V^{3}H: View Variation and View Heredity for Incomplete Multiview Clustering

Xiang Fang , Yuchong Hu , *Member, IEEE*, Pan Zhou , *Senior Member, IEEE*, and Dapeng Oliver Wu , *Fellow, IEEE*

Abstract—Real data often appear in the form of multiple incomplete views. Incomplete multiview clustering is an effective method to integrate these incomplete views. Previous methods only learn the consistent information between different views and ignore the unique information of each view, which limits their clustering performance and generalizations. To overcome this limitation, we propose a novel View Variation and View Heredity approach (V^{3}H). Inspired by the variation and the heredity in genetics, V^{3}H first decomposes each subspace into a variation matrix for the corresponding view and a heredity matrix for all the views to represent the unique information and the consistent information respectively. Then, by aligning different views based on their cluster indicator matrices, V^{3}H integrates the unique information from different views to improve the clustering performance. Finally, with the help of the adjustable low-rank representation based on the heredity matrix, V^{3}H recovers the underlying true data structure to reduce the influence of the large incompleteness. More importantly, V^{3}H presents possibly the first work to introduce genetics to clustering algorithms for learning simultaneously the consistent information and the unique information from incomplete multiview data. Extensive experimental results on fifteen benchmark datasets validate its superiority over other state-of-the-arts.

Impact Statement—Incomplete multiview clustering is a popular technology to cluster incomplete datasets from multiple sources. The technology is becoming more significant due to the absence of the expensive requirement of labeling these datasets. However, previous algorithms cannot fully learn the information of each view. Inspired by variation and heredity in genetics, our proposed algorithm V^{3}H fully learns the information of each view. Compared with the state-of-the-art algorithms, V^{3}H improves clustering performance by more than 20% in representative cases. With the large improvement on multiple datasets, V^{3}H has wide potential applications including the analysis of pandemic, financial and election datasets. The DOI of our codes is 10.24 433/CO.2 119 636.v1.

Index Terms—Incomplete multiview clustering, view heredity, view variation.

I. INTRODUCTION

IN MOST real-world applications, the collected data always appear in multiple views or come from different sources [1]–[4], which are called *multiview data* [5]. As an illustration, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the strain of coronavirus that causes coronavirus disease 2019 (COVID-19) [6]–[8]. When we analyze the structural proteins of SARS-CoV-2 to study vaccines, the S (spike), E (envelope), M (membrane), and N (nucleocapsid) proteins can serve as four views [9]–[12]. Also, the financial data often come from multiple sources, and each data source corresponds to a view. Multiview clustering provides a natural way to integrate these views [13]–[16].

Recently, many multiview clustering algorithms have been proposed [17]–[19]. Most of them assume that each view is complete. But real-world data often suffer from incompleteness [20]–[22]. For example, when detecting COVID-19, the blood test, the temperature measurement, and the neuroimage can be regarded as three views of the detection. But when the disease first broke out, most individuals only perform one or two tests due to the lack of detection conditions. As such, this incompleteness may lead to the lack of columns or rows in the view matrix, which fails previous algorithms.

To cluster incomplete multiview data, some efforts have been made. PVC [23] aligns the same samples in different views by constructing a latent subspace. To solve the incomplete multi-modality clustering problem, IMG [20] transforms the collected incomplete multiview data to a complete representation in a latent space. To extend PVC to more than two views, MIC [24] extends MultiNMF [25] based on weighted nonnegative matrix factorization (NMF) [26] and $L_{2,1}$-norm regularization. To decrease the impact of a large missing rate, DAIMC [27] extends MIC by combining semi-nonnegative matrix factorization (semi-NMF) [28] and $L_{2,1}$-norm regularization regression. To solve the multiview co-clustering with incomplete data problem, [5] integrates complex patterns of incomplete multiview data. To explore the local structure, UEAF [29] learns a consensus representation for all views.

However, these incomplete multiview clustering methods still have three main drawbacks. First, these methods only consider the consistent information between all the views and ignore the unique information of each view. For example, DAIMC performs clustering by aligning the consistent information. When clustering the data with little alignment information, DAIMC

cannot achieve satisfactory performance due to ignoring the unique information. The unique information from each view can help us better analyze the geometry of the corresponding view. Second, these methods are difficult to learn nonlinear information between samples. Most of these methods are based on NMF. NMF is a linear operation, which only extracts the linear structure information among samples from the data. When processing some datasets with nonlinear structures, these methods may ignore much important nonlinear information among samples, which limits the application of these methods. Third, these methods do not perform well in some datasets with relatively large missing rates. Based on these samples presented in all views, these algorithms learn the structure information of datasets for clustering. As the missing rate increases, the number of presented samples will decrease significantly. Thus, these algorithms cannot learn enough information and will obtain poor clustering performance. If we directly use these methods to process important data, it is difficult to learn accurate data information because of the above drawbacks. As an illustration, if we use these methods to directly analyze the data of SARS-CoV-2, these methods are difficult to obtain satisfactory results, which may lead to slow research progress on COVID-19. Therefore, incomplete multiview clustering still contains significant issues.

To address these issues, we propose a novel **V**iew **V**ariation and **V**iew **H**eredity approach (V^3H). By introducing biological heredity and biological variation, the theory of genetics can effectively analyze the consistent trait information and the unique trait information in the biological world [30]–[32]. Inspired by the theory, V^3H first learns a subspace from each view and decomposes each subspace into a heredity matrix shared by all the views and a variation matrix of the corresponding view. The shared heredity matrix can extract the consistent information between all the views, while each variation matrix can learn the unique information of the corresponding view. Based on each variation matrix, V^3H constructs a graph Laplacian to obtain a corresponding cluster indicator matrix. Then, V^3H aligns different views by minimizing the disagreement between each cluster indicator matrix and the consensus cluster indicator matrix. To measure the disagreement, V^3H introduces the linear kernel into the Laplacian for spectral clustering. Finally, instead of learning the low-rank representation of all the subspaces, V^3H designs an adjustable low-rank representation model via the η-norm of the variation matrix and the τ-norm of error matrices. V^3H's contributions are mainly summarized as follows:

- To our best knowledge, V^3H is a **pioneering** work to introduce genetics into the clustering algorithm, which will promote the intersection between the clustering algorithm and other disciplines. Moreover, it is also the **first** attempt to learn both the consistent information and the unique information of incomplete views simultaneously based on the subspace decomposition.
- By minimizing the disagreement between each cluster indicator matrix and the consensus, V^3H can learn a satisfactory cluster indicator matrix for each view and integrate the unique information of these views, which improves its clustering performance. By introducing the linear kernel into the Laplacian, V^3H learns the nonlinear structure in

the dataset, which guarantees its applicability in datasets with nonlinear structure.
- Based on the adjustable low-rank representation model, V^3H can recover the underlying true data structure as needed, which helps us cluster the multiview data with a relatively large missing rate.
- Experimental results on fifteen benchmark datasets demonstrate the superiority of V^3H over other state-of-the-arts. Impressively, in terms of three evaluation metrics, V^3H improves the clustering performance by more than 20% in representative cases.

The rest of the paper is organized as follows. Section II presents some related works. Section III describes the notation and the background. Section IV first motivates V^3H's main idea, then proposes our V^3H approach, and finally solves it efficiently. Section V evaluates V^3H's performance. Section VI concludes the paper.

II. RELATED WORKS

The most relevant work of this paper is incomplete multiview clustering, and we present some related works in this section. Recently, many incomplete multiview clustering methods have been proposed [20], [23], [24], [27], [29]. Based on the number of views clustered, we can divide these methods into the following two categories.

(i) Incomplete two-view clustering (e.g., PVC and IMG). Incomplete two-view clustering methods can only cluster incomplete data with two views. PVC [23] learns the common and private latent spaces based on NMF [26], [33] and L_1-norm regularization. But PVC simply projects samples from each view into a common subspace and overlooks the global information among the two views. To obtain better clustering performance on multi-modal visual datasets, IMG [20] extends PVC and removes the nonnegative constraint to simplify optimization. But both PVC and IMG can only solve the incomplete two-view clustering problem, which limits their application to incomplete data with more than two views.

(ii) Incomplete multiview clustering (e.g., MIC, DAIMC and UEAF). Incomplete multiview clustering methods can cluster incomplete data with more than two views. As the first method for incomplete multiview clustering, MIC [24] first fills the missing samples in each incomplete view with average feature values, then learns a common latent subspace based on weighted NMF and $L_{2,1}$-norm regularization. But MIC only simply fills the missing samples with average feature values and if we cluster the data with a relatively large missing rate, this simply filling may result in a serious deviation. To align the information of the presented samples, DAIMC [27] extends MIC via weighted semi-NMF [28] and $L_{2,1}$-norm regularized regression. To obtain the robust clustering results, UEAF [29] performs the unified common embedding aligned with incomplete views inferring framework. Both DAIMC and UEAF rely too much on alignment information. When clustering the dataset without enough alignment information, DAIMC and UEAF always obtain unsatisfactory performance because the loss of alignment information will reduce the availability of their models.

Note that besides three main drawbacks in Section I, the previous methods have the above drawbacks. These drawbacks always result in unsatisfactory clustering performance, which limits the real-world applications of these methods.

III. NOTATION AND BACKGROUND

For convenience, we define some notations through the paper. All the matrices are written in uppercase. $[n] \stackrel{\text{def}}{=} \{1, 2, \ldots, n\}$. For a matrix $\boldsymbol{A}$, its ij-th element and i-th column are denoted by $\boldsymbol{A}_{i,j}$ and $\boldsymbol{A}_i$ separately; its trace is denoted by $\text{Tr}(\boldsymbol{A})$; its Frobenius norm is denoted by $||\boldsymbol{A}||_F$; its $L_{2,1}$-norm is denoted by $||\boldsymbol{A}||_{2,1}$; its nuclear norm is denoted by $||\boldsymbol{A}||_*$. $|\cdot|$ is the absolute operator; $<\cdot, \cdot>$ is the inner product operator; $\mathbf{1}$ is a column vector with all elements as 1; $\boldsymbol{I}$ is an identity matrix. For a complete multiview dataset, it is denoted by $\{\boldsymbol{U}^{(v)}\}_{v=1}^{n_v} \in \mathbb{R}^{n \times d_v}$, where d_v is the feature dimension of the v-th view, n is the sample number, and n_v is the view number. When suffering from incompleteness, $\{\boldsymbol{U}^{(v)}\}_{v=1}^{n_v}$ can be updated to $\{\boldsymbol{X}^{(v)}\}_{v=1}^{n_v} \in \mathbb{R}^{m_v \times d_v}$, where m_v is the number of presented samples in the v-th view. For the v-th view, $\boldsymbol{Z}^{(v)}$ is its subspace matrix; $\boldsymbol{N}^{(v)}$ is its variation matrix; $\boldsymbol{L}^{(v)}$ denotes its Laplacian graph; $\boldsymbol{E}^{(v)}$ is its error matrix; $\boldsymbol{F}^{(v)}$ is its cluster indicator matrix. For all the views of $\{\boldsymbol{X}^{(v)}\}_{v=1}^{n_v}$, $\boldsymbol{H}$ is the consensus cluster indicator matrix; $\boldsymbol{M}$ is the heredity matrix; c is the cluster number. α, β and η are nonnegative hyper-parameters.

A. Incomplete Multiview Data

The v-th original view matrix (including missing and presented samples) is represented as $\boldsymbol{U}^{(v)} \in \mathbb{R}^{n \times d_v}$, n and d_v are the number of samples and features, respectively. By removing the missing samples, we can update the v-th original view matrix to a new view matrix $\boldsymbol{X}^{(v)} \in \mathbb{R}^{m_v \times d_v}$, where m_v is the number of presented samples ($m_v < n$). To indicate the update, we define an incomplete index matrix $\boldsymbol{W}^{(v)} \in \mathbb{R}^{m_v \times n}$ [29]

$$\boldsymbol{W}_{i,j}^{(v)} = \begin{cases} 1, & \text{if the } i\text{-th sample is the } j\text{-th presented} \\ & \text{sample in the } v\text{-th view}; \\ 0, & \text{otherwise}. \end{cases} \quad (1)$$

B. Multiview Subspace Clustering

As an effective complete multiview clustering method, multiview subspace clustering (MVSC) integrates different views by first performing subspace clustering on each view and then unifying these subspaces to learn a cluster indicator matrix [34]. The framework of MVSC is as follows

$$\min_F \sum_v (||\boldsymbol{U}^{(v)} - \boldsymbol{U}^{(v)}\boldsymbol{Z}^{(v)} - \boldsymbol{E}^{(v)}||_F^2 + \beta||\boldsymbol{E}^{(v)}||_1$$

$$+ \eta\text{Tr}(\boldsymbol{F}^T\boldsymbol{L}_Z^{(v)}\boldsymbol{F}))$$

$$\text{s.t. } \boldsymbol{F}^T\boldsymbol{F} = \boldsymbol{I}, \boldsymbol{Z}^{(v)^T}\mathbf{1} = \mathbf{1}, \boldsymbol{Z}_{i,i}^{(v)} = 0, i \in [n], \quad (2)$$

where $\boldsymbol{F}$ is the cluster indicator matrix; for the v-th view, $\boldsymbol{L}_Z^{(v)}$ is its Laplacian graph. $\boldsymbol{Z}_{i,i}^{(v)} = 0$ means that all diagonal elements of $\boldsymbol{Z}^{(v)}$ are 0.

