Multiview Spectral Clustering With Bipartite Graph

Haizhou Yang, Quanxue Gao, Wei Xia, *Graduate Student Member, IEEE*, Ming Yang, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*— Multi-view spectral clustering has become appealing due to its good performance in capturing the correlations among all views. However, on one hand, many existing methods usually require a quadratic or cubic complexity for graph construction or eigenvalue decomposition of Laplacian matrix; on the other hand, they are inefficient and unbearable burden to be applied to large scale data sets, which can be easily obtained in the era of big data. Moreover, the existing methods cannot encode the complementary information between adjacency matrices, *i.e.*, similarity graphs of views and the low-rank spatial structure of adjacency matrix of each view. To address these limitations, we develop a novel multi-view spectral clustering model. Our model well encodes the complementary information by Schatten p-norm regularization on the third tensor whose lateral slices are composed of the adjacency matrices of the corresponding views. To further improve the computational efficiency, we leverage anchor graphs of views instead of full adjacency matrices of the corresponding views, and then present a fast model that encodes the complementary information embedded in anchor graphs of views by Schatten p-norm regularization on the tensor bipartite graph. Finally, an efficient alternating algorithm is derived to optimize our model. The constructed sequence was proved to converge to the stationary KKT point. Extensive experimental results indicate that our method has good performance.**

***Index Terms*— Multi-view clustering, graph fusion, large scale data.**

I. Introduction

MULTI-VIEW clustering has attracted more and more attention in artificial intelligence and image recognition due to the facts that multi-view data are ubiquitous in practical applications and help provide some complementary information embedded in multi-views for improving clustering performances [1]–[7]. It divides samples into several clusters such that the samples in the same cluster have high similarity to each other. During the last decade, many multi-view clustering methods have been developed and achieved impressive clustering performance, among which spectral clustering (SC) has attracted more and more attention.

SC has been widely used for clustering due to the fact that adjacency matrix may well encode the relationship between data points with arbitrary shape and spectral theory [8]–[12]. The key step of SC is to compute eigenvectors of the graph Laplacian matrix which is obtained by affinity matrix. For multi-view data, Kumar and Rai presented a well-known co-regularized spectral clustering [13]. It learns indicator matrices of views by leveraging SC on the corresponding views and then obtains the common indicator matrix by minimizing the divergence between them. To well encode the cluster structure and improve stableness of spectral clustering algorithm, Cai *et al.* [14] added non-negative relaxation on indicator matrix. However, both of them treat each indicator matrix equally and ignore the salient difference between views which is important for multi-view clustering. To take full advantage of this information, an auto-weighted multiple graph learning (AMGL) model was presented [15]. It adaptively assigns reasonable weights for Laplacian matrices of different views.

It is well known that adjacency matrix, which encodes relationship between data, is fixed in the aforementioned methods. Thus, the quality of adjacency matrix has a large impact on the algorithm. When the input graphs are of poor quality, their clustering performance degrades remarkably. However, in practical applications, it is still an open problem to design a suitable graph for each view manually due to the unknown and complex distribution of data. To improve the flexibility of algorithm, Nie *et al.* [16] proposed a Laplacian rank constrained graph model, which is called self-weighted multiview clustering (SwMC), for spectral clustering. SwMC leverages the weighted mean square error to minimize the divergence between similarity graphs of different views.

Although the aforementioned multi-view spectral methods have achieved impressive performance in most experiments, they usually involve two time consuming steps. The first step is to construct the affinity graph whose size is $N \times N$, where N denotes the number of samples. It usually takes $\mathcal{O}(VN^2d)$ time, where V denotes the view number and d is dimension of view. The second step is to compute the eigenvalue decomposition, which takes $\mathcal{O}(N^3)$. This makes the aforementioned methods inefficient and inherently difficult to be applied to large scale data, which are ubiquitous in the era of big data [17]–[21]. To this end, Cai and Chen [22] proposed a large scale spectral clustering approach for single-view data. Although preliminary clustering performance is good, they still have the following deficiency:

1) In their methods, the similarity graphs is predefined. Thus, their performance relies heavily on predefined similarity graphs, resulting in inferior results.

Manuscript received August 13, 2021; revised March 16, 2022; accepted April 12, 2022. Date of publication May 13, 2022; date of current version May 26, 2022. This work was supported in part by the National Natural Science Foundation of China (NFSC) under Grant 62176203 and Grant 61773302, in part by the Natural Science Basic Research Plan in Shaanxi Province under Grant 2020JZ-19, in part by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) under Grant 202200035, in part by the Fundamental Research Funds for the Central Universities, in part by the Innovation Fund of Xidian University, and in part by the Special Projects for Key Fields in Higher Education of Guangdong under Grant 2020ZDZX3077. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sebastian Bosse. *(Corresponding author: Quanxue Gao.)*

Haizhou Yang, Quanxue Gao, and Wei Xia are with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: qxgao@xidian.edu.cn).

Ming Yang is with the Departments of Mathematics and Computer and Information Science, Westfield State University, Westfield, MA 01086 USA.

Xinbo Gao is with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China.

Digital Object Identifier 10.1109/TIP.2022.3171411

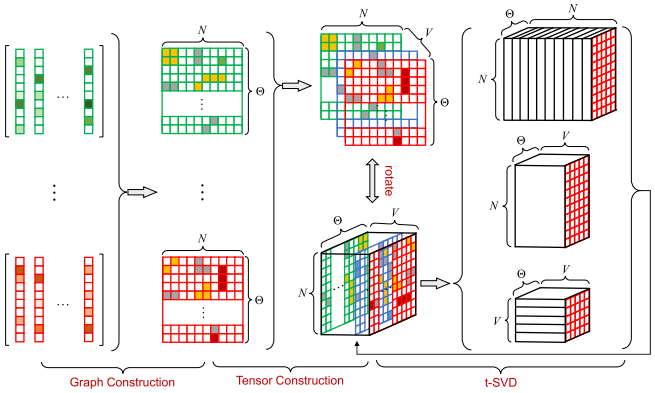


Fig. 1. The flowchart of multi-view spectral clustering with bipartite graph (When $\Theta = N$, we construct full size graph, and when $\Theta = M$, we construct bipartite graph).

2) All of them need to use post-processing to get discrete labels, which limits their performances.
3) They failed to encode the complementary information between adjacency matrices of views and the low-rank spatial structure of adjacency matrix of each view. However, these information help improve the performance of multi-view clustering.

To solve the above limitations, considering the advantage of tensor Schatten p-norm [23]–[27], we present an effective multi-view spectral clustering model which encodes both the complementary information in graphs of views and low-rank spatial structure of each adjacency matrix. To be specific, we take $N \times N$ adjacency matrices of views as input and construct a 3-order tensor that is composed of adjacency matrices, and then minimize the devergence between them by using tensor singular value decomposition (t-SVD) based tensor Schatten p-norm (Seen Fig. 1). To further improve the computational efficiency of our proposed method, considering the advantage of bipartite graph [28], instead of directly constructing the full $N \times N$ graph, for each view, we learn a $N \times M (M \ll N)$ anchor graph which encodes the relationship between N data points and M anchor points, and then use tensor Schatten p-norm regularization on tensor bipartite, which is composed of anchor graphs of views, to minimize divergence and encode complementary information of views. To avoid the selection of hyper-parameter and encode discriminative information, we present free-parameter weighted strategy, which well takes into account the salient difference between views, to learn a common shared indicator matrix that has K connected components. What's more, we propose an fast method to calculate eigenvalue decomposition of Laplacian matrix. The contributions are summarized as follows:

- We reduce the divergence between anchor graphs, which well preserve manifold structure of each view, via the minimization of tensor Schatten p-norm, which helps encode the complementary information in graphs and low-rank spatial structure of each graph.
- Our model adaptively assigns the weighted parameters for different views by joint taking advantage of the salient difference between views and connected components. This indicates that the learned common shared graph directly characterizes the cluster structure of data. So, our method does not need post-processing to obtains the discrete labels of data.
- An efficient algorithm associated with augmented Lagrange multiplier is developed to tackle the multi-view clustering problem. Moreover, the proposed approach is shown to be tractable and closed-form solutions for all sub-problems of minimizing different variables alternatively are obtained. And we mathematically prove that the proposed algorithm always converges to the KKT stationary point.
- Our proposed algorithm reduces the main computational complexity from $\mathcal{O}(N^3 + VN^2d)$ to $\mathcal{O}(M^2N + VNMd)$, compared with our non-bipartite model. Thus, our proposed tensor bipartite model is time-economical and can be applied to large-scale multi-view clustering.

II. Related Work

In recent years, a great deal of effort has been devoted to the study of multi-view graph clustering. Graph-based Multi-view Spectral Clustering has been more and more popular because of the use of dimensionality reduction, so it is more suitable for common high-dimensional data. Due to the advantages of the graph-based methods, it is also widely used in related fields as graph signal processing (GSP) [29]–[31] and graph neural networks (GNN) [32], [33]. Graph-based clustering methods usually find a compatible fusion graph of multiple views, and then use other algorithms based on this fusion graph to generate the desired final clusters. In order to learn the good common fusion graph, Zhan *et al.* [34] presented a graph learning model for multi-view clustering (MVGL). MVGL learns a common shared graph by adaptively linear combination of similarity graphs of different views with Laplacian rank constraint. To well exploit local geometric structure, Nie *et al.* [35] leveraged the idea of laplacian embedding to learn a common shared graph and presented MLAN which encodes the local intrinsic geometric structure of data. However, MLAN indicates that all views have the same local intrinsic geometric structure. This constraint is unreasonable in practical applications, resulting in suboptimal performance.

In order to solve large scale data, Li *et al.* [36] used a small anchor graph instead of full graph to learn the real-valued indicator matrix and presented an efficient spectral clustering model which is called MVSC for clustering. Lin *et al.* [37] proposed a multi-view attributed graph (MAGC) framework for clustering. It exploits both node attributes and graph structure. motivated by this, Kang *et al.* [38] first constructed anchor graphes of views, which encode the similarity between nodes and anchor points, and then learned a common shared bipartite graph with the connectivity constraint. In Auto-Encoder (AE)-based deep subspace clustering, Lv *et al.* [39] applied pairwise similarity to weigh the reconstruction loss to capture local structure information, while a similarity is learned by the self-expression layer in deep neural networks.

Besides above, it is worth mentioning that, different from the aforementioned clustering methods, Xie *et al.* [40], [41] used self-representation coefficients of views to construct a tensor and minimized the tensor nuclear norm based on t-SVD to learn the view-consensus adjacency matrix for clustering. It has been noted by many authors that high-dimensional

data sets are more compressible when treated as tensors and compressed via t-SVD [42]. Kilmer and Martin *et al.* replaced the tensor rank function by its tensor nuclear norm via t-SVD [43] so that the rank approximation becomes tractable. But it is well known that the tensor nuclear norm regularization may over-penalize the larger singular values (the larger ones may carry undesirable information) when there is a large gap between adjacent singular values (*e.g.*, see [44]). So, Gao *et al.* [23] introduced tensor Schatten p-norm and showed that solving the tensor rank minimization problem by using tensor Schatten p-norm can avoid over-penalizing the larger singular values and will be more beneficial to exploit low-rank structural information of tensor. So in this paper, the tensor Schatten p-norm (to the power p) will be used as a better rank approximation.

III. Notations

In this paper, we use bold calligraphy letters for third-order tensors, *e.g.*, $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, bold upper case letters for matrices, *e.g.*, $\mathbf{A}$, bold lower case letters for vectors, *e.g.*, $\mathbf{a}$, and lower case letters such as a_{ijk} for the entries of $\boldsymbol{\mathcal{A}}$. The i-th frontal slice of $\boldsymbol{\mathcal{A}}$ is $\boldsymbol{\mathcal{A}}^{(i)}$. $\overline{\boldsymbol{\mathcal{A}}}$ is the discrete Fast Fourier Transform (FFT) of $\boldsymbol{\mathcal{A}}$ along the third dimension, *i.e.*, $\overline{\boldsymbol{\mathcal{A}}} = \text{fft}(\boldsymbol{\mathcal{A}}, [], 3)$. Thus, $\boldsymbol{\mathcal{A}} = \text{ifft}(\overline{\boldsymbol{\mathcal{A}}}, [], 3)$. The trace of matrix $\mathbf{A}$ is denoted by $\text{tr}(\mathbf{A})$. $\mathbf{I}$ is an identity matrix. The Frobenius norm of $\boldsymbol{\mathcal{A}}$ is defined as $\|\boldsymbol{\mathcal{A}}\|_F = \sqrt{\sum_{i,j,k} |a_{ijk}|^2}$.

Definition 1 (t-product [43]): Suppose $\mathcal{X} \in \mathbb{R}^{n_1 \times m \times n_3}$ and $\mathcal{Y} \in \mathbb{R}^{m \times n_2 \times n_3}$, then the t-product $\mathcal{X} * \mathcal{Y} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is defined as

$$\mathcal{X} * \mathcal{Y} = \text{ifft}(\text{bdiag}(\overline{XY}), [], 3),$$

where we use $\overline{X}$ =bdiag($\overline{\mathcal{X}}$) to denote the block diagonal matrix whose blocks are frontal slices of $\overline{\mathcal{X}}$.

And by using the t-product, we have the following new product decompositions of tensors (To save spaces, the definitions of orthogonal tensor, f-diagonal tensor and tensor transpose are omitted (*e.g.*, see [43])):

Definition 2 (t-SVD [43]): The tensor Singular Value Decomposition (t-SVD) of $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is given by $\boldsymbol{\mathcal{A}} = \mathcal{U} * \mathcal{S} * \mathcal{V}^{\mathrm{T}}$, where $\mathcal{U}$ and $\mathcal{V}$ are orthogonal tensors of size $n_1 \times n_1 \times n_3$ and $n_2 \times n_2 \times n_3$ respectively. $\mathcal{S}$ is an f-diagonal tensor of size $n_1 \times n_2 \times n_3$, and $*$ denotes the t-product.

Definition 3: Given $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, $h = \min(n_1, n_2)$, tensor Schatten p-norm of tensor $\boldsymbol{\mathcal{A}}$ is defined as

$$\begin{aligned} \|\boldsymbol{\mathcal{A}}\|_{\circledS_p} &= \left(\sum_{i=1}^{n_3} \left\| \overline{\boldsymbol{\mathcal{A}}}^{(i)} \right\|_{\circledS_p}^{p} \right)^{\frac{1}{p}} \\ &= \left(\sum_{i=1}^{n_3} \sum_{j=1}^{h} \sigma_j \left(\overline{\boldsymbol{\mathcal{A}}}^{(i)} \right)^{p} \right)^{\frac{1}{p}} \end{aligned} \tag{1}$$

where $0 \leq p \leq 1$, $\sigma_j(\overline{\boldsymbol{\mathcal{A}}}^{(i)})$denotes the j-th singular value of $\overline{\boldsymbol{\mathcal{A}}}^{(i)}$.

