of each image, we form the multi-view ORL dataset with four views, in which their feature dimensions are 1024, 512, 1024, and 1024, respectively.

(6) **3 Sources database**[5]: This database contains 948 news articles collected from three online news sources: BBC, Reuters, and The Guardian. In our experiments, we select a subset which contains 169 stories reported in all of the three sources to compare different methods. The 169 stories were categorized into six topical labels: business, entertainment, health, politics, sport, and technology.

(7) **BBCSport database** [59]: The original BBCSport database contains 737 documents about the sport news articles collected from the BBC Sport website. These documents are described by 2-4 views and categorized into five classes. In our experiments, we choose a subset[6] with 116 samples described by all of the four views to validate the effectiveness of our method.

[2]http://lig-membres.imag.fr/grimal/data.html.

[3]https://www.dropbox.com/s/ulvatoo8gepcdfk/Caltech101-7.mat.

[4]The original ORL face database is available at: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[5]http://erdos.ucd.ie/datasets/3sources.html.

[6]https://github.com/GPMVCDummy/GPMVC/tree/master/partialMV/PVC/recreateResults/data.

TABLE I
DESCRIPTION OF THE USED BENCHMARK DATASETS.

Database	# Class	# View	# Samples	# Features
handwritten	10	2	2000	240/76
BUAA	10	2	90	100/100
Cornell	5	2	195	195/1703
Caltech7	7	2	1474	512/928
ORL	40	4	400	1024/512/1024/1024
3 Sources	6	3	169	3560/3631/3068
BBCSport	5	4	116	1991/2063/2113/2158

3) Incomplete multi-view data construction: In our experiments, we construct two types of incomplete multi-view data. (i) *Incomplete case that few samples have complete views*: Following the experimental settings in [27], for handwritten, BUAA, Cornell, and Caltech7 databases, we randomly select 10%, 30%, 50%, 70%, and 90% samples from the database as the paired samples. For the half of corresponding remaining samples, we remove their first view, and for the other half of samples, we remove their second view to form the incomplete multi-view scenarios. Similarly, for the ORL multi-view database, we also randomly select 10%, 30%, and 50% samples as the paired samples, and then we follow the previous strategy to randomly select 25%, 25%, 25%, and 25% samples of the corresponding remaining samples as the single view samples. (ii) *Incomplete case that all samples have missing views*: In our experiments, we exploit the ORL, 3 Sources, and BBCsport databases to construct the incomplete multi-view data with no paired samples, where about 53%, 55%, and 55% instances are randomly removed from each view of the three databases, respectively. For fairness, we repeatedly perform all compared methods 5 times on these databases and report their average clustering results.

B. Experimental results and analysis

Experimental results of different methods on the above two types of incomplete multi-view databases are enumerated in Table II-Table VII and shown in Fig.3, respectively. From the experimental results, we can obtain the following points:

(1) In most cases, Concat and BSV perform the 1st worst and 2nd worst in comparison with the other methods. This demonstrates that concatenating all views into one long view is not a good approach in dealing with the multi-view clustering tasks. This is mainly because that Concat not only ignores the differences of different views in the feature scales and distributions, but also cannot exploit the complementary information across different views. While for BSV, the missing instances in each view will obviously be clustered into the same group since they are filled in the same average instance, which makes BSV achieve very bad performance especially for the case with large number of missing instances in every view. The experimental results of Concat and BSV in the tables and figures also prove that filling in the missing instances with the corresponding average instances is not a good choice to address the incomplete problem of multi-view clustering.

(2) From Table II-Table V and Fig.3, we can find that PMVC, IMG, DCNMF, GPMVC, and the proposed method

can achieve much better performance than BSV and Concat in most cases, which proves that exploiting the complementary information of multi-view features to learn a common representation is an effective approach in dealing with the incomplete problem.

(3) From Table II-Table V and Fig.3, we can find that DCNMF and GPMVC perform better than PMVC on the handwritten digit database and BUAA database, while perform worse than PMVC in some cases especially on the Caltech7 and ORL databases. Compared with PMVC, DCNMF and GPMVC all try to exploit the geometric structure of each view to guide the common representation learning. Thus these experimental results indicate that exploiting the intrinsic geometric structure of data has the potential to learn a more discriminative and compact common representation for clustering. However, if the constructed geometric structure is not the intrinsic structure, it will lead to the opposite effect. Thus capture the intrinsic structure of each view is crucial for these methods.