C. Biological Variation and Biological Heredity in Genetics

In genetics, biological heredity denotes the passing on of traits from parents to their children; biological variation represents the unique trait information of their children. By introducing biological heredity and biological variation, genetics provides some theories to analyze the consistent trait information and the unique trait information in the biological world [35], [36]. In genetics, $\boldsymbol{P}$ denotes the observed trait information. Based on the theory for quantitative traits influenced by maternal [35]–[37], $\boldsymbol{P}$ is partitioned as

$$\boldsymbol{P} = \boldsymbol{B}_H + \boldsymbol{B}_V, \quad (3)$$

where $\boldsymbol{B}_H$ denotes the biological heredity representation (i.e., consistent trait information) and $\boldsymbol{B}_V$ denotes the biological variation representation (i.e., unique trait information). Based on Eq. (3), genetics can explain the observed biological traits from the genetic level.

IV. PROPOSED V³H APPROACH

By showing the characteristics of incomplete multiview data, we first present the motivation of our proposed V³H approach. Then we model V³H as the joint of the view variation and the view heredity. Finally, we design a seven-step procedure to optimize V³H.

A. Motivation

Real-world incomplete multiview data have the two main characteristics [38]: (i) the common samples presented in all views can be used to extract the consistent information from different views and to integrate these views; (ii) these samples existing in partial views can learn the unique information of the corresponding views. Therefore, our motivation is to simultaneously learn consistent information and unique information from incomplete multiview data for clustering. Borrowing the idea of genetics in Section III-C, we propose some definitions in Definition 1, which relies on the following assumptions: (i) for an incomplete multiview dataset $\{\boldsymbol{X}^{(v)}\}_{v=1}^{n_v} \in \mathbb{R}^{m_v \times d_v}$, each of its subspaces is the perturbation of a consensus subspace; (ii) each subspace can represent the data structure of the corresponding view.

Definition 1: (Parent Subspace, Child Subspace, View Heredity, View Variation). For the $\{v\}_{v=1}^{n_v}$-th view matrix $\boldsymbol{X}^{(v)} \in \mathbb{R}^{m_v \times d_v}$, $\boldsymbol{Z}^{(v)} \in \mathbb{R}^{m_v \times m_v}$ denotes its subspace, which is calculated by Eq. (2). Assume that all the subspaces originate from a consensus subspace $\boldsymbol{Z}^*$. Thus, $\boldsymbol{Z}^*$ is defined as the parent subspace and $\boldsymbol{Z}^{(v)}$ is defined as the child subspace. $\boldsymbol{M}$ is defined as the heredity matrix, which represents the consistent information shared by all the views. $\boldsymbol{N}^{(v)}$ is defined as the variation matrix of the v-th view, which represents the unique information of the v-th view. The view heredity is the phenomenon that the consistent information exists in both parent subspace and child subspace. The view variation is the phenomenon that the unique information only exists in the corresponding child subspace. In general, different subspaces $\{\boldsymbol{Z}^{(v)}\}_{v=1}^{n_v}$ have different variation matrices $\{\boldsymbol{Z}^{(v)}\}_{v=1}^{n_v}$, but share the same heredity matrix $\boldsymbol{M}$.

Ideally, we want to obtain the optimal representation by learning the parent subspace $\boldsymbol{Z}^*$ because $\boldsymbol{Z}^*$ contains most of the available information on the data. However, it is difficult to learn an available $\boldsymbol{Z}^*$ due to missing samples and noises.

Therefore, we adopt a clever method to avoid learning $\boldsymbol{Z}^*$ directly. Instead, we can learn both the heredity matrix $\boldsymbol{M}$ and the variation matrix $\boldsymbol{N}^{(v)}$ as an alternative to learning $\boldsymbol{Z}^*$. For $\boldsymbol{Z}^{(v)}$, it can be decomposed into a heredity matrix $\boldsymbol{M}$ and a variation matrix $\boldsymbol{N}^{(v)}$. We assume that for a specific incomplete multiview dataset, when the missing rate changes, the dimensions of its heredity matrix and variation matrices do not change. This assumption can guarantee that we can integrate the subspaces with different dimensions.

In Eq. (3), the genetics analyzes the influence of biological heredity and biological variation on biological traits. Inspired by this, to integrate the heredity matrix and variation matrices, we design a subspace decomposition model as follows

$$\boldsymbol{W}^{(v)^T}\boldsymbol{Z}^{(v)}\boldsymbol{W}^{(v)} = \boldsymbol{M} - p\boldsymbol{N}^{(v)}, \qquad (4)$$

where p is adjustable as needed; $\boldsymbol{W}^{(v)^T}\boldsymbol{Z}^{(v)}\boldsymbol{W}^{(v)}$ is composed of $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{N}^{(v)} \in \mathbb{R}^{n \times n}$. For Eq. (3) and Eq. (4), $\boldsymbol{W}^{(v)^T}\boldsymbol{Z}^{(v)}\boldsymbol{W}^{(v)}$ corresponds to $\boldsymbol{P}$; $\boldsymbol{M}$ corresponds to $\boldsymbol{B}_H$; $-p\boldsymbol{N}^{(v)}$ corresponds to $\boldsymbol{B}_V$. Compared with $\boldsymbol{Z}^{(v)} \in \mathbb{R}^{m_v \times m_v}$, both $\boldsymbol{M}$ and $\boldsymbol{N}^{(v)}$ have larger matrix sizes ($n > m_v$). Since each column of $\boldsymbol{Z}^{(v)}$ corresponds to a sample, $\boldsymbol{Z}^{(v)}$ only contains the information of m_v presented samples. Both $\boldsymbol{M}$ and $\boldsymbol{N}^{(v)}$ contain the information of n samples (i.e., all the samples including the missing samples and the presented samples). Eq. (4) can learn three kinds of information: the consistent information between views (learned by $\boldsymbol{M}$), the unique information (learned by $\boldsymbol{N}^{(v)}$), and the relationship information between samples (learned by $\boldsymbol{W}^{(v)^T}\boldsymbol{Z}^{(v)}\boldsymbol{W}^{(v)}$). Since these three kinds of information are exactly expected in the clustering process, solving Eq. (4) can become a feasible alternative to learning $\boldsymbol{Z}^*$. Note that we have three variables ($\boldsymbol{Z}^{(v)}$, $\boldsymbol{M}$, and $\boldsymbol{N}^{(v)}$), but we only have one equation (Eq. (4)). Therefore, in the next section, we add some constraints on $\boldsymbol{M}$ and $\boldsymbol{N}^{(v)}$ to solve Eq. (4).

B. View Variation for View Alignment

Inspired by the phenomenon that the expression traits of parents and children are similar [39], [40], we attempt to align the expression traits of parent subspace and child subspace.[1] We treat the cluster indicator matrices as the expression traits of the corresponding views because we can obtain better clustering results after aligning these cluster indicator matrices. To formulate the alignment, we design the following view alignment model

$$\min_{\boldsymbol{F}^{(v)}, \boldsymbol{H}} \sum_v (\mathrm{Tr}(\boldsymbol{F}^{(v)^T}\boldsymbol{L}_N^{(v)}\boldsymbol{F}^{(v)}) + \gamma \mathrm{Dis}(\boldsymbol{F}^{(v)}, \boldsymbol{H})), \qquad (5)$$

where γ is a nonnegative hyper-parameter, $\boldsymbol{F}^{(v)} \in \mathbb{R}^{n \times c}$ is the cluster indicator matrix of the v-th view, and $\boldsymbol{H} \in \mathbb{R}^{n \times c}$ is the consensus cluster indicator matrix of all the

---

[1]In Section IV-A we can note that one view matrix $\boldsymbol{X}^{(v)}$ and its subspace matrix $\boldsymbol{Z}^{(v)}$ are in one-to-one correspondence. For ease of description, we still refer to this subspace-based alignment as view alignment.

views. $\mathrm{Tr}(\boldsymbol{F}^{(v)^T}\boldsymbol{L}_N^{(v)}\boldsymbol{F}^{(v)})$ is used to learn the cluster indicator matrix of each view; $\mathrm{Dis}(\boldsymbol{F}^{(v)}, \boldsymbol{H})$ means the disagreement between $\boldsymbol{F}^{(v)}$ and $\boldsymbol{H}$; $\boldsymbol{L}_N^{(v)} \in \mathbb{R}^{n \times n}$ is the normalized graph Laplacian, where $\boldsymbol{L}_N^{(v)} = \boldsymbol{G}^{(v)} - \boldsymbol{S}^{(v)}$, $\boldsymbol{S}^{(v)} = (|\boldsymbol{N}^{(v)}| + |\boldsymbol{N}^{(v)}|^T)/2$, and $\boldsymbol{G}_{i,i}^{(v)} = \sum_j \boldsymbol{S}_{i,j}^{(v)}$ ($i \in [n]$, $j \in [n]$).

For the measure of $\mathrm{Dis}(\boldsymbol{F}^{(v)}, \boldsymbol{H})$, a popular method is using the linear kernel function, which is simple and widely used. However, the linear kernel function can only learn linear structural information in most cases, and it is difficult to learn nonlinear information.

An ideal method is to design a linear kernel function that can learn nonlinear structures. Fortunately, the linear kernel used in the Laplacian for spectral clustering can learn the nonlinear structure of the data [41]. Thus, we choose the linear kernel function to measure the disagreement $\mathrm{Dis}(\boldsymbol{F}^{(v)}, \boldsymbol{H})$. Therefore, $\mathrm{Dis}(\boldsymbol{F}^{(v)}, \boldsymbol{H})$ is defined as follows

$$\mathrm{Dis}(\boldsymbol{F}^{(v)}, \boldsymbol{H}) = ||\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T} - \boldsymbol{H}\boldsymbol{H}^T||_F^2, \qquad (6)$$

where $\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T}$ is the linear kernel of $\boldsymbol{F}^{(v)}$ and $\boldsymbol{H}\boldsymbol{H}^T$ is the linear kernel of $\boldsymbol{H}$. Therefore, we rewrite Eq. (5) as

$$\min_{\boldsymbol{F}^{(v)}, \boldsymbol{H}} \sum_v (\mathrm{Tr}(\boldsymbol{F}^{(v)^T}\boldsymbol{L}_N^{(v)}\boldsymbol{F}^{(v)}) + \gamma ||\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T} - \boldsymbol{H}\boldsymbol{H}^T||_F^2). \qquad (7)$$

Combining Eq. (2), (4), and (7), we can obtain

$$\min_{\boldsymbol{F}^{(v)}} \sum_v (||\boldsymbol{X}^{(v)} - \boldsymbol{X}^{(v)}\boldsymbol{Z}^{(v)} - \boldsymbol{E}^{(v)}||_F^2 + \beta ||\boldsymbol{E}^{(v)}||_1$$
$$+ \alpha \mathrm{Tr}(\boldsymbol{F}^{(v)^T}\boldsymbol{L}_N^{(v)}\boldsymbol{F}^{(v)}) + \gamma ||\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T} - \boldsymbol{H}\boldsymbol{H}^T||_F^2)$$
$$\text{s.t. } \boldsymbol{W}^{(v)^T}\boldsymbol{Z}^{(v)}\boldsymbol{W}^{(v)} = \boldsymbol{M} - p\boldsymbol{N}^{(v)}, \boldsymbol{F}^{(v)^T}\boldsymbol{F}^{(v)} = \boldsymbol{I},$$
$$\boldsymbol{H}^T\boldsymbol{H} = \boldsymbol{I}, \boldsymbol{N}^{(v)}\mathbf{1} = \mathbf{1}, \boldsymbol{N}_{i,i}^{(v)} = 0, i \in [n], \qquad (8)$$

where the constraint $\boldsymbol{N}^{(v)}\mathbf{1} = \mathbf{1}$ treats all samples equally, which can learn the unique information of the v-th view; the constraint $\boldsymbol{N}_{i,i}^{(v)} = 0$ can ensure that each sample can only be represented as the combination of other samples [42].

C. View Heredity for Adjustable Low-Rank Representation

Most real-world multiview data often have the low-rank subspace representations, which can be used to recover the underlying true data structure [43]–[45]. Thus, the reasonable low-rank representation can improve the performance of subspace-based clustering. Outlier Pursuit [46] is a popular technology to obtain the proper low-rank representation, which is formulated as

$$\min_{\boldsymbol{M}, \boldsymbol{E}^{(v)}} ||\boldsymbol{M}||_* + \beta \sum_v ||\boldsymbol{E}^{(v)}||_{2,1}, \qquad (9)$$

where $||\boldsymbol{M}||_*$ is used to approximate the rank of $\boldsymbol{M}$; $||\boldsymbol{E}^{(v)}||_{2,1}$ can learn the low-rank representation of $\boldsymbol{E}^{(v)}$.