Remark 1: It is easy to see that, let $p = 1$, the tensor Schatten p-norm of tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ becomes tensor nuclear norm [45]: $\|\boldsymbol{\mathcal{A}}\|_* = \sum_{i=1}^{n_3} \sum_{j=1}^{h} \sigma_j \left(\overline{\boldsymbol{\mathcal{A}}}^{(i)} \right)$. For easy of representation, we leverage matrix instead of tensor to explain the Schatten p-norm. Given matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ and its singular values $\sigma_1, \ldots, \sigma_h$ which are sorted in descending order. Then for $p > 0$, we have $\|\mathbf{A}\|_{\circledS_p}^p = \sigma_1^p + \cdots + \sigma_h^p$. When $p \to 0$, we have $\lim_{p \to 0} \|\mathbf{A}\|_{\circledS_p}^p = \#\{i : \sigma_i \neq 0\} = \text{rank}(A)$. Hence, in literature, for $0 \leq p \leq 1$, the Schatten p-norm (which is a quasi-norm [46]) is introduced for the rank approximation.

The main concepts of our work is summarized in the Table I.

TABLE I
Tensor Notations

Notation	Explanation
$\mathcal{X} * \mathcal{Y}$	The t-product of 3rd-order tensors $\mathcal{X} \in \mathbb{R}^{n_1 \times m \times n_3}$, $\mathcal{Y} \in \mathbb{R}^{m \times n_2 \times n_3}$.
$\|\boldsymbol{\mathcal{A}}\|_F$	The Frobenius norm of a tensor $\boldsymbol{\mathcal{A}}$.
$\overline{\boldsymbol{\mathcal{A}}}$	The Fourier transformed tensor of $\boldsymbol{\mathcal{A}}$.
$\overline{\boldsymbol{\mathcal{A}}}^{(i)}$	The i-th frontal slice of $\overline{\boldsymbol{\mathcal{A}}}$.
$\|\boldsymbol{\mathcal{A}}\|_{\circledS_p}$	The tensor Schatten p-norm of a tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$.

IV. The Proposed Method

A. Problem Formulation and Objective

Given multi-view data $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(V)}$, where $\mathbf{X}^{(v)} \in \mathbb{R}^{d_v \times N}$ denotes data matrix of the v-th view, d_v and N denote the feature dimension and number of samples in the v-th view, respectively. Denote by $\tilde{\mathbf{L}}_{\mathbf{B}^{(v)}}$, $\mathbf{B}^{(v)}$, $\mathbf{D}^{(v)}$ the normalized Laplacian matrix, similarity matrix and degree matrix of the v-th view data, respectively, then $\tilde{\mathbf{L}}_{\mathbf{B}^{(v)}} = \mathbf{I} - \mathbf{D}_{(v)}^{-\frac{1}{2}} \mathbf{B}^{(v)} \mathbf{D}_{(v)}^{-\frac{1}{2}}$, $\mathbf{D}^{(v)}$ is a diagonal matrix whose i-th diagonal element is $\mathbf{D}^{(v)}(i,i) = \sum_{j=1}^{i_{max}} \mathbf{B}^{(v)}(i,j)$, where i_{max} is the number of rows or columns of square matrix $\mathbf{D}_{(v)}$. One of the most representative multi-view spectral clustering methods, which take into account salient difference between views, is

$$\begin{aligned} &\min_{\mathbf{F}, \alpha^{(v)}} \sum_{v=1}^{V} \alpha^{(v)^r} \text{tr}\left(\mathbf{F}^{\mathrm{T}} \tilde{\mathbf{L}}_{\mathbf{B}^{(v)}} \mathbf{F} \right) \\ &\text{s.t. } \mathbf{F}^{\mathrm{T}} \mathbf{F} = \mathbf{I}, \sum_{v=1}^{V} \alpha^{(v)} = 1, \alpha^{(v)} \geq \quad 0 \end{aligned} \tag{2}$$

where $\mathbf{F}$ is a $N \times K$ matrix, $\alpha^{(v)}$ is the non-negative normalized weight factor for the v-th view and r is a scalar to control the distribution of different weights among different views.

Despite impressive performance, it still has the following deficiency. **First**, hyper-parameter r affects the final performance, this reduces the flexibility of model. In practical applications, it is still an open problem to manually choose a suitable value due to the large divergence between views and complex structure of data. **Second**, it needs post-processing such as K-means to learn discrete labels. This makes the performance suboptimal. Performance of Eq. (2) heavily depends on the predefined graphs $\mathbf{B}^{(v)}$. As the aforementioned analysis, it is impossible to construct a good graph manually in real-world applications. Thus, it does not get clustering performance from $\mathbf{B}^{(v)}$'s and usually leverages k-means to obtain final clustering performance. **Third**, it doesn't encode both the complementary information in adjacency matrices of views and low-rank spatial structure of each adjacency

matrix. Eq. (2) obtains the weighted Laplacian matrix by linear combination of all graphs, which is element by element and ignores the spatial structure and complementary information between them. Thus, it is inefficient and inherently difficult to be applied to large scale data.

To tackle the above deficiency, we present a new tensorized spectral clustering model. Specifically, to remove hyper-parameter r, we design a reasonable weighted strategy to exploit the salient difference between views. Our objective is

$$\begin{aligned}&\min_{\mathbf{F},\alpha^{(v)}} \sum_{v=1}^{V} \frac{1}{\alpha^{(v)}} \mathrm{tr}\left(\mathbf{F}^{\mathrm{T}} \tilde{\mathbf{L}}_{\mathbf{B}^{(v)}} \mathbf{F}\right) \\ &\text{s.t. } \mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}, \sum_{v=1}^{V} \alpha^{(v)} = 1, \alpha^{(v)} \geq \quad 0 \end{aligned} \tag{3}$$

Eq. (3) requires to predefine a similarity graph for each view and calculate eigen-vectors of the weighted Laplacian matrix. Inspired by tensor nuclear norm, which well exploits complementary information and spatial structure embedded in tensor, we leverage tensor Schatten p-norm (See Definition 1) to measure the similarity between $\mathbf{B}^{(v)}$'s. Thus, we have

$$\begin{aligned}&\min_{\mathbf{F},\alpha^{(v)},\boldsymbol{\mathcal{B}}} \|\boldsymbol{\mathcal{B}}\|_{\circledS p}^{p} + \sum_{v=1}^{V} \frac{1}{\alpha^{(v)}} \mathrm{tr}\left(\mathbf{F}^{\mathrm{T}} \tilde{\mathbf{L}}_{\mathbf{B}^{(v)}} \mathbf{F}\right) \\ &\text{s.t. } \mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}, \sum_{v=1}^{V} \alpha^{(v)} = 1, \alpha^{(v)} \geq \quad 0, \\ &\mathbf{B}^{(v)}\mathbf{1} = \mathbf{1}, \mathbf{B}^{(v)} \geq \quad 0 \end{aligned} \tag{4}$$

where $\boldsymbol{\mathcal{B}}(:, v, :) = \mathbf{B}^{(v)}$ denotes a $N \times N$ graph of tensor $\boldsymbol{\mathcal{B}}$, the normalized Laplacian matrix $\tilde{\mathbf{L}}_{\mathbf{B}^{(v)}} = \mathbf{I} - \mathbf{D}_{(v)}^{-\frac{1}{2}} \mathbf{B}^{(v)} \mathbf{D}_{(v)}^{-\frac{1}{2}}$ is a $N \times N$ matrix and $\mathbf{D}^{(v)}$ is a $N \times N$ diagonal matrix, whose diagonal entries are $\mathbf{D}^{(v)}(i, i) = \sum_{j=1}^{i_{max}} \mathbf{B}^{(v)}(i, j)$. $\mathbf{F}$ denotes the $N \times K$ cluster indicator matrix, the optimal solution $\mathbf{F}$ are composed of the eigenvectors corresponding to the smallest eigenvalues of $\tilde{\mathbf{L}}_{B^{(v)}}$.

Lemma 1: [47] The multiplicity K of zero eigenvalues of $\tilde{\mathbf{L}} = \sum_{v=1}^{V} (\frac{\tilde{\mathbf{L}}_{\mathbf{B}^{(v)}}}{\alpha^{(v)}})$ is equals to the number of connected components in the graph associated with $\mathbf{G}$.

In Eq. (4), $\mathbf{B}^{(v)}$ usually has no exact K-connected components, where K denotes cluster number. This results in the need for post-process to get final labels. To handle this problem, motivated by Lemma 1, we use the Laplacian rank constraint to ensure that the common shared weighted graph $\mathbf{G}$ has exact K-connected components. Thus, we adaptively tune the **hidden parameter** β such that the weighted graph $\mathbf{G} = (\sum_{v=1}^{V} \frac{\mathbf{B}^{(v)}}{\alpha^{(v)}})/(\sum_{v=1}^{V} \frac{1}{\alpha^{(v)}})$ has exact K-connected components. Thus, we have

$$\begin{aligned}&\min_{\mathbf{F},\alpha^{(v)},\boldsymbol{\mathcal{B}}} \|\boldsymbol{\mathcal{B}}\|_{\circledS p}^{p} + \beta \sum_{v=1}^{V} \frac{1}{\alpha^{(v)}} \mathrm{tr}\left(\mathbf{F}^{\mathrm{T}} \tilde{\mathbf{L}}_{\mathbf{B}^{(v)}} \mathbf{F}\right) \\ &s.t. \ \mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}, \sum_{v=1}^{V} \alpha^{(v)} = 1, \alpha^{(v)} \geq \quad 0, \\ &\mathbf{B}^{(v)}\mathbf{1} = \mathbf{1}, \mathbf{B}^{(v)} \geq \quad 0 \end{aligned} \tag{5}$$

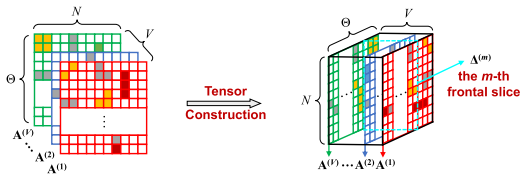


Fig. 2. Construction of tensor $\boldsymbol{\mathcal{A}}$ schatten p-norm, $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N \times V \times \Theta}$.

Note that, hidden parameter means that it cannot be tuned manually in real applications. β can be adaptively updated as follows. We first initialize β with a small value, and update it according to the number of eigenvalue zero of $\tilde{\mathbf{L}}$ after each iteration. If this number is smaller than K, β is multiplied by 2; or if it is greater than $K + 1$, β is divided by 2, otherwise we terminate the iterations.

Remark 2: As shown in Fig. 2, when $\Theta = N$, we construct full size graph so tensor $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{A}} \in \mathbb{R}^{N \times V \times N}$. We use the complete $N \times N$ adjacency matrix to characterizes the relationship between the i-th data point and the j-th data point in the v-th view. Then the tensor $\boldsymbol{\mathcal{B}}$'s m-th frontal slice $\boldsymbol{\Delta}^{(m)}$ represents the relationship between the N data points and the m-th data point in different views. Considering that different views usually have different cluster structures, we add a tensor multi-rank minimization constraint on the tensor Schatten-p norm to ensure that each $\boldsymbol{\Delta}^{(m)}$ has a spatial low-rank structure. Thus $\boldsymbol{\Delta}^{(m)}$ can better characterize the complementary information embedded in different views.

B. Optimization

To optimize Eq. (5, we use Augmented Lagrange Multiplier (ALM) method [48] to iteratively solve the optimal solution. First, auxiliary variable $\boldsymbol{\mathcal{J}}$ is introduced to replace tensor $\boldsymbol{\mathcal{B}}$ in our model and rewrite (5) as the following problem:

$$\begin{aligned}&\min_{\mathbf{F},\boldsymbol{\mathcal{J}},\boldsymbol{\mathcal{B}},\alpha^{(v)}} \|\boldsymbol{\mathcal{J}}\|_{\circledS p}^{p} + \frac{\mu}{2} \left\| \boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{J}} - \frac{\boldsymbol{\mathcal{Q}}}{\mu} \right\|_F^2 \\ &\qquad + \beta \sum_{v=1}^{V} \frac{1}{\alpha^{(v)}} \mathrm{tr}\left(\mathbf{F}^{\mathrm{T}} \tilde{\mathbf{L}}_{B^{(v)}} \mathbf{F}\right) \\ &s.t. \ \mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}, \sum_{v=1}^{V} \alpha^{(v)} = 1, \alpha^{(v)} \geq \quad 0, \\ &\mathbf{B}^{(v)}\mathbf{1} = \mathbf{1}, \mathbf{B}^{(v)} \geq \quad 0 \end{aligned} \tag{6}$$

where tensor $\boldsymbol{\mathcal{Q}}$ is Lagrange multiplier, μ is the penalty parameter. To optimize the model (6), we have the following four subproblems:

Solving F with fixed $\mathbf{B}^{(v)}$, $\boldsymbol{\mathcal{Q}}$, $\alpha^{(v)}$ and $\boldsymbol{\mathcal{J}}$. In this case, the optimization w.r.t $\mathbf{F}$ in Eq. (6) becomes

$$\arg\min_{\mathbf{F}^{\mathrm{T}}\mathbf{F}=\mathbf{I}} \sum_{v=1}^{V} \frac{1}{\alpha^{(v)}} \mathrm{tr}\left(\mathbf{F}^{\mathrm{T}} \tilde{\mathbf{L}}_{\mathbf{B}^{(v)}} \mathbf{F}\right) \tag{7}$$

We can solve (7) by using the eigenvalue decomposition, where the optimal solution to variable $\mathbf{F}$ are the first K eigenvectors corresponding to the first K largest eigenvalues of matrix

$$\tilde{\mathbf{L}} = \sum_{v=1}^{V} \frac{1}{\alpha^{(v)}} \tilde{\mathbf{L}}_{\mathbf{B}^{(v)}} \tag{8}$$

Since $\tilde{\mathbf{L}}_{\mathbf{B}^{(v)}}$ is normalized Laplacian matrix of the v-th view and $\sum_{v=1}^{V}\alpha^{(v)}=1, \alpha^{(v)}\geq 0$, the matrix $\tilde{\mathbf{L}}$ is a real symmetric matrix. Thus, we can obtain the optimal solution of matrix $\mathbf{F}$ by directly performing eigenvalue decomposition on Laplacian matrix $\tilde{\mathbf{L}}$.

Solving $\mathcal{J}$ with fixed $\mathbf{B}^{(v)}$, $\mathcal{Q}$, $\mathbf{F}$ and $\alpha^{(v)}$. Thus, the model (6) becomes

$$\arg\min_{\mathcal{J}} \|\mathcal{J}\|_{\circledS p}^{p}+\frac{\mu}{2}\left\|\left(\mathcal{B}+\frac{\mathcal{Q}}{\mu}\right)-\mathcal{J}\right\|_F^2 \tag{9}$$

We resort to the following Theorem 1 to solve Eq. (9) (*e.g.*, see [23]).