(4) The proposed method significantly outperforms the other methods on the above multi-view databases with all kinds of incomplete cases. For instance, on the handwritten digit database (Table II), the proposed method achieves 3 percent and 6 percent improvement of ACC and NMI in comparison with the second best method, *i.e.*, DCNMF, respectively. In Table VII, the NMI, ACC, and purity of the proposed method are about 33%, 26%, and 31% higher than those of BSV and Concat on the BBCSport database. These good experimental results strongly prove the effectiveness the proposed method in dealing with all kinds of the incomplete multi-view clustering tasks.

TABLE VI
MEAN NMIS (%), ACCS (%), AND PURITIES (%) OF DIFFERENT METHODS ON THE FIRST INCOMPLETE CASE OF THE ORL DATABASE. BOLD NUMBERS DENOTE THE BEST RESULT.

	NMI			ACC			Purity		
Method \ PER	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
BSV	30.55	43.73	55.55	28.50	39.20	47.74	30.55	41.70	51.44
Concat	35.44	35.03	46.93	27.72	28.97	41.67	23.58	30.83	44.27
PMVC	64.50	75.50	82.36	47.12	59.38	67.97	49.09	62.14	71.11
IMG	65.75	76.78	83.41	48.17	63.88	72.65	50.13	66.03	73.62
DCNMF	60.63	70.76	75.25	41.86	54.11	60.93	48.38	59.83	65.08
GPMVC	61.63	71.26	77.21	42.13	57.30	63.61	49.56	61.75	70.39
IMSC_AGL	**70.87**	**82.83**	**86.65**	**53.93**	**72.50**	**76.31**	**56.73**	**74.52**	**78.72**

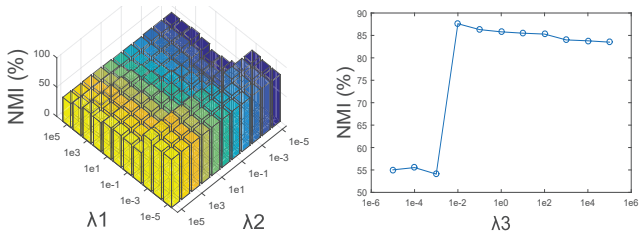
TABLE VII
MEAN NMIS (%), ACCS (%), AND PURITIES (%) OF DIFFERENT METHODS ON THE SECOND INCOMPLETE MULTI-VIEW DATABASES. BOLD NUMBERS DENOTE THE BEST RESULT.

	ORL			3 Sources			BBCSport		
Method	NMI	ACC	Purity	NMI	ACC	Purity	NMI	ACC	Purity
BSV	43.04	37.95	40.73	19.97	37.37	47.03	18.60	43.21	47.00
Concat	55.78	40.56	43.40	26.98	41.89	53.37	19.48	42.90	46.34
IMSC_AGL	**85.41**	**74.23**	**76.53**	**49.03**	**59.93**	**69.16**	**53.76**	**69.14**	**78.10**

C. Analysis of the parameter sensitivity

In this section, we mainly focus on analyzing the sensitivity of the three tunable parameters, *i.e.*, $\lambda_1, \lambda_2, \lambda_3$, in model (11). We first define a candidate set, *i.e.*, $\left\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4, 10^5\right\}$, for the three parameters, and then perform the proposed method with different combinations of the three parameters [60]. In Fig.4 and Fig.5, we show the relationships of NMI (%) and the three parameters on the ORL and BUAA face datasets with 50% and 70% paired samples, respectively. From these figures, it is obvious that the proposed method can obtain the stable and satisfied NMIs when the three parameters are located in some feasible areas. For example, on the ORL dataset, when parameters $\lambda_1, \lambda_2, \lambda_3$ locates in the range of $\left[10^{-5}, 10^{-2}\right]$, $\left[10^{1}, 10^{5}\right]$, and $\left[10^{-1}, 10^{2}\right]$, respectively, a satisfactory clustering performance can be obtained. This demonstrates that the proposed method is insensitive to the three parameters to some extent.