In fact, on the one hand, the rank approximation of $\boldsymbol{M}$ by the nuclear function will lead to a large deviation, which may result in unusable clustering results. On the other hand, $L_{2,1}$-norm is indifferentiable at the point of zero, which renders
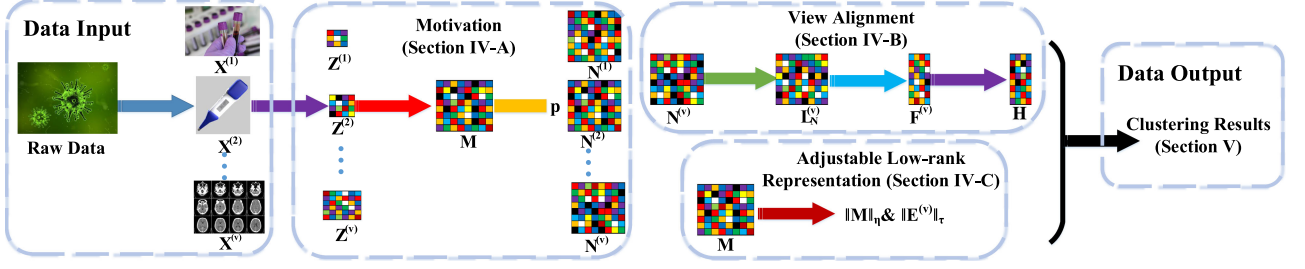
Fig. 1. Data flow of our proposed V³H approach. Given an incomplete dataset $\{X^{(v)}\}_{v=1}^{n_v}$, V³H learns the corresponding subspace matrices $\{Z^{(v)}\}_{v=1}^{n_v}$. In Section IV-A, V³H decomposes $Z^{(v)}$ into a heredity matrix M and a variation matrix $N^{(v)}$. In Section IV-B, by constructing a normalized graph Laplacian $L_N^{(v)}$, V³H aligns the cluster indicator matrix $F^{(v)}$ with the consensus cluster indicator matrix H. In Section IV-C, V³H leverages η-norm and τ-norm for adjustable low-rank representation. Section V shows the clustering results.

(a) η-norm

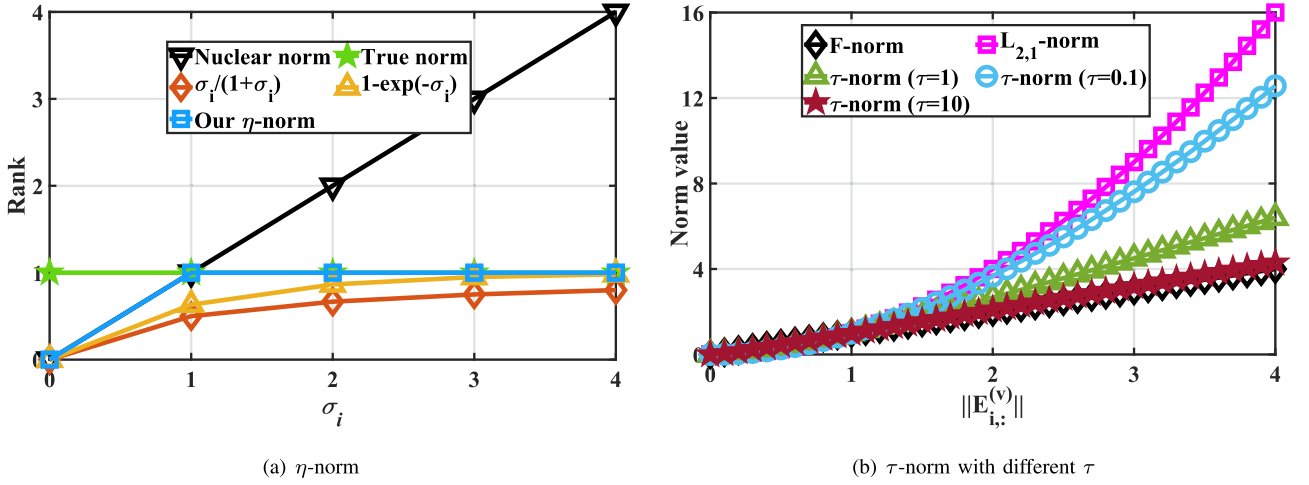(b) τ-norm with different τ

Fig. 2. Performance of η-norm and τ-norm.

the derivate of L_2-norm at sparse rows senseless. Besides, two cases fail Eq. (9): (i) If a dataset suffers from incompleteness or noises, its heredity matrix M is often unavailable. When the unavailable heredity matrix is used directly in Eq. (9), we will have difficulty obtaining the satisfactory low-rank representation because calculating the rank of the unavailable heredity matrix will produce a large deviation. (ii) When handling different clustering tasks, we often need different $E^{(v)}$ with different sparsity [47]. The adjustable sparsity of $E^{(v)}$ is necessary. Therefore, Eq. (9) also has two drawbacks: unavailable heredity matrix and non-adjustable sparsity.

The nonconvex relaxation of matrix rank is a popular technique. Evoked by [48]–[50], to learn available M, we propose η-norm defined by

$$||M||_\eta = \sum_i \frac{(\eta + w_i)\sigma_i(M)}{\eta + w_i\sigma_i(M)}, \qquad (10)$$

where η is adjustable as needed ($\eta > 0$); w is the weight vector ($w_i > 0$); $\sigma_i(M)$ is the i-th singular value of M.

Note that different from the common norm (e.g., $L_{2,1}$-norm and L_F-norm, etc.), our proposed η-norm is not a real norm. η-norm has the following characteristics:

1) η-norm is unitarily invariant, and $||M||_\eta = ||UMV||_w$ for any orthonormal $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$;

2) when $\eta \to \infty$, we have $||M||_\eta \to ||M||_{w,*}$, where $|| \cdot ||_{w,*}$ is the weighted nuclear norm [51];

3) when $\eta \to 0$, we have $||M||_\eta \to \text{rank}(M)$.

To show the advantage of our proposed η-norm, we compare η-norm with several rank relaxation approaches (i.e., $\sigma_i/(1 + \sigma_i)$ and $1 - exp(-\sigma_i)$ in Fig. 2(a)). As shown in Fig. 2(a), η-norm ($\eta = 0.001$ in this figure) is closer to the true rank than other approaches. Thus, by learning the satisfactory low-rank representation, η-norm can ensure the availability of the heredity matrix.

Similar to Eq. (10), to learn an adjustable sparse representation, we adopt the τ-norm of matrix $E^{(v)}$ defined by

$$||E^{(v)}||_\tau = \sum_i \frac{(1 + \tau)||E_{i,:}^{(v)}||^2}{\tau + ||E_{i,:}^{(v)}||}, \qquad (11)$$

where τ is adjustable for different tasks. Based on matrix $E^{(v)}$, we design a diagonal matrix $D_E^{(v)}$ defined by

$$D_{E_{i,i}}^{(v)} = \frac{(1 + \tau)(||E_i^{(v)}|| + 2\tau)}{(||E_i^{(v)}|| + \tau)^2}. \qquad (12)$$

Theorem 1: For any matrix $E^{(v)}$, we have

$$\frac{\partial ||E^{(v)}||_\tau}{\partial E^{(v)}} = D_E^{(v)} E^{(v)}. \qquad (13)$$

Proof:

$$\frac{\partial ||\boldsymbol{E}^{(v)}||_\tau}{\partial \boldsymbol{E}_{i,:}^{(v)}} = \frac{\partial ((1+\tau)||\boldsymbol{E}_{i,:}^{(v)}||_2^2/(1+\tau||\boldsymbol{E}_{i,:}^{(v)}||_2))}{\partial \boldsymbol{E}_{i,:}^{(v)}}$$

$$= \frac{(1+\tau)(||\boldsymbol{E}_i^{(v)}||+2\tau)}{(||\boldsymbol{E}_i^{(v)}||+\tau)^2}\boldsymbol{E}_{i,:}^{(v)}$$

$$\boldsymbol{D}_{i,i}^{(v)}\boldsymbol{E}_{i,:}^{(v)} = \frac{(1+\tau)(||\boldsymbol{E}_i^{(v)}||+2\tau)}{(||\boldsymbol{E}_i^{(v)}||+\tau)^2}\boldsymbol{E}_{i,:}^{(v)}. \qquad (14)$$

Obviously, $\partial ||\boldsymbol{E}^{(v)}||_\tau/\partial \boldsymbol{E}_{i,:}^{(v)} = \boldsymbol{D}_{i,i}^{(v)}\boldsymbol{E}_{i,:}^{(v)}$ in all the cases. ∎

Similar to η-norm, τ-norm has the following characteristics:

1) τ-norm is nonnegative and global differentiable;

2) when $\tau \to \infty$, we have $||\boldsymbol{E}^{(v)}||_\tau \to ||\boldsymbol{E}^{(v)}||_F^2$ and $\boldsymbol{D}_E^{(v)} \to \boldsymbol{I}$;

3) when $\tau \to 0$, we have $||\boldsymbol{E}^{(v)}||_\tau \to ||\boldsymbol{E}^{(v)}||_{2,1}$ and $\boldsymbol{D}_{E_{i,i}}^{(v)} \to 1/||\boldsymbol{E}_{i,:}^{(v)}||_2$.

To verify the adjustability of τ-norm, we compare τ-norm with $L_{2,1}$-norm and F-norm in Fig. 2(b). When τ is relatively small ($\tau = 0.1$), τ-norm is near to F-norm. As τ decreases, $||\boldsymbol{E}^{(v)}||_\tau$ is closer to $||\boldsymbol{E}^{(v)}||_{2,1}$, and $\boldsymbol{E}^{(v)}$ becomes more sparse. Since we can choose different τ to adjust the sparsity, τ-norm has wider applications than $L_{2,1}$-norm and F-norm.

Therefore, considering both the available low-rank subspaces (Eq. (10)) and the adjustable sparsity (Eq. (11)), we can obtain the adjustable low-rank representation as follows

$$\min_{\boldsymbol{M},\boldsymbol{E}^{(v)}} ||\boldsymbol{M}||_\eta + \beta \sum_v ||\boldsymbol{E}^{(v)}||_\tau. \qquad (15)$$

D. Objective Function

Combining the view alignment (Eq. (8)) and the adjustable low-rank representation (Eq. (15)), we have

$$\min_{\boldsymbol{M},\boldsymbol{N}^{(v)},\boldsymbol{Z}^{(v)},\boldsymbol{E}^{(v)},\boldsymbol{F}^{(v)},\boldsymbol{F}^*} ||\boldsymbol{M}||_\eta + \sum_v (\beta||\boldsymbol{E}^{(v)}||_\tau$$

$$+ \alpha\mathrm{Tr}(\boldsymbol{F}^{(v)^T}\boldsymbol{L}_N^{(v)}\boldsymbol{F}^{(v)}) + \gamma||\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T} - \boldsymbol{H}\boldsymbol{H}^T||_F^2)$$

s.t. $\boldsymbol{F}^{(v)^T}\boldsymbol{F}^{(v)} = \boldsymbol{I}, \boldsymbol{H}^T\boldsymbol{H} = \boldsymbol{I}, \boldsymbol{N}^{(v)}\boldsymbol{1} = \boldsymbol{1}, \boldsymbol{N}_{i,i}^{(v)} = 0,$

$$\boldsymbol{W}^{(v)^T}\boldsymbol{Z}^{(v)}\boldsymbol{W}^{(v)} = \boldsymbol{M} - p\boldsymbol{N}^{(v)}, \boldsymbol{X}^{(v)} = \boldsymbol{X}^{(v)}\boldsymbol{Z}^{(v)} - \boldsymbol{E}^{(v)}. \qquad (16)$$

Eq. (16) is a nonconvex function, which is often difficult to optimize directly. In the next section, we will design an iteration procedure to optimize it.

E. Optimization

To optimize Eq. (16), we design the following augmented Lagrangian function

$$J = ||\boldsymbol{M}||_\eta + \sum_v (\alpha\mathrm{Tr}(\boldsymbol{F}^{(v)^T}\boldsymbol{L}_N^{(v)}\boldsymbol{F}^{(v)}) + \beta||\boldsymbol{E}^{(v)}||_\tau$$

$$+ \gamma||\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T} - \boldsymbol{H}\boldsymbol{H}^T||_F^2 + \frac{\omega}{2}(||\boldsymbol{W}^{(v)^T}\boldsymbol{Z}^{(v)}\boldsymbol{W}^{(v)}$$

Algorithm 1: V³H.

Input: $\{\boldsymbol{X}^{(v)}\}_{v=1}^{n_v}, \{\boldsymbol{W}^{(v)}\}_{v=1}^{n_v}, \alpha, \beta, \gamma, \lambda_{\max}$, and c.