Theorem 1: [23] Suppose $\mathcal{Z}\in\mathbb{R}^{n_1\times n_2\times n_3}$, let $\mathcal{Z}=\mathcal{U}*\mathcal{S}*\mathcal{V}^{\mathrm{T}}$. For the following model:

$$\arg\min_{\mathcal{X}} \frac{1}{2}\|\mathcal{X}-\mathcal{Z}\|_F^2+\tau\|\mathcal{X}\|_{\circledS p}^{p} \tag{10}$$

the optimal solution $\mathcal{X}^*$ is

$$\mathcal{X}^*=\Gamma_{\tau\cdot n_3}(\mathcal{Z})=\mathcal{U}*\mathrm{ifft}\left(\mathbf{P}_{\tau\cdot n_3}\left(\overline{\mathcal{Z}}\right)\right)*\mathcal{V}^{\mathrm{T}} \tag{11}$$

where $P_{\tau\cdot n_3}(\bar{\mathcal{Z}})$ (with frontal slices $P_{\tau\cdot n_3}(\bar{\mathcal{Z}}^{(i)})$) is a f-diagonal tensor, whose elements are found via using the generalized shrinkage-thresholding (GST) algorithm introduced in Lemma 1 of [23].

Using Theorem 1, we easily get the optimal solution of the model (9), *i.e.*,

$$\mathcal{J}^*=\Gamma_{\frac{1}{\mu}}\left(\mathcal{B}+\frac{\mathcal{Q}}{\mu}\right) \tag{12}$$

Solving $\mathbf{B}^{(v)}$ with fixed $\mathcal{J}$, $\mathcal{Q}$, $\mathbf{F}$ and $\alpha^{(v)}$. The optimization w.r.t $\mathbf{B}^{(v)}$ in Eq. (6) becomes

$$\begin{aligned}\arg\min_{\mathbf{B}^{(v)}} &\frac{\mu}{2}\sum_{v=1}^{V}\left\|\mathbf{B}^{(v)}-\mathbf{J}^{(v)}-\frac{\mathbf{Q}^{(v)}}{\mu}\right\|_F^2\\ &+\beta\sum_{v=1}^{V}\frac{1}{\alpha^{(v)}}\mathrm{tr}\left(\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{B}^{(v)}}\mathbf{F}\right)\\ s.t.\ &\mathbf{F}^{\mathrm{T}}\mathbf{F}=\mathbf{I}, \sum_{v=1}^{V}\alpha^{(v)}=1, \alpha^{(v)}\geq\ \ 0,\\ &\mathbf{B}^{(v)}\mathbf{1}=\mathbf{1}, \mathbf{B}^{(v)}\geq\ \ 0\end{aligned} \tag{13}$$

The first term in Eq. (13) can be rewritten as

$$\begin{aligned}\frac{\mu}{2}\sum_{v=1}^{V}\left\|\mathbf{B}^{(v)}-\mathbf{J}^{(v)}-\frac{\mathbf{Q}^{(v)}}{\mu}\right\|_F^2&=\mathrm{Const}\\ &+\frac{\mu}{2}\sum_{v=1}^{V}\left\{\mathrm{tr}\left(\mathbf{B}^{(v)}\mathbf{B}^{(v)^{\mathrm{T}}}\right)-2\mathrm{tr}\left(\mathbf{B}^{(v)^{\mathrm{T}}}\mathbf{P}^{(v)}\right)\right\}\end{aligned} \tag{14}$$

where $\mathbf{P}^{(v)}=\mathbf{J}^{(v)}-\frac{1}{\mu}\mathbf{Q}^{(v)}$.

The second term in Eq. (13) can be rewritten as

$$\mathrm{tr}\left(\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{B}^{(v)}}\mathbf{F}\right)=\mathrm{Const}-\sum_{v=1}^{V}\mathrm{tr}\left(\mathbf{B}^{(v)^{\mathrm{T}}}\mathbf{W}^{(v)^{\mathrm{T}}}\right) \tag{15}$$

where $\mathbf{W}^{(v)}=\frac{1}{\alpha^{(v)}}\mathbf{D}_{(v)}^{-\frac{1}{2}}\mathbf{F}\mathbf{F}^{T}\mathbf{D}_{(v)}^{-\frac{1}{2}}$.

Substituting Eq. (14) and Eq. (15) into Eq. (13), we rewrite Eq. (13) as the following optimization problem.

$$\begin{aligned}\arg\min_{\mathbf{B}^{(v)}} &\frac{\mu}{2}\sum_{v=1}^{V}\left\|\mathbf{B}^{(v)}-\frac{\mathbf{O}^{(v)}}{\mu}\right\|_F^2\\ s.t.\ &\mathbf{B}^{(v)}\mathbf{1}=\mathbf{1}, \mathbf{B}^{(v)}\geq 0\end{aligned} \tag{16}$$

where $\mathbf{O}^{(v)}=\mu\mathbf{J}^{(v)}-(\mathbf{Q}^{(v)}-\beta(\mathbf{W}^{(v)})^{T}))$.

In Eq. (16), all $\mathbf{B}^{(v)}(v=1,\cdots,V)$ are independent. For each $\mathbf{B}^{(v)}$, the closed-form solution $\mathbf{B}^{(v)^*}$ is $\mathbf{B}^{(v)^*}(i,:)=(\frac{\mathbf{A}^{(v)}(i,:)}{\mu}+\gamma\mathbf{1})_+$ [49], where γ is the Lagrangian multiplier.

Solving $\alpha^{(v)}$ with fixed other variables. By simple algebra, the model (6) becomes

$$\begin{aligned}\arg\min_{\alpha^{(v)}} &\sum_{v=1}^{V}\frac{1}{\alpha^{(v)}}\mathrm{tr}\left(\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{B}^{(v)}}\mathbf{F}\right)\\ s.t.&\sum_{v=1}^{V}\alpha^{(v)}=1, \alpha^{(v)}\geq 0\end{aligned} \tag{17}$$

Let $h^{(v)}=\mathrm{tr}(\mathbf{F}^{T}\tilde{\mathbf{L}}_{\mathbf{B}^{(v)}}\mathbf{F})$, the Eq. (17) becomes

$$\begin{aligned}\arg\min_{\alpha^{(v)}} &\sum_{v=1}^{V}\frac{h^{(v)}}{\alpha^{(v)}}\\ s.t.&\sum_{v=1}^{V}\alpha^{(v)}=1, \alpha^{(v)}\geq 0\end{aligned} \tag{18}$$

According to the method of Lagrange multipliers, with the condition that $\sum_{v=1}^{V}\alpha^{(v)}=1$ and $\alpha^{(v)}\geq 0$, the Eq. (18) can be rewritten as

$$\arg\min_{\alpha^{(v)},\lambda}\sum_{v=1}^{V}\frac{h^{(v)}}{\alpha^{(v)}}+\lambda\left(\sum_{v=1}^{V}\alpha^{(v)}-1\right) \tag{19}$$

where λ is the Lagrange multiplier.

Take the derivative of each $\alpha^{(v)}$ in Eq. (19), we can get

$$-\frac{h^{(v)}}{(\alpha^{(v)})^2}+\lambda=0 \tag{20}$$

where $v=1,2,3,\cdots,V$.

Combining Eq. (20) and expression $\sum_{v=1}^{V}\alpha^{(v)}=1$, we can obtain a saddle point of the Lagrangian function and the Lagrange multiplier λ is

$$\lambda=\left(\sum_{v}\sqrt{h^{(v)}}\right)^2 \tag{21}$$

Substituting λ into Eq. (20) with simple algebraic calculation, the optimal $\alpha^{(v)}$ can be obtained by

$$\alpha^{(v)}=\frac{\sqrt{h^{(v)}}}{\sum_{v=1}^{V}\sqrt{h^{(v)}}} \tag{22}$$

The whole algorithm is summarized in Algorithm 1.

Algorithm 1 Algorithm to Solve (5)

Input: Data matrices: $\{\mathbf{X}^{(v)}\}_{v=1}^{V} \in \mathbb{R}^{N\times d_v}$, anchors number M, and cluster number K.
Output: Graph $\mathbf{G}$ with K-connected components.
Construct graphs $\mathbf{B}^{(v)} \in \mathbb{R}^{N\times N}$ by [17];
Initialize $\mathcal{Q} = \mathcal{J} = \mathbf{0}$, $\mu = 10^{-5}$, $max\_\mu = 10^{10}$, $\eta = 1.1$, $\alpha^{(v)} = \frac{1}{V}$;
while *not converge* **do**
  Calculate $\mathbf{F}$ via solving Eq. (7);
  Calculate $\mathcal{J}$ by using Eq. (12);
  Calculate $\{\mathbf{B}^{(v)}\}_{v=1}^{V}$ via solving Eq. (16);
  Calculate $\{\alpha^{(v)}\}_{v=1}^{V}$ by using Eq. (22);
  Calculate $\mathcal{Q}$ and μ: $\mathcal{Q} = \mathcal{Q} + \mu(\mathcal{B} - \mathcal{J})$, $\mu = \min(\eta \times \mu, max\_\mu)$;
end
Directly achieve the K clusters based on the connectivity of $\mathbf{G} = (\sum_{v=1}^{V} \frac{\mathbf{B}^{(v)}}{\alpha^{(v)}})/(\sum_{v=1}^{V} \frac{1}{\alpha^{(v)}})$;
return Clustering results.

V. Fast Method With Bipartite Graph

A. Motivation and Objective

The model (5) has a good performance, but it is time consuming to directly solve the eigenvalue decomposition of weighted Laplacian matrix in (5). The computational complexity is $\mathcal{O}(N^3)$. Moreover, it requires to predefine $N \times N$ graph $\mathbf{B}^{(v)}(v = 1, \cdots, V)$, this stage takes $\mathcal{O}(VN^2d)$ time, where $d = \sum_{v=1}^{V} d_v$. Thus, it is inefficient and inherently difficult to be applied to large scale data.

As the aforementioned analysis, the Eq. (3) in (5) requires to predefine a similarity graph for each view and calculate eigenvectors of the weighted Laplacian matrix, so it is inefficient and inherently difficult to be applied to large scale multi-view data. To reduce the computational complexity, we construct an effective bipartite graph $\mathbf{C}^{(v)} \in \mathbb{R}^{N\times M}$, which exploits the relationship between N data points and $M(M \ll N)$ anchors, instead of $N \times N$ global graph $\mathbf{B}^{(v)}$, inspired by [28]. This remarkably reduces the main computational complexity from $\mathcal{O}(VN^2d)$ to $\mathcal{O}(VNMd)$. Thus, we have

$$\begin{aligned}\min_{\mathbf{F},\alpha^{(v)},\mathcal{C}} &\|\mathcal{C}\|_{\mathcal{S}p}^{p} + \sum_{v=1}^{V}\frac{1}{\alpha^{(v)}}\mathrm{tr}\left(\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{S}^{(v)}}\mathbf{F}\right)\\ s.t.\ &\mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}, \sum_{v=1}^{V}\alpha^{(v)} = 1, \alpha^{(v)} \geq 0,\\ &\mathbf{C}^{(v)}\mathbf{1} = \mathbf{1}, \mathbf{C}^{(v)} \geq 0\end{aligned} \tag{23}$$

where $\mathcal{C}(:, v, :) = \mathbf{C}^{(v)}$ denotes a $N \times M$ graph of tensor $\mathcal{C}$, $\tilde{\mathbf{L}}_{S^{(v)}} = \mathbf{I} - \mathbf{D}_{(v)}^{-\frac{1}{2}}\mathbf{S}^{(v)}\mathbf{D}_{(v)}^{-\frac{1}{2}}$ is the normalized Laplacian matrix of $\mathbf{S}^{(v)} \in \mathbb{R}^{(N+M)\times(N+M)}$ with $\mathbf{S}^{(v)} = \begin{bmatrix} & \mathbf{C}^{(v)} \\ \mathbf{C}^{(v)^{\mathrm{T}}} & \end{bmatrix}$. $\mathbf{D}^{(v)}$ is a diagonal matrix whose diagonal elements are $\mathbf{D}^{(v)}(i, i) = \sum_{j=1}^{N+M}\mathbf{S}^{(v)}(i, j)$.

In Eq. (23), $\mathbf{S}^{(v)}$ is intrinsically comprised of double $\mathbf{C}^{(v)}$, the K-connected $\mathbf{S}^{(v)}$ certainly guarantees the K-connected $\mathbf{C}^{(v)}$, so we can also tune the **hidden parameter** β to ensure the weighted graph $\mathbf{G} = (\sum_{v=1}^{V} \frac{\mathbf{C}^{(v)}}{\alpha^{(v)}})/(\sum_{v=1}^{V} \frac{1}{\alpha^{(v)}})$ has exact K-connected components. Thus, **our final objective** is

$$\begin{aligned}\min_{\mathbf{F},\alpha^{(v)},\mathcal{C}} &\|\mathcal{C}\|_{\mathcal{S}p}^{p} + \beta\sum_{v=1}^{V}\frac{1}{\alpha^{(v)}}\mathrm{tr}\left(\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{S}^{(v)}}\mathbf{F}\right)\\ s.t.\ &\mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}, \sum_{v=1}^{V}\alpha^{(v)} = 1, \alpha^{(v)} \geq 0,\\ &\mathbf{C}^{(v)}\mathbf{1} = \mathbf{1}, \mathbf{C}^{(v)} \geq 0\end{aligned} \tag{24}$$

Both of the weighted graphs in Eq. (5) and Eq. (24) have exact K-connected components, so the **hidden parameter** β in Eq. (24) can be updated with the same method in Eq. (5).

Remark 3: According to the construction of tensor $\mathcal{C}$, we set $\Theta = M(M \ll N)$ so that tensor $\mathcal{C} = \mathcal{A} \in \mathbb{R}^{N\times V\times M}$, which is show in Fig. 2. Thus we have that, for tensor $\mathcal{C}$, adjacency matrix $\mathbf{C}^{(v)} = \mathbf{A}^{(v)} \in \mathbb{R}^{N\times M}$ characterizes the similarity between N sample points and M anchors in the v-th view, if the i-th sample and m-th anchor belong to the same cluster, then $C_{m,i}^{(v)}$ is high, otherwise $C_{m,i}^{(v)}$ is low or zero. The m-th frontal slice $\mathbf{\Delta}^{(m)}$ of tensor $\mathcal{C}$ is a matrix whose columns are composed of vectors $\mathbf{C}_{:,i}^{(v)}(v = 1, 2, \ldots, V)$, $\mathbf{C}_{:,i}^{(v)}$ denotes the i-th column of indicator matrix $\mathbf{C}^{(v)}$, which characterizes the relationship between $\mathbf{X}^{(v)}$ and the i-th cluster. The purpose of multi-view clustering is that $\mathbf{C}_{:,i}^{(1)}, \mathbf{C}_{:,i}^{(2)}, \ldots, \mathbf{C}_{:,i}^{(V)}$ are as similar practical applications, there are exactly equal. Moreover, in practical applications, there has a large difference between cluster structures of different views. Thus, the first term in (24), *i.e.*, tensor multi-rank minimization constraint on $\mathcal{C}$ can make sure that $\mathbf{\Delta}^{(m)}$ has spatial low-rank structure. It helps exploit the complementary information embedded in inter-views and get the view-consensus indicator matrix.