For different databases, it is still an open problem to adaptively select the optimal values for these parameters to our best knowledge. In this section we provide a simple strategy to find the optimal combination of the three parameters for experiments. From Fig.4 and Fig.5, we can find that the proposed method is relatively insensitive to the selection of parameter λ_1 in the small parameter range $\left[10^{-5}, 10^{-2}\right]$ to some extent, thus we can set λ_1 with a fixed value like 10^{-3} at first, and then focus on finding the best combination of parameters λ_2 and λ_3. According to the experimental results shown in Fig.4 and Fig.5, we define two candidate sets $\left\{10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}\right\}$ and $\left\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\right\}$ for parameters λ_2 and λ_3, respectively. Then perform the proposed method with different values of the two parameters. In this way, we can find the best combination of parameters λ_2 and λ_3 in the 2D space formed by their candidate parameters. To obtain the optimal value of parameter λ_1, we can fix λ_2 and λ_3 with the obtained optimal combination and then perform the proposed method with different values of λ_1. As a result, the optimal combination of these three values can be achieved. Then we use the selected parameters to conduct experiments and report the results for comparison.



(a) Parameters λ_1 and λ_2 (b) Parameter λ_3

Fig. 4. NMI (%) versus (a) parameters λ_1 and λ_2 by fixing parameter λ_3, (b) parameter λ_3 by fixing parameters λ_1 and λ_2, on the ORL face dataset with 50% paired samples.

D. Convergence analysis based on experiments

In this section, two experiments are conduct to prove the convergence property of the proposed optimization approach listed in Algorithm 1. In Fig. 6, we show the objective function value versus the iteration steps, in which the objective function value is calculated as $obj = \sum_v \left(\gamma_1^{(v)} + \gamma_2^{(v)}\right) \Big/ \sum_v \left\|X^{(v)}\right\|_F^2$ according to the original model (11), where $\gamma_1^{(v)} = \left\|Z^{(v)}\right\|_* + \lambda_1 Tr\left(F^{(v)T} G^{(v)T} L_w^{(v)} G^{(v)} F^{(v)}\right) + \lambda_2 \left\|E^{(v)}\right\|_1 + \lambda_3 \left(c - Tr\left(F^{(v)} F^{(v)T} U U^T\right)\right)$ and

TABLE II
MEAN NMIS (%) AND ACCS (%) OF DIFFERENT METHODS ON THE HANDWRITTEN DIGIT DATABASE. BOLD NUMBERS DENOTE THE BEST RESULT.

	NMI					ACC				
Method \ PER	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
BSV	37.04	44.48	51.50	58.61	66.26	43.08	50.46	57.39	64.44	69.29
Concat	47.71	54.43	61.12	70.30	79.34	46.01	57.46	66.45	78.64	86.63
PMVC	55.13	60.85	64.88	68.54	72.83	63.81	70.90	73.44	75.20	77.82
IMG	58.04	62.38	64.91	68.21	73.57	69.22	75.41	76.36	77.54	81.78
DCNMF	54.23	65.56	74.41	78.14	80.90	51.21	76.63	80.61	86.16	89.16
GPMVC	60.99	63.99	72.23	73.68	75.24	65.60	74.04	76.94	79.06	81.08
IMSC_AGL	**76.60**	**79.05**	**82.46**	**83.05**	**86.27**	**80.76**	**84.81**	**87.41**	**89.77**	**92.57**

TABLE III
MEAN NMIS (%) AND ACCS (%) OF DIFFERENT METHODS ON THE BUAA DATABASE. BOLD NUMBERS DENOTE THE BEST RESULT.