Initialize $\boldsymbol{E}^{(v)} = \boldsymbol{Z}^{(v)} = \boldsymbol{M} = \boldsymbol{N}^{(v)} = \boldsymbol{0},$

$\boldsymbol{C}_1^{(v)} = \boldsymbol{C}_2^{(v)} = \boldsymbol{0}, \boldsymbol{F}^{(v)}, \boldsymbol{F}^*$, and ω.

repeat

  for $v = 1$ to n_v **do**

    Update $\boldsymbol{Z}^{(v)}$ by Eq. (26);

    Update $\boldsymbol{N}^{(v)}$ and ζ based on Eq. (30), Eq. (31), and

    the constraint $\boldsymbol{N}_{i,:}^{(v)}\boldsymbol{1} - 1 = 0$;

    Update $\boldsymbol{E}^{(v)}$ by Eq. (34);

    Update $\boldsymbol{F}^{(v)}$ by Eq. (36);

    Update $\boldsymbol{C}_1^{(v)}, \boldsymbol{C}_2^{(v)}$, and ω by Eq. (38);

  end for

  Update $\boldsymbol{M}$ by Eq. (18) and DC programming;

  Update $\boldsymbol{H}$ by Eq. (37);

Until converges

Output: $\boldsymbol{F}^{(v)}, \boldsymbol{H}, \boldsymbol{M}, \boldsymbol{N}^{(v)}$ and clustering results.

$$- \boldsymbol{M} + p\boldsymbol{N}^{(v)} + \frac{\boldsymbol{C}_2^{(v)}}{\omega}||_F^2 + ||\boldsymbol{X}^{(v)} - \boldsymbol{X}^{(v)}\boldsymbol{Z}^{(v)} - \boldsymbol{E}^{(v)}$$

$$+ \frac{\boldsymbol{C}_1^{(v)}}{\omega}||_F^2 - \zeta^{(v)^T}(\boldsymbol{N}^{(v)}\boldsymbol{1} - \boldsymbol{1}))), \qquad (17)$$

where matrices $\boldsymbol{C}_1^{(v)}, \boldsymbol{C}_2^{(v)}$ and vector $\zeta^{(v)}$ are Lagrange multipliers, ω is a nonnegative penalty parameter. Eq. (17) is not convex for all variables simultaneously, and it is difficult to solve Eq. (17) in one step. Thus, we design the following seven-step procedure to update each variable iteratively [52].

Step 1: Update $\boldsymbol{M}$. Fixing the other variables, the problem to update $\boldsymbol{M}$ is degraded to solve the following problem

$$\boldsymbol{M} = \arg\min_{\boldsymbol{M}} \frac{\omega}{2}||\boldsymbol{W}^{(v)^T}\boldsymbol{Z}^{(v)}\boldsymbol{W}^{(v)} - \boldsymbol{M} + p\boldsymbol{N}^{(v)}$$

$$+ \frac{\boldsymbol{C}_2^{(v)}}{\omega}||_F^2 + ||\boldsymbol{M}||_\eta. \qquad (18)$$

To solve Eq. (18), we first develop the following theorem.

Theorem 2: We first set $\boldsymbol{A} = \boldsymbol{W}^{(v)^T}\boldsymbol{Z}^{(v)}\boldsymbol{W}^{(v)} + p\boldsymbol{N}^{(v)} + \boldsymbol{C}_2^{(v)}/\omega$. The SVD operation of $\boldsymbol{A}$ is $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}_A\boldsymbol{V}^T$, where $\boldsymbol{\Sigma}_A = diag(\sigma_A)$. Set $H(\boldsymbol{M}) = h \circ \sigma_M$ be a unitarily invariant function, where $h(\sigma) = \Sigma_i(\eta + w_i)\sigma_i(\boldsymbol{M})/(\eta + w_i\sigma_i(\boldsymbol{M}))$ and $\omega > 0$. Based on Eq. (18), we have

$$\min_{\boldsymbol{M}} H(\boldsymbol{M}) + \frac{\omega}{2}||\boldsymbol{M} - \boldsymbol{A}||_F^2. \qquad (19)$$

Then an optimal solution to Eq. (19) is $\boldsymbol{M}^* = \boldsymbol{U}\boldsymbol{\Sigma}_M^*\boldsymbol{V}^T$, where $\boldsymbol{\Sigma}_M^* = diag(\sigma^*)$ and $\sigma^* = prox_{h,\omega}(\sigma_A)$. $prox_{h,\omega}(\sigma_A)$ is the *Moreau-Yosida operator*, which is defined as

$$prox_{h,\omega}(\sigma_A) = \arg\min h(\sigma) + \frac{\omega}{2}||\sigma - \sigma_A||_2^2. \qquad (20)$$

Proof: Since $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}_A\boldsymbol{V}^T$, we have $\boldsymbol{\Sigma}_A = \boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V}$. Denoting $\boldsymbol{Q} = \boldsymbol{U}^T\boldsymbol{M}\boldsymbol{V}$ which has the same singular values as $\boldsymbol{M}$, we have

$$H(\boldsymbol{M}) + \frac{\omega}{2}||\boldsymbol{M} - \boldsymbol{A}||_F^2 \qquad (21a)$$

$$= H(\boldsymbol{Q}) + \frac{\omega}{2}||\boldsymbol{Q} - \boldsymbol{\Sigma}_A||_F^2 \qquad (21b)$$

$$\geq H(\Sigma_Q) + \frac{\omega}{2}||\Sigma_Q - \Sigma_A||_F^2 \qquad (21c)$$

$$= H(\Sigma_M) + \frac{\omega}{2}||\Sigma_M - \Sigma_A||_F^2 \qquad (21d)$$

$$= h(\sigma) + \frac{\omega}{2}||\sigma - \sigma_A||_2^2$$

$$\geq h(\sigma^*) + \frac{\omega}{2}||\sigma^* - \sigma_A||_2^2. \qquad (21e)$$

Eq. (21b) holds because the Frobenius norm is unitarily invariant, Eq. (21c) holds due to the Hoffman-Wielandt inequality, and Eq. (21d) holds based on $\Sigma_Q = \Sigma_M$. Therefore, Eq. (21d) is a lower bound of Eq. (21a). Due to $\Sigma_M = \Sigma_Q = Q = U^T M V$, the singular value decomposition (SVD) of M is $M = U^T \Sigma_M V$. By minimizing Eq. (21e), we learn σ^*. Hence $M^* = U diag(\sigma^*) V^T$, which is the optimal solution of Eq. (19). Thus, Theorem 2 is proved. ■

Note that Eq. (20) is a combination of concave and convex functions, which motivates us to leverage the difference of convex (DC) programming algorithm [53]. The algorithm decomposes a nonconvex function as the difference of two convex functions and iteratively optimizes it by linearizing the concave term at each iteration. For the i-th inner iteration, $g_i = \partial h(\sigma^i)$ denotes the gradient of $h(\cdot)$ at σ^i, where σ^i is the updated value of $\sigma(M)$ after the i-th inner iteration. $U diag(\sigma_A) V^T$ is the SVD of $W^{(v)^T} Z^{(v)} W^{(v)} + p N^{(v)} + C_2^{(v)}/\omega$. For the $(i+1)$-th inner iteration, we have

$$\sigma^{i+1} = \arg\min <g_i, \sigma^i> + \frac{\omega}{2}||\sigma^i - \sigma_A||_2^2, \qquad (22)$$

which admits the following closed-form solution

$$\sigma^{i+1} = \left(\sigma_A - \frac{g_i}{\omega^i}\right). \qquad (23)$$

After several iterations, it at least converges to a locally optimal point σ^*. Then $M = U diag(\sigma^*) V^T$.

Step 2: Update $Z^{(v)}$. Fixing the other variables, the problem to update $Z^{(v)}$ is degraded to minimize

$$J(Z^{(v)}) = \left\|W^{(v)^T} Z W^{(v)} - M + p N^{(v)} + \frac{C_2^{(v)}}{\omega}\right\|_F^2$$

$$+ \left\|X^{(v)} - X^{(v)} Z^{(v)} - E^{(v)} + \frac{C_1^{(v)}}{\omega}\right\|_F^2. \qquad (24)$$

Setting the derivative $J(Z^{(v)})$ w.r.t $Z^{(v)}$ to $\mathbf{0}$, we have

$$2 W^{(v)} \left(W^{(v)^T} Z W^{(v)} - M + p N^{(v)} + \frac{C_2^{(v)}}{\omega}\right) W^{(v)^T}$$

$$+ 2 X^{(v)^T} \left(X^{(v)} - X^{(v)} Z^{(v)} - E^{(v)} + \frac{C_1^{(v)}}{\omega}\right) = \mathbf{0}. \qquad (25)$$

Based on the definition of $W^{(v)}$, we can find $W^{(v)} W^{(v)^T} = I$. By solving Eq. (25), we can update $Z^{(v)}$ by

$$Z^{(v)} = (I + X^{(v)^T} X^{(v)})^{-1} \left(X^{(v)^T} \left(X^{(v)} - E^{(v)}\right)\right.$$

$$\left. + \frac{C_1^{(v)}}{\omega}\right) + W^{(v)} \left(M - p N^{(v)} - \frac{C_2^{(v)}}{\omega}\right) W^{(v)^T}\right). \qquad (26)$$

Step 3: Update $N^{(v)}$ and ζ. Fixing the other variables, the problem to update $N^{(v)}$ is degraded to minimize

$$J(N^{(v)}) = \frac{\omega}{2} \left\|W^{(v)^T} Z^{(v)} W^{(v)} - M + p N^{(v)} + \frac{C_2^{(v)}}{\omega}\right\|_F^2$$

$$+ \alpha \text{Tr}(F^{(v)^T} L_N^{(v)} F^{(v)}) - \zeta^{(v)^T} (N^{(v)} \mathbf{1} - \mathbf{1}). \qquad (27)$$

We define $K^{(v)} = W^{(v)^T} Z^{(v)} W^{(v)} - M + C_2^{(v)}/\omega$, and Eq. (27) can be equivalent to

$$J(N^{(v)}) = \frac{\omega p}{2}||N^{(v)} - \frac{1}{p}K^{(v)}||_F^2 + \frac{\alpha}{2} \sum_{i,j}^m ||F_{i,:}^{(v)}$$

$$- F_{j,:}^{(v)}||_2^2 N^{(v)} - \zeta^{(v)^T} (N^{(v)} \mathbf{1} - \mathbf{1}). \qquad (28)$$

Note that Eq. (28) is independent to each row. Defining $T_{i,j}^{(v)} = ||F_{i,:}^{(v)} - F_{j,:}^{(v)}||_2^2$, we transform minimizing Eq. (28) into

$$\min_{N_{i,:}^{(v)} \geq 0, N_{i,i}^{(v)} = 0} \frac{\omega p}{2}||N_{i,:}^{(v)} - \frac{1}{p}K_{i,:}^{(v)}||_2^2 + \frac{\alpha}{2} N_{i,:}^{(v)} T_{i,:}^{(v)^T}$$

$$- \zeta_i^{(v)^T} (N_{i,:}^{(v)} \mathbf{1} - 1)$$

$$\Leftrightarrow \min_{N_{i,:}^{(v)} \geq 0, N_{i,i}^{(v)} = 0} \frac{\omega p}{2}||N_{i,:}^{(v)} + \frac{\alpha}{2\omega p} T_{i,:}^{(v)} - \frac{1}{p}K_{i,:}^{(v)}||_2^2$$

$$- \zeta_i^{(v)^T} (N_{i,:}^{(v)} \mathbf{1} - 1). \qquad (29)$$

Defining $Y_{i,j}^{(v)} = K_{i,j}^{(v)}/p - \alpha/(2p\omega) T_{i,j}^{(v)} + \zeta_i^{(v)}/(p\omega)$, we can obtain the optimal $N^{(v)}$ by

$$N_{i,j}^{(v)} = \begin{cases} \max(Y_{i,j}^{(v)}, 0), & i \neq j; \\ 0, & \text{otherwise,} \end{cases} \qquad (30)$$

where $\max(Y_{i,j}^{(v)}, 0)$ is used to ensure that all elements of $N^{(v)}$ are not less than 0. Based on the constraint $N_{i,:}^{(v)} \mathbf{1} - 1 = 0$, we can update $\zeta_i^{(v)}$ by

$$\zeta_i^{(v)} = \frac{\omega p}{m-1} \left(1 - \sum_{j=1, j \neq i}^m \left(\frac{1}{p}K_{i,j}^{(v)} - \frac{\alpha}{2p\omega}T_{i,j}^{(v)}\right)\right). \qquad (31)$$

Step 4: Update $E^{(v)}$. Fixing the other variables, the problem to update $E^{(v)}$ is degraded to minimize

$$J(E^{(v)}) = \frac{\omega}{2}||X^{(v)} - X^{(v)} Z^{(v)} - E^{(v)}$$

$$+ \frac{C_1^{(v)}}{\omega}||_F^2 + \beta||E^{(v)}||_\tau. \qquad (32)$$

Deriving $J(E^{(v)})$ w.r.t. $E^{(v)}$, we can have

$$\frac{\partial J(E^{(v)})}{\partial E^{(v)}}$$

$$= \omega \left(E^{(v)} + X^{(v)} Z^{(v)} - X^{(v)} - \frac{C_1^{(v)}}{\omega}\right) + \beta D_E^{(v)} E^{(v)}. \qquad (33)$$

Setting $\partial J(\boldsymbol{E}^{(v)})/\partial \boldsymbol{E}^{(v)} = 0$, we can update $\boldsymbol{E}^{(v)}$ by

$$\boldsymbol{E}^{(v)} = \left(\boldsymbol{I} + \frac{\beta}{\omega}\boldsymbol{D}_E^{(v)}\right)^{-1}\left(\boldsymbol{X}^{(v)} - \boldsymbol{X}^{(v)}\boldsymbol{Z}^{(v)} + \frac{\boldsymbol{C}_1^{(v)}}{\omega}\right). \tag{34}$$

Step 5: Update $\boldsymbol{F}^{(v)}$. Fixing the other variables, we can update $\boldsymbol{F}^{(v)}$ by

$$\min \alpha \mathrm{Tr}(\boldsymbol{F}^{(v)^T}\boldsymbol{L}_N^{(v)}\boldsymbol{F}^{(v)}) + \gamma||\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T} - \boldsymbol{H}\boldsymbol{H}^T||_F^2)$$

$$\Leftrightarrow \min \alpha \mathrm{Tr}(\boldsymbol{F}^{(v)^T}\boldsymbol{L}_N^{(v)}\boldsymbol{F}^{(v)} + \gamma(||\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T}||_F^2$$

$$+ ||\boldsymbol{H}\boldsymbol{H}^T||_F^2 - \mathrm{Tr}(\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T}\boldsymbol{H}\boldsymbol{H}^T)). \tag{35}$$

Note that $||\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T}||_F^2 = ||\boldsymbol{H}\boldsymbol{H}^T||_F^2 = c$, and c is a constant for a specific dataset. Thus, we transfer Eq. (35) into

$$\min \mathrm{Tr}(\boldsymbol{F}^{(v)^T}(\alpha\boldsymbol{L}_N^{(v)} - \gamma\boldsymbol{H}\boldsymbol{H}^T)\boldsymbol{F}^{(v)}). \tag{36}$$

Obviously, we can solve Eq. (36) by the eigenvalue decomposition. The optimal $\boldsymbol{F}^{(v)}$ is the eigenvector set that corresponds to the first c smallest eigenvalues of matrix $(\alpha\boldsymbol{L}_N^{(v)} - \gamma\boldsymbol{H}\boldsymbol{H}^T)$.