B. Optimization

We can introduce an auxiliary variable $\mathcal{J}$ and rewrite (24) as the following problem:

$$\begin{aligned}\min_{\mathbf{F},\mathcal{J},\mathcal{C},\alpha^{(v)}} &\|\mathcal{J}\|_{\mathcal{S}p}^{p} + \frac{\mu}{2}\left\|\mathcal{C} - \mathcal{J} - \frac{\mathcal{Q}}{\mu}\right\|_F^2 +\\ &\beta\sum_{v=1}^{V}\frac{1}{\alpha^{(v)}}\mathrm{tr}\left(\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{S}^{(v)}}\mathbf{F}\right)\\ s.t.\ &\mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}, \sum_{v=1}^{V}\alpha^{(v)} = 1, \alpha^{(v)} \geq 0,\\ &\mathbf{C}^{(v)}\mathbf{1} = \mathbf{1}, \mathbf{C}^{(v)} \geq 0\end{aligned} \tag{25}$$

The optimization process of Eq. (25) contains the following steps:

Solving F with fixed $\mathbf{C}^{(v)}$, $\mathcal{Q}$, $\alpha^{(v)}$ and $\mathcal{J}$. In this case, the optimization w.r.t $\mathbf{F}$ in Eq. (25) becomes

$$\arg\min_{\mathbf{F}^{\mathrm{T}}\mathbf{F}=\mathbf{I}} \mathrm{tr}\left(\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{L}}\mathbf{F}\right) \tag{26}$$

where $\tilde{\mathbf{L}} = \sum_{v=1}^{V}\frac{1}{\alpha^{(v)}}(\mathbf{I} - \mathbf{D}_{(v)}^{-\frac{1}{2}}\mathbf{S}^{(v)}\mathbf{D}_{(v)}^{-\frac{1}{2}})$.

To directly optimize (26), the computational complexity is $\mathcal{O}((N + M)^2K)$. We herein provide an effective algorithm.

By simple matrix algebra, we have

$$\begin{aligned}\operatorname{tr}\left(\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{L}}\mathbf{F}\right) &= \sum_{v=1}^{V}\frac{1}{\alpha^{(v)}}\operatorname{tr}\left(\mathbf{F}^{\mathrm{T}}\mathbf{F}\right)\\ &\quad -\sum_{v=1}^{V}\frac{1}{\alpha^{(v)}}\operatorname{tr}\left(\mathbf{F}^{\mathrm{T}}\mathbf{D}_{(v)}^{-\frac{1}{2}}\mathbf{S}^{(v)}\mathbf{D}_{(v)}^{-\frac{1}{2}}\mathrm{F}\right)\\ &\operatorname{tr}\left(\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{L}}\mathbf{F}\right) = \mathrm{Const}\\ &\quad -\sum_{v=1}^{V}\frac{1}{\alpha^{(v)}}\operatorname{tr}\left(\mathbf{F}^{\mathrm{T}}\mathbf{D}_{(v)}^{-\frac{1}{2}}\mathbf{S}^{(v)}\mathbf{D}_{(v)}^{-\frac{1}{2}}\mathbf{F}\right) \qquad (27)\end{aligned}$$

Let us denote $\mathbf{F} = \left[\mathbf{F}_N^{\mathrm{T}}\ \mathbf{F}_M^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\mathbf{D}^{(v)} = diag(\mathbf{D}_N^{(v)}, \mathbf{D}_M^{(v)})$, where $\mathbf{F}_N \in \mathbb{R}^{N\times K}$ is the first N rows of $\mathbf{F}$ and $\mathbf{F}_M \in \mathbb{R}^{M\times K}$ is the remaining M rows of $\mathbf{F}$, $\mathbf{D}_N^{(v)} \in \mathbb{R}^{N\times N}$ and $\mathbf{D}_M^{(v)} \in \mathbb{R}^{M\times M}$ are diagonal matrices whose diagonal elements are $\mathbf{D}_N^{(v)}(i,i) = \sum_{j=1}^{M}\mathbf{C}^{(v)}(i,j)$ and $\mathbf{D}_M^{(v)}(j,j) = \sum_{i=1}^{N}\mathbf{C}^{(v)}(i,j)$. Substituting the above identities and Eq. (27) into Eq. (26), and by simple linear algebra, the optimal solution of Eq. (26) can be obtained by solving Eq. (28).

$$\underset{\mathbf{F}_N^{\mathrm{T}}\mathbf{F}_N+\mathbf{F}_M^{\mathrm{T}}\mathbf{F}_M=\mathbf{I}}{\arg\max}\ \operatorname{tr}\left(\mathbf{F}_N^{\mathrm{T}}\sum_{v=1}^{V}\frac{2\mathbf{C}^{(v)}\mathbf{D}_M^{(v)^{-\frac{1}{2}}}}{\alpha^{(v)}}\mathbf{F}_M\right) \qquad (28)$$

To solve the (28), we first introduce Theorem 2.

Theorem 2: Given $\mathbf{E}\in\mathbb{R}^{N\times M}$, $\mathbf{F}_N\in\mathbb{R}^{N\times K}$, $\mathbf{F}_M\in\mathbb{R}^{M\times K}$. The optimal solutions of

$$\underset{\mathbf{F}_N^{\mathrm{T}}\mathbf{F}_N+\mathbf{F}_M^{\mathrm{T}}\mathbf{F}_M=\mathbf{I}}{\arg\max}\ \operatorname{tr}\left(\mathbf{F}_N^{\mathrm{T}}\mathbf{E}\mathbf{F}_M\right) \qquad (29)$$

are $\mathbf{F}_N = \frac{\sqrt{2}}{2}\mathbf{U}_1$, $\mathbf{F}_M = \frac{\sqrt{2}}{2}\mathbf{V}_1$, where $\mathbf{U}_1$ and $\mathbf{V}_1$ are the leading K left and right singular vectors of $\mathbf{E}$, respectively.

Proof: According to Eq. (29), we have

$$\begin{aligned}\operatorname{tr}(\mathbf{F}_N^{\mathrm{T}}\mathbf{E}\mathbf{F}_M) &= \frac{1}{2}(\operatorname{tr}(\mathbf{F}_N^{\mathrm{T}}\mathbf{E}\mathbf{F}_M) + \operatorname{tr}(\mathbf{F}_M^{\mathrm{T}}\mathbf{E}^{\mathrm{T}}\mathbf{F}_M))\\ &= \frac{1}{2}\operatorname{tr}\left(\begin{bmatrix}\mathbf{F}_N\\ \mathbf{F}_M\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{0} & \mathbf{E}\\ \mathbf{E}^{\mathrm{T}} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{F}_N\\ \mathbf{F}_M\end{bmatrix}\right) \qquad (30)\end{aligned}$$

Then, (29) is equivalent to

$$\begin{aligned}&\underset{\mathbf{F}_N,\mathbf{F}_M}{\arg\max}\ \frac{1}{2}\operatorname{tr}\left\{\begin{bmatrix}\mathbf{F}_N\\ \mathbf{F}_M\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{0} & \mathbf{E}\\ \mathbf{E}^{\mathrm{T}} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{F}_N\\ \mathbf{F}_M\end{bmatrix}\right\}\\ &\text{s.t. } [\mathbf{F}_N\ \mathbf{F}_M]^{\mathrm{T}}[\mathbf{F}_N\ \mathbf{F}_M] = \mathbf{I} \qquad (31)\end{aligned}$$

The optimal solution of (31) can be solved by

$$\frac{1}{2}\begin{bmatrix}\mathbf{0} & \mathbf{E}\\ \mathbf{E}^{\mathrm{T}} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{F}_N\\ \mathbf{F}_M\end{bmatrix} = \begin{bmatrix}\mathbf{F}_N\\ \mathbf{F}_M\end{bmatrix}\Lambda \qquad (32)$$

where Λ is a diagonal matrix and its elements are composed of eigenvalues of $\frac{1}{2}\begin{bmatrix}\mathbf{0} & \mathbf{E}\\ \mathbf{E}^{\mathrm{T}} & \mathbf{0}\end{bmatrix}$.

Some simple block matrix multiplication yields

$$\begin{cases}\frac{1}{2}\mathbf{E}\mathbf{F}_M = \mathbf{F}_N\Lambda\\ \frac{1}{2}\mathbf{E}^{\mathrm{T}}\mathbf{F}_N = \mathbf{F}_M\Lambda\end{cases} \qquad (33)$$

Then it follows,

$$\begin{cases}(\frac{\sqrt{2}}{2}\mathbf{E})^{\mathrm{T}}(\frac{\sqrt{2}}{2}\mathbf{E})\mathbf{F}_M = \mathbf{F}_M(\sqrt{2}\Lambda)^2\\ (\frac{\sqrt{2}}{2}\mathbf{E})(\frac{\sqrt{2}}{2}\mathbf{E})^{\mathrm{T}}\mathbf{F}_N = \mathbf{F}_N(\sqrt{2}\Lambda)^2\end{cases} \qquad (34)$$

According to (34), $\mathbf{F}_N$ and $\mathbf{F}_M$ are composed of the leading K left and right singular vectors of $\frac{\sqrt{2}}{2}\mathbf{E}$. Let us use $\mathbf{U}_1$ and $\mathbf{V}_1$ to denote the leading K left and right singular vectors of $\mathbf{E}$, respectively. Thus, we have $\mathbf{F}_M = \frac{\sqrt{2}}{2}\mathbf{V}_1$, $\mathbf{F}_N = \frac{\sqrt{2}}{2}\mathbf{U}_1$. ■

Denote by $\mathbf{E} = \sum_{v=1}^{V}\frac{\mathbf{C}^{(v)}\mathbf{D}_M^{(v)^{-\frac{1}{2}}}}{\alpha^{(v)}}$, and according to Theorem 2, the optimal $\mathbf{F}^*$ in Eq. (28) is $\mathbf{F}^* = \frac{\sqrt{2}}{2}\left[\mathbf{U}_1^{\mathrm{T}}\ \mathbf{V}_1^{\mathrm{T}}\right]^{\mathrm{T}}$. Here $\mathbf{U}_1$ and $\mathbf{V}_1$ can be obtained by performing SVD on $\mathbf{E}$, which takes the computational complexity $\mathcal{O}(VNM + M^2N)$. So tackling Eq. (28) instead of directly solving Eq. (26) is much more efficient because the number of anchors $M \ll N$ toward large-scale clustering.

Solving $\mathcal{J}$ with fixed $\mathbf{C}^{(v)}$, $\mathcal{Q}$, $\mathbf{F}$ and $\alpha^{(v)}$. In this case, $\mathcal{J}$ can be solved by

$$\underset{\mathcal{J}}{\arg\min}\ \|\mathcal{J}\|_{\circledS p}^{p} + \frac{\mu}{2}\left\|\left(\mathcal{C}+\frac{\mathcal{Q}}{\mu}\right)-\mathcal{J}\right\|_F^2 \qquad (35)$$

Same as the Eq. (9), we use Theorem 1 to obtain the optimal solution of tensor $\mathcal{J}$ in Eq. (35):

$$\mathcal{J}^* = \Gamma_{\frac{1}{\mu}}\left(\mathcal{C}+\frac{\mathcal{Q}}{\mu}\right) \qquad (36)$$

Solving $\mathbf{C}^{(v)}$ with fixed $\mathcal{J}$, $\mathcal{Q}$, $\mathbf{F}$ and $\alpha^{(v)}$. The optimization w.r.t $\mathbf{C}^{(v)}$ in Eq. (25) becomes

$$\begin{aligned}&\underset{\mathbf{C}^{(v)}}{\arg\min}\ \frac{\mu}{2}\sum_{v=1}^{V}\left\|\mathbf{C}^{(v)}-\mathbf{J}^{(v)}-\frac{\mathbf{Q}^{(v)}}{\mu}\right\|_F^2\\ &\qquad +\beta\sum_{v=1}^{V}\frac{1}{\alpha^{(v)}}\operatorname{tr}\left(\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{S}^{(v)}}\mathbf{F}\right)\\ &\text{s.t.}\quad \mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}, \sum_{v=1}^{V}\alpha^{(v)} = 1, \alpha^{(v)} \geq 0,\\ &\qquad \mathbf{C}^{(v)}\mathbf{1} = \mathbf{1}, \mathbf{C}^{(v)} \geq 0 \qquad (37)\end{aligned}$$

The first term in Eq. (37) can be rewritten as Eq. (14) using $\mathbf{C}^{(v)}$ instead of $\mathbf{B}^{(v)}$, the second term in Eq. (37) can be rewritten as

$$\operatorname{tr}\left(\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{S}^{(v)}}\mathbf{F}\right) = \mathrm{Const} - 2\sum_{v=1}^{V}\operatorname{tr}\left(\mathbf{C}^{(v)^{\mathrm{T}}}\mathbf{H}^{(v)^{\mathrm{T}}}\right) \qquad (38)$$

where $\mathbf{H}^{(v)} = \frac{1}{\alpha^{(v)}}\mathbf{D}_{(v)}^{-\frac{1}{2}}\mathbf{F}_M\mathbf{F}_N^{T}\mathbf{D}_{(v)}^{-\frac{1}{2}}$. By simple linear algebra, Eq. (37) becomes

$$\begin{aligned}&\underset{\mathbf{C}^{(v)}}{\arg\min}\ \frac{\mu}{2}\sum_{v=1}^{V}\left\|\mathbf{C}^{(v)}-\frac{\mathbf{O}^{(v)}}{\mu}\right\|_F^2\\ &s.t.\ \ \mathbf{C}^{(v)}\mathbf{1} = \mathbf{1}, \mathbf{C}^{(v)} \geq 0 \qquad (39)\end{aligned}$$

where $\mathbf{O}^{(v)} = \mu\mathbf{J}^{(v)} - (\mathbf{Q}^{(v)} - 2\beta(\mathbf{H}^{(v)})^{\mathrm{T}})$.

In Eq. (39), all $\mathbf{C}^{(v)}(v = 1, \cdots, V)$ are also independent, so we can solve $\mathbf{C}^{(v)}$ using the same method as solving $\mathbf{B}^{(v)}$ in Eq. (16).

Algorithm 2 Algorithm to Solve (24)

Input: Data matrices: $\{\mathbf{X}^{(v)}\}_{v=1}^{V} \in \mathbb{R}^{N\times d_v}$, anchors number M, and cluster number K.
Output: Graph $\mathbf{G}$ with K-connected components.
Construct graphs $\mathbf{C}^{(v)} \in \mathbb{R}^{N\times M}$ like [17];
Initialize $\mathcal{Q} = \mathcal{J} = \mathbf{0}$, $\mu = 10^{-5}$, $max\_\mu = 10^{10}$, $\eta = 1.1$, $\alpha^{(v)} = \frac{1}{V}$;
while *not converge* **do**
  Calculate $\mathbf{F}$ by solving Eq. (28);
  Calculate $\mathcal{J}$ by using Eq. (36);
  Calculate $\{\mathbf{C}^{(v)}\}_{v=1}^{V}$ by solving Eq. (39);
  Calculate $\{\alpha^{(v)}\}_{v=1}^{V}$ by using Eq. (41);
  Calculate $\mathcal{Q}$ and μ: $\mathcal{Q} = \mathcal{Q} + \mu(\mathcal{C} - \mathcal{J})$, $\mu = \min(\eta \times \mu, max\_\mu)$;
end
Directly achieve the K clusters based on the connectivity of $\mathbf{G} = (\sum_{v=1}^{V} \frac{\mathbf{C}^{(v)}}{\alpha^{(v)}})/(\sum_{v=1}^{V} \frac{1}{\alpha^{(v)}})$;
return Clustering results.