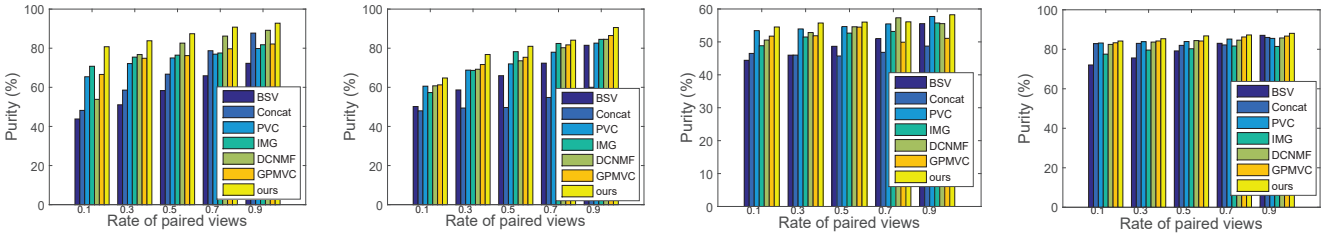
	NMI					ACC				
Method \ PER	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
BSV	43.10	53.03	61.78	69.91	82.56	48.33	56.96	64.26	70.81	80.16
Concat	51.22	51.95	52.43	56.51	62.66	45.62	46.61	47.46	52.34	57.58
PMVC	61.35	67.07	71.97	78.70	84.22	57.41	66.46	70.01	75.92	80.73
IMG	54.72	67.53	76.74	82.83	85.90	53.95	67.39	76.14	79.36	80.78
DCNMF	61.78	68.75	72.05	79.66	86.42	58.36	67.58	72.15	76.58	82.42
GPMVC	62.12	70.25	74.33	81.63	86.78	58.98	68.75	74.28	78.28	84.24
IMSC_AGL	**63.52**	**75.16**	**80.29**	**84.52**	**89.84**	**65.72**	**78.16**	**80.76**	**82.77**	**89.54**

TABLE IV
MEAN NMIS (%) AND ACCS (%) OF DIFFERENT METHODS ON THE CORNELL DATABASE. BOLD NUMBERS DENOTE THE BEST RESULT.

	NMI					ACC				
Method \ PER	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
BSV	8.66	8.19	8.89	12.69	19.34	42.41	43.93	44.84	46.32	47.66
Concat	8.07	7.56	8.30	10.21	13.47	38.80	38.06	36.96	36.79	38.48
PMVC	15.76	16.00	18.21	19.76	21.03	42.56	42.56	43.79	42.56	43.03
IMG	12.56	16.62	19.24	20.89	22.98	**45.13**	45.79	47.08	45.51	44.76
DCNMF	13.59	17.72	19.17	21.69	23.98	39.94	43.29	43.18	45.74	45.52
GPMVC	13.90	16.07	18.99	15.03	17.07	40.39	43.86	46.53	44.56	44.35
IMSC_AGL	**18.52**	**20.28**	**21.46**	**22.84**	**24.35**	43.25	**46.02**	**47.35**	**46.37**	**48.56**

TABLE V
MEAN NMIS (%) AND ACCS (%) OF DIFFERENT METHODS ON THE CALTECH7 DATABASE. BOLD NUMBERS DENOTE THE BEST RESULT.

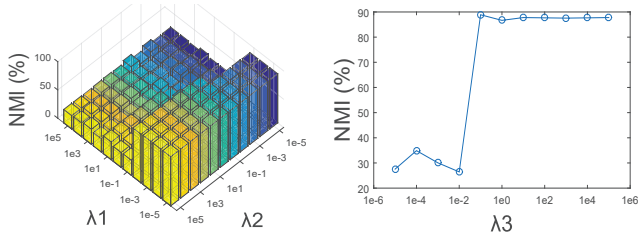
	NMI					ACC				
Method \ PER	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
BSV	29.04	32.11	35.13	38.83	44.16	42.66	40.97	39.83	42.92	46.99
Concat	33.82	34.44	34.56	38.15	45.44	36.83	31.74	36.36	43.38	47.08
PMVC	38.99	40.26	40.17	41.60	41.94	43.46	43.96	44.46	44.76	44.34
IMG	32.38	33.29	35.05	35.96	37.64	42.05	42.36	42.23	41.17	43.23
DCNMF	33.86	38.19	41.40	41.24	44.04	40.63	44.53	45.62	48.50	50.74
GPMVC	40.05	40.96	41.83	42.61	46.02	45.57	47.19	46.99	46.99	49.10
IMSC_AGL	**40.71**	**42.02**	**43.35**	**45.15**	**47.28**	**50.89**	**51.20**	**52.14**	**53.05**	**56.30**



(a) handwritten (b) BUAA (c) Cornell (d) Caltech7

Fig. 3. Purities (%) of different methods on the (a) handwritten digit database, (b) BUAA database, (c) Cornell database, and (d) Caltech7 database with different rates of paired samples.