Step 6: Update $\boldsymbol{H}$. Similar to Step 5, we update $\boldsymbol{H}$ by

$$\min -\gamma \sum_v \mathrm{Tr}(\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T}\boldsymbol{H}\boldsymbol{H}^T)$$

$$\Leftrightarrow \max \mathrm{Tr}\left(\boldsymbol{H}^T\left(\sum_v(\gamma\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T})\right)\boldsymbol{H}\right). \tag{37}$$

By the eigenvalue decomposition, we can learn the optimal $\boldsymbol{H}$, which is also the eigenvector set corresponding to the first c largest eigenvalues of matrix $(\sum_v(\gamma\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T}))$.

Step 7: Update $\boldsymbol{C}_1^{(v)}, \boldsymbol{C}_2^{(v)}$ and ω. We can update them by

$$\boldsymbol{C}_1^{(v)} = \boldsymbol{C}_1^{(v)} + \omega(\boldsymbol{X}^{(v)} - \boldsymbol{X}^{(v)}\boldsymbol{Z}^{(v)} - \boldsymbol{E}^{(v)})$$

$$\boldsymbol{C}_2^{(v)} = \boldsymbol{C}_2^{(v)} + \omega(\boldsymbol{W}^{(v)^T}\boldsymbol{Z}\boldsymbol{W}^{(v)} - \boldsymbol{M} + p\boldsymbol{N}^{(v)})$$

$$\omega = \min(\varphi\omega, \omega_{max}), \tag{38}$$

where φ and ω_{max} are constants.

The V^{3}H algorithm is shown in Algorithm 1. We provide its codes in Code Ocean (DOI:10.24 433/CO.2 119 636.v1) and Github (https://github.com/ZeusDavide/TAI_V3H.git).

F. Convergence and Complexity

1) Convergence Analysis: To optimize our proposed V^{3}H, we need to solve seven subproblems in Algorithm 1. Each subproblem has a closed solution w.r.t the corresponding variable. The objective function is bounded, and all the above seven steps do not increase the objective function value. Thus, the objective function can reduce monotonically to a stationary value, and V^{3}H can at least find a locally optimal solution.

2) Complexity Analysis: From Section IV-E, the major computational costs of our proposed V^{3}H mainly come from the operations like matrix inverse, SVD, and eigenvalue decomposition. Therefore, Steps 1, 5, and 6 are the main computational costs. For Step 2, its major computational costs are from the inverse operation $(\boldsymbol{I} + \boldsymbol{X}^{(v)^T}\boldsymbol{X}^{(v)})^{-1}$. Since both $\boldsymbol{I}$ and $\boldsymbol{X}^{(v)}$

TABLE I
STATISTICS OF THE DATASETS, WHERE "#. FEATURES" MEANS THE TOTAL
FEATURE DIMENSION OF ALL THE VIEWS

Dataset	3-Sources	20-NGs	100-Ls	BBC (3v)	BBC (4v)
#. samples	169	500	1600	282	685
#. views	3	3	3	3	4
#. features	10259	1500	192	7591	18641
#. clusters	6	5	100	5	5
Dataset	BS (2v)	BS (4v)	BUAA	Coil	Digit
#. samples	544	116	180	1440	2000
#. views	2	4	2	3	5
#. features	6386	8345	200	11078	585
#. clusters	5	5	3	20	10
Dataset	NUS	ORL	Scene	Yale	YaleB
#. samples	2400	400	2688	165	650
#. views	6	3	4	3	3
#. features	1134	14150	1248	14150	12554
#. clusters	12	40	8	15	10

are not updated in each iteration, we can pre-compute the inverse operation before the iteration for simplicity. For Steps 1, 5, and 6, they have the same computational complexity $O(n^3)$. Therefore, the whole computational complexity of V^{3}H is about $O(in_v n^3)$, where i is the iteration number, n_v is the view number, and n is the sample number.

Note that the complexity of V^{3}H has nothing to do with the feature dimension d_v. Since most real-world data are high-dimensional [54], V^{3}H will have wide applications.

V. PERFORMANCE EVALUATION

We first illustrate the clustering performance of the proposed V^{3}H, then verify V^{3}H's convergence, and finally analyze the sensitivity of V^{3}H's parameters.

A. Datasets

We conduct experiments on fifteen well-known popular datasets: 3-Sources,[2] 20 New Groups (20-NGs),[3] 100 Leaves (100-Ls),[4] BBC with 3 views (BBC (3v)),[5] BBC with 4 views (BBC (4v)),[6] BBCSport with 2 views (BS (2v)),[7] BBCSport with 4 views (BS (4v)),[8] BUAA [55], Coil [56], Digit,[9] NUS [57], ORL [58], Outdoor Scene (Scene) [59], Yale,[10] and Extended YaleB (YaleB) [60].

For these datasets, most multiview clustering algorithms often cluster their common subsets for simplicity. To compare fairly with these algorithms, we use the same size datasets for clustering. The important statistics of used datasets are shown in Table I, and the detailed statistics are as follows:

*1) **3-Sources**: it is a news dataset that has 948 samples collected from 3 views: BBC with 3560 features, Reuters with 3631 features and The Guardian with 3068 features. Following [29],

[2][Online]. Available: http://mlg.ucd.ie/datasets/3sources.html
[3][Online]. Available: http://lig-membres.imag.fr/grimal/data.html
[4][Online]. Available: https://archive.ics.uci.edu/ml/datasets/One-hundred+plant+species+leaves+data+set
[5][Online]. Available: http://mlg.ucd.ie/datasets/segment.html
[6][Online]. Available: http://mlg.ucd.ie/datasets/segment.html
[7][Online]. Available: http://mlg.ucd.ie/datasets/segment.html
[8][Online]. Available: http://mlg.ucd.ie/datasets/segment.html
[9][Online]. Available: http://archive.ics.uci.edu/ml/datasets.html
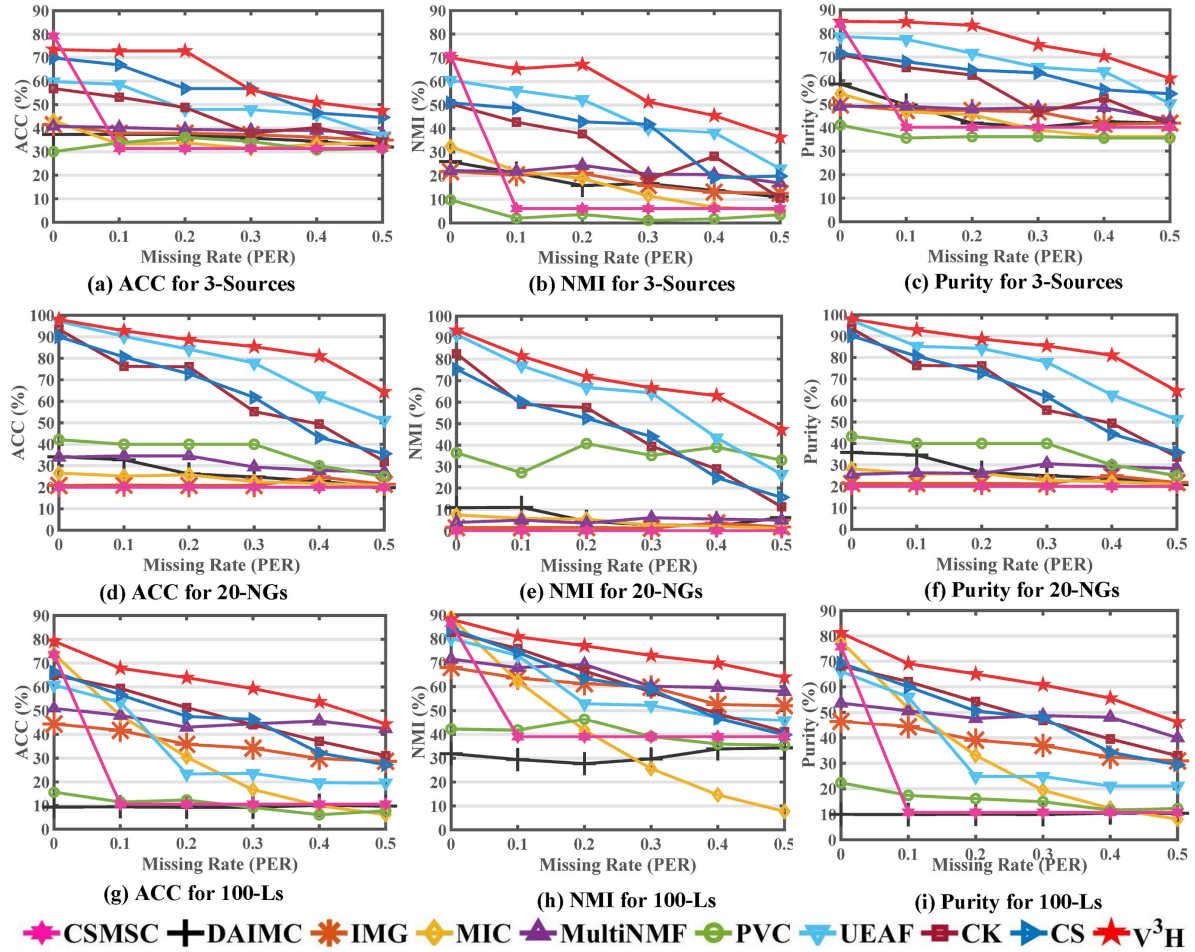[10][Online]. Available: http://vision.ucsd.edu/content/yale-face-database

Fig. 3. Results on 3-Sources, 20-NGs, and 100-Ls datasets.

we select a subset with 169 samples, which are categorized into 6 clusters.

2) **20-NGs**: it is a document dataset that has 500 samples from 3 views. Each view has 500 features. These documents are categorized into 5 clusters.

3) **100-Ls**: it has 1600 samples from 100 clusters. Each sample appears in 3 views, and each view has 64 features.

4) **BBC (3v)** and 5) **BBC (4v)**: the original BBC dataset has 685 samples, which are described by 3-4 views categorized into 5 clusters. Following [61], we choose a subset with 282 samples described by 3 views. These views include 2582 features, 2544 features, 2465 features, respectively. Following [62], we also choose the full dataset described by 4 views. These views include 4659 features, 4633 features, 4665 features, and 4684 features, respectively.

6) **BS (2v)** and 7) **BS (4v)**: the original BBCSport dataset has 737 samples, which are described by 2-4 views and categorized into 5 clusters. Following [63], we select a subset with 544 samples described by 2 views. These views include 3183 features and 3203 features, respectively. Following [29], we also use a subset with 116 samples described by 4 views. These views include 1991 features, 2063 features, 2113 features, and 2158 features, respectively.

8) **BUAA**: it has 180 image samples, which are categorized into 10 clusters. Each image is described by 2 views, and each view has 100 features.

9) **Coil**: it is an image dataset that has 1440 samples consisting of 20 clusters. Each image appears in three different views: Intensity with 1024 features, LBP feature with 3304 features, and Gabor feature with 6750 features.

10) **Digit**: it has 2000 samples categorized into 10 clusters. It has 5 views: FOU with 76 features, FAC with 216 features, KAR with 64 features, PIX with 240 features, and ZER with 47 features.