Solving $\alpha^{(v)}$ with fixed other variables. In this case, the optimization w.r.t $\alpha^{(v)}$ in Eq. (25) becomes

$$\begin{aligned}&\arg\min_{\alpha^{(v)}} \sum_{v=1}^{V} \frac{1}{\alpha^{(v)}} \mathrm{tr}\left(\mathbf{F}^{\mathrm{T}} \tilde{\mathbf{L}}_{\mathbf{S}^{(v)}} \mathbf{F}\right)\\ &s.t. \sum_{v=1}^{V} \alpha^{(v)} = 1, \alpha^{(v)} \geq 0\end{aligned} \tag{40}$$

Using the same method as in Eq. (17), let $g^{(v)} = \sqrt{\mathrm{tr}(\mathbf{F}^{(v)} \tilde{\mathbf{L}}_{\mathbf{S}^{(v)}} \mathbf{F})}$, the optimal $\alpha^{(v)}$ is

$$\alpha^{(v)} = \frac{\sqrt{g^{(v)}}}{\sum_{v=1}^{V} \sqrt{g^{(v)}}} \tag{41}$$

The whole algorithm is summarized in Algorithm 2.

C. Convergence Analysis

The convergence analysis of Algorithm 1 and 2 are quite similar, for simplicity, we only conduct the convergence analysis of Algorithm 2.

Lemma 2 (Proposition 6.2 of [50]): Suppose $F : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}$ is represented as $F(X) = f \circ \sigma(X)$, where $X \in \mathbb{R}^{n_1\times n_2}$ with SVD $X = U\mathrm{diag}(\sigma_1, \ldots, \sigma_n)V^{\mathrm{T}}$, $n = \min(n_1, n_2)$, and f is differentiable. The gradient of $F(X)$ at X is

$$\frac{\partial F(X)}{\partial X} = U\mathrm{diag}(\theta)V^{\mathrm{T}}, \tag{42}$$

where $\theta = \frac{\partial f(y)}{\partial y}|_{y=\sigma(X)}$.

Theorem 3: [Convergence Analysis of **Algorithm 2**] Let $P_k = \{\mathcal{C}_k, \mathcal{J}_k, \mathcal{Q}_k\}$, $1 \leq k < \infty$ in (25) be a sequence generated by **Algorithm 2**, then

1) P_k is bounded;
2) Any accumulation point of P_k is a stationary KKT point of (25).

1) Proof of the 1st Part: To minimize $\mathcal{J}$ at step $k+1$ in (25), the optimal $\mathcal{J}_{k+1}$ needs to satisfy the first-order optimal condition $\lambda\nabla_{\mathcal{J}}\|\mathcal{J}_{k+1}\|_{\circledS p}^{p} + \mu_k(\mathcal{J}_{k+1} - \mathcal{C}_{k+1} - \frac{1}{\mu_k}\mathcal{Q}_k) = 0$.

Recall that when $0 < p < 1$, in order to overcome the singularity of $(|\eta|^p)' = p\eta/|\eta|^{2-p}$ near $\eta = 0$, we consider for $0 < \epsilon \ll 1$ the approximation

$$\partial|\eta|^p \approx \frac{p\eta}{\max\{\epsilon^{2-p}, |\eta|^{2-p}\}}.$$

Letting $\overline{\mathcal{J}}^{(i)} = \overline{\mathcal{U}}^{(i)}\mathrm{diag}\left(\sigma_j(\overline{\mathcal{J}}^{(i)})\right)\overline{\mathcal{V}}^{(i)\mathrm{H}}$, then it follows from Defn. 1 and Lemma 2 that

$$\frac{\partial\|\overline{\mathcal{J}}^{(i)}\|_{\circledS p}^{p}}{\partial\overline{\mathcal{J}}^{(i)}} = \overline{\mathcal{U}}^{(i)}\mathrm{diag}\left(\frac{p\sigma_j(\overline{\mathcal{J}}^{(i)})}{\max\{\epsilon^{2-p}, |\sigma_j(\overline{\mathcal{J}}^{(i)})|^{2-p}\}}\right)\overline{\mathcal{V}}^{(i)\mathrm{H}}.$$

And then one can obtain

$$\begin{aligned}&\frac{p\sigma_j(\overline{\mathcal{J}}^{(i)})}{\max\{\epsilon^{2-p}, |\sigma_j(\overline{\mathcal{J}}^{(i)})|^{2-p}\}} \leq \frac{p}{\epsilon^{1-p}}\\ &\Longrightarrow \left\|\frac{\partial\|\overline{\mathcal{J}}^{(i)}\|_{\circledS p}^{p}}{\partial\overline{\mathcal{J}}^{(i)}}\right\|_F^2 \leq \sum_{i=1}^{N} \frac{p^2}{\epsilon^{2(1-p)}}.\end{aligned}$$

So $\frac{\partial\|\overline{\mathcal{J}}\|_{\circledS p}^{p}}{\partial\overline{\mathcal{J}}}$ is bounded.

Let us denote $\widetilde{\mathbf{F}}_V = \frac{1}{\sqrt{V}}\mathbf{F}_V$, $\mathbf{F}_V$ is the discrete Fourier transform matrix of size $V \times V$, $\mathbf{F}_V^{\mathrm{H}}$ denotes its conjugate transpose. For $\mathcal{J} = \overline{\mathcal{J}} \times_3 \widetilde{\mathbf{F}}_V$ and using the chain rule in matrix calculus, one can obtain that

$$\nabla_{\mathcal{J}}\|\mathcal{J}\|_{\circledS p}^{p} = \frac{\partial\|\mathcal{J}\|_{\circledS p}^{p}}{\partial\overline{\mathcal{J}}} \times_3 \widetilde{\mathbf{F}}_V^{\mathrm{H}}$$

is bounded.

And it follows that

$$\begin{aligned}&\mathcal{Q}_{k+1} = \mathcal{Q}_k + \mu_k(\mathcal{C}_{k+1} - \mathcal{J}_{k+1})\\ &\Longrightarrow \lambda\nabla_{\mathcal{J}}\|\mathcal{J}_{k+1}\|_{\circledS p}^{p} = \mathcal{Q}_{k+1},\end{aligned}$$

thus $\mathcal{Q}_{k+1}$ appears to be bounded.

Moreover, by using the updating rule

$$\mathcal{Q}_k = \mathcal{Q}_{k-1} + \mu_{k-1}(\mathcal{C}_k - \mathcal{J}_k),$$

we can deduce

$$\begin{aligned}\mathcal{L}_{\mu_k}(\mathcal{C}_{k+1}, \mathcal{J}_{k+1}, \mathcal{Q}_k) &\leq \mathcal{L}_{\mu_k}(\mathcal{C}_k, \mathcal{J}_k, \mathcal{Q}_k)\\ &= \mathcal{L}_{\mu_{k-1}}(\mathcal{C}_k, \mathcal{J}_k; \mathcal{Q}_{k-1})\\ &\quad + \frac{\mu_k + \mu_{k-1}}{2\mu_{k-1}^2}\|\mathcal{Q}_k - \mathcal{Q}_{k-1}\|_F^2\\ &\quad + \frac{\|\mathcal{Q}_k\|_F^2}{2\mu_k} - \frac{\|\mathcal{Q}_{k-1}\|_F^2}{2\mu_{k-1}}.\end{aligned} \tag{43}$$

Thus, summing two sides of (43) from $k = 1$ to n, we have

$$\begin{aligned}&\mathcal{L}_{\mu_n}(\mathcal{C}_{n+1}, \mathcal{J}_{n+1}, \mathcal{Q}_n)\\ &\leq \mathcal{L}_{\mu_0}(\mathcal{C}_1, \mathcal{J}_1, \mathcal{Q}_0)\\ &\quad + \frac{\|\mathcal{Q}_n\|_F^2}{2\mu_n} - \frac{\|\mathcal{Q}_0\|_F^2}{2\mu_0}\\ &\quad + \sum_{k=1}^{n}\left(\frac{\mu_k + \mu_{k-1}}{2\mu_{k-1}^2}\|\mathcal{Q}_k - \mathcal{Q}_{k-1}\|_F^2\right).\end{aligned} \tag{44}$$

Observe that

$$\sum_{k=1}^{\infty} \frac{\mu_k + \mu_{k-1}}{2\mu_{k-1}^2} < \infty,$$

we have the right-hand side of (44) is finite and thus $\mathcal{L}_{\mu_n}(\mathcal{C}_{n+1}, \mathcal{J}_{n+1}, \mathcal{Q}_n)$ is bounded. Notice

$$\mathcal{L}_{\kappa_n}(\mathcal{C}_{n+1}, \mathcal{J}_{n+1}, \mathcal{Q}_n) = \beta \sum_{v=1}^{V} \frac{1}{\alpha^{(v)}} \mathrm{tr}\left(\mathbf{F}^T \tilde{\mathbf{L}}_{\mathbf{S}^{(v)}} \mathbf{F}\right) + \lambda \|\mathcal{J}_{n+1}\|_{\circledS p}^p + \frac{\mu_n}{2} \|\mathcal{C}_{n+1} - \mathcal{J}_{n+1} + \frac{\mathcal{Q}_n}{\mu_n}\|_F^2, \quad (45)$$

and each term of (45) is nonnegative, following from the boundedness of $\mathcal{L}_{\mu_n}(\mathcal{C}_{n+1}, \mathcal{J}_{n+1}, \mathcal{Q}_n)$, we can deduce each term of (45) is bounded. And $\|\mathcal{J}_{n+1}\|_{\circledS p}^p$ being bounded implies that all singular values of $\mathcal{J}_{n+1}$ are bounded and hence $\|\mathcal{J}_{n+1}\|_F^2$ (the sum of squares of singular values) is bounded. Therefore, the sequence $\{\mathcal{J}_k\}$ is bounded. Because

$$\mathcal{Q}_{k+1} = \mathcal{Q}_k + \mu_k(\mathcal{C}_k - \mathcal{J}_k) \Longrightarrow \mathcal{C}_k = \mathcal{J}_k + \frac{\mathcal{Q}_{k+1} - \mathcal{Q}_k}{\mu_k},$$

and in light of the boundedness of $\mathcal{J}_k$, $\mathcal{Q}_k$, it is clear that $\mathcal{C}_k$ is also bounded.

2) Proof of the 2nd Part: From Weierstrass-Bolzano theorem, there exists at least one accumulation point of the sequence P_k. We denote one of the points $P^* = \{\mathcal{C}^*, \mathcal{J}^*, \mathcal{Q}^*\}$. Without loss of generality, we assume $\{\mathcal{P}_k\}_{k=1}^{+\infty}$ converge to P^*.

Note that from the updating rule for $\mathcal{Q}$, we have

$$\mathcal{Q}_{k+1} = \mathcal{Q}_k + \mu_k(\mathcal{C}_k - \mathcal{J}_k) \Longrightarrow \mathcal{J}^* = \mathcal{C}^*.$$

In the $\mathcal{J}$-subproblem, we have

$$\lambda \nabla_{\mathcal{J}} \|\mathcal{J}_{k+1}\|_{\circledS p}^p = \mathcal{Q}_{k+1} \Longrightarrow \mathcal{Q}^* = \lambda \nabla_{\mathcal{J}} \|\mathcal{J}^*\|_{\circledS p}^p.$$

In the $\mathbf{C}^{(v)}$-subproblem, one has

$$2\beta\partial \mathrm{tr}(\mathbf{C}_{k+1}^{(v)} \mathbf{H}^{(v)T}) - \alpha^{(v)} \mu_k (\mathbf{J}_{k+1}^{(v)} - \mathbf{C}_{k+1}^{(v)} + \frac{\mathbf{Q}_k^{(v)}}{\mu_k}) = 0.$$

Now by the updating rule $\mathcal{Q}=\mathcal{Q}+\mu(\mathcal{C}-\mathcal{J})$, it is simple to show that

$$\alpha^{(v)} \mathbf{Q}_{k+1}^{(v)} = 2\beta \mathbf{H}^{(v)} \Longrightarrow \alpha^{(v)} \mathbf{Q}^{(v)*} = 2\beta \mathbf{H}^{(v)},$$

Therefore, one can see that the sequences $\mathcal{C}^*$, $\mathcal{J}^*$, $\mathcal{Q}^*$ satisfy the KKT conditions of the Lagrange function (25).

D. Complexity

There are two parts of our method that are time consuming: (1) Construction of graphs $\{\mathbf{C}^{(v)}\}_{v=1}^V$, same to [17], (2) Optimization by iteratively solving Eq.(6). The first stage takes $\mathcal{O}(VNMd + VNM\log(M))$ time, where $d = \sum_{v=1}^V d_v$, V, M and N are the number of views, anchors and samples, respectively. The second stage mainly focuses on three variables ($\mathbf{C}^{(v)}$, $\mathcal{J}$ and $\mathbf{F}$). For $\mathcal{J} \in \mathbb{R}^{N \times M \times V}$, solving $\mathcal{J}$-subproblem involves calculating the 3D FFT and 3D inverse FFT of an $N \times V \times M$ tensor and N SVDs of $M \times V$ matrices in the Fourier domain, both of which are with the complexity of $\mathcal{O}(2VNM\log(VM))$ and $\mathcal{O}(V^2MN)$. So the complexity in updating these variables iteratively are $\mathcal{O}(VNM(K+1) + VNM\log(M))$, $\mathcal{O}(2VNM\log(VM) + V^2MN)$ and $\mathcal{O}(VNM + M^2N)$, where K and t are the number of clusters and iteration, respectively. Due to $M \ll N$, the main complexity in this stage is $\mathcal{O}(M^2Nt + 2VNMt\log(VM))$. Therefore, the main computational complexity of our method is actually $\mathcal{O}(M^2Nt + VNMd)$, which is linear to N. The computational complexity of MVSC and our proposed method are summarized in Table II.

VI. Experiments

A. Evaluations on Synthetic Data

We have verified the effectiveness and superiority of our method through experiments.

We added visualizations results of the learned graph on synthetic noisy data in Fig. 3. Since features from different views are embedded with different data features, different clustering structures can be expressed by graphs constructed from different views. Thus, three different bipartite graphs are designed as the input of the single-view graphs, as shown in Fig. 3(a-c), where Fig. 3(a) and Fig. 3(b) have two catty-corner blocks showing different cluster distributions, Fig. 3(c) doesn't have any cluster structures because of the filled Gaussian noise. All these single-view graphics are uniformly added with random noise. The joint graph learned by our method (see Fig. 3(d)) has accurate 3 connected components, and the information in both Graph #1 and Graph #2 are absorbed into the final clusters.

We can conclude that: both of the informative structure within each view and the compatible structure between multiple views are explored by our model. Moreover, our model adaptively assigns the weighted parameters for different views by joint taking into account salient difference between views and connected components. Thus, the result will not be affected by the severely ill-views. In conclusion, our method can well handle the data with high amounts of noise.

B. Evaluations on Real Datasets

1) Experimental Setup: In this subsection, we will introduce datasets, comparisons and metrics adopted for evaluation.