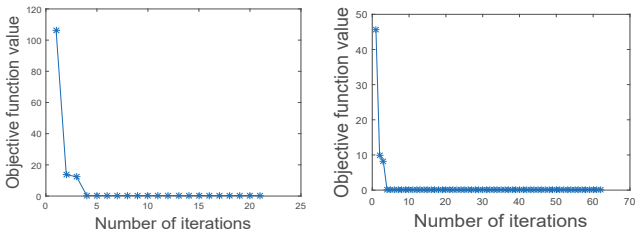
$\gamma_2^{(v)} = \left\|Y^{(v)} - Y^{(v)}Z^{(v)} - E^{(v)}\right\|_F^2 + \left\|Z^{(v)} - S^{(v)}\right\|_F^2 + \left\|Z^{(v)} - W^{(v)}\right\|_F^2$. From Fig. 6, it is obvious that the objective function curve is monotonically decreasing till to the stable level, which demonstrates that our provided optimization approach monotonically decreases the objective problem. This also proves that the proposed method will converge to the local optimal point after a few iterations. Moreover, from the objective function value, we can find that the proposed



(a) Parameters λ_1 and λ_2 (b) Parameter λ_3

Fig. 5. NMI (%) versus (a) parameters λ_1 and λ_2 by fixing parameter λ_3, (b) parameter λ_3 by fixing parameters λ_1 and λ_2, on the BUAA face dataset with 70% paired samples.

method has very promising convergence efficiency.



(a) BUAA (b) ORL

Fig. 6. Objective function value versus the iteration step of the proposed method on the BUAA face and ORL face databases, in which 70% and 50% samples are randomly selected as the paired samples.

VI. Conclusion

In this paper, a novel incomplete multi-view clustering framework is proposed. Different from the conventional methods that exploit the matrix factorization technique for incomplete multi-view clustering, the proposed method is the first one that integrates the spectral clustering and adaptive graph learning technique to tackle this problem. Compared with the conventional methods, the proposed method is more flexible since it is able to handle all kinds of incomplete cases. We have conducted several experiments on the two types of incomplete multi-view data. Experimental results commonly show that the proposed method can achieve a better performance than some state-of-the-art methods, which proves its effectiveness in dealing with the incomplete multi-view clustering tasks.

References

[1] X. Yuan, J. Yu, Z. Qin, and T. Wan, "A sift-lbp image retrieval model based on bag of features," in *IEEE International Conference on Image Processing*, 2011.

[2] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Annual Conference on Computational Learning Theory*. ACM, 1998, pp. 92–100.

[3] S.-Y. Zhi and H. Zhou, "Partial multi-view clustering," in *AAAI Conference on Artificial Intelligence*, 2014, pp. 1968–1974.

[4] H. Zhao, Z. Ding, and Y. Fu, "Multi-view clustering via deep matrix factorization." in *AAAI Conference on Artificial Intelligence*, 2017, pp. 2921–2927.

[5] Z. Zhang, L. Liu, J. Qin, F. Zhu, F. Shen, Y. Xu, L. Shao, and H. T. Shen, "Highly-economized multi-view binary compression for scalable image clustering," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 717–732.

[6] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection." in *AAAI Conference on Artificial Intelligence*, 2017, pp. 4147–4153.

[7] J. Li, H. Yong, B. Zhang, M. Li, L. Zhang, and D. Zhang, "A probabilistic hierarchical model for multi-view and multi-feature classification," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 3498–3505.

[8] Y. Wang and L. Wu, "Beyond low-rank representations: Orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering," *Neural Networks*, vol. 103, pp. 1–8, 2018.

[9] K. Zhan, C. Zhang, J. Guan, and J. Wang, "Graph learning for multiview clustering," *IEEE Transactions on Cybernetics*, no. 99, pp. 1–9, 2017.

[10] F. Nie, G. Cai, J. Li, and X. Li, "Auto-weighted multi-view learning for image clustering and semi-supervised classification," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1501–1511, 2018.

[11] J. Li, B. Zhang, G. Lu, and D. Zhang, "Generative multi-view and multi-feature learning for classification," *Information Fusion*, vol. 45, pp. 215–226, 2019.

[12] G. Luo, S. Dong, K. Wang, W. Zuo, S. Cao, and H. Zhang, "Multi-views fusion cnn for left ventricular volumes estimation on cardiac mr images," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 1924–1934, 2018.

[13] F. Nie, G. Cai, and X. Li, "Multi-view clustering and semi-supervised classification with adaptive neighbours." in *AAAI Conference on Artificial Intelligence*, 2017, pp. 2408–2414.

[14] Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4833–4843, 2018.