11) **NUS**: it consists of 2400 samples categorized into 12 clusters. Besides, each image is described by 6 views: CH with 65 features, CM with 226 features, CORR with 145 features, ED with 74 features and WV with 129 features.

12) **ORL**: it is an image dataset, which is described by 3 views. The dataset has 400 samples categorized into 40 clusters. These views include 4096 features, 3304 features and 6750 features, respectively.

13) **Scene**: it is an image dataset that has 2688 samples consisting of 8 clusters. Each image is described by 4 views: GIST with 512 features, color moment with 432 features, HOG with 256 features and LBP with 48 features.
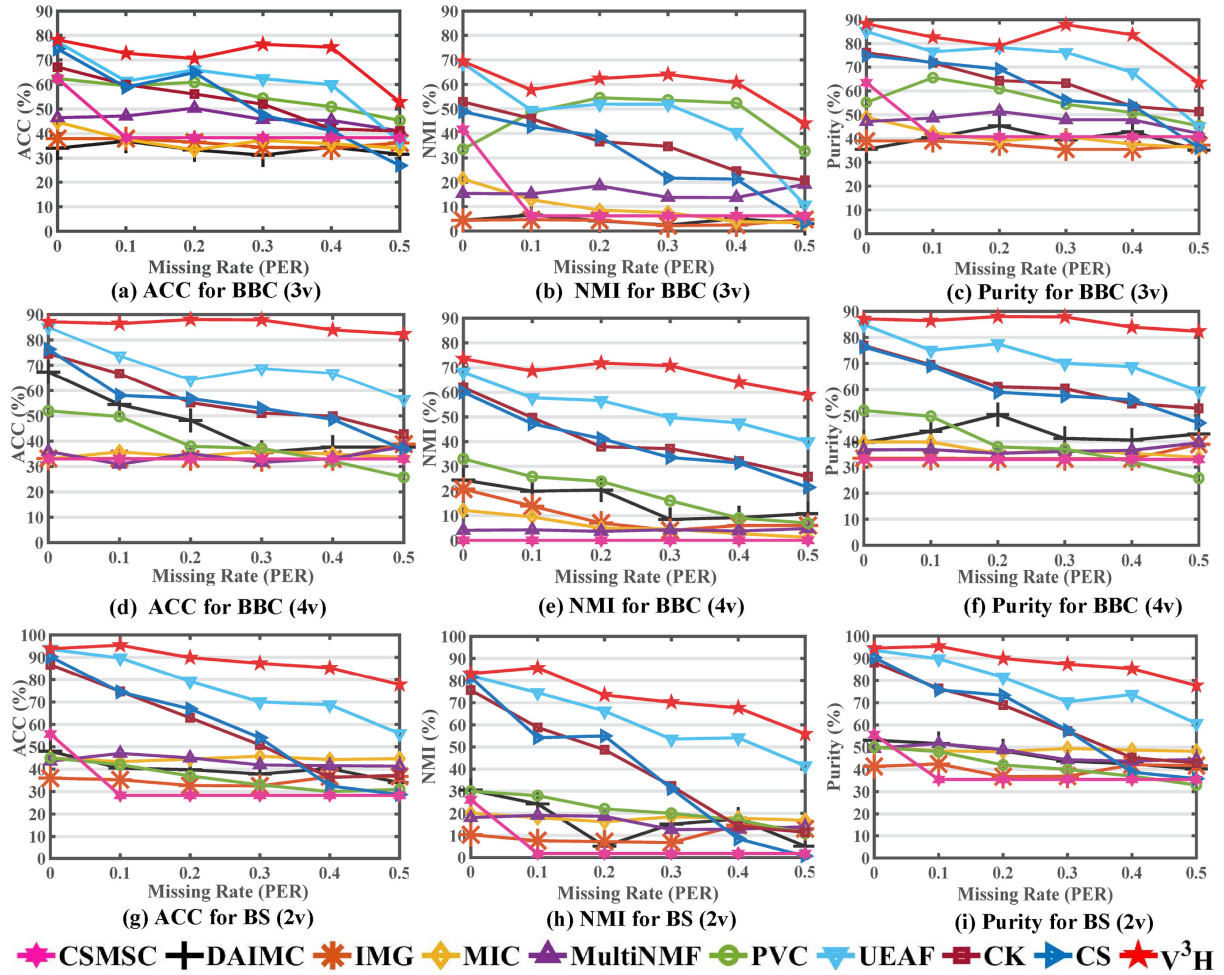
Fig. 4.   Results on BBC (3v), BBC (4v) and BS (2v) datasets.

14) **Yale**: it has 165 samples categorized into 15 clusters. It is described by 3 views: Intensity with 4096 features, LBP with 3304 features and Gabor with 6750 features.

15) **YaleB**: it is a face image dataset that has 2414 samples described by 3 views. These views include 2500 features, 3304 features and 6750 features, respectively. Following [64], we choose a subset with 650 samples.

B. Compared Methods

We compare our proposed $\mathbf{V^3H}$ with the following state-of-the-art methods, which are most relevant to our work:

1) **CSMSC** [1] learns a consistent representation and a set of specific representations from complete multiview data;

2) **DAIMC** [27] extends MIC based on weighted semi-NMF and $L_{2,1}$-Norm regularization regression;

3) **IMG** [20] transforms the collected incomplete multiview data to a complete representation in a latent space;

4) **MIC** [24] extends MultiNMF based on weighted NMF and $L_{2,1}$-Norm regularization;

5) **MultiNMF** [25] extends NMF to multiview scenes by jointing these views;

6) **PVC** [23] aligns the same samples in different views by constructing a latent subspace;

7) **UEAF** [29] learns a consensus representation for all views by extending MIC;

8) **Concat-K-means clustering (CK)**, and 9) **Concat-Spectral clustering (CS)**. CK and CS are two baselines that concatenate all views into one single view for clustering. (i) CK: we first fill the missing samples with the average features for each view. Then we concatenate the features of all the views, and perform K-means clustering on the concatenated view. (ii) CS: similar to CK, we perform spectral clustering on the concatenated view.

For CSMSC and MultiNMF, they cannot directly handle incomplete multiview data. Following [27], we first fill the missing samples with average feature values. Then we perform CSMSC and MultiNMF. Since IMG and PVC cannot perform clustering on the incomplete data with more than two views, we use these methods on all the two-views combinations and report average results for fairness. Since our proposed V^3H has three parameters, α, β, and γ, we adjust them to get the best performance (see Section V-D). Since p, η, τ, and w_i are adjustable as needed, we set $p = 1$, $\eta = 10^{-3}$, $\tau = 10^{-2}$, and $w_i = 1$ in our experiment for simplicity.
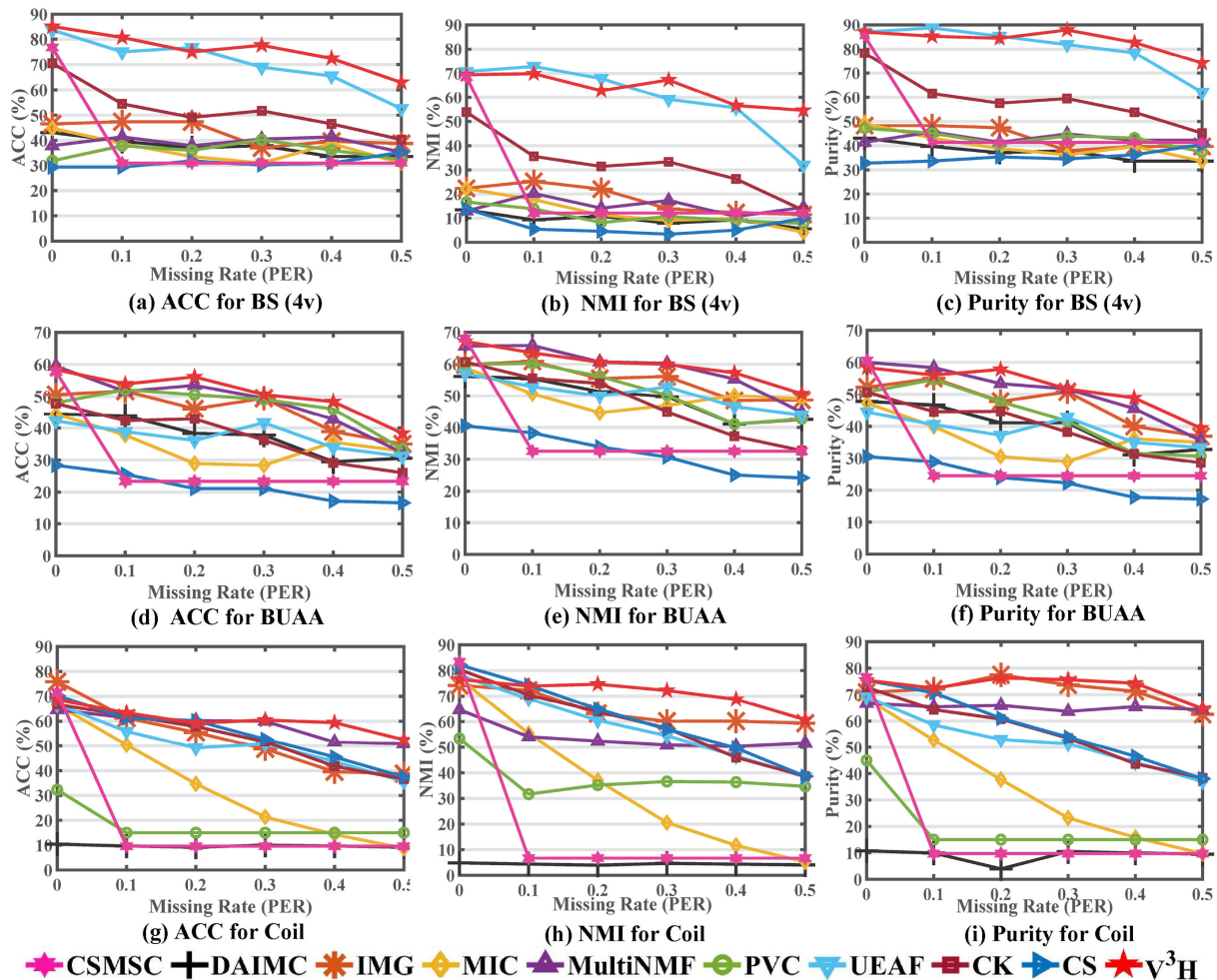
Fig. 5.   Results on BS (4v), BUAA, and Coil datasets.

Following [20], we repeat each incomplete multiview clustering experiment 10 times to obtain the average performance. Following [27] and [24], we randomly delete some samples from each view to get incomplete views. We set the missing rate (PER) from 0 (each view is complete) to 0.5 (each view has 50% samples missing) with 0.1 as an interval.

Evaluation Metric: Following [29], we evaluate the experimental results by three popular metrics: Accuracy (ACC), Normalized Mutual Information (NMI) and Purity. For these metrics, the larger value represents better performance.

C. Clustering Performance and Analysis

Fig. 3–7 show the clustering results on these real-world datasets. Obviously, our proposed V³H significantly performs better than other state-of-the-art methods in most cases. Especially, when we cluster the YaleB dataset with PER = 0.2 (Fig. 7(g), 7(h), and 7(i)), V³H gains large improvements at least 23.24% in ACC, 22.42% in NMI, and 22.46% in Purity over the best performing compared method PVC.

For convenience, we first divide the compared methods into 3 groups: single-view methods (CK and CS), two-view methods (IMG and PVC) and multiview methods (CSMSC, DAIMC,

MIC, MultiNMF and UEAF). Then, based on the experimental results, we compare and analyze the performance of different groups. Finally, we critically analyze each method.

V³H versus single-view methods: compared with single-view methods, V³H achieves better performance on all the datasets. For instance, when clustering the Scene dataset with PER=0.5 (Fig. 7(a), 7(b), and 7(c)), compared with CK and CS, V³H raises the clustering results at least 31.10% in ACC, 25.73% in NMI, and 30.54% in Purity, respectively. It is because the Scene dataset has 4 views. CK and CS only simply concatenate these views, which cannot learn the relationship information between different views. On the contrary, V³H can extract the relationship information by integrating different views. Therefore, integrating effectively different views is necessary for multiview clustering.

V³H versus two-view methods: compared with two-view methods, V³H obtains better clustering results in all the cases. For the datasets with more than 2 views (e.g., BS (4v), 3-Sources, 20-NGs, etc.), V³H performs better than PVC and IMG. When we cluster the BS (4v) dataset with PER=0.5 (Fig. 5(a), 5(b) and 5(c)), compared with PVC and IMG, V³H raises the clustering results at least 24.14% in ACC, 43.10% in NMI, and 34.47% in Purity, respectively. The reason is that PVC and IMG can
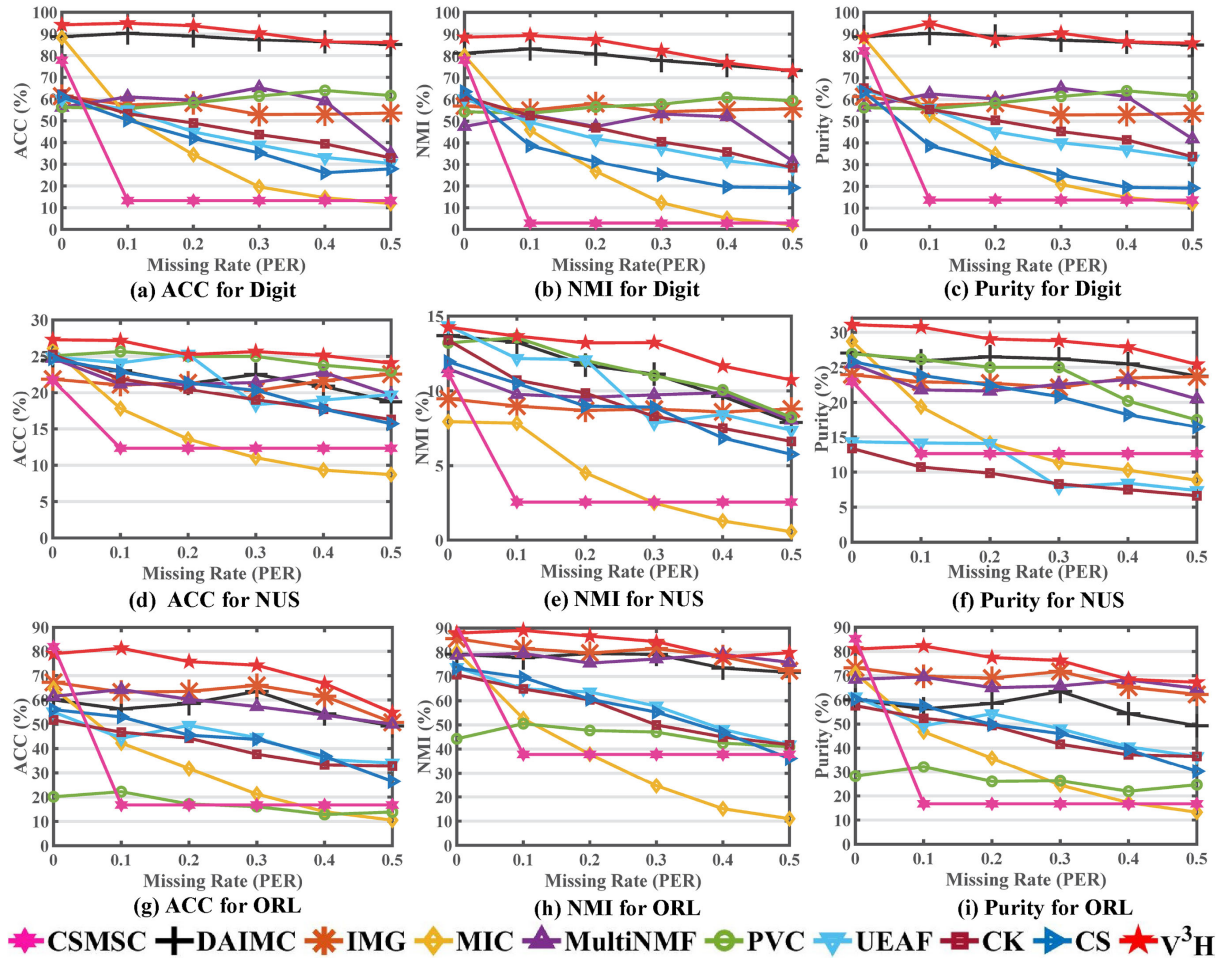
Fig. 6.   Results on Digit, NUS, and ORL datasets.

only integrate two views, and structural information of the rest views is not learned, which illustrates the necessity of integrating all the views. As for the two-view datasets (e.g., BUAA and BS (2v)), V³H still outperforms PVC and IMG. For the BUAA dataset with PER=0.2 (Fig. 5(d), 5(e) and 5(f)), compared with PVC and IMG, V³H improves the performance at least 3.51% in ACC, 1.07% in NMI, and 5.97% in Purity, respectively. The main reason is that the BUAA dataset contains the nonlinear structural information, V³H can learn the information based on the linear kernel used in the Laplacian for spectral clustering.

V³H versus multiview methods: compared with multiview methods, V³H also achieves better clustering performance in most cases. When we cluster the BBC (4v) dataset with PER=0.5 (Fig. Fig. 4(d), 4(e), and 4(f)), compared with multiview methods, V³H raises the performance at least 25.84% in ACC, 18.95% in NMI, and 22.92% in Purity, respectively. It is because each view of the BBC (4v) dataset includes the unique information. V³H can learn the unique information through the corresponding variation matrix, while the unique information is ignored by multiview methods. More impressively, as PER on the dataset increases, V³H achieves satisfactory and relatively stable clustering results, while the performance of all multiview methods drops significantly. It is because based on the subspace decomposition in Eq. (4), V³H can learn the information from

both the presented samples and the missing samples. But these multiview methods can only learn information from the presented samples.

In summary, all the methods are analyzed as follows:

1) **CK and CS**: for most clustering tasks, the performance of CK and CS are close because they simply concatenate all views. Although this concatenation is easy to operate, CK and CS always perform poorly due to ignoring the relationship between different views. Therefore, CK and CS often have limited applications in multiview datasets.

2) **PVC and IMG**: for the incomplete two-view datasets, PVC can obtain pretty clustering performance by establishing a latent subspace from two views. Similarly, IMG also performs well in incomplete two-view clustering by introducing manifold learning into PVC. But when clustering the data with more than 2 incomplete views, PVC and IMG cannot obtain an optimal latent subspace due to ignoring the global structure of the multiview data. Thus, it is difficult for them to obtain satisfactory results in incomplete multiview clustering tasks.

3) **CSMSC, DAIMC, MIC, MultiNMF, and UEAF**: when clustering complete multiview datasets, CSMSC, MIC, and MultiNMF can perform well by learning a subspace from each view. As the missing rate increases, the clustering results of these three methods drop significantly. The reason is that these three
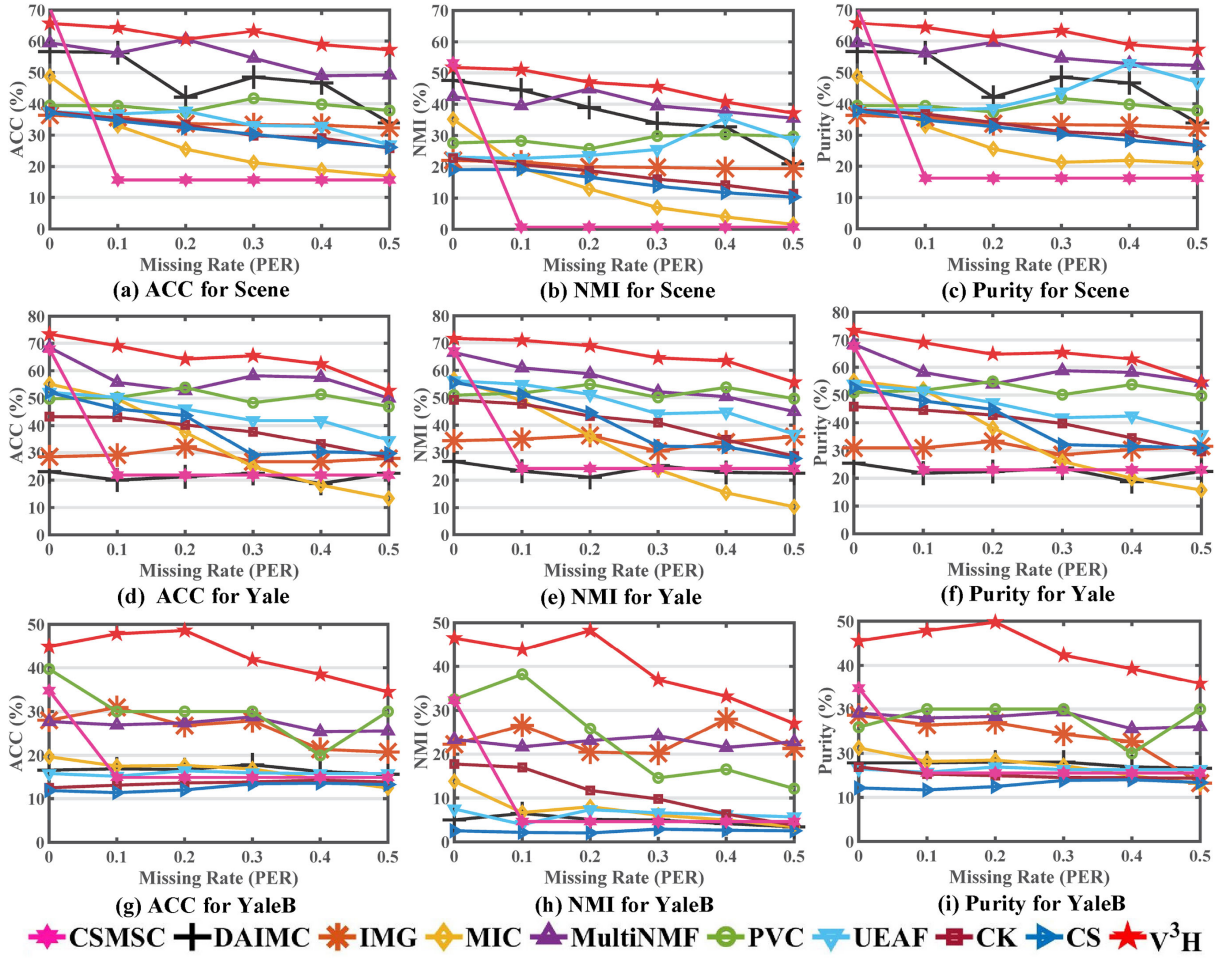
Fig. 7. Results on Scene, Yale, and YaleB datasets.

methods simply fill the missing samples with the average feature values, which neglects the hidden information of the missing samples. Since real-world data are often incomplete, these three methods are difficult to be widely used. For these datasets with little alignment information, DAIMC and UEAF always obtain unsatisfactory clustering results because they learn the consensus representation by aligning these views. These drawbacks limit the application of these methods.

4) **Our proposed V³H**: by aligning cluster indicator matrices from different views and learning the low-rank representation, V³H can perform satisfactorily in most cases, which shows its wide application. Moreover, when the missing rate is relatively large (e.g., PER=0.4 or PER=0.5), V³H has more obvious superiority over other state-of-the-arts, which illustrates its effectiveness in high-incompleteness applications.

D. Parameter Sensitivity

In terms of $\{\alpha, \beta, \gamma\}$, we conduct the hyper-parameter experiments on 3-Sources dataset. Similar to [27], we set PER=0.5 and report V³H's NMI versus α, β, and γ within the set of $\{10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$.

As shown in Fig. 8(a) and 8(b) our proposed V³H obtains stable and satisfactory clustering performance across a wide

range of these parameters. Thus, V³H is insensitive to the variation of the parameters. Also, V³H obtains the best clustering results when we set $\alpha=10^{-3}$, $\beta=10^{-4}$ and $\gamma=10^{-1}$, which are the recommended values.

E. Convergence Study

Based on the recommended values of these hyper-parameters, we study the convergence by conducting the experiments on the Coil dataset with different PERs, i.e., PER=0.1, PER=0.3, PER=0.5. Fig. 8(c) show that the convergence curve versus the iteration number, and "Obj fun val" represents "objective function value," which is calculated by $(||\boldsymbol{M}||_\eta + \sum_v(\beta||\boldsymbol{E}^{(v)}||_\tau + \alpha\mathrm{Tr}(\boldsymbol{F}^{(v)^T}\boldsymbol{L}_N^{(v)}\boldsymbol{F}^{(v)}) - \gamma\mathrm{Tr}(\boldsymbol{F}^{(v)}\boldsymbol{F}^{(v)^T}\boldsymbol{F}^*\boldsymbol{F}^{*^T}))) / (||\boldsymbol{X}^{(v)} - \boldsymbol{X}^{(v)}\boldsymbol{Z}^{(v)} - \boldsymbol{E}^{(v)}||_F^2 + ||\boldsymbol{W}^{(v)^T}\boldsymbol{Z}^{(v)}\boldsymbol{W}^{(v)} - \boldsymbol{M} + p\boldsymbol{N}^{(v)}||_F^2)$, similar to [65]. Obviously, our proposed V³H has converged just after 10 iterations for all PERs, which shows its fast convergence.

Note that the convergence curves under different PERs are close to each other. This is because, based on Eq. (4), we can learn the information of all samples (including presented samples and missing samples). When PER changes, the dimensions of $\boldsymbol{M}$ and $\boldsymbol{N}^{(v)}$ do not change.

(a) Study for α and β (b) Study for γ (c) Convergence study on Coil
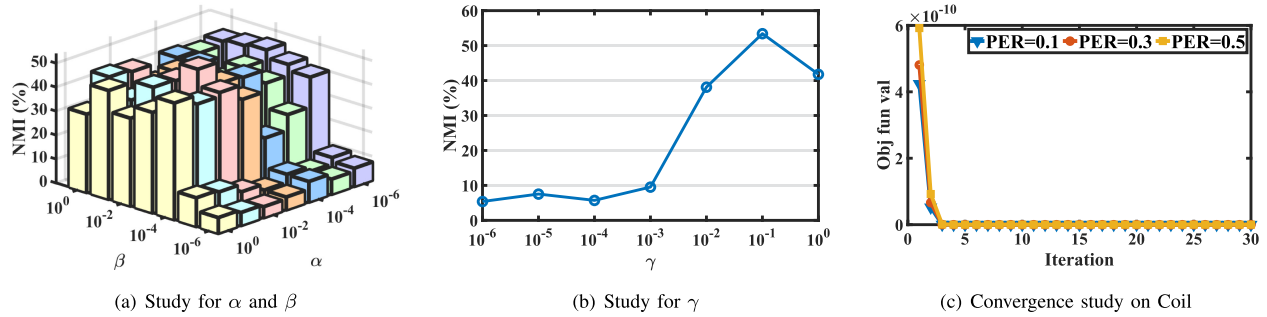
Fig. 8.   Convergence and parameter studies.

Besides, when the iteration number is the same, $obj(PER = 0.5) > obj(PER = 0.3) > obj(PER = 0.1)$. The reason is as follows: in the objective function, only the dimensions of $\boldsymbol{E}$ and $\boldsymbol{Z}^{(v)}$ will change as the missing rate increases. Thus, only the values of $||\boldsymbol{E}^{(v)}||_\tau$ and $||\boldsymbol{X}^{(v)} - \boldsymbol{X}^{(v)}\boldsymbol{Z}^{(v)} - \boldsymbol{E}^{(v)}||_F^2$ will rely on missing rate. In fact, for a robust algorithm, its error matrix $\boldsymbol{E}$ will generally be small. Therefore, we can approximate the numerator of obj as a constant. As the missing rate increases, the value of $||\boldsymbol{X}^{(v)} - \boldsymbol{X}^{(v)}\boldsymbol{Z}^{(v)} - \boldsymbol{E}^{(v)}||_F^2$ will decrease, and the objective function value will increase.

VI. CONCLUSION

In this paper, we propose a novel **V**iew **V**ariation and **V**iew **H**eredity approach (V^{3}H) for incomplete multiview clustering. As far as we know, V^{3}H is the **first** attempt to introduce genetics into the clustering method. Also, it can learn the consistent information and the unique information based on view variation and view heredity respectively. Extensive experiments on fifteen datasets demonstrate the superiority of V^{3}H over other state-of-the-art methods. Impressively, when clustering the YaleB dataset with the missing rate of 0.2, V^{3}H improves at least 23.24% in ACC, 22.42% in NMI, and 22.46% in Purity over the best performing compared method.

Our proposed V^{3}H is an offline approach for high-dimensional incomplete multiview clustering. A larger challenge is to cluster large-scale high-dimensional data. In the future, we will introduce online learning into V^{3}H for the large-scale high-dimensional data about COVID-19. We collect a large amount of data about COVID-19 every day, and online learning is an effective way to process these data. Based on online learning, we will process these data.

REFERENCES

[1] S. Luo, C. Zhang, W. Zhang, and X. Cao, "Consistent and specific multi-View subspace clustering," *Assoc. Adv. Artif. Intell.*, 2018.

[2] X. Li, H. Zhang, R. Wang, and F. Nie, "Multi-view clustering: A scalable and parameter-free bipartite graph fusion method,"*IEEE Trans. Pattern Anal. Mach. Intell.*, early access, doi: 10.1109/TPAMI.2020.3011148.

[3] X. Liu *et al.*, "Efficient and effective regularized incomplete multi-view clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be Published, doi: 10.1109/TPAMI.2020.2974828.

[4] G. Chao, S. Sun, and J. Bi, "Survey on multi-view clustering," 2018, *arXiv:1712.06246*.

[5] G. Chao *et al.*, "Multi-view cluster analysis with incomplete data to understand treatment effects," *Inf. Sci.*, vol. 494, pp. 278–293, 2019, doi: 10.1016/j.ins.2019.04.039.

[6] A. Remuzzi and G. Remuzzi, "Covid-19 and Italy: what next?" *Lancet*, vol. 395, no. 10231, pp. 1225–1228, Mar. 2020.

[7] Y. Bai *et al.*, "Presumed asymptomatic carrier transmission of COVID-19," *Jama*, vol. 323, no. 14, pp. 1406–1407, 2020, doi: 10.1001/jama.2020.2565.

[8] H. A. Rothan and S. N. Byrareddy, "The epidemiology and pathogenesis of coronavirus disease (covid-19) outbreak," *J. Autoimmun.*, vol. 109, 2020, Art. no. 102433.

[9] A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, and D. Veesler, "Structure, function, and antigenicity of the SARS-COV-2 spike glycoprotein," *Cell*, vol. 181, no. 2, pp. 281–292, 2020.

[10] W. Wang *et al.*, "Detection of SARS-CoV-2 in different types of clinical specimens," *Jama*, vol. 323, no. 18, pp. 1843–1844, 2020.

[11] L. Zou *et al.*, "Sars-cov-2 viral load in upper respiratory specimens of infected patients," *New England J. Med.*, vol. 382, no. 12, pp. 1177–1179, 2020.

[12] R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, and Q. Zhou, "Structural basis for the recognition of SARS-COV-2 by full-length human ace2," Sci., pp. 1444–1448, 2020.

[13] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proc. Int. Conf. Data Mining Ser.*, 2004, pp. 19–26.

[14] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 129–136.

[15] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 2598–2604.

[16] M.-S. Chen, L. Huang, C.-D. Wang, and D. Huang, "Multi-View clustering in latent embedding space," *Assoc. Adv. Artif. Intell.*, 2020.

[17] F. Nie, S. Shi, and X. Li, "Auto-weighted multi-view co-clustering via fast matrix factorization," *Pattern Recognit.*, 2020, Art. no. 107207.

[18] J. Han, J. Xu, F. Nie, and X. Li, "Multi-view k-means clustering with adaptive sparse memberships and weight allocation," *IEEE Trans. Knowl. Data Eng.*, early acess, doi: 10.1109/TKDE.2020.2986201.

[19] Z. Yang, N. Liang, W. Yan, Z. Li, and S. Xie, "Uniform distribution non-negative matrix factorization for multiview clustering." *IEEE Trans. Cybern.*, early access, doi: 10.1109/TCYB.2020.2984552.

[20] H. Zhao, H. Liu, and Y. Fu, "Incomplete multi-modal visual data grouping." in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2392–2398.

[21] N. Xu, Y. Guo, X. Zheng, Q. Wang, and X. Luo, "Partial multi-view subspace clustering," in *ACM Multimedia*, 2018, pp. 1794–1801.

[22] W. Shao, L. He, C. Lu, and P. S. Yu, "Online multi-view clustering with incomplete views," *Big Data*, pp. 1012–1017, Dec. 2016, doi: 10.1109/BigData.2016.7840701.

[23] S.-Y. Li, Y. Jiang, and Z.-H. Zhou, "Partial multi-view clustering," *Assoc. Adv. Artif. Intell.*, 2014.

[24] W. Shao, L. He, and S. Y. Philip, "Multiple incomplete views clustering Via weighted nonnegative matrix factorization with $l_{2,1}$ regularization," in *Proc. Euro. Conf. Mach. Learn. Princ. Pract. Knowl. Dis. Databases.*, 2015, pp. 318–334.

[25] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proc. SIAM Ind. Conf. Data Mining*, 2013, pp. 252–260.

[26] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, pp. 788–791, 1999.

[27] M. Hu and S. Chen, "Doubly aligned incomplete multi-view clustering." *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 2262–2268.

[28] C. H. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.

[29] J. Wen, Z. Zhang, Y. Xu, B. Zhang, L. Fei, and H. Liu, "Unified embedding alignment with missing views inferring for incomplete multi-View clustering," *Assoc. Adv. Artif. Intell.*, 2019.

[30] T. Dobzhansky and T. G. Dobzhansky, *Genetics of the Evolutionary Process*. New York, NY, USA: Columbia Univ. Press, 1970.

[31] E. Mayr, *Populations, Species, and Evolution: An Abridgment of Animal Species and Evolution*. Cambridge, MA, USA: Harvard Univ. Press, 1970, vol. 19.

[32] P. Hedrick, Genetics of Populations. Jones & Bartlett Learning, 2011.

[33] D. Cai, X. He, and J. Han, "Sparse projections over graph." in *AAAI*, 2008, pp. 610–615.

[34] H. Gao, F. Nie, X. Li, and H. Huang, "Multi-view subspace clustering," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 4238–4246.

[35] R. Willham, "The covariance between relatives for characters composed of components contributed by related individuals," Biometrics, pp. 18–27, 1963.

[36] J. Zhu and B. Weir, "Analysis of cytoplasmic and maternal effects i. a genetic model for diploid plant seeds and animals," *Theor. Appl. Genet.*, pp. 153–159, 1994.

[37] P. M. Visscher, W. G. Hill, and N. R. Wray, "Heritability in the genomics era - concepts and misconceptions," *Nat. Rev. Genet.*, pp. 255–266, 2008.

[38] J. Guo and J. Ye, "Anchors bring ease: An embarrassingly simple approach to partial multi-view clustering," *Assoc. Adv. Artif. Intell.*, 2019, pp. 118–125.

[39] L. L. Cavalli-Sforza and W. F. Bodmer, *Genetics of Human Populations*. Courier Corporation, 1999.

[40] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nat. Rev. Genet.*, pp. 85–97, 2006.

[41] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Proc. Conf. Neural Infor. Process. Syst.*, 2011, pp. 1413–1421.

[42] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, Mar. 2011.

[43] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[44] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. Altern. Complement. Med.*, pp. 1–37, 2011.

[45] H. He and X. Shen, "Ranked Subspace Learning Method for Gene Expression Data Classification," in *Proc. Int. Conf. Artif. Intell.*, 2007, pp. 25–28.

[46] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," in *Proc. Conf. Neural Infor. Process. Syst.*, 2010, pp. 2496–2504.

[47] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[48] lldiko E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, pp. 109–135, 1993.

[49] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Stat. Assoc.*, pp. 1348–1360, 2001.

[50] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Process.*, vol. 4, no. 7, pp. 932–946, Jul. 1995.

[51] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2862–2869.

[52] Z. Kang, C. Peng, and Q. Cheng, "Robust pca via nonconvex rank approximation," in *Proc. Int. Conf. Data Mining*, 2015, pp. 211–220.

[53] P. D. Tao and L. T. H. An, "Convex analysis approach to Dc programming: Theory, algorithms and applications," *Acta Math. Vietnamica*, pp. 289–355, 1997.

[54] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2227–2240, Nov. 2014.

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

[56] S. A. NENE, "Columbia object image library(coil-20)," Tech. Rep., 1996.

[57] T. S. Chua, J. Tang, R. Hong, H. Li, and Z. Luo, "Nus-Wide: A real-world web image database from National University of Singapore," in *Proc. Conf. Image Video Retrieval*, 2009, pp. 1–9.

[58] A. C. Harter, "Harter, A: Parameterisation of a stochastic model for human face identification," *Proc. Workshop Appl. Comput. Vis.*, 1994, pp. 138–142.

[59] A. Monadjemi, B. Thomas, and M. Mirmehdi, "Experiments on High Resolution Images Towards Outdoor Scene Classification," May 2002.

[60] D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[61] W. Weng, W. Zhou, J. Chen, H. Peng, and H. Cai, "Enhancing multiview clustering through common subspace integration by considering both global similarities and local structures," *Neurocomputing*, pp. 375–386, 2020.

[62] H. Wang, Y. Yang, and B. Liu, "Gmc: Graph-based multi-view clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 6, pp. 1116–1129, Jun. 2020.

[63] S. Zhou *et al.*, "Multiple kernel clustering with neighbor-kernel subspace segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 4, pp. 1351–1362, Apr. 2020.

[64] X. Wang, X. Guo, Z. Lei, C. Zhang, and S. Z. Li, "Exclusivity-Consistency Regularized Multi-View Subspace Clustering," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1–9.

[65] J. Wen, Y. Xu, and H. Liu, "Incomplete multiview spectral clustering with adaptive graph learning," *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1418–1429, Apr. 2020.

Xiang Fang received the M.E. degree from the School of Computer Science and Technology, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2020. He is currently a Research Assistant with the School of Computer Science and Technology, HUST. His research interests include image processing, pattern recognition, data mining, and machine learning.
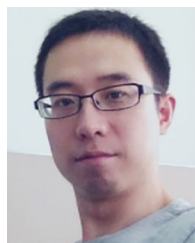
Yuchong Hu (Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2010. He is currently an Associate Professor with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China. His research interests include fault tolerance and the repair and read or write performance of storage systems.

Pan Zhou (Senior Member, IEEE) received the Ph.D. degree from the Georgia Institute of Technology, Atlanta, GA, USA, in 2011. He is currently a Full Professor and a Ph.D. Advisor with the School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan, China. His research interests include security and privacy, big data analytics, machine learning, and information networks. He is currently an Associate Editor for the IEEE Transactions on Network Science and Engineering.

Dapeng Oliver Wu (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2003. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA. His research interests include networking, communications, signal processing, computer vision, machine learning, smart grid, and information and network security.