2) Datasets: The following 5 multi-view datasets are selected to investigate the dominance of our proposed method:

- MSRC-v5 [54] contains 7 varieties of objects with 210 images. Same to [17], we choose 24-dimension (D) CM feature, 576-D HOG feature, 512-D GIST feature, 256-D LBP feature, 254-D CENT feature as 5 views.
- Handwritten4 [55] consists of 10 digits with 2,000 images created from UCI machine learning repository. 76-D FOU feature, 216-D FAC feature, 47-DZER feature and 6-D MOR feature are employed as 4 views.
- Mnist4 [56] involves 4 handwritten digits, from digit 0 to digit 3, with 4,000 images. We utilize 30-D ISO feature, 9-D LDA feature and 30-D NPE feature as 3 views.
- Caltech101-20 [57] is a subsets of Caltech101 datasets, and comprises 20 categories with 2,386 images. We employ 48-D GABOR feature, 40-D WM feature,

TABLE II

COMPUTATIONAL COMPLEXITY ANALYSIS, WHERE V, M, N AND K ARE THE NUMBER OF VIEWS, ANCHORS, DATA POINTS AND CLUSTERS, RESPECTIVELY. t IS THE ITERATION NUMBER. $M \ll N$, N/A MEANS NOT APPLICABLE

Method	Construction of Graphs	Solving $\mathbf{C}^{(v)}$	Solving $\mathcal{J}$	Solving $\mathbf{F}$	Total
MVSC [36]	$\mathcal{O}(VN^2d))$	N/A	N/A	$\mathcal{O}(N^3)$	$\mathcal{O}(N^3 + VN^2d)$
MSC-FG	$\mathcal{O}(VN^2d)$	$\mathcal{O}(VN^2(K+1) + VN^2\log(N))$	$\mathcal{O}(2VN^2\log(VN) + V^2N^2)$	$\mathcal{O}(N^3)$	$\mathcal{O}(VN^2d + (N^3 + VN^2\log(N))t)$
MSC-BG	$\mathcal{O}(VNMd + VNM\log(M))$	$\mathcal{O}(VNM(K+1) + VNM\log(M))$	$\mathcal{O}(2VNM\log(VM) + V^2MN)$	$\mathcal{O}(VNM + M^2N)$	$\mathcal{O}(VNMd + M^2Nt)$

TABLE III

THE CLUSTERING RESULTS ON MSRC-V5 AND HANDWRITTEN4 DATASETS

Dataset	**MSRC-v5**						
Metric	ACC	NMI	Purity	PER	REC	F-score	ARI
Co-reg [13]	0.635±0.007	0.578±0.006	0.659±0.006	0.511±0.008	0.535±0.007	0.522±0.007	0.425±0.030
SwMC [16]	0.776±0.000	0.774±0.000	0.805±0.000	0.687±0.000	0.831±0.000	0.752±0.000	0.708±0.000
MVGL [34]	0.690±0.000	0.663±0.000	0.733±0.000	0.466±0.000	0.715±0.000	0.564±0.000	0.476±0.000
MVSC [36]	0.794±0.075	0.672±0.058	0.756±0.071	0.585±0.091	0.779±0.035	0.664±0.062	0.600±0.079
SMSC [51]	0.766±0.000	0.717±0.000	0.804±0.000	0.672±0.000	0.718±0.000	0.694±0.000	0.643±0.000
AMGL [15]	0.751±0.078	0.704±0.044	0.789±0.056	0.621±0.090	0.744±0.026	0.674±0.063	0.615±0.079
MLAN [35]	0.681±0.000	0.630±0.000	0.733±0.000	0.494±0.000	0.718±0.000	0.694±0.000	0.643±0.000
SFMC [17]	0.810±0.000	0.721±0.000	0.810±0.000	0.657±0.000	0.782±0.000	0.714±0.000	0.663±0.000
RMSC [52]	0.762±0.040	0.663±0.026	0.769±0.030	0.640±0.030	0.660±0.034	0.650±0.031	0.592±0.036
CSMSC [53]	0.758±0.007	0.735±0.010	0.793±0.008	0.736±0.014	0.673±0.008	0.703±0.010	0.653±0.012
MSC-FG	0.843±0.000	0.761±0.000	0.843±0.000	0.738±0.000	0.756±0.000	0.747±0.000	0.706±0.000
MSC-BG	**0.981±0.000**	**0.960±0.000**	**0.981±0.000**	**0.961±0.000**	**0.963±0.000**	**0.962±0.000**	**0.956±0.000**
Dataset	**Handwritten4**						
Metric	ACC	NMI	Purity	PER	REC	F-score	ARI
Co-reg [13]	0.784±0.010	0.758±0.004	0.795±0.008	0.698±0.010	0.724±0.005	0.710±0.007	0.667±0.037
SwMC [16]	0.758±0.000	0.833±0.000	0.792±0.000	0.686±0.000	0.867±0.000	0.766±0.000	0.737±0.000
MVGL [34]	0.811±0.000	0.809±0.000	0.831±0.000	0.721±0.000	0.826±0.000	0.770±0.000	0.743±0.000
MVSC [36]	0.796±0.059	0.820±0.030	0.808±0.044	0.715±0.082	0.838±0.035	0.769±0.046	0.741±0.053
SMSC [51]	0.742±0.000	0.781±0.000	0.759±0.000	0.675±0.000	0.767±0.000	0.717±0.000	0.685±0.000
AMGL [15]	0.704±0.045	0.762±0.040	0.732±0.042	0.591±0.081	0.781±0.022	0.670±0.060	0.628±0.070
MLAN [35]	0.778±0.045	0.832±0.027	0.812±0.045	0.706±0.053	0.871±0.017	0.779±0.039	0.752±0.044
SFMC [17]	0.853±0.000	0.871±0.000	0.873±0.000	0.775±0.000	**0.910±0.000**	0.837±0.000	0.817±0.000
RMSC [52]	0.681±0.043	0.661±0.022	0.713±0.037	0.582±0.035	0.617±0.026	0.599±0.030	0.553±0.034
CSMSC [53]	0.806±0.001	0.793±0.001	0.867±0.001	0.778±0.001	0.743±0.001	0.760±0.001	0.733±0.001
MSC-FG	0.740±0.000	0.766±0.000	0.745±0.000	0.666±0.000	0.724±0.000	0.694±0.000	0.658±0.000
MSC-BG	**0.889±0.000**	**0.922±0.000**	**0.889±0.000**	**0.871±0.000**	0.893±0.000	**0.882±0.000**	**0.869±0.000**

254-D CENT feature, 1,984-D HOG feature, 512-D GIST feature and 928-D LBP feature as 6 views.

- Reuters [58] includes 6 categories with 18,757 documents, which are described in five languages including English, France, German, Italian and Spanish. We employ 21,531-D EN feature, 24,892-D FR feature, 24,892-D GR feature, 15,506-D IT feature and 11,547-D SP feature as 5 views.

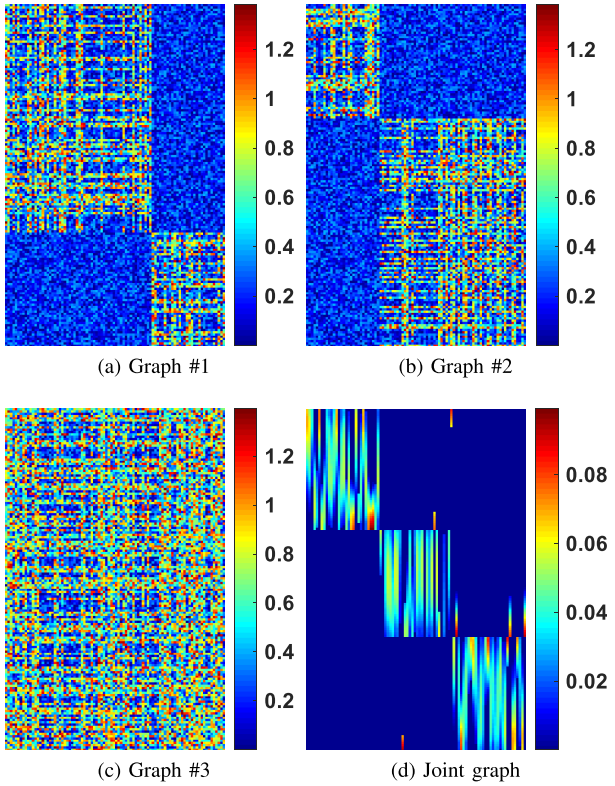
3) Comparisons and Metrics: We select 10 multi-view clustering algorithms as our comparison methods including Co-reg [13], SwMC [16], MVGL [34], MVSC [36], SMSC [51], AMGL [15], MLAN [35], SFMC [17], RMSC [52], CSMSC [53], and then evaluate performance by 7 indicators, including (1) Accuracy (ACC); (2) Normalized Mutual Information (NMI); (3) Purity; (4) Precision (PRE); (5) Recall (REC); (6) Fscore and (7) Adjusted Rand Index (ARI). The clustering performance is positively related to the value of all metrics. For more detailed definitions about each of the metrics, please refer to [41].

4) Comparisons With State-of-the-Art Methods: The clustering results of our model and comparison methods on the relevant datasets are shown in Tables III, IV. In order to verify the superiority of our algorithm on large-scale datasets, we test our algorithm on Reuters dataset, and the relevant results obtained are shown in Table V. All the algorithms run on a standard Windows 10 Server with an Intel (R) Xeon (R) Gold 6230 CPU and 128 GB RAM. In Tables III, IV, V, **MSC-FG** means the result of *the proposed method with full size graph*, **MSC-BG** means the result of *the proposed method with bipartite graph*. To ensure the accuracy of the experimental results, we repeat 20 times experiments for each comparison algorithm independently and then calculate the averages with corresponding standard deviations as the final results. From Tables III, IV, V, we have the following interesting observations:

- MSC-BG outperforms our chosen comparison algorithm in the majority of cases, demonstrating its superior performance in multi-view clustering.
- Our method without anchor selection (MSC-FG) is inferior to our final method (MSC-BG) and MSC-FG cannot handle large-scale datasets. The reason may be that anchor selection can reduce memory consumption and time consumption significantly.
- The performance of Co-reg is worse than that of the other multi-view methods. One of the reasons probably is that Co-reg neglects the significant difference among different views for clustering, and another reason may be that the performance heavily depends on the predefined graphs manually. However, since the data in real-world applications are very complex, it is difficult to select the appropriate graph.
- Compared with MVSC, MSC-BG has better clustering performance. For example, our method obtains improvement around 18.7%, 28.8%, 22.5%, 37.6%, 18.4%,

TABLE IV
THE CLUSTERING RESULTS ON MNIST4 AND CALTECH101-20 DATASETS

Dataset	Mnist4						
Metric	ACC	NMI	Purity	PER	REC	F-score	ARI
Co-reg [13]	0.785±0.003	0.602±0.001	0.786±0.002	0.670±0.002	0.696±0.002	0.682±0.001	0.575±0.002
SwMC [16]	0.914±0.000	0.799±0.000	0.912±0.000	0.844±0.000	0.852±0.000	0.848±0.000	0.799±0.000
MVGL [34]	0.912±0.000	0.785±0.000	0.910±0.000	0.795±0.000	0.804±0.000	0.800±0.000	0.733±0.000
MVSC [36]	0.733±0.115	0.651±0.069	0.780±0.070	0.650±0.092	0.773±0.041	0.704±0.066	0.592±0.096
SMSC [51]	0.913±0.000	0.789±0.000	0.913±0.000	0.843±0.000	0.850±0.000	0.846±0.000	0.795±0.000
AMGL [15]	0.910±0.000	0.785±0.000	0.910±0.000	0.836±0.000	0.843±0.000	0.840±0.000	0.786±0.000
MLAN [35]	0.744±0.001	0.659±0.001	0.744±0.000	0.643±0.001	**0.921±0.001**	0.757±0.001	0.656±0.001
SFMC [17]	0.917±0.000	0.801±0.000	0.917±0.000	0.846±0.000	0.855±0.000	0.852±0.000	0.802±0.000
RMSC [52]	0.705±0.000	0.486±0.000	0.705±0.000	0.590±0.000	0.606±0.000	0.598±0.000	0.462±0.001
CSMSC [53]	0.643±0.000	0.645±0.010	0.832±0.008	0.776±0.014	0.612±0.008	0.684±0.010	0.562±0.012
MSC-FG	0.919±0.000	0.802±0.000	0.919±0.000	0.852±0.000	0.859±0.000	0.855±0.000	0.807±0.000
MSC-BG	**0.938±0.000**	**0.861±0.000**	**0.938±0.000**	**0.884±0.000**	0.891±0.000	**0.888±0.000**	**0.850±0.000**
Dataset	**Caltech101-20**						
Metric	ACC	NMI	Purity	PER	REC	F-score	ARI
Co-reg [13]	0.412±0.006	0.587±0.003	0.754±0.004	0.712±0.008	0.243±0.004	0.363±0.006	0.295±0.025
SwMC [16]	0.599±0.000	0.493±0.000	0.700±0.000	0.509±0.000	0.625±0.000	0.431±0.000	0.265±0.000
MVGL [34]	0.600±0.000	0.474±0.000	0.696±0.000	0.325±0.000	0.653±0.000	0.440±0.000	0.282±0.000
MVSC [36]	0.595±0.000	0.613±0.000	0.717±0.000	0.542±0.000	0.546±0.000	0.541±0.000	0.451±0.000
SMSC [51]	0.582±0.000	0.590±0.000	0.748±0.000	0.701±0.000	0.473±0.000	0.565±0.000	0.485±0.000
AMGL [15]	0.557±0.047	0.552±0.061	0.677±0.058	0.480±0.093	0.539±0.015	0.503±0.054	0.397±0.080
MLAN [35]	0.526±0.007	0.474±0.003	0.666±0.000	0.279±0.003	0.559±0.020	0.372±0.007	0.198±0.007
SFMC [17]	0.642±0.000	0.595±0.000	0.748±0.000	0.586±0.000	**0.677±0.000**	0.628±0.000	0.461±0.000
RMSC [52]	0.385±0.024	0.512±0.012	0.742±0.013	0.692±0.038	0.231±0.019	0.346±0.026	0.288±0.027
CSMSC [53]	0.474±0.037	0.648±0.011	0.563±0.031	0.290±0.034	0.730±0.037	0.415±0.039	0.356±0.040
MSC-FG	0.565±0.000	0.595±0.000	0.716±0.000	0.559±0.000	0.476±0.000	0.514±0.000	0.428±0.000
MSC-BG	**0.667±0.000**	**0.727±0.000**	**0.794±0.000**	**0.772±0.000**	0.543±0.000	**0.637±0.000**	**0.581±0.000**



(a) Graph #1 (b) Graph #2

(c) Graph #3 (d) Joint graph

Fig. 3. Experiments on synthetic multi-view data.

29.8%, and 35.6% in terms of ACC, NMI, Purity, PER, REC, F-score, and ARI on MSRC-v5 dataset. The reason may be that our method makes good use of the complementary information and spatial structure information between different views, enabling better clustering of the data.