[15] L. Zhao, Z. Chen, Y. Yang, Z. J. Wang, and V. C. Leung, "Incomplete multi-view clustering via deep semantic mapping," *Neurocomputing*, vol. 275, pp. 1053–1062, 2018.

[16] Z. Zhang, L. Liu, F. Shen, H. T. Shen, and L. Shao, "Binary multi-view clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[17] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data." in *International Joint Conferences on Artificial Intelligence*, 2013, pp. 2598–2604.

[18] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Annual International Conference on Machine Learning*. ACM, 2009, pp. 129–136.

[19] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Advances in Neural Information Processing Systems*, 2011, pp. 1413–1421.

[20] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao, "Low-rank tensor constrained multiview subspace clustering," in *IEEE International Conference on Computer Vision*, 2015, pp. 1582–1590.

[21] Y. Wang, W. Zhang, L. Wu, X. Lin, M. Fang, and S. Pan, "Iterative views agreement: an iterative low-rank based structured optimization method to multi-view spectral clustering," in *International Joint Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 2153–2159.

[22] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, and X. Huang, "Robust subspace clustering for multi-view data by exploiting correlation consensus," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3939–3949, 2015.

[23] A. Trivedi, P. Rai, H. Daumé III, and S. L. DuVall, "Multiview clustering with incomplete views," in *Advances in Neural Information Processing Systems Workshop*, 2010.

[24] J. Wen, Z. Zhang, Y. Xu, and Z. Zhong, "Incomplete multi-view clustering via graph regularized matrix factorization," in *European Conference on Computer Vision Workshop*, 2018.

[25] W. Shao, L. He, and S. Y. Philip, "Multiple incomplete views clustering via weighted nonnegative matrix factorization with l_ {2, 1} regularization," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2015, pp. 318–334.

[26] H. Gao, Y. Peng, and S. Jian, "Incomplete multi-view clustering," in *International Conference on Intelligent Information Processing*. Springer, 2016, pp. 245–255.

[27] H. Zhao, H. Liu, and Y. Fu, "Incomplete multi-modal visual data grouping." in *International Joint Conferences on Artificial Intelligence*, 2016, pp. 2392–2398.

[28] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[29] J. Chen and J. Yang, "Robust subspace segmentation via low-rank representation," *IEEE Transactions on Cybernetics*, vol. 44, no. 8, pp. 1432–1445, 2014.

[30] H. Gao, F. Nie, X. Li, and H. Huang, "Multi-view subspace clustering," in *IEEE International Conference on Computer Vision*, 2015, pp. 4238–4246.

[31] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, 2002, pp. 849–856.

[32] Y. Yang, F. Shen, Z. Huang, H. T. Shen, and X. Li, "Discrete nonnegative spectral clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1834–1845, 2017.

[33] L. Hagen and A. B. Kahng, "New spectral methods for ratio cut partitioning and clustering," *IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems*, vol. 11, no. 9, pp. 1074–1085, 1992.

[34] C. Tian, Y. Xu, J. Wang, N. Luo, and H. Zou, "Enhanced cnn for image denoising," *arXiv preprint arXiv:1810.11834*, 2018.

[35] S. Zeng, J. Gou, and X. Yang, "Improving sparsity of coefficients for robust sparse and collaborative representation-based image classification," *Neural Computing and Applications*, vol. 30, no. 10, pp. 2965–2978, 2018.

[36] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu, "Non-negative low rank and sparse graph for semi-supervised learning," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2328–2335.

[37] X. Li, G. Cui, and Y. Dong, "Graph regularized non-negative low-rank matrix factorization for image clustering," *IEEE Transaction on Cybernetics*, vol. 47, no. 11, pp. 3840–3853, 2017.

[38] Y. Lu, C. Yuan, W. Zhu, and X. Li, "Structurally incoherent low-rank nonnegative matrix factorization for image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5248–5260, 2018.

[39] L. Fei, Y. Xu, X. Fang, and J. Yang, "Low rank representation with adaptive distance penalty for semi-supervised subspace classification," *Pattern Recognition*, vol. 67, pp. 252–262, 2017.

[40] J. Wen, B. Zhang, Y. Xu, J. Yang, and N. Han, "Adaptive weighted nonnegative low-rank representation," *Pattern Recognition*, vol. 81, pp. 326–340, 2018.

[41] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[42] J. Wen, N. Han, X. Fang, L. Fei, K. Yan, and S. Zhan, "Low-rank preserving projection via graph regularized reconstruction," *IEEE Transactions on Cybernetics*, 2018.

[43] Y. Lu, Z. Lai, X. Li, W. K. Wong, C. Yuan, and D. Zhang, "Low-rank 2-d neighborhood preserving projection for enhanced robust image representation," *IEEE Transactions on Cybernetics*, 2018.

[44] Z. Zhang, Y. Xu, L. Shao, and J. Yang, "Discriminative block-diagonal representation learning for image recognition," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 7, pp. 3111–3125, 2018.

[45] F. Nie, X. Wang, M. I. Jordan, and H. Huang, "The constrained laplacian rank algorithm for graph-based clustering." in *AAAI Conference on Artificial Intelligence*, 2016, pp. 1969–1976.

[46] J. Wen, X. Fang, J. Cui, L. Fei, K. Yan, Y. Chen, and Y. Xu, "Robust sparse linear discriminant analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[47] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[48] Y. Lu, C. Yuan, X. Li, Z. Lai, D. Zhang, and L. Shen, "Structurally incoherent low-rank 2dlpp for image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 6147–6158, 2018.

[49] Y. Zhang, "An alternating direction algorithm for nonnegative matrix factorization," *Technical Report, Rice University*, 2010.

[50] B. Qian, X. Shen, Y. Gu, Z. Tang, and Y. Ding, "Double constrained nmf for partial multi-view clustering," in *International Conference on Digital Image Computing: Techniques and Applications*. IEEE, 2016, pp. 1–7.

[51] N. Rai, S. Negi, S. Chaudhury, and O. Deshmukh, "Partial multi-view clustering using graph regularized nmf," in *International Conference on Pattern Recognition*. IEEE, 2016, pp. 2192–2197.

[52] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Information retrieval*, vol. 12, no. 4, pp. 461–486, 2009.

[53] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph." in *AAAI Conference on Artificial Intelligence*, 2015, pp. 2750–2756.

[54] D. Huang, J. Sun, and Y. Wang, "The buaa-visnir face database instructions," *School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001*, 2012.

[55] Y. Guo, "Convex subspace representation learning from multi-view data." in *AAAI Conference on Artificial Intelligence*, vol. 1, 2013, p. 2.

[56] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[57] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[58] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[59] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in *International Conference on Machine Learning*. ACM, 2006, pp. 377–384.

[60] J. Wen, Y. Xu, Z. Li, Z. Ma, and Y. Xu, "Inter-class sparsity based discriminative least square regression," *Neural Networks*, vol. 102, pp. 36–47, 2018.

Jie Wen received the M.S. degree at Harbin Engineering University, China in 2015. He is currently pursuing the Ph.D degree in the the School of Computer Science and Technology at Harbin Institute of Technology, Shenzhen. His research interests include, image and video processing, pattern recognition, data clustering, and machine learning.

Yong Xu received his B.S. degree, M.S. degree in 1994 and 1997, respectively. He received the Ph.D. degree in Pattern Recognition and Intelligence system at NUST (China) in 2005. Now he works at Harbin Institute of Technology, Shenzhen. His current interests include pattern recognition, biometrics, machine learning and video analysis. More information please refer to http://www.yongxu.org/lunwen.html.

Hong Liu received Bachelor degree in computer science in 1990, Master degree in computer science in 1993 and Doctor degree in electrical control and automation engineering in 1996, post-doctoral in computer science and technology in 1996. Professor Liu is currently the supervisor of doctoral students, the director of Research Department of Shenzhen Graduate School and director of Intelligent Robot Laboratory of Peking University. He is also an IEEE member, an executive director and vice secretary-Intelligent Automation Committee of Chinese Automation Association (IACAA). His expertise is in the areas of image processing and pattern recognition, intelligent robots and computer vision, intelligent micro-systems hardware and software co-design. He has published more than 100 papers in the important scholarly journals and international conferences, and access to subsidized Korean Academy of an Jung-geun Awards, Department of Space Science and Technology Progress Award, and Peking University Teaching Excellence Award, Aetna award and candidates of Peking University Top Ten Teachers. He has done exchange visits in many famous universities and research institutions in several countries and regions, including the United States, Canada, France, the Netherlands, Japan, Korea, Singapore, Hong Kong and so on.