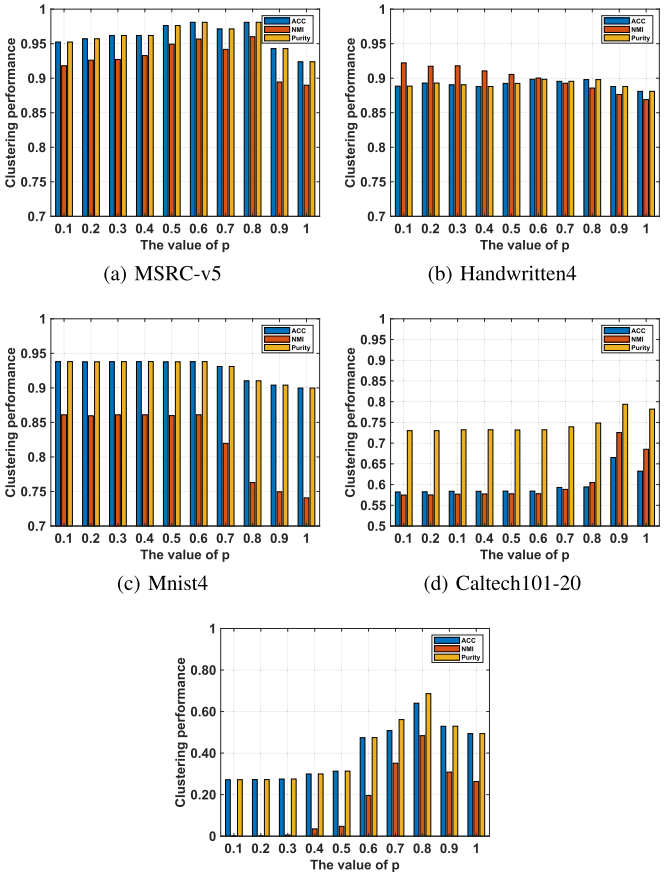
TABLE V
THE CLUSTERING RESULTS ON REUTERS DATASET, "OM" IS "OUT-OF-MEMORY", RUNNING TIME(IN SECONDS)

Dataset	Reuters			
Metric	ACC	NMI	Purity	Running Time
Co-reg [13]	0.563	0.326	0.552	2672.33
SwMC [16]	OM	OM	OM	OM
MVGL [34]	OM	OM	OM	OM
MVSC [36]	0.596	0.347	0.574	333.68
SMSC [51]	OM	OM	OM	OM
AMGL [15]	OM	OM	OM	OM
MLAN [35]	OM	OM	OM	OM
SFMC [17]	0.602	0.354	0.552	310.00
RMSC [52]	OM	OM	OM	OM
CSMSC [53]	OM	OM	OM	OM
LMVSC [18]	0.589	0.335	0.615	**160.88**
MSC-FG	OM	OM	OM	OM
MSC-BG	**0.640**	**0.484**	**0.686**	462.33

- MSC-BG is more stable than MVSC. The reason may be that our proposed method can directly obtain the clustering results based on the connected components without any post-processing, while MVSC cannot. MVSC still need to employ K-means to compute clustering labels.
- MSC-BG obtains the best clustering result with taking a little more time cost. Comparing with MSC-FG, which cannot handle large-scale datasets, the proposed fast method scales linearly with the data size. Thus, the proposed method can well handle large-scale datasets and is time-economical.

C. Further Evaluation

1) Effect of Parameter p: We analyzed the impact of p in the tensor Schatten p-norm on the clustering results on MSRC-v5, Handwritten4, Mnist4, Caltech101-20 and Reuters

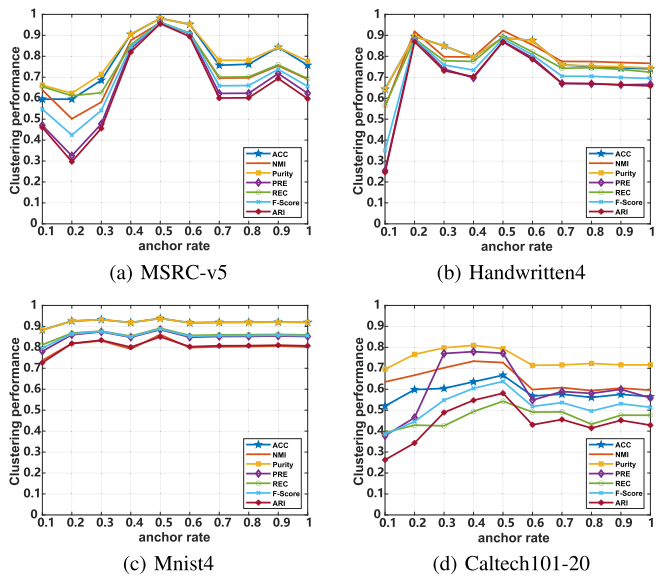


(a) MSRC-v5 (b) Handwritten4

(c) Mnist4 (d) Caltech101-20

(e) Reuters

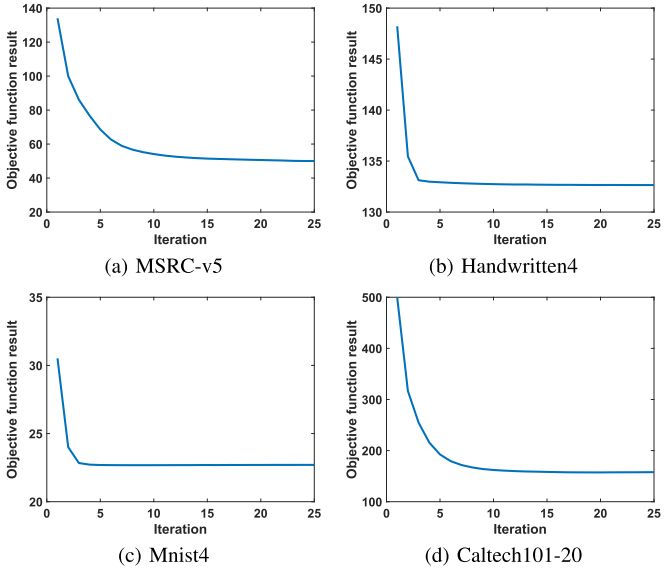
Fig. 4. ACC, NMI and purity scores as a function of p on MSRC-v5, Handwritten4, Mnist4, Caltech101-20 and reuters datasets.



(a) MSRC-v5 (b) Handwritten4

(c) Mnist4 (d) Caltech101-20

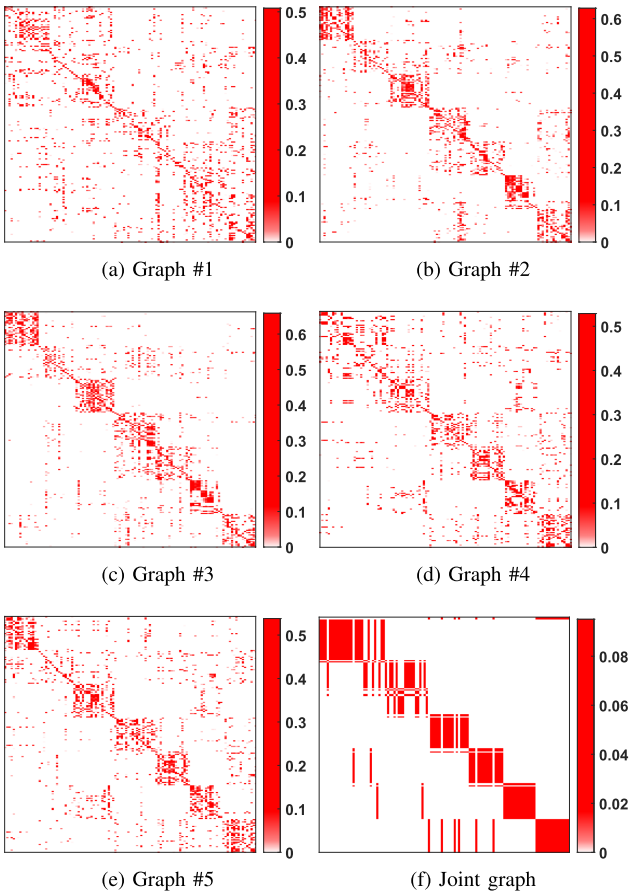
Fig. 5. The clustering results as a function of anchor rate on MSRC-v5, Handwritten4, Mnist4 and Caltech101-20 datasets.



(a) MSRC-v5 (b) Handwritten4

(c) Mnist4 (d) Caltech101-20

Fig. 6. Convergence experiment on MSRC-v5, Handwritten4, Mnist4 and Caltech101-20 datasets.



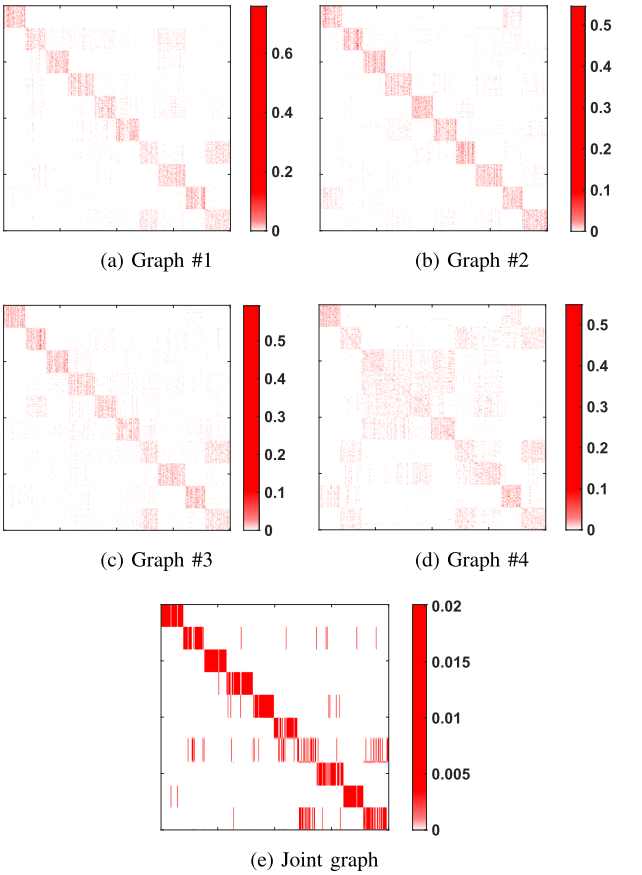
(a) Graph #1 (b) Graph #2

(c) Graph #3 (d) Graph #4

(e) Graph #5 (f) Joint graph

Fig. 7. The graphs visualizations on MSRC-v5 dataset.

datasets. In particular, we take the value of p to increase sequentially by 0.1 from 0.1 until $p = 1$, and give the ACC, NMI and Purity for each value of p. Then the relationship between the parameter p and the clustering results is plotted through all the experimental results, which is show in Fig. 4. It can be found that the results under different p are distinguishing mostly and when $p = 0.6$, $p = 0.1$, $p = 0.6$, $p = 0.9$ and $p = 0.8$, we obtain the best clustering results on MSRC-v5, Handwritten4, Mnist4, Caltech101-20 and Reuters dataset, respectively. This is probably mainly due to the fact that tensor Schatten p-norm makes good use of the

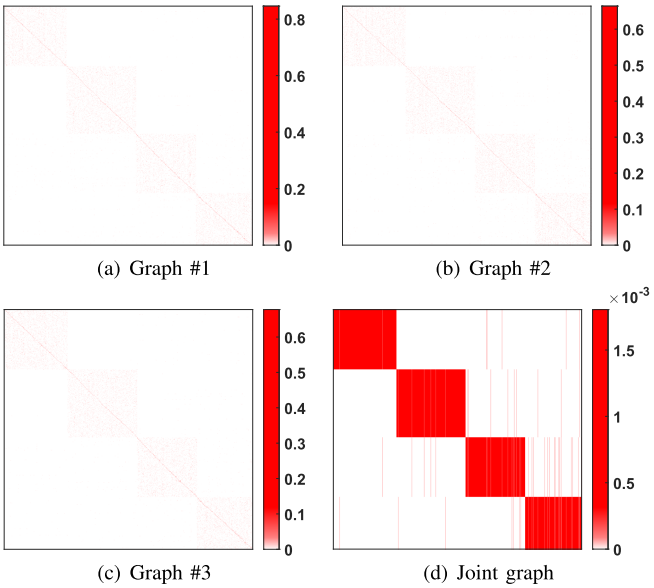


(a) Graph #1 (b) Graph #2

(c) Graph #3 (d) Graph #4

(e) Joint graph

Fig. 8. The graphs visualizations on Handwritten4 dataset.



(a) Graph #1 (b) Graph #2

(c) Graph #3 (d) Joint graph

Fig. 9. The graphs visualizations on Mnist4 dataset.

complementary information and spatial structure embedded in graphs of different views, enabling better clustering of the data.

2) Effect of the Anchor Rate: We analyzed the effect of the anchor rate(the proportion of anchors to all samples) on the clustering results on MSRC-v5, Handwritten4, Mnist4 and Caltech101-20 datasets. Therefore, we take the value of anchor rate to increase sequentially by 0.1 from 0.1 to 1, and give results by conducting experiments with different anchor rates.

TABLE VI

THE RESULTS ON DIFFERENT SCALE REUTERS DATASET, "OM" IS "OUT-OF-MEMORY", RUNNING TIME(IN SECONDS)

Dataset scale	Metric	ACC	NMI	Purity	Running Time
3000	MSC-FG	0.688	0.496	0.688	278.66
4000	MSC-FG	0.511	0.627	0.516	337.08
5000	MSC-FG	0.408	0.603	0.413	633.50
6000	MSC-FG	0.462	0.606	0.462	1014.13
7000	MSC-FG	0.538	0.614	0.538	1585.86
8000	MSC-FG	0.596	0.622	0.596	2371.08
9000	MSC-FG	0.538	0.409	0.539	3414.15
10000	MSC-FG	0.485	0462	0.486	3799.12
11000	MSC-FG	0.441	0.422	0.442	5092.37
12000	MSC-FG	0.405	0.321	0.405	5446.97
13000+	MSC-FG	OM	OM	OM	OM
18757	MSC-BG	0.640	0.484	0.686	462.33

The experimental results are shown in Fig. 5. It is easy to find that there is huge difference in the clustering results with different anchor rate. When the anchor rate is set to 0.5, MSC-BG obtains the best performance on MSRC-v5, Handwritten4 and Mnist4 datasets. In addition to this we can also find that the relationship curves in Fig. 5 are not monotonously increasing, this shows that we do not need to set a larger anchor rate to get better clustering results. In summary, we fix the anchor rate at 0.5 uniformly for the experiments on selected four datasets.

3) Convergence Experiment Analysis: We obtain our clustering results by minimizing Eq. (24). Therefore, we calculated the results of Eq. (24) for different iteration as a way to analyze the convergence of the model. Taking MSRC-v5, Handwritten4, Mnist4 and Caltech101-20 datasets as example, we draw the relationship between Eq. (24) value and Iteration in Fig. 6. According to Fig.6, we can see that the proposed method obtains relatively stable objective function value within a few iterations. These experimental results show that MSC-BG can converge quickly and satisfy our previous theoretical analysis.

4) Graph Visualization Analysis: To show the clustering performance of MSC-BG more clearly, we draw the input graphs and the learned view-consensus graph on MSRC-v5, Handwritten4 and Mnist4 datasets in Fig 7, 8, 9, respectively. The input graphs corresponding to all views on three datasets are shown in Fig. 7(a-e), 8(a-d), 9(a-c), and Fig. 7(f), 8(e), 9(d) are the view-consensus graph corresponding to MSRC-v5, Handwritten4 and Mnist4 datasets, respectively. We can observe the connected components in both the learned view-consensus graph and the input graphs of all views, but the connected components in the input graphs are much less clear than them in the learned view-consensus graph. And there are exact K-connected components in all learned view-consensus graphs ($K = 7, 10, 4$ for MSRC-v5, Handwritten4 and Mnist4 datasets, respectively). It indicates that our method can use the hidden information embedded between different views to characterize the cluster structure. The results of the this experiment demonstrate that MSC-BG helps to ensure the learned view-consensus graph's rank close to the target rank.

5) Dataset Scale Analysis: In order to show the effectiveness of the base framework, we performed clustering using MSC-FG on Reuters dataset of different size. In order to show the effectiveness of the base framework, we performed clustering using MSC-FG on Reuters dataset of different size.

To be specifical, we divide the Reuters dataset, taking values from 3,000 in order 1,000 up to 18,000 (Reuters has a total of 18,757 data), and use MSC-FG in turn for clustering. The clustering results on different scale Reuters dataset are shown in Table VI. In Table VI, we can see that when the data scale reaches 13,000, MSC-FG cannot handle this dataset, but MSC-BG can handle 18,757 data, which is enough to see the superiority of MSC-BG to handle large-scale datasets.

VII. Conclusion

We propose an effective multi-view spectral clustering model. Our method uses the method of minimizing tensor Schatten p-norm to learn the common graph, which well characterizes the spatial structure and complementary information embedded in views. We also propose an efficient algorithm to solve the proposed model in an alternating way. Our method learns a good graph which has K-connected components by employing the connectivity constraint. Moreover, our method learns the $N \times M (M \ll N)$ graph instead of the $N \times N$ graph, where N and M are the number of data points and anchors, so our method is time-economical. Extensive experimental results indicate that our method has good performance on real-world datasets.

References

[1] 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 Trans. Image Process.*, vol. 24, no. 11, pp. 3939–3949, Nov. 2015.

[2] J. Wu, Z. Lin, and H. Zha, "Essential tensor learning for multi-view spectral clustering," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5910–5922, Dec. 2019.

[3] C. Zhang *et al.*, "Generalized latent multi-view subspace clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 86–99, Jan. 2020.

[4] W. Xia, X. Zhang, Q. Gao, X. Shu, J. Han, and X. Gao, "Multiview subspace clustering by an enhanced tensor nuclear norm," *IEEE Trans. Cybern.*, early access, Feb. 26, 2021, doi: 10.1109/TCYB.2021.3052352.

[5] Y. Chen, S. Wang, C. Peng, Z. Hua, and Y. Zhou, "Generalized nonconvex low-rank tensor approximation for multi-view subspace clustering," *IEEE Trans. Image Process.*, vol. 30, pp. 4022–4035, 2021.

[6] Y. Yun, W. Xia, Y. Zhang, Q. Gao, and X. Gao, "Self-representation and class-specificity distribution based multi-view clustering," *Neurocomputing*, vol. 437, pp. 9–20, May 2021.

[7] X. Sun, Y. Wang, M. Yang, and X. Zhang, "Robust multiview subspace clustering of images via tighter rank approximation," *IEEE Access*, vol. 9, pp. 81173–81188, 2021.

[8] C. Alzate and J. A. K. Suykens, "Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 335–347, Feb. 2010.

[9] C. Lu, S. Yan, and Z. Lin, "Convex sparse spectral clustering: Single-view to multi-view," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2833–2843, Jun. 2016.

[10] H. Xu, X. Zhang, W. Xia, Q. Gao, and X. Gao, "Low-rank tensor constrained co-regularized multi-view spectral clustering," *Neural Netw.*, vol. 132, pp. 245–252, Dec. 2020.

[11] K. K. Sharma and A. Seal, "Multi-view spectral clustering for uncertain objects," *Inf. Sci.*, vol. 547, pp. 723–745, Feb. 2021.

[12] Z. Huang, J. T. Zhou, H. Zhu, C. Zhang, J. Lv, and X. Peng, "Deep spectral representation learning from multi-view data," *IEEE Trans. Image Process.*, vol. 30, pp. 5352–5362, 2021.

[13] A. Kumar and P. Rai, "Co-regularized multi-view spectral clustering," in *Proc. NeurIPS*, 2011, pp. 1413–1421.

[14] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 1977–1984.

[15] F. Nie *et al.*, "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification," in *Proc. IJCAI*, 2016, pp. 1881–1887.

[16] F. Nie, J. Li, and X. Li, "Self-weighted multiview clustering with multiple graphs," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2564–2570.

[17] X. Li, H. Zhang, R. Wang, and F. Nie, "Multiview clustering: A scalable and parameter-free bipartite graph fusion method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 330–344, Jan. 2022, doi: 10.1109/TPAMI.2020.3011148.

[18] Z. Kang, W. Zhou, Z. Zhao, J. Shao, M. Han, and Z. Xu, "Large-scale multi-view subspace clustering in linear time," in *Proc. 34th AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 4412–4419.

[19] B. Yang, X. Zhang, F. Nie, F. Wang, W. Yu, and R. Wang, "Fast multi-view clustering via nonnegative and orthogonal factorization," *IEEE Trans. Image Process.*, vol. 30, pp. 2575–2586, 2020.

[20] X. Chen, Y. Ye, Q. Wu, and F. Nie, "Fast manifold ranking with local bipartite graph," *IEEE Trans. Image Process.*, vol. 30, pp. 6744–6756, 2021.

[21] Q. Qiang, B. Zhang, F. Wang, and F. Nie, "Fast multi-view discrete clustering with anchor graphs," in *Proc. AAAI*, 2021, pp. 9360–9367.

[22] D. Cai and X. Chen, "Large scale spectral clustering via landmark-based sparse representation," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1669–1680, Aug. 2015.

[23] Q. Gao, P. Zhang, W. Xia, D. Xie, X. Gao, and D. Tao, "Enhanced tensor RPCA and its application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2133–2140, Jun. 2021.

[24] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, "Tensor robust principal component analysis with a new tensor nuclear norm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 925–938, Jan. 2020.

[25] H. Kong, X. Xie, and Z. Lin, "t-Schatten-p norm for low-rank tensor recovery," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 6, pp. 1405–1419, Dec. 2018.

[26] Y. Zhao, Y. Yun, X. Zhang, Q. Li, and Q. Gao, "Multi-view spectral clustering with adaptive graph learning and tensor schatten p-norm," *Neurocomputing*, vol. 468, pp. 257–264, Jan. 2022.

[27] W. Xia, Q. Gao, Q. Wang, and X. Gao, "Tensor completion-based incomplete multiview clustering," *IEEE Trans. Cybern.*, early access, Jan. 25, 2022, doi: 10.1109/TCYB.2021.3140068.

[28] W. Liu, J. He, and S. Chang, "Large graph construction for scalable semi-supervised learning," in *Proc. ICML*, 2010, pp. 679–686.

[29] A. Ortega, P. Frossard, J. Kovačevič, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018.

[30] Y. Tanaka, Y. C. Eldar, A. Ortega, and G. Cheung, "Sampling signals on graphs: From theory to applications," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 14–30, Nov. 2020.

[31] G. Cheung, E. Magli, Y. Tanaka, and M. K. Ng, "Graph spectral image processing," *Proc. IEEE*, vol. 106, no. 5, pp. 907–930, May 2018.

[32] J. H. Giraldo, S. Javed, N. Werghi, and T. Bouwmans, "Graph CNN for moving object detection in complex environments from unseen videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 225–233.

[33] A. Mondal, S. R, J. H. Giraldo, T. Bouwmans, and A. S. Chowdhury, "Moving object detection for event-based vision using graph spectral clustering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 876–884.

[34] K. Zhan, C. Zhang, J. Guan, and J. Wang, "Graph learning for multi-view clustering," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 2887–2895, Oct. 2018.

[35] F. Nie, G. Cai, J. Li, and X. Li, "Auto-weighted multi-view learning for image clustering and semi-supervised classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1501–1511, Sep. 2017.

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

[37] Z. Lin, Z. Kang, L. Zhang, and L. Tian, "Multi-view attributed graph clustering," *IEEE Trans. Knowl. Data Eng.*, early access, Aug. 6, 2021, doi: 10.1109/TKDE.2021.3101227.

[38] Z. Kang, Z. Lin, X. Zhu, and W. Xu, "Structured graph learning for scalable subspace clustering: From single view to multiview," *IEEE Trans. Cybern.*, early access, Mar. 17, 2021, doi: 10.1109/TCYB.2021.3061660.
[39] J. Lv, Z. Kang, X. Lu, and Z. Xu, "Pseudo-supervised deep subspace clustering," *IEEE Trans. Image Process.*, vol. 30, pp. 5252–5263, 2021.
[40] Y. Xie, D. Tao, W. Zhang, L. Zhang, and Y. Qu, "On unifying multi-view self-representations for clustering by tensor multi-rank minimization," *Int. J. Comput. Vis.*, vol. 126, no. 11, pp. 1157–1179, 2018.
[41] Y. Xie, W. Zhang, Y. Qu, L. Dai, and D. Tao, "Hyper-Laplacian regularized multilinear multiview self-representations for clustering and semi-supervised learning," *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 572–586, Feb. 2020.
[42] M. E. Kilmer, L. Horesh, H. Avron, and E. Newman, "Tensor-tensor algebra for optimal representation and compression of multiway data," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 28, Jul. 2021, Art. no. e2015851118.
[43] M. E. Kilmer and C. D. Martin, "Factorization strategies for third-order tensors," *Linear Algebra Appl.*, vol. 435, no. 3, pp. 641–658, 2011.
[44] Q. Gao, W. Xia, Z. Wan, D. Xie, and P. Zhang, "Tensor-SVD based graph learning for multi-view subspace clustering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 4, 2020, pp. 3930–3937.
[45] M. Signoretto, L. D. Lathauwer, and J. A. K. Suykens, "Nuclear norms for tensors and their use for convex multilinear estimation," ESAT-SISTA, KU Leuven, Leuven, Belgium, Tech. Rep. 10-186, 2010.
[46] F. Nie, H. Wang, X. Cai, H. Huang, and C. Ding, "Robust matrix completion via joint schatten p-norm and lp-norm minimization," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 566–574.
[47] F. R. Chung and F. C. Graham, *Spectral Graph Theory*, no. 92. Providence, RI, USA: American Mathematical Society, 1997.
[48] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. NeurIPS*, 2011, pp. 612–620.
[49] F. Nie, X. Wang, M. I. Jordan, and H. Huang, "The constrained Laplacian rank algorithm for graph-based clustering," in *Proc. AAAI*, 2016, pp. 1969–1976.
[50] A. S. Lewis and H. S. Sendov, "Nonsmooth analysis of singular values—Part I: Theory," *Set-Valued Anal.*, vol. 13, no. 3, pp. 213–241, Sep. 2005.
[51] Z. Hu, F. Nie, R. Wang, and X. Li, "Multi-view spectral clustering via integrating nonnegative embedding and spectral embedding," *Inf. Fusion*, vol. 55, pp. 251–259, Mar. 2020.
[52] K. R. Radhika, C. N. Pushpa, J. Thriveni, and K. R. Venugopal, "RMSC: Robust modeling of subspace clustering for high dimensional data," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2017, pp. 1535–1539.
[53] S. Luo, C. Zhang, W. Zhang, and X. Cao, "Consistent and specific multi-view subspace clustering," in *Proc. AAAI*, 2018, pp. 3730–3737.
[54] J. Winn and N. Jojic, "LOCUS: Learning object classes with unsupervised segmentation," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 756–763.
[55] D. Dua and C. Graff, "UCI machine learning repository," School Inf. Comput. Sci., Univ. California, Irvine, Irvine, CA, USA, 2017. [Online]. Available: http://archive.ics.uci.edu/ml
[56] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.
[57] 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," *Comput. Vis. Image Understand.*, vol. 106, no. 1, pp. 59–70, Jan. 2007.
[58] C. Apté, F. Damerau, and S. M. Weiss, "Automated learning of decision rules for text categorization," *ACM Trans. Inf. Syst.*, vol. 12, no. 3, pp. 233–251, Jul. 1994.

Haizhou Yang received the B.Eng. degree in communication engineering from Xidian University, Xi'an, China, in 2021, where he is currently pursuing the master's degree in communication and information system. His research interests include pattern recognition and deep learning.

Quanxue Gao received the B.Eng. degree from Xi'an Highway University, Xi'an, China, in 1998, the M.S. degree from the Gansu University of Technology, Lanzhou, China, in 2001, and the Ph.D. degree from Northwestern Polytechnical University, Xi'an, in 2005. He was an Associate Researcher with the Biometrics Center, The Hong Kong Polytechnic University, Hong Kong, from 2006 to 2007. From 2015 to 2016, he was a Visiting Scholar with the Department of Computer Science, The University of Texas at Arlington, Arlington, USA. He is currently a Professor with the School of Telecommunications Engineering, Xidian University, and a Key Member of the State Key Laboratory of Integrated Services Networks. He has authored around 80 technical articles in refereed journals and proceedings, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, CVPR, AAAI, and IJCAI. His current research interests include pattern recognition and machine learning.

Wei Xia (Graduate Student Member, IEEE) received the B.Eng. degree in communication engineering from the Lanzhou University of Technology, Lanzhou, China, in 2018. He is currently pursuing the Ph.D. degree in communication and information system with Xidian University, Xi'an, China. His research interests include multi-modal learning, representation learning, and graph neural networks.

Ming Yang received the B.S. degree in math from Jilin University, Changchun, China, in 2007, and the Ph.D. degree in math from Texas A&M University, College Station, USA, in 2012. Currently, he is an Assistant Professor of data science with the Departments of Mathematics and Computer and Information Science, Westfield State University, MA, USA. He has published several research articles in top-tier journals, including *SIAM Journal on Imaging Sciences*, *Pattern Recognition* (Elsevier), IEEE SIGNAL PROCESSING LETTERS, *Neural Networks* (Elsevier), *Journal of Dynamics and Differential Equations*, *Linear and Multilinear Algebra*, *PLOS One*, and *Applied Sciences-Basel*. His research interests include machine learning, image processing, and tensor decomposition.

Xinbo Gao (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in electronic engineering, signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Postdoctoral Research Fellow at the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. Since 2001, he has been at the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of the Ministry of Education, China; a Professor of pattern recognition and intelligent system with Xidian University; and a Professor of computer science and technology with Chongqing University of Posts and Telecommunications. He has published six books and around 300 technical articles in refereed journals and proceedings. His current research interests include image processing, computer vision, multimedia analysis, machine learning, and pattern recognition. He is a fellow of the Institute of Engineering and Technology and the Chinese Institute of Electronics. He served as the general chair/co-chair, the program committee chair/co-chair, and a PC member for around 30 major international conferences. He is on the editorial boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier).