Marginalized Multiview Ensemble Clustering

Zhiqiang Tao, Hongfu Liu, *Member, IEEE*, Sheng Li, *Senior Member, IEEE*, Zhengming Ding, *Member, IEEE*, and Yun Fu, *Fellow*, *IEEE*

***Abstract*— Multiview clustering (MVC), which aims to explore the underlying cluster structure shared by multiview data, has drawn more research efforts in recent years. To exploit the complementary information among multiple views, existing methods mainly learn a common latent subspace or develop a certain loss across different views, while ignoring the higher level information such as basic partitions (BPs) generated by the single-view clustering algorithm. In light of this, we propose a novel marginalized multiview ensemble clustering (M^2VEC) method in this paper. Specifically, we solve MVC in an EC way, which generates BPs for each view individually and seeks for a consensus one. By this means, we naturally leverage the complementary information of multiview data upon the same partition space. In order to boost the robustness of our approach, the marginalized denoising process is adopted to mimic the data corruptions and noises, which provides robust partition-level representations for each view by training a single-layer autoencoder. A low-rank and sparse decomposition is seamlessly incorporated into the denoising process to explicitly capture the consistency information and meanwhile compensate the distinctness between heterogeneous features. Spectral consensus graph partitioning is also involved by our model to make M^2VEC as a unified optimization framework. Moreover, a multilayer M^2VEC is eventually delivered in a stacked fashion to encapsulate nonlinearity into partition-level representations for handling complex data. Experimental results on eight real-world data sets show the efficacy of our approach compared with several state-of-the-art multiview and EC methods. We also showcase our method performs well with partial multiview data.**

***Index Terms*— Multiview clustering, ensemble clustering, low-rank representation, auto-encoders.**

I. Introduction

Multiview data are widely used throughout the fields of machine learning, data mining, and computer vision, which mainly refer to the data collected from multiple sources, captured by various sensors or represented with different feature descriptors. For instance, the same news story is usually reported by different articles, human activity could be recorded by RGB video and depth camera, and images are encoded with kinds of hand-crafted and deep features. As multiview data always provide complementary information among different views, it has attracted great research efforts on many tasks, such as time series classification [1], subspace learning [2], dimensionality reduction [3]–[5], outlier detection [6], [7], and cross-domain adaptation [8]. Among these interesting topics, we devote to solve multiview clustering (MVC) problem [9] in this paper.

MVC aims to discover the underlying cluster structure shared by all the views, where the key problem is to exploit the complementary information among multiview data. To this end, most existing methods either develop a certain loss [9]–[11] to fuse multiview data during the clustering process or learn a common latent space [12]–[14] to explore the consistency information across views before the clustering. Although these methods have achieved effective performance, they mainly perform clustering with raw multiview data, yet ignore to utilize the higher level information to bridge the distinct gap between heterogeneous feature spaces/domains. Hence, one promising way [15], [16] is to transform multiview data into the same partition space and finally solve MVC via cluster ensembles.

Ensemble clustering (EC) [17]–[19] integrates multiple basic partitions (BPs) as the consensus clustering result, which naturally has the ability of leveraging complementary information from heterogeneous sources [20]. However, there exist two limitations of existing EC methods to handle multiview data. First, generic EC methods treat each BP equally, which neglects to explicitly consider the connection between different views. Second, the disagreements and outliers among multiple BPs can heavily mislead the clustering process. Thus, it is not reasonable to directly adopt EC for tackling the MVC problem.

To address the above-mentioned challenges, we propose a novel marginalized multiview EC (M^2VEC) algorithm (see Fig. 1) in this paper. Specifically, our method takes as input a set of view-specific coassociation matrices, each of which is summarized from BPs in an individual view and works as a pairwise affinity matrix [18]. To alleviate the "noises" (i.e., disagreements and outliers) in BPs, marginalized denoising autoencoder (mDA) [21] is leveraged to deliver partition-level representations for each view. A low-rank and sparse decomposition is seamlessly incorporated into the denoising process to seek for the consensus representation

Manuscript received January 18, 2018; revised December 16, 2018; accepted March 11, 2019. This research is supported in part by the NSF IIS Award 1651902 and U.S. Army Research Office Award W911NF-17-1-0367. *(Corresponding author: Zhiqiang Tao.)*

Z. Tao is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: zqtao@ece.neu.edu).

H. Liu is with the Michtom School of Computer Science, Brandeis University, Waltham, MA 02453 USA (e-mail: hongfuliu@brandeis.edu).

S. Li is with the Department of Computer Science, University of Georgia, Athens, GA 30602 USA (e-mail: sheng.li@uga.edu).

Z. Ding is with the Department of Computer, Information and Technology, Indiana University-Purdue University Indianapolis, Indianapolis IN 46202 USA (e-mail: zd2@iu.edu).

Y. Fu is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA, and also with the Khoury College of Computer and Information Sciences, Northeastern University, Boston, MA 02115 USA (e-mail: yunfu@ece.neu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2906867

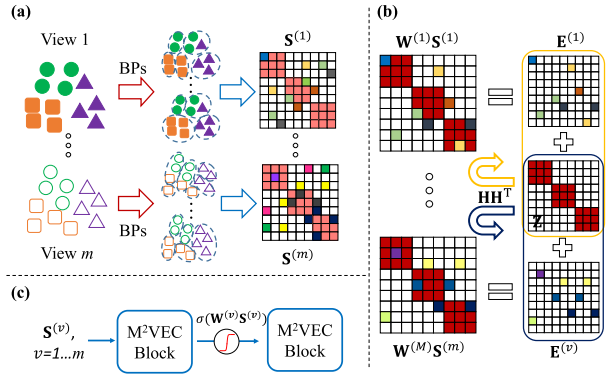


Fig. 1. Illustration of the proposed M^{2}VEC framework. (a) Multiview BPs generation. $\mathbf{S}^{(v)}$ denotes the coassociation matrix of the vth view. (b) Single-layer M^{2}VEC model which jointly learns marginalized denoiser $\mathbf{W}^{(v)}$, low-rank representation $\mathbf{Z}$, consensus partition $\mathbf{H}$, and sparse residual matrix $\mathbf{E}^{(v)}$. (c) Example of multilayer M^{2}VEC built with two blocks.

shared by all the views and meanwhile compensate the distinctness between heterogeneous features. Moreover, spectral graph partitioning is also involved by our model with a carefully designed constraint, which eventually makes the proposed M^{2}VEC as a unified optimization framework. By using a stacked fashion, a multilayer M^{2}VEC is developed to perform marginalized denoising from coarse to fine, and thus provides more robust and rich representations for the clustering task.

A. Related Work

MVC methods could be roughly divided into three categories: 1) cotraining; 2) common subspace; and 3) late fusion based methods. Specifically, cotraining-based methods [9], [10], [22] alternatively maximize the mutual consistency across two distinct views, and usually design a certain loss to directly perform clustering with multiview features. For example, a coregularized constraint is provided in [10] for the clustering purpose. Unlike the first category of methods, which combine multiview features during the clustering process, the common subspace-based methods [12]–[14] generally target to learn a latent low-dimensional subspace from multiple views simultaneously, and conduct existing clustering algorithms on the learned common representations for the final result. This kind of methods could be further classified as canonical correlation analysis (CCA)-based [12], [23], nonnegative matrix factorization (NMF)-based [13], [24] and low-rank subspace-based ones [14], [25], [26]. The late fusion approaches [15], [16] try to solve MVC through fusing BPs of multiple views. Nevertheless, these early attempts neglect to fully utilize the connection across multiview data. In addition, multiview weight learning [27], [28] is also an effective direction for solving the MVC task.

EC [17] has been an important alternative to the traditional clustering task, where utility function [19] and coassociation matrix [18] are the two representative directions. The utility-function-based methods directly measure the consensus between BPs with a predefined utility function. For example, Wu *et al.* [20] provided a family of K-means (KM)-based utility functions and linked the consensus clustering to a KM problem. On the other hand, the coassociation matrix-based methods compute an affinity matrix upon BPs and transform EC as a graph partitioning problem. Along this line, lots of efforts have been made, such as hierarchical consensus clustering [18] and spectral EC (SEC) [29]. In addition, some other representative EC methods include the framework using NMF [30], linked-based [31], bipartite graph [32], and wisdom-of-crowds [33]. Two recent works [34], [35] also learn robust representations from BPs to boost the EC performance, where [34] imposes a low-rank constraint on the coassociation matrix, and [35] feeds BPs into the stacked mDAs. However, these two methods focus on the generic EC problem, which are not specifically designed for multiview data.

B. Motivation

MVC and EC share a similar motivation, which both target at boosting the performance by fusing information from multiple sources. Compared with traditional MVC methods, EC provides higher level information (i.e., coassociation matrix) to handle multiview data, which bridges the gap among distinct feature spaces. However, it is not straightforward to directly solve MVC problem as an EC task, since the latter treats each BP equally yet without explicitly considering the complementary information among different views. Hence, it is more reasonable to learn a consensus representation shared by coassociation matrices from multiple views. On the other hand, each coassociation matrix suffers from the outliers and disagreements among BPs [34] due to the nature of unsupervised learning. Thus, how to obtain robust representations for multiview coassociation matrices also remains challenging.

The single-layer M^{2}VEC [Fig. 1(b)] is proposed to handle the above-mentioned challenges. Inspired by the stacked denoising autoencoders [36], [37], a multilayer M^{2}VEC [Fig. 1(c)] is also provided with the following reasons. 1) Stacking multiple M^{2}VEC blocks enables to perform marginalized denoising process from coarse to fine, which provides robust partition-level feature representations for each view. 2) By going deeper, we explore the consensus information among multiple views layer by layer, and thus build more discriminative representations shared by all the views than single layer [38], [39].

C. Our Contributions

This paper is a substantial extension of our previous work [40]. Compared with [40], which mainly learns a common low-rank representation shared by multiview partitions, we incorporate an mDA into our model to further boost the robustness. Specifically, we jointly perform *infinity* denoising and view-consensus representation learning, and extend our model to a multilayer architecture. More theoretical analyses, model discussions, and experimental evaluations are also provided. Moreover, we showcase our approach can work well with partial multiview data.

The contributions of this paper are highlighted in four folds.

1) A novel M^{2}VEC model is proposed to exploit the higher level information of multiview data for the clustering task.
2) A marginalized denoiser is leveraged by our model to deliver robust partition-level representation of each view.
3) We provide a unified optimization framework to jointly learn the marginalized denoiser, low-rank representation, and consensus partition among multiple views.
4) By using the stacked strategy, multilayer M^{2}VEC is developed to obtain robust and rich representations in a *deep* fashion.

The remainder of this paper is organized as follows. The proposed M^{2}VEC algorithm is elaborated in Section II from single layer to multilayer. Extensive experimental results and discussions are reported in Section III, and a final conclusion is given by Section IV.

II. Methodology

In this section, we first introduce some preliminary knowledge of our work, and then present the proposed M^{2}VEC method and its optimization solution. After that, we give the architecture of our multilayer model.

A. Preliminary

1) Problem Formulation: Given a set of n data points with m views (i.e., feature representations or modalities), we denote the data set of each view as $\mathcal{X}^{(v)} = \{x_1^{(v)}, \ldots, x_n^{(v)}\}$, $1 \le v \le m$. For any v, we assume $\mathcal{X}^{(v)}$ is sampled from K crispy clusters, denoted as $\mathcal{C} = \{C_1, \ldots, C_K\}$. Let $\Pi^{(v)} = \{\pi_1^{(v)}, \ldots, \pi_r^{(v)}\}$ be a group of r BPs for $\mathcal{X}^{(v)}$, where each BP $\pi_i^{(v)}$ partitions $\mathcal{X}^{(v)}$ into K_i clusters, i.e., $\pi_i^{(v)} = \{\pi_i^{(v)}(x_1^{(v)}), \ldots, \pi_i^{(v)}(x_n^{(v)})\}$ is a set of categorical data, $1 \le \pi_i^{(v)}(x_j^{(v)}) \le K_i$, $1 \le i \le r$, and $1 \le j \le n$. It is worthy to note that, K_i is set to be different from K (e.g., $K \le K_i \le \sqrt{n}$) to ensure the diversity among multiple BPs, which has shown to be an effective way to uncover various cluster structures [18], [41]. To exploit the higher level information from BPs, each $\Pi^{(v)}$ is summarized to be a coassociation matrix $\mathbf{S}^{(v)} \in \mathbb{R}^{n\times n}$ as

$$\mathbf{S}^{(v)}\big(x_p^{(v)}, x_q^{(v)}\big) = \frac{1}{r}\sum_{i=1}^{r}\delta\big(\pi_i^{(v)}\big(x_p^{(v)}\big), \pi_i^{(v)}\big(x_q^{(v)}\big)\big) \tag{1}$$

where $x_p^{(v)}, x_q^{(v)} \in \mathcal{X}^{(v)}$, and $\delta(a, b) = 1$ if $a = b$; 0 otherwise. Equation (1) computes the pairwise affinity upon categorical data. By this means, $\mathbf{S}^{(v)}$ could be used as the vth view's feature representations, which encode the relationship between data points in a partition space. In this paper, we aim to reveal the cluster structure shared by multiview data in an EC way, by taking as input a set of coassociation matrices (i.e., $\mathbf{S}^{(1)} \ldots \mathbf{S}^{(m)}$).

Remark 1: For each individual view, the coassociation matrix encodes higher lever information than raw features, as it summarizes BPs obtained from features. The benefits of using coassociation matrix lies at handling sample variations [20]; preserving local similarity [42]; and capturing various cluster structures [18]. In particular, by using coassociation matrix, our approach naturally transforms multiview data into the same partition space, which bridges the gap between heterogeneous feature spaces of different views.

2) Marginalized Denoising Autoencoders: By reconstructing noisy samples to the original ones, stacked denoising autoencoders method [36] has shown great success to learn robust feature representation in many fields, such as domain adaptation and cluster analysis. Generally, more corrupted samples yield more stable and better performance, which however, inevitably burdens the training process. In light of this, the mDA [21], [37] is employed to mimic the *infinity* data corruption process and achieve robust feature representations in a highly efficient way. For any v, let $\bar{\mathbf{S}}^{(v)} = [\mathbf{S}^{(v)} \ldots \mathbf{S}^{(v)}]$ be the composited samples by repeating $\mathbf{S}^{(v)}$ τ times and $\tilde{\mathbf{S}}^{(v)}$ be the corrupted samples corresponding to $\bar{\mathbf{S}}^{(v)}$. A single-layer mDA is given by

$$\min_{\mathbf{W}^{(v)}} \mathrm{tr}[(\bar{\mathbf{S}}^{(v)} - \mathbf{W}^{(v)}\tilde{\mathbf{S}}^{(v)})^{\mathrm{T}}(\bar{\mathbf{S}}^{(v)} - \mathbf{W}^{(v)}\tilde{\mathbf{S}}^{(v)})] \tag{2}$$

where $\mathbf{W}^{(v)} \in \mathbb{R}^{n\times n}$ is in essence a linear feature transformation matrix, and τ is expected to be $\tau \to \infty$. Note that when τ is set as a specific number, (2) performs the denoising process τ times and works as ordinary least square problems [43], which enjoys a closed-form solution. Nevertheless, when $\tau \to \infty$, mDA marginalizes the corruption process, and gives the solution as

$$\mathbf{W}^{(v)} = \mathbb{E}[\mathbf{P}^{(v)}]\mathbb{E}[\mathbf{Q}^{(v)}]^{-1} \tag{3}$$

where $\mathbf{P}^{(v)} = \tilde{\mathbf{S}}^{(v)}\tilde{\mathbf{S}}^{(v)\mathrm{T}}$ and $\mathbf{Q}^{(v)} = \tilde{\mathbf{S}}^{(v)}\tilde{\mathbf{S}}^{(v)\mathrm{T}}$. Following [21], $\mathbb{E}[\mathbf{P}^{(v)}]$ and $\mathbb{E}[\mathbf{Q}^{(v)}]$ are computed by

$$\begin{aligned}\mathbb{E}[\mathbf{P}^{(v)}] &= (\mathbf{S}^{(v)}\mathbf{S}^{(v)\mathrm{T}}) \otimes \Theta_{\mathbf{P}}\\ \mathbb{E}[\mathbf{Q}^{(v)}] &= (\mathbf{S}^{(v)}\mathbf{S}^{(v)\mathrm{T}}) \otimes \Theta_{\mathbf{Q}}\end{aligned} \tag{4}$$

where $\Theta_{\mathbf{P}} = (1-p) \times \mathbf{1}^{n\times n}$, $\Theta_{\mathbf{Q}} = \Theta_{\mathbf{P}} \otimes \mathrm{diag}(1-p)$, and $\otimes$ denotes Hadamard product. $\mathrm{diag}(1-p) \in \mathbb{R}^{n\times n}$ represents a diagonal matrix consisting of values $(1-p)$, whereas p is the probability of occurring corruption for each element in $\mathbf{S}^{(v)}$.

B. Marginalized Multiview Ensemble Clustering

In this paper, we propose to learn a consensus representation shared by multiple views via low-rank and sparse decomposition, with jointly performing marginalized denoising and spectral clustering (SC) in a unified optimization framework. We formulate the proposed M^{2}VEC method as

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{E}^{(v)},\mathbf{W}^{(v)},\mathbf{H}} \; & \mathrm{tr}(\mathbf{H}^{\mathrm{T}}\mathbf{L}_z\mathbf{H}) + \lambda_1\|\mathbf{Z}\|_* + \lambda_2\sum_{v=1}^{m}\|\mathbf{E}^{(v)}\|_1\\ & + \frac{\lambda_3}{2}\sum_{v=1}^{m}\mathrm{tr}[(\bar{\mathbf{S}}^{(v)} - \mathbf{W}^{(v)}\tilde{\mathbf{S}}^{(v)})^{\mathrm{T}}\\ & \qquad (\bar{\mathbf{S}}^{(v)} - \mathbf{W}^{(v)}\tilde{\mathbf{S}}^{(v)})]\\ \text{s.t. } & \forall v, \mathbf{W}^{(v)}\mathbf{S}^{(v)} + \mathbf{H}\mathbf{H}^{\mathrm{T}} = \mathbf{W}^{(v)}\mathbf{S}^{(v)}\mathbf{Z} + \mathbf{E}^{(v)},\\ & \mathbf{Z} \ge 0, \mathbf{Z}\mathbf{1} = \mathbf{1}, \mathbf{H}^{\mathrm{T}}\mathbf{H} = \mathbf{I},\end{aligned} \tag{5}$$

where $\mathbf{Z} \in \mathbb{R}^{n\times n}$ represents the consensus low-rank representation, $\mathbf{H} \in \mathbb{R}^{n\times K}$ denotes the partition result, $\mathbf{E}^{(v)} \in \mathbb{R}^{n\times n}$ and $\mathbf{W}^{(v)} \in \mathbb{R}^{n\times n}$ are the sparse residual matrix and feature transformation matrix for the vth view, respectively. In (5), $\bar{\mathbf{S}}^{(v)}$

and $\tilde{\mathbf{S}}^{(v)}$ are the composition of τ times $\mathbf{S}^{(v)}$ and its corruption, $\mathbf{I} \in \mathbb{R}^{n\times n}$ is the identical matrix, $\mathbf{1} \in \mathbb{R}^n$ is the vector of all ones, and $\lambda_1, \lambda_2, \lambda_3 > 0$ are three balancing parameters. $\mathbf{L}_z = \mathbf{D}_z - \mathbf{Z}$ is the Laplacian matrix [44] built on $\mathbf{Z}$, where $\mathbf{D}_z$ is a diagonal matrix consisting of the sum of each row in $\mathbf{Z}$. Following [45], [46], the nuclear norm $\|\mathbf{Z}\|_*$ is employed to measure the rank while the ℓ_1 norm $\|\mathbf{E}^{(v)}\|_1$ is used to characterize the sparseness.

In our model, the marginalized denoising process is conducted for each view to learn a *denoiser* to handle the noises existing in BPs, which provides robust feature representations for the following task. To highlight the same cluster structure shared by different views, we seek for a low-rank representation $\mathbf{Z}$ to reveal the membership between data points through all the views. Meanwhile, to compensate the "conflict" between heterogeneous features, we learn a sparse residual matrix $\mathbf{E}^{(v)}$ for each single view. The partition result $\mathbf{H}$ is obtained by performing spectral graph partitioning on $\mathbf{Z}$, which is characterized as the trace minimization term with $\mathbf{L}_z$.

Taking a close look at (5), a *self-boost* constraint is carefully developed to iteratively enhance the cluster structure of each view's representation, i.e., $\mathbf{W}^{(v)}\mathbf{S}^{(v)} + \mathbf{H}\mathbf{H}^{\mathrm{T}} = \mathbf{W}^{(v)}\mathbf{S}^{(v)}\mathbf{Z} + \mathbf{E}^{(v)}$. Upon this constraint, we first obtain a high-quality consensus partition $\mathbf{H}$ from $\mathbf{Z}$, then in return, we leverage $\mathbf{H}$ to further guide the learning of $\mathbf{Z}$. In addition, the probabilistic simplex constraint ($\mathbf{Z} \geq 0$, $\mathbf{Z}\mathbf{1} = \mathbf{1}$) [47] is also involved in our model to keep the probability property of $\mathbf{Z}$.

C. Optimization

A unified optimization framework that jointly considers three learning tasks and two constraints is provided by (5), which could be divided it into several subproblems and solve them iteratively. In detail, the augmented Lagrange multiplier (ALM) algorithm with alternating direction minimizing (ADM) strategy [48], [49] is applied to address our M^{2}VEC problem. To facilitate the optimization process, we first introduce an auxiliary variable $\mathbf{J} \in \mathbb{R}^{n\times n}$ with $\mathbf{Z} = \mathbf{J}$ to make (5) separable.

Let $\Omega(\mathbf{W}) \equiv \frac{1}{2}\sum_{v=1}^{m} \mathrm{tr}[(\bar{\mathbf{S}}^{(v)} - \mathbf{W}^{(v)}\tilde{\mathbf{S}}^{(v)})^{\mathrm{T}}(\bar{\mathbf{S}}^{(v)} - \mathbf{W}^{(v)}\tilde{\mathbf{S}}^{(v)})]$, then our problem is equivalently converted as

$$\begin{aligned}&\min_{\theta}\ \mathrm{tr}(\mathbf{H}^{\mathrm{T}}\mathbf{L}_z\mathbf{H}) + \lambda_1\|\mathbf{J}\|_* + \lambda_2\sum_{v=1}^{m}\|\mathbf{E}^{(v)}\|_1 + \lambda_3\Omega(\mathbf{W})\\ &\text{s.t. } \forall v, \mathbf{W}^{(v)}\mathbf{S}^{(v)} + \mathbf{H}\mathbf{H}^{\mathrm{T}} = \mathbf{W}^{(v)}\mathbf{S}^{(v)}\mathbf{Z} + \mathbf{E}^{(v)},\\ &\quad \mathbf{Z} = \mathbf{J}, \mathbf{Z} \geq 0, \mathbf{Z}\mathbf{1} = \mathbf{1}, \mathbf{H}^{\mathrm{T}}\mathbf{H} = \mathbf{I},\end{aligned} \tag{6}$$

where $\theta = \{\mathbf{J}, \mathbf{Z}, \mathbf{E}^{(v)}, \mathbf{W}^{(v)}, \mathbf{H}\}$ represents the set of optimization variables. Denoting $\mathbf{R}^{(v)} \equiv \mathbf{S}^{(v)} - \mathbf{S}^{(v)}\mathbf{Z}$, the augmented Lagrange function of (6) is written as

$$\begin{aligned}\mathcal{L} =\ & \mathrm{tr}(\mathbf{H}^{\mathrm{T}}\mathbf{L}_z\mathbf{H}) + \lambda_1\|\mathbf{J}\|_* + \lambda_2\sum_{v=1}^{m}\|\mathbf{E}^{(v)}\|_1 + \lambda_3\Omega(\mathbf{W})\\ &+ \sum_{v=1}^{m}\Phi(\mathbf{W}^{(v)}\mathbf{R}^{(v)} + \mathbf{H}\mathbf{H}^{\mathrm{T}} - \mathbf{E}^{(v)}, \mathbf{Y}^{(v)})\\ &+ \Phi(\mathbf{Z} - \mathbf{J}, \Lambda) + \langle \mathbf{Z}\mathbf{1} - \mathbf{1}, \mathbf{u}\rangle + \frac{\mu}{2}\|\mathbf{Z}\mathbf{1} - \mathbf{1}\|_2^2\end{aligned} \tag{7}$$

where $\mathbf{Z} \geq 0$, $\forall v\ \mathbf{Y}^{(v)} \in \mathbb{R}^{n\times n}$, $\Lambda \in \mathbb{R}^{n\times n}$, and $\mathbf{u} \in \mathbb{R}^n$ refer to Lagrange multipliers, $\Phi(\mathbf{A}, \mathbf{B}) \equiv \langle \mathbf{A}, \mathbf{B}\rangle + \frac{\mu}{2}\|\mathbf{A}\|_{\mathrm{F}}^2$, and $\mu > 0$ denotes a penalty parameter.

In the following, we will address $\mathbf{J}$, $\mathbf{Z}$, $\mathbf{E}^{(v)}$, $\mathbf{W}^{(v)}$, and $\mathbf{H}$ at $(t+1)$ iteration in sequence by fixing the others.

Subproblem of $\mathbf{J}$*:* It is equivalent to solve $\mathcal{L}$ with respect to $\mathbf{J}$ by

$$\min_{\mathbf{J}} \frac{\lambda_1}{\mu}\|\mathbf{J}\|_* + \frac{1}{2}\|\mathbf{J} - \left(\mathbf{Z}_{(t)} + \frac{\Lambda_{(t)}}{\mu}\right)\|_{\mathrm{F}}^2. \tag{8}$$

As shown in the previous work [46], (8) could be effectively solved by a closed-form solution as

$$\mathbf{J}_{(t+1)} = \mathcal{S}_{\frac{\lambda 1}{\mu}}\left(\mathbf{Z}_{(t)} + \frac{\Lambda_{(t)}}{\mu}\right) \tag{9}$$

where $\mathcal{S}(\cdot)$ represents the singular value threshold operator [50], and is defined by

$$\begin{aligned}\mathcal{S}_{\epsilon}(\mathbf{X}) &= \mathbf{U}\mathcal{D}_{\epsilon}(\Sigma)\mathbf{V}^{\mathrm{T}}\\ \mathcal{D}_{\epsilon}(\mathbf{A}) &= [\mathrm{sgn}(\mathbf{A}_{ij})(|\mathbf{A}_{ij}| - \epsilon)_+]\end{aligned} \tag{10}$$

with $\epsilon > 0$ and $(\cdot)_+ = \max\{\cdot, 0\}$. $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$ is the singular value decomposition (SVD) decomposition and $\mathcal{D}(\cdot)$ denotes the elementwise soft-thresholding shrinkage operator [48].

Subproblem of $\mathbf{Z}$*:* The solution of $\mathbf{Z}_{(t+1)}$ is generally given by taking derivate of $\mathcal{L}$ with respect to $\mathbf{Z}$ and setting it as zero. However, it is nontrivial to compute $\mathrm{tr}(\mathbf{H}^{\mathrm{T}}\mathbf{L}_z\mathbf{H})$ with a matrix form, as $\mathbf{L}_z = \mathbf{D}_z - \mathbf{Z}$ and $\mathbf{D}_z$ is the degree matrix of $\mathbf{Z}$. To simplify this term, an auxiliary matrix $\mathbf{G} \in \mathbb{R}^{n\times n}$ is introduced as

$$\mathbf{G} = \left[\mathbf{G}_1 \ldots \mathbf{G}_j \ldots \mathbf{G}_n\right],\ \mathbf{G}_j = \begin{bmatrix}\|\mathbf{H}_1 - \mathbf{H}_j\|_2^2\\ \vdots\\ \|\mathbf{H}_n - \mathbf{H}_j\|_2^2\end{bmatrix} \tag{11}$$

where $\mathbf{H}_j$ represents the jth row vector in $\mathbf{H}$. According to (11) and the property of Laplacian matrix [44], a simple deduction could be given as

$$\mathrm{tr}(\mathbf{H}^{\mathrm{T}}\mathbf{L}_z\mathbf{H}) = \frac{1}{2}\sum_{i,j}^{n}\|\mathbf{H}_i - \mathbf{H}_j\|^2\mathbf{Z}_{ij} = \frac{1}{2}\mathrm{tr}(\mathbf{G}^{\mathrm{T}}\mathbf{Z}).$$

Then, we can equivalently solve $\mathbf{Z}_{(t+1)}$ by

$$\begin{aligned}\min_{\mathbf{Z}\geq 0}\ & \frac{1}{2\mu}\mathrm{tr}(\mathbf{G}_{(t)}^{\mathrm{T}}\mathbf{Z}) + \frac{1}{2}\sum_{v=1}^{m}\|\mathbf{W}_{(t)}^{(v)}\mathbf{R}^{(v)} + \mathbf{H}_{(t)}\mathbf{H}_{(t)}^{\mathrm{T}} - \mathbf{E}_{(t)}^{(v)}\\ &+ \frac{\mathbf{Y}_{(t)}^{(v)}}{\mu}\|_{\mathrm{F}}^2 + \frac{1}{2}\|\mathbf{Z} - \mathbf{J}_{(t)} + \frac{\Lambda_{(t)}}{\mu}\|_{\mathrm{F}}^2 + \frac{1}{2}\|\mathbf{Z}\mathbf{1} - \mathbf{1} + \frac{\mathbf{u}_{(t)}}{\mu}\|_2^2.\end{aligned} \tag{12}$$

Inspired by Lin *et al.* [49] and Zhuang *et al.* [51], we linearize the (12) at $\mathbf{Z}_{(t)}$ as

$$\min_{\mathbf{Z}\geq 0}\ \langle \mathbf{Z} - \mathbf{Z}_{(t)}, \mathbf{F}_{(t)}\rangle + \frac{\eta_{(t)}}{2}\|\mathbf{Z} - \mathbf{Z}_{(t)}\|_{\mathrm{F}}^2 \tag{13}$$

where $\eta_{(t)} = \|\mathbf{G}_{(t)}\|_2^2 + \sum_v \|\mathbf{W}_{(t)}^{(v)}\mathbf{S}^{(v)}\|_2^2 + 1 + \|\mathbf{1}\|_2^2$, and

$$\mathbf{F}_{(t)} = \frac{\mathbf{G}_{(t)}}{2\mu} + \sum_{v=1}^{m} \mathbf{S}^{(v)\mathrm{T}}\mathbf{W}_{(t)}^{(v)\mathrm{T}}\left(\mathbf{E}_{(t)}^{(v)} - \mathbf{W}_{(t)}^{(v)}\mathbf{R}_{(t)}^{(v)} - \mathbf{H}_{(t)}\mathbf{H}_{(t)}^{\mathrm{T}} - \frac{\mathbf{Y}_{(t)}^{(v)}}{\mu}\right) + \left(\mathbf{Z}_{(t)} - \mathbf{J}_{(t+1)} + \frac{\Lambda_{(t)}}{\mu}\right) + \left(\mathbf{Z}_{(t)}\mathbf{1} - \mathbf{1} + \frac{\mathbf{u}_{(t)}}{\mu}\right)\mathbf{1}^{\mathrm{T}}$$

with $\mathbf{R}_{(t)}^{(v)} \equiv \mathbf{S}^{(v)} - \mathbf{S}^{(v)}\mathbf{Z}_{(t)}$. Taking all the deductions above into consideration, the solution of $\mathbf{Z}_{(t+1)}$ is finally given by

$$\begin{aligned}\mathbf{Z}_{(t+1)} &= \underset{\mathbf{Z}\geq 0}{\operatorname{argmin}} \|\mathbf{Z} - \left(\mathbf{Z}_{(t)} - \frac{1}{\eta_{(t)}}\mathbf{F}_{(t)}\right)\|_{\mathrm{F}}^2 \\ &= \left(\mathbf{Z}_{(t)} - \eta_{(t)}^{-1}\mathbf{F}_{(t)}\right)_+ . \end{aligned} \tag{14}$$

Subproblem of $\mathbf{E}^{(v)}$*:* For each view v, we update $\mathbf{E}_{(t+1)}^{(v)}$ by

$$\min_{\mathbf{E}^{(v)}} \frac{\lambda_2}{\mu}\|E^{(v)}\|_1 + \frac{1}{2}\left\|\mathbf{E}^{(v)} - \left(\mathbf{W}_{(t)}^{(v)}\mathbf{R}_{(t+1)}^{(v)} + \mathbf{H}_{(t)}\mathbf{H}_{(t)}^{\mathrm{T}} + \mathbf{Y}_{(t)}^{(v)}/\mu\right)\right\|_{\mathrm{F}}^2. \tag{15}$$

Following [48], we have

$$\mathbf{E}_{(t+1)}^{(v)} = \mathcal{D}_{\frac{\lambda_2}{\mu}}\left(\mathbf{W}_{(t)}^{(v)}\mathbf{R}_{(t+1)}^{(v)} + \mathbf{H}_{(t)}\mathbf{H}_{(t)}^{\mathrm{T}} + \frac{\mathbf{Y}_{(t)}^{(v)}}{\mu}\right). \tag{16}$$

Subproblem of $\mathbf{W}^{(v)}$*:* For any v, we obtain $\mathbf{W}_{(t+1)}^{(v)}$ by

$$\begin{aligned}\min_{\mathbf{W}^{(v)}} \ & \frac{\lambda_3}{2}\mathrm{tr}[(\bar{\mathbf{S}}^{(v)} - \mathbf{W}^{(v)}\tilde{\mathbf{S}}^{(v)})^{\mathrm{T}}(\bar{\mathbf{S}}^{(v)} - \mathbf{W}^{(v)}\tilde{\mathbf{S}}^{(v)})] \\ & + \frac{\mu}{2}\left\|\mathbf{W}^{(v)}\mathbf{R}_{(t+1)}^{(v)} + \mathbf{H}_{(t)}\mathbf{H}_{(t)}^{\mathrm{T}} - \mathbf{E}_{(t+1)}^{(v)} + \frac{\mathbf{Y}_{(t)}^{(v)}}{\mu}\right\|_{\mathrm{F}}^2. \end{aligned} \tag{17}$$

By setting the derivative of (17) with respect to $\mathbf{W}^{(v)}$ as zero, we have the following solution as:

$$\mathbf{W}_{(t+1)}^{(v)} = [\hat{\mathbf{P}}^{(v)}][\hat{\mathbf{Q}}^{(v)}]^{-1} \tag{18}$$

with

$$\begin{aligned}\hat{\mathbf{P}}^{(v)} &= \lambda_3\mathbf{P}^{(v)} + \left(\mu\mathbf{E}_{(t+1)}^{(v)} - \mu\mathbf{H}_{(t)}\mathbf{H}_{(t)}^{\mathrm{T}} - \mathbf{Y}_{(t)}^{(v)}\right)\mathbf{R}_{(t+1)}^{(v)\mathrm{T}} \\ \hat{\mathbf{Q}}^{(v)} &= \lambda_3\mathbf{Q}^{(v)} + \mu\mathbf{R}_{(t+1)}^{(v)}\mathbf{R}_{(t+1)}^{(v)\mathrm{T}}\end{aligned}$$

where $\mathbf{P}^{(v)}$ and $\mathbf{Q}^{(v)}$ are defined by (3). Here, we also expect to learn a robust feature transformation matrix via *infinity* denoising process. Thus, by following the similar strategy of mDA [21], we obtain $\mathbf{W}^{(v)}$ with the expectations of $\hat{\mathbf{P}}^{(v)}$ and $\hat{\mathbf{Q}}^{(v)}$, upon the weak law of large numbers. As $\tau \to \infty$, we rewrite (18) as

$$\begin{aligned}\mathbf{W}_{(t+1)}^{(v)} &= \mathbb{E}[\hat{\mathbf{P}}^{(v)}]\mathbb{E}[\hat{\mathbf{Q}}^{(v)}]^{-1} \\ &= \left(\lambda_3\mathbb{E}[\mathbf{P}^{(v)}] + \left(\mu\mathbf{E}_{(t+1)}^{(v)} - \mu\mathbf{H}_{(t)}\mathbf{H}_{(t)}^{\mathrm{T}} - \mathbf{Y}_{(t)}^{(v)}\right)\mathbf{R}_{(t+1)}^{(v)\mathrm{T}}\right) \\ &\quad \left(\lambda_3\mathbb{E}[\mathbf{Q}^{(v)}] + \mu\mathbf{R}_{(t+1)}^{(v)}\mathbf{R}_{(t+1)}^{(v)\mathrm{T}}\right)^{-1}\end{aligned} \tag{19}$$

where $\mathbb{E}[\mathbf{P}^{(v)}]$ and $\mathbb{E}[\mathbf{Q}^{(v)}]$ are given by (4).

Subproblem of $\mathbf{H}$*:* In (7), we have two parts with respect to the partition $\mathbf{H}$, corresponding to SEC and the *self-boost* constraint, respectively. Note that this constraint not only enables an interaction between learning $\mathbf{Z}$ and finding $\mathbf{H}$ but also employs the term $\mathbf{H}\mathbf{H}^{\mathrm{T}}$ to enhance the cluster structure of $\mathbf{Z}_{(t+1)}$. However, when computing $\mathbf{H}$ from $\mathbf{Z}$, this term will bring into the clustering process some unnecessary "noises" induced by $\mathbf{E}^{(v)}$ and $\mathbf{Y}^{(v)}$. Thus, we omit the part of $\mathcal{L}$ containing $\mathbf{H}\mathbf{H}^{\mathrm{T}}$ and recast the subproblem of $\mathbf{H}$ as

$$\mathbf{H}_{(t+1)} = \underset{\mathbf{H}^{\mathrm{T}}\mathbf{H}=\mathbf{I}}{\operatorname{argmin}} \mathrm{tr}(\mathbf{H}^{\mathrm{T}}\mathbf{L}_z\mathbf{H}) \tag{20}$$

where $\mathbf{L}_z$ is updated by $\mathbf{Z}_{(t+1)}$. As following [52], [53], one popular solution for (20) is to set $\mathbf{H}_{(t+1)}$ as the first K smallest eigenvectors of $\mathbf{L}_z$.

Multipliers: Totally, we have $m+2$ multipliers, which could be updated as the following:

$$\begin{aligned}\Delta_{(t+1)}^{(v)} &= \mathbf{W}_{(t+1)}^{(v)}\mathbf{R}_{(t+1)}^{(v)} + \mathbf{H}_{(t+1)}\mathbf{H}_{(t+1)}^{\mathrm{T}} - \mathbf{E}_{(t+1)}^{(v)} \\ \mathbf{Y}_{(t+1)}^{(v)} &= \mathbf{Y}_{(t)}^{(v)} + \mu\Delta^{(v)}, \forall v = 1, \ldots, m \\ \Lambda_{(t+1)} &= \Lambda_{(t)} + \mu(\mathbf{Z}_{(t+1)} - \mathbf{J}_{(t+1)}) \\ \mathbf{u}_{(t+1)} &= \mathbf{u}_{(t)} + \mu(\mathbf{Z}_{(t+1)}\mathbf{1} - \mathbf{1})\end{aligned} \tag{21}$$

where $\Delta^{(v)}$ is introduced for the conciseness. The entire solution for M^{2}VEC is summarized in Algorithm 1.

Algorithm 1 M^{2}VEC

Input: Co-association matrices of m views $\mathbf{S}^{(1)}, \ldots, \mathbf{S}^{(m)}$, cluster number K, parameters $\lambda_1, \lambda_2, \lambda_3 > 0$, the corruption ratio p.

Initial: $\mathbf{J}_{(0)} = \mathbf{Z}_{(0)} = \Lambda_{(0)} = \mathbf{0} \in \mathbb{R}^{n\times n}$, $\mathbf{H}_{(0)} = \mathbf{0} \in \mathbb{R}^{n\times K}$
$\mathbf{E}_{(0)}^{(v)} = \mathbf{Y}_{(0)}^{(v)} = \mathbf{0} \in \mathbb{R}^{n\times n}$, $v = 1, \ldots, m$,
$\mathbf{u} = \mathbf{0} \in \mathbb{R}^{n\times 1}$
$\mu = 10^{-3}$, $\mu_{\max} = 10^{10}$, $\epsilon = 10^{-4}$, $\rho > 1$, $t = 0$.

Initialize $\mathbf{W}^{(v)}$ via Eq. (3) for each view;
while *not converged* **do**
  Update $\mathbf{J}_{(t+1)}$ via Eq. (9);
  Update $\mathbf{Z}_{(t+1)}$ via Eq. (14);
  Update $\mathbf{E}_{(t+1)}^{(v)}$ via Eq. (16);
  Update $\mathbf{W}_{(t+1)}^{(v)}$ via Eq. (19);
  Update $\mathbf{L}_z$ and $\mathbf{D}_z$ by $\mathbf{Z}_{(t+1)}$;
  Set $\mathbf{H}_{(t+1)}$ as the smallest K eigenvectors of $\mathbf{L}_z$;
  Update the Lagrangian multipliers via Eq. (21);
  Check the convergence condition:
  $(\max\{\|\Delta_{(t+1)}^{(1)}\|_\infty, \ldots, \|\Delta_{(t+1)}^{(m)}\|_\infty\} < \epsilon) \wedge$
  $(\|\mathbf{J}_{(t+1)} - \mathbf{Z}_{(t+1)}\|_\infty < \epsilon) \wedge (\|\mathbf{Z}_{(t+1)}\mathbf{1} - \mathbf{1}\|_\infty < \epsilon)$;
  $\mu = \min\{\rho\mu, \mu_{\max}\}$, $t = t + 1$;
end while

Output: $\mathbf{Z}$, $\mathbf{H}$ and $\mathbf{W}^{(1)} \ldots \mathbf{W}^{(m)}$

D. Stacked M^{2}VEC

Stacked deep neural networks generally lead to rich and discriminative feature representations [36], [37], which is able to facilitate downstream tasks such as clustering. In light of this, the proposed M^{2}VEC is stacked as an individual building block in a multilayer architecture, to perform MVC in a *deep* fashion.

Let $\mathbf{S}^{(v,h)}$ denote the hidden representations given by the h-layer's M^{2}VEC of the vth view and l be the number of

stacked M^{2}VEC blocks, then we have

$$\mathbf{S}^{(v,h)} = \sigma(\mathbf{W}^{(v,h)}\mathbf{S}^{(v,h-1)}) \tag{22}$$

where $\sigma(\cdot)$ represents the elementwise nonlinear activation function (e.g., sigmoid and tanh), $\mathbf{W}^{(v,h)}$ is the learned feature transformation matrix given by the hth M^{2}VEC block, $1 \leq h \leq l$, and $\mathbf{S}^{(v,0)} = \mathbf{S}^{(v)}$ is the original coassociation matrix. Upon (22), we employ the greedy layerwise training strategy and take $\mathbf{S}^{(v,h)}$ as input for the $(h+1)$th M^{2}VEC block.

Remark 2: Compared with the single-layer M^{2}VEC, which only adopts a linear marginalized denoiser, (22) integrates nonlinear property into the partition-level representation for the next layer. By this means, the multilayer M^{2}VEC enhances the expressive ability for handling complex data [39].

We obtain the final clustering result by running SC on consensus low-rank representations from the last M^{2}VEC block. The entire procedure of our stacked M^{2}VEC is summarized by Algorithm 2.

Algorithm 2 MVC by Stacked M^{2}VEC

Input: Basic partitions $\Pi^{(1)}, \ldots, \Pi^{(m)}$, layer number l

1: Derive $\mathbf{S}^{(v)}$ from $\Pi^{(v)}$ via Eq. (1) for each view;
2: **for** $h = 1$ **to** $l-1$ **do**
3: Conduct Algorithm 1 with $\mathbf{S}^{(v,h-1)}$ for learning $\mathbf{W}^{(v,h)}$;
4: Obtain $\mathbf{S}^{(v,h)}$ with $\mathbf{W}^{(v,h)}$ and $\mathbf{S}^{(v,h-1)}$ via Eq. (22);
5: **end for**
6: Conduct Algorithm 1 with $\mathbf{S}^{(v,l-1)}$ for learning for $\mathbf{Z}$;
7: Run spectral clustering on $\mathbf{Z}$.

Output: Clustering result π

E. Model Discussion

1) Convergence: In general, it is challenging to guarantee a global convergence for solving the optimization problem with more than two variables. However, as shown in [48] and [49], the ALM solver with ADM strategy could be an effective solution for the problem such as (5), since each subproblem of the proposed M^{2}VEC has a closed-form solution. Moreover, empirical evidence on real-world data sets exhibits a stable convergence behavior of our model.

2) Complexity Analysis: The major computation parts of Algorithm 1 are: 1) the SVD decomposition in step 3; 2) matrix multiplication and inverse in steps 4 and 6; and 3) the soft-thresholding operation in step 5. The first two parts roughly cost $\mathcal{O}(n^3)$. For the third part, (16) takes $\mathcal{O}(n^2)$ for each view, resulting the complexity of $\mathcal{O}(mn^2)$. Hence, the total computing cost of Algorithm 1 is $\mathcal{O}(T(mn^2+n^3))$, where T denotes the iteration number. To make Algorithm 1 scalable for large-scale data sets, several off-the-shell acceleration methods could be used such as *divide-and-conquer* [54] and the skinny SVD based ones [55], [56]. However, in this paper, we focus on improving robustness and effectiveness for MVC, and thus we leave the scalability in our future work.

The time complexity given in Algorithm 2 is roughly linear to the layer number l, since we learn a linear marginalized denoiser with closed-form solution in each single layer and adopt a stacked fashion to build the multilayer M^{2}VEC.

III. Experiment

A. Experimental Setting

1) Data Sets: Eight real-world data sets including text and image data are used in the experiment. We characterize each data set as the follows.

a) Text Data: 3-Sources[1] data set is collected from three online news sources [i.e., British Broadcasting Corporation (BBC), Guardian, and Reuter], which consists of 169 stories and is divided into six categories. *four-Areas*[2] data set covers the papers from 20 conferences throughout four areas of database, data mining, machine learning, and information retrieval. We remove the cross-domain authors and describe the remaining 4236 authors with two views of conference and abstract term. *BBCSport*[1] contains 737 news articles from the BBC Sport website corresponding to five sport topics from 2004 to 2005. Here, we use a subset of this data set (i.e., 544 articles) provided in [14]. *WebKB* data set [57] represents 1051 web documents through two views of *content* and *link*, which is a subset of the four universities data set[3] and with two categories.

b) Image Data: Caltech101 data set [58] provides an image set of 101 categories for the object recognition task. Following [11], we employ a 7-class subset (termed *Caltech101-7* with 1474 images) and 20-class subset (termed *Caltech101-20* with 2386 images) in the experiment, each of which encodes images with six different feature descriptors such as 48-D Gabor feature, 40-D wavelet moments, 254-D CENTRIST feature, 1984-D HOG feature, 512-D GIST feature, and 928-D LBP feature. University of California at Irvine (UCI) *Digit*[4] data set is an image data set of 0–9 handwritten digits, where each digit is depicted by 76 Fourier coefficients of character shape and 240-D pixel averages in 2×3 window, respectively. *Notting-Hill* data set [59] is widely used for video face clustering, which collects 4660 faces of 76 tracks from 5 main cast in the movie "Notting Hill." Here, we use the data set provided in [26], including 550 images with three features of intensity, LBP, and Gabor features.

2) Compared Methods: The compared methods used in our experiment can be divided into three groups.

a) Baseline Methods: As following [14], we implement three baseline methods based on SC [60]: 1) Spectral$_{\text{BSV}}$ returns SC result of the best single view (BSV); 2) Spectral$_{\text{CON}}$ performs SC with the concatenated features of multiple views; and 3) Spectral$_{\text{SUM}}$ sums the Gaussian kernel matrices of each view and conducts SC on the averaged one.

b) MVC Methods: Four representative MVC methods are used in the experiment, including coregularized SC (CRSC) [10], multiview NMF (MultiNMF) [24], robust multiview SC (RMVSC) [14], and diversity-induced multiview subspace clustering (DiMSC) [26].

[1] http://mlg.ucd.ie/datasets
[2] http://web.cs.ucla.edu/~yzsun/data/four_area.zip
[3] http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/
[4] http://archive.ics.uci.edu/ml/datasets.html

TABLE I

CLUSTERING PERFORMANCE ON EIGHT REAL-WORLD DATA SETS BY NMI% (RED BOLD FONT FOR THE BEST AND BLUE ITALIC FOR THE SECOND)

Datasets	*3-Sources*	*4-Areas*	*BBCSport*	*WebKB*	*Caltech101-7*	*Caltech101-20*	*Digit*	*Notting-Hill*	score
Spectral$_{BSV}$	47.14±2.45	43.27±8.53	71.76±0.32	53.80±0.00	48.29±2.89	59.65±1.35	63.92±2.40	71.58±2.90	5.88
Spectral$_{CON}$	51.66±2.22	0.87±0.09	54.77±1.52	71.77±0.00	34.92±1.17	38.19±0.84	65.09±2.37	64.28±5.66	4.80
Spectral$_{SUM}$	46.40±4.09	37.06±10.45	60.56±2.48	71.77±0.00	40.78±2.23	52.65±1.58	76.60±3.07	77.59±4.68	5.87
CRSC [10]	51.55±2.93	32.97±2.39	63.71±2.48	72.85±0.00	47.76±2.61	56.29±1.36	72.23±3.05	75.20±4.74	6.03
MultiNMF [24]	39.69±3.95	9.55±1.92	22.60±3.43	53.30±0.00	47.88±3.27	59.23±1.43	72.81±2.55	78.53±2.90	4.97
RMVSC [14]	37.66±4.20	50.81±0.04	80.26±2.91	72.36±0.00	45.49±2.20	54.52±1.47	76.71±1.95	67.17±4.93	6.15
DiMSC [26]	73.68±1.16	48.85±0.08	58.61±1.73	53.24±0.00	37.92±0.16	28.21±0.67	34.50±1.25	79.86±0.22	5.24
KCC$_{BSV}$ [20]	60.42±2.81	49.03±6.18	84.25±2.33	33.29±0.00	55.01±2.19	60.67±1.10	74.59±2.78	77.08±7.10	6.32
KCC$_{SUM}$ [20]	64.75±3.58	53.09±5.93	87.33±2.77	51.22±9.61	57.10±1.56	60.51±1.87	82.98±3.29	77.15±6.62	6.81
SEC$_{BSV}$ [29]	55.23±1.87	55.22±5.09	87.26±2.08	33.21±0.23	53.46±1.62	61.91±1.07	76.73±1.09	73.80±1.42	6.35
SEC$_{SUM}$ [29]	59.21±3.50	74.06±0.00	86.86±4.47	34.52±0.00	57.32±0.72	59.56±1.28	66.90±2.89	69.73±0.04	6.52
RSEC$_{BSV}$ [34]	63.60±0.00	43.09±4.42	64.95±0.00	25.53±0.00	60.52±0.72	62.13±1.27	**88.20±0.05**	78.61±0.00	6.26
RSEC$_{SUM}$ [34]	41.65±0.00	46.67±0.00	62.24±0.00	75.75±0.00	43.10±0.65	57.92±1.24	85.72±1.68	80.50±0.20	6.26
IEC$_{BSV}$ [35]	65.49±5.53	64.33±10.07	77.10±5.65	65.81±11.16	54.75±4.75	58.97±1.96	72.21±4.18	66.92±8.59	6.73
IEC$_{SUM}$ [35]	67.68±4.48	72.37±8.60	79.87±7.50	35.89±0.00	57.94±4.54	59.53±2.22	64.12±4.42	72.96±6.88	6.56
MVEC [40]	*75.69±0.00*	*77.66±0.00*	*90.83±0.00*	*79.46±0.00*	*61.28±0.00*	**62.87±0.94**	86.66±0.99	*84.17±0.91*	*7.88*
M^{2}VEC	**77.22±0.54**	**77.74±0.00**	**91.01±0.00**	**81.43±0.00**	**63.18±0.41**	*62.10±1.10*	*87.41±0.07*	**86.49±0.00**	**7.98**

TABLE II

CLUSTERING PERFORMANCE ON EIGHT REAL-WORLD DATA SETS BY RN% (RED BOLD FONT FOR THE BEST AND BLUE ITALIC FOR THE SECOND)

Datasets	*3-Sources*	*4-Areas*	*BBCSport*	*WebKB*	*Caltech101-7*	*Caltech101-20*	*Digit*	*Notting-Hill*	score
Spectral$_{BSV}$	35.53±3.30	30.66±12.59	69.57±0.14	61.82±0.00	28.67±2.84	30.79±3.22	53.77±3.80	71.78±5.82	5.08
Spectral$_{CON}$	33.77±4.01	-0.02±0.00	46.67±3.71	84.50±0.00	21.31±1.54	16.51±1.05	55.30±3.54	62.85±8.11	4.11
Spectral$_{SUM}$	34.60±4.92	21.53±14.84	55.19±2.31	84.50±0.00	25.87±0.99	26.69±2.46	69.50±5.50	74.17±7.65	5.11
CRSC [10]	35.78±5.12	17.73±4.16	61.92±0.83	83.71±0.00	28.14±1.81	29.43±2.43	66.79±5.05	69.61±7.94	5.17
MultiNMF [24]	17.20±5.58	0.37±0.54	13.58±2.75	66.08±0.00	32.30±4.86	34.15±4.92	64.59±4.79	77.43±6.84	4.28
RMVSC [14]	25.05±4.94	46.21±0.05	78.65±8.40	84.91±0.00	31.23±2.11	26.06±2.18	70.73±3.73	58.63±8.10	5.47
DiMSC [26]	*67.34±0.51*	50.01±0.21	53.25±2.89	57.18±0.00	26.01±0.13	10.88±0.62	14.08±0.76	78.68±0.08	4.60
KCC$_{BSV}$ [20]	47.24±6.84	33.57±10.23	86.18±2.38	33.11±0.00	33.63±3.72	29.46±2.84	60.98±6.25	73.38±11.33	5.31
KCC$_{SUM}$ [20]	53.38±7.93	47.58±8.92	88.44±3.14	57.38±13.29	32.72±2.74	29.60±3.37	75.93±6.75	70.10±11.61	5.96
SEC$_{BSV}$ [29]	37.09±2.89	44.47±7.98	89.73±3.12	32.50±0.01	32.67±2.59	30.43±1.68	66.33±2.03	65.64±2.03	5.31
SEC$_{SUM}$ [29]	43.24±6.53	77.35±0.00	86.96±8.20	32.78±0.00	31.50±1.75	29.24±2.70	55.09±3.82	57.81±0.11	5.48
RSEC$_{BSV}$ [34]	54.52±0.00	25.42±8.85	40.74±0.00	25.18±0.00	44.15±1.08	33.65±2.18	**87.35±0.04**	73.20±0.00	5.35
RSEC$_{SUM}$ [34]	16.39±0.00	41.56±0.00	36.93±0.00	87.22±0.00	36.54±1.15	36.54±2.11	77.00±3.82	76.95±0.75	5.53
IEC$_{BSV}$ [35]	57.62±10.93	60.07±15.77	73.48±10.92	77.31±13.80	34.93±8.43	28.96±4.61	57.62±8.09	54.69±14.55	5.86
IEC$_{SUM}$ [35]	62.45±7.58	74.20±13.49	73.84±13.13	36.80±0.00	36.77±5.58	29.85±3.98	51.19±6.70	66.11±12.09	5.77
MVEC [40]	62.21±0.00	*83.25±0.00*	**92.06±0.00**	*89.65±0.00*	*42.72±0.00*	**43.06±3.63**	81.55±4.12	*80.76±1.19*	*7.62*
M^{2}VEC	**79.40±1.88**	**83.37±0.00**	*92.05±0.00*	**90.89±0.00**	**45.55±0.56**	*41.07±2.96*	*86.14±0.05*	**82.36±0.00**	**7.94**

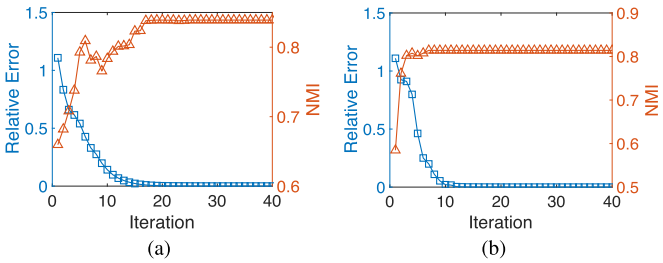


Fig. 2. Convergence analysis of the proposed M^{2}VEC method. (a) and (b) Relative error (NavyBlueblue line) and NMI (RedOrange orange line) curves with respect to iterations on *Digit* and *WebKB* data sets.

c) EC Methods: We compare the proposed M^{2}VEC with four state-of-the-art EC algorithms, such as K-means-based consensus clustering (KCC) [20], SEC [29], robust SEC (RSEC) [34], and Infinite EC (IEC) [35]. We report all the EC methods with BPs from BSV and BPs from all the views (subscripted by SUM), respectively.

3) Validation Criteria: Two widely used clustering validation criteria are used to evaluate the clustering performance of all the methods, which are normalized mutual information (NMI) [61] and Normalized Rand Index (Rn) [62]. These two metrics are both positive measures and ranged from 0 to 1, where NMI will drop to zero for a random partition and *Rn* might be negative to the extremely poor clustering result.

4) Implementation Details: We adopt the random parameter selection (RPS) strategy [18], [20] to obtain a set of $r = 100$ BPs for each view individually. In detail, we run KM with cosine similarity and a random cluster number from $[K, \sqrt{n}]$ on each single-view data r times. These multiview BPs are fed as the default input to all the EC methods. On the other hand, we use the original multiview data for traditional MVC methods and follow their preprocessing steps. In the experiment, we set the true cluster number for all the compared methods, and run the authors' released codes with recommended parameters. We test each method 20 times and report the average result along with the *standard deviation* (*std*). For the proposed M^{2}VEC, we set $\lambda_1 = 1$, $\lambda_2 = 0.01$, $\lambda_3 = 1$ and corruption ratio $p = 0.1$, and employ a two-layer architecture (i.e., $l = 2$) with tanh activation function as the default setting.

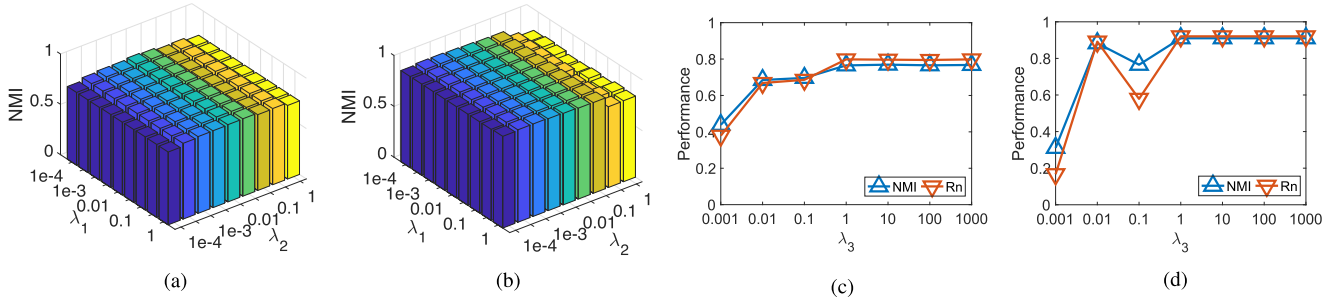


Fig. 3. Parameter study on *3-Sources* and *BBCSport* data sets. (a) and (b) Analysis in terms of λ_1 and λ_2. (c) and (d) Impact of λ_3 to our approach.

B. Clustering Performance

1) Overview: Tables I and II summarize the clustering performance of the proposed M^{2}VEC and other methods by NMI and *Rn*, respectively. In general, our algorithm achieves the best performance on all the data sets in terms of two evaluation metrics. To give an overview of the evaluation result, we calculate a measurement score as the following: $\text{score}(A_i) = \sum_j (f(A_i, D_j))/(\max_i f(A_i, D_j))$, where $f(A_i, D_j)$ denotes the NMI or *Rn* value of A_i method on the D_j data set. This score gives an overall comparison between different methods on all the data sets [20]. As can be seen, our proposed M^{2}VEC outperforms three different types of compared methods with a substantial improvement.

2) Ensemble Versus Multiview: Tables I and II also give a comparison result between (EC and traditional MVC algorithms, where EC integrates BPs, while MVC directly performs with raw data. Generally, EC methods enjoy a better clustering performance than MVC, which fully demonstrates the significant superiority and great potentiality of exploiting higher level information (i.e., BPs) to address the MVC task. This is mainly due to two reasons: 1) BPs are a group of clustering results in nature, and thus the partition-level representations can alleviate the noises in raw data to some extent and 2) transforming multiview data into the partition space bridges the gap between heterogeneous features. The proposed M^{2}VEC inherits these good properties from EC, and hence outperforms traditional MVC methods with a clear improvement.

3) Single-View Versus Multiview: Although EC methods appear to be a very promising direction for MVC, it can be seen that they have a similar, sometimes even worse performance on multiview data compared with their BSV results. For instance, SEC_{SUM} [29] performs slightly better (1.31% NMI and 0.28% *Rn* higher) than SEC_{BSV} on the *WebKB* data set but degrades badly (lowering round 10% NMI and 11% *Rn*) on the *Digit* data set. Similar observations could be found by RSEC [34] on *Notting-Hill* and *3-Sources*, and IEC [35] on *Caltech101-20* and *WebKB*, respectively. This is mainly because exiting EC methods treat BPs of each view equally, yet neglect to exploit the complementation information among multiview data. Different from EC methods, our approach explicitly considers the connection between multiple views through learning a consensus low-rank representation, which results in a better clustering performance.

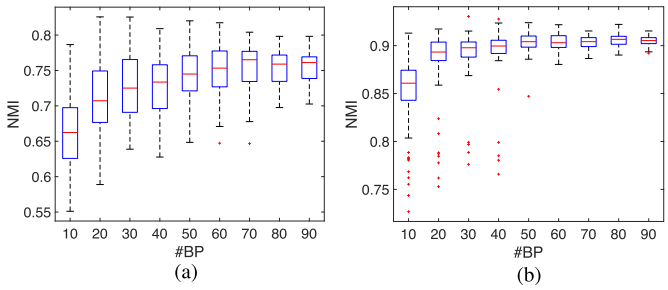


Fig. 4. Exploration to the number of BPs. We evaluate our approach with different number BPs on the *3-Sources* and *BBCSport* data sets. (a) *3-Sources* by NMI. (b) *BBCSport* by NMI.

4) MVEC Versus M^{2}VEC: Similar to other EC methods, our previous work MVEC [40] directly takes input as BPs without considering the noises of BPs. To alleviate this noise issue, the proposed M^{2}VEC leverages the mDA to obtain a robust partition-level representation of each view. Moreover, a multilayer architecture is also developed to further boost the robustness of the representation. As can be seen, we perform better than MVEC on the majority data sets and achieve comparable results on the remainder, which indicates M^{2}VEC as a substantial extension.

C. Model Discussion

1) Convergence: To show the convergence property given in Algorithm 1, we calculate the relative error for the top M^{2}VEC block by $\max\{\|\Delta^{(v)}\|_{\text{F}}/\|\mathbf{S}^{(v)}\|_{\text{F}}\}_{v=1}^{m}$. As shown in Fig. 2, our method converges steadily within 40 iterations. Moreover, during the optimization process, the NMI curve generally goes up and achieves stable after several reasonable fluctuations, which shows our algorithm has a strong convergence behavior.

2) Parameter Study: There are three parameters λ_1, λ_2, and λ_3 in our model, where λ_1 controls the rank of view-consensus representation, λ_2 balances the sparseness of view-specific residual matrix, and λ_3 corresponds to the marginalized denoiser. Fig. 3(a) and (b) reports the NMI values of M^{2}VEC by ranging λ_1 and λ_2 from the set of $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 0.01, 0.05, 0.1, 0.5, 1\}$ with the fixed λ_3, whereas Fig. 3(c) and (d) shows the NMI of M^{2}VEC by setting λ_3 from $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ with the fixed λ_1, λ_2. As can be seen, our approach is insensitive to λ_1 and λ_2 with a relatively small value range, i.e., $[0.01, 1]$. This is mainly due to two reasons: 1) the coassociation matrix

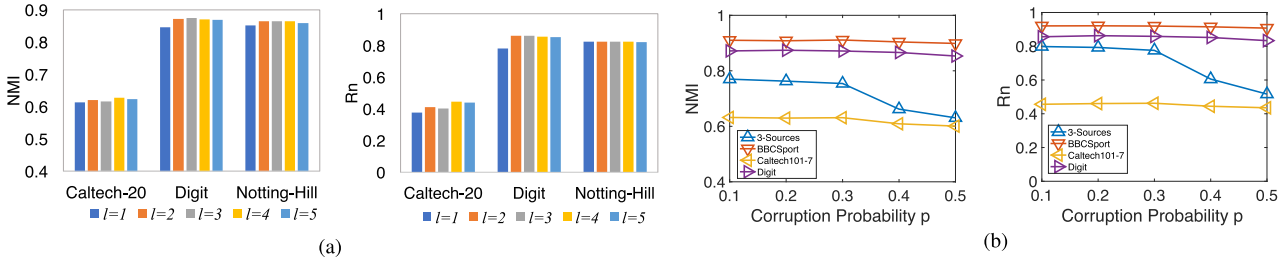


Fig. 5. Performance of M^{2}VEC with different layers (l) and corruption ratios (p). (a) We test M^{2}VEC from single layer to five layers on the *Caltech101-20* (shortened by *Caltech20*), *Digit* and *Notting-Hill* data sets. (b) We report the performance of two-layer M^{2}VEC by varying p from 0.1 to 0.5 on four data sets.

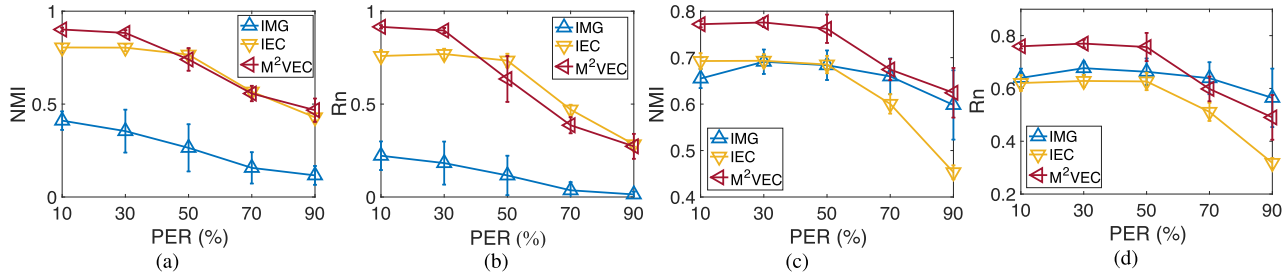


Fig. 6. Performance of M^{2}VEC under the case of incomplete view data. PER indicates the percentage of partial instances. We compare with IMG [65] and IEC [35] on two data sets, where IMG is specifically designed for the PVC problem.

usually has a good low-rank property and exhibits a nice block-diagonal structure; 2) coassociation matrices transform multiview data into the same partition space and are well normalized by (1).

We also investigate the impact of λ_3 to our model. As shown in Fig. 3(c) and (d), setting λ_3 a small value (i.e., $\lambda_3 < 1$) lowers the importance of the mDA in (5), which leads to a poor optimization result for the denoiser and thus degrades the clustering performance. However, the performance is quite stable when we set $\lambda_3 \geq 1$. This fully shows the effectiveness of our marginalized denoiser for the clustering task.

3) Impact of Basic Partitions Number: Our approach generates 100 BPs (denoted by $\Pi^{(v)}$) with each single-view data in advance, and obtains the final clustering result by feeding multiview BPs (i.e., $\Pi^{(1)} \dots \Pi^{(m)}$) to Algorithm 2. Here, to explore the impact of BPs number, we randomly select r BPs from $\Pi^{(v)}$ of each view and test M^{2}VEC with all the sampled sets, where r is ranged from 10 to 90. For each r, the sampling and testing process is repeated 100 times. As shown in Fig. 4, the NMI value generally goes up with a reducing variance when #BP increases. This justifies that a relatively large BPs number can improve stableness of our method.

4) Network Analysis: As the default setting to our model, we employ a two-layer architecture ($l = 2$) and set the corruption ratio as $p = 0.1$ for the marginalized denoising process. Here, we explore the impact of different layer numbers and corruption ratios to our method. Fig. 5(a) shows the clustering performance of M^{2}VEC with different layers. In general, multilayer M^{2}VEC performs better than the single layer, especially on the *Digit* data set, which shows the effectiveness of using stacked M^{2}VEC to enhance the expressive ability of our model. However, as multilayer M^{2}VEC falls into a simplified deep fully connected network, it is usually hard to be optimized as the depth of network increases [63], [64], leading to the performance of M^{2}VEC may slightly degrade when $l \geq 3$. On the other side, as shown in Fig. 5(b), our method is robust to different corruption ratios in the majority of data sets. Thus, we may suggest to set p from $[0.1, 0.3]$.

5) Component Analysis: Marginalized stacked denoising autoencoders (mSDA) [21], [37] and coassociation matrix are two core components in our model. To fully show the superiority of M^{2}VEC over these two parts, we implement a strong baseline by combing mSDA and multiview coassociation matrices together. For fairness comparison, we employ a two-layer mSDA with the same corruption ratio ($p = 0.1$) and activation function to our model. We perform mSDA on the coassociation matrix in each view and summarize all these views' feature representations from the last layer. The final clustering result is obtained by conducting KM or SC [60] on the averaged representation [i.e., $1/m \sum_{v=1}^{m} \mathbf{S}^{(v,2)}$, where $\mathbf{S}^{(v,2)}$ is given in (22) with $\mathbf{W}$ learned from mDA], respectively. As given in Table III, our proposed M^{2}VEC consistently outperforms mSDA+KM and mSDA+SC. It actually works as an important "sanity check" for our model, which shows: 1) marginalized denoising performs well with partition-level features for the MVC task and 2) learning a consensus representation across different views could significantly boost the performance.

D. Partial Multiview Clustering

Partial multiview clustering (PVC) [65], [66] is a practical problem in the real-world applications, as incomplete view data are more common than the complete ones. For example, in the social media network, it might be easy to access a lot of

TABLE III
COMPONENT ANALYSIS ON EIGHT REAL-WORLD DATA SETS BY NMI AND *Rn* (%)

	Datasets	*3-Sources*	*4-Areas*	*BBCSport*	*WebKB*	*Caltech101-7*	*Caltech101-20*	*Digit*	*Notting-Hill*
NMI	mSDA+KM	72.30±3.56	76.80±5.46	83.23±5.47	78.36±0.28	56.03±3.61	61.69±1.53	65.35±4.18	75.89±4.34
	mSDA+SC	70.90±0.26	76.58±0.00	88.96±0.17	56.67±0.00	56.42±0.00	60.53±0.84	74.93±0.46	70.08±1.97
	M^{2}VEC	77.22±0.54	77.74±0.00	91.01±0.00	81.43±0.00	63.18±0.41	62.10±1.10	87.41±0.07	86.49±0.00
Rn	mSDA+KM	69.12±6.91	81.15±8.42	79.47±11.04	88.92±0.17	36.09±6.17	37.64±4.20	52.08±6.15	70.74±8.61
	mSDA+SC	64.50±0.45	81.29±0.00	90.87±0.17	67.15±0.00	31.96±0.00	30.65±1.61	68.38±0.67	58.75±3.85
	M^{2}VEC	79.40±1.88	83.37±0.00	92.05±0.00	90.89±0.00	45.55±0.56	41.07±2.96	86.14±0.05	82.36±0.00

posted images or comments individually, whereas it is hard to collect the images with their corresponding comments simultaneously. Previous works define the instances with multiple accessible features or modalities as *complete view* data, while the instances with only one view being available as *partial view* data. Here, we follow the same setting of [65] and [66], which only considers the case of two-view data and evenly distributes the partial examples across each view. Specifically, let ϵ be partial example ratio (PER) to indicate the percentage of partial view instances, then we have $(1-\epsilon)\%$ complete view data, $0.5\epsilon\%$ view-1 data and $0.5\epsilon\%$ view-2 data.

EC is able to handle incomplete BPs by labeling the missing positions in each BP as zeros. By this means, traditional EC methods could also handle the PVC problem by taking into consideration all the incomplete BPs from multiple views jointly. However, in our case, we expect to learn the connection between different views, and thus compute a view-specific *partial coassociation* matrix. In detail, we generate BPs by using complete and partial examples [$(1-0.5\epsilon)\%$ in total] within each view and leaving the missing part as zeros. This may result in a large amount of missing values in the partial coassociation matrix as PER increases too high. To alleviate such detrimental effect, we employ view-2's partial coassociation matrix to complete the missing values in view-1, and vice versa. By playing this simple trick, we use the relationship between partial examples in one view to compensate the missing values in other view.

We test the proposed M^{2}VEC with partial multiview data on the *BBCSport* and *Notting-Hill* data sets. Following the PVC setting, we use the best two views of *Notting-Hill*. We compare with incomplete multimodality grouping (IMG) [65] and IEC [35] to evaluate the effectiveness of our approach on PVC problem, where IMG is one of the most recent PVC methods. Fig. 6 shows the clustering performance of M^{2}VEC by increasing PER from 10% to 90% with a step of 20%. As can be seen, our model generally outperforms other methods when PER is less than 50%, and achieve comparable results when the partial examples get aggravated (i.e., PER $\geq$ 70%). This demonstrates our M^{2}VEC works well under the scenario of incomplete view data.

IV. CONCLUSION

A novel M^{2}VEC algorithm was proposed in this paper, which jointly performed marginalized denoising, consensus representation learning, and spectral graph partitioning in a unified optimization framework. A multilayer model is provided by stacking M^{2}VEC blocks to deliver robust and rich partition-level representations for the clustering purpose. Experimental results on eight real-world data sets demonstrated the superiority of the proposed M^{2}VEC over several state-of-the-art multiview and EC methods. We also presented extensive analyses to discuss our model from several aspects, and gave an example of using M^{2}VEC with partial multiview data.

REFERENCES

[1] S. Li, Y. Li, and Y. Fu, "Multi-view time series classification: A discriminative bilinear projection approach," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 989–998.

[2] Z. Ding and Y. Fu, "Low-rank common subspace for multi-view learning," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 110–119.

[3] L. Xie, D. Tao, and H. Wei, "Joint structured sparsity regularized multiview dimension reduction for video-based facial expression recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 2, pp. 28:1–28:21, 2017.

[4] Y. Liu, Z. Gu, Y.-M. Cheung, and K. A. Hua, "Multi-view manifold learning for media interestingness prediction," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2017, pp. 308–314.

[5] X. Dong, L. Zhu, X. Song, J. Li, and Z. Cheng, "Adaptive collaborative similarity learning for unsupervised multi-view feature selection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 2064–2070.

[6] S. Li, M. Shao, and Y. Fu, "Multi-view low-rank analysis for outlier detection," in *Proc. SIAM Int. Conf. Data Mining*, 2015, pp. 748–756.

[7] H. Zhao and Y. Fu, "Dual-regularized multi-view outlier detection," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, Jul. 2015, pp. 4077–4083.

[8] S. Wang, Z. Ding, and Y. Fu, "Coupled marginalized auto-encoders for cross-domain multi-view learning," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, Jul. 2016, pp. 2125–2131.

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

[10] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Proc. Adv Neural Inf. Process. Syst. (NIPS)*, 2011, pp. 1413–1421.

[11] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2750–2756.

[12] M. B. Blaschko and C. H. Lampert, "Correlational spectral clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[13] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 650–658.

[14] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2149–2155.

[15] D. Greene and P. Cunningham, "A matrix factorization approach for integrating multiple data views," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2009, pp. 423–438.

[16] E. Bruno and S. Marchand-Maillet, "Multiview clustering: A late fusion approach using latent models," in *Proc. 32nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2009, pp. 736–737.

[17] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Dec. 2003.

[18] A. L. N. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 835–850, Jun. 2005.

[19] A. Topchy, A. K. Jain, and W. Punch, "Clustering ensembles: Models of consensus and weak partitions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1866–1881, Dec. 2005.

[20] J. Wu, H. Liu, H. Xiong, J. Cao, and J. Chen, "K-means-based consensus clustering: A unified view," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 1, pp. 155–169, Jan. 2015.

[21] M. Chen, Z. E. Xu, K. Q. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 1627–1634.

[22] A. Kumar and H. Daumé, III, "A co-training approach for multi-view spectral clustering," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 393–400.

[23] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 129–136.

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

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

[26] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang, "Diversity-induced multi-view subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 586–594.

[27] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2598–2604.

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

[29] H. Liu, T. Liu, J. Wu, D. Tao, and Y. Fu, "Spectral ensemble clustering," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 715–724.

[30] T. Li, C. Ding, and M. Jordan, "Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, Oct. 2007, pp. 577–582.

[31] N. Iam-On, T. Boongoen, S. M. Garrett, and C. J. Price, "A link-based approach to the cluster ensemble problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2396–2409, Dec. 2011.

[32] X. Z. Fern and C. E. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning," in *Proc. 21st Int. Conf. Mach. Lear.*, 2004, p. 36.

[33] M. Yousefnezhad, S.-J. Huang, and D. Zhang, "Woce: A framework for clustering ensemble by exploiting the wisdom of crowds theory," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 486–499, Feb. 2018.

[34] Z. Tao, H. Liu, S. Li, and Y. Fu, "Robust spectral ensemble clustering," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 367–376.

[35] H. Liu, M. Shao, S. Li, and Y. Fu, "Infinite ensemble for image clustering," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1745–1754.

[36] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.

[37] M. Chen, K. Q. Weinberger, Z. Xu, and F. Sha, "Marginalizing stacked linear denoising autoencoders," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 3849–3875, Dec. 2015.

[38] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[39] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[40] Z. Tao, H. Liu, S. Li, Z. Ding, and Y. Fu, "From ensemble clustering to multi-view clustering," in *Proc. 26th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 2843–2849.

[41] H. Jia and Y.-M. Cheung, "Subspace clustering of categorical and numerical data with an unknown number of clusters," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3308–3325, Apr. 2018.

[42] D. Luo, C. Ding, H. Huang, and F. Nie, "Consensus spectral clustering in near-linear time," in *Proc. IEEE 27th Int. Conf. Data Eng.*, Apr. 2011, pp. 1079–1090.

[43] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. New York, NY, USA: Springer-Verlag, 2006.

[44] U. Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[45] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 663–670.

[46] 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.

[47] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the ℓ_1-ball for learning in high dimensions," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 272–279.

[48] Z. Lin, M. Chen, and Y. Ma. (Sep. 26, 2010). "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices." [Online]. Available: https://arxiv.org/abs/1009.5055

[49] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 612–620.

[50] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[51] L. Zhuang, J. Wang, Z. Lin, A. Y. Yang, Y. Ma, and N. Yu, "Locality-preserving low-rank representation for graph construction from nonlinear manifolds," *Neurocomputing*, vol. 175, pp. 715–722, Jan. 2016.

[52] H. Zha, X. He, C. Ding, H. D. Simon, and M. Gu, "Spectral relaxation for K-means clustering," in *Proc. Adv. neural Inf. Process. Syst.*, 2002, pp. 1057–1064.

[53] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: Spectral clustering and normalized cuts," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 551–556.

[54] A. Talwalkar, L. Mackey, Y. Mu, S.-F. Chang, and M. I. Jordan, "Distributed low-rank subspace segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3543–3550.

[55] X. Zhang, F. Sun, G. Liu, and Y. Ma, "Fast low-rank subspace segmentation," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1293–1297, May 2014.

[56] S. Xiao, W. Li, D. Xu, and D. Tao, "FaLRR: A fast low rank representation solver," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4612–4620.

[57] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point cloud: From transductive to semi-supervised learning," in *Proc. 22nd Int. Conf. Mach. learn.*, 2005, pp. 824–831.

[58] 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, 2007.

[59] Y.-F. Zhang, C. Xu, H. Lu, and Y.-M. Huang, "Character identification in feature-length films using global face-name matching," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1276–1288, Nov. 2009.

[60] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. neural Inf. Process. Syst.*, 2001, pp. 849–856.

[61] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.

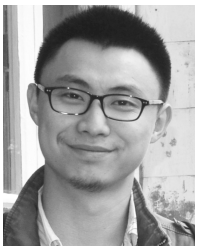
[62] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for k-means clustering," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 877–886.

[63] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, vol. 15, 2011, pp. 315–323.

[64] J. Li, T. Zhang, W. Luo, J. Yang, X.-T. Yuan, and J. Zhang, "Sparseness analysis in the pretraining of deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 6, pp. 1425–1438, Jun. 2017.

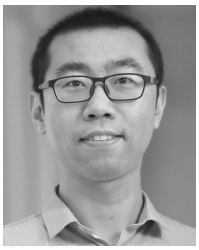
[65] H. Zhao, H. Liu, and Y. Fu, "Incomplete multi-modal visual data grouping," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 2392–2398.

[66] S.-Y. Li, Y. Jiang, and Z.-H. Zhou, "Partial multi-view clustering," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1968–1974.



Zhiqiang Tao received the B.E. degree in software engineering from the School of Computer Software and the M.S. degree in computer science from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree with Northeastern University, Boston, MA, USA.

His current research interests include data cluster analysis, subspace learning, ensemble clustering, and unsupervised deep representation learning.



Hongfu Liu (M'18) received the bachelor's and master's degrees in management information systems from the School of Economics and Management, Beihang University, Beijing, China, in 2011 and 2014, respectively, and the Ph.D. degree in computer engineering from Northeastern University, Boston, MA, USA, in 2018.

He is currently a Tenure-Track Assistant Professor with the Michtom School of Computer Science, Brandeis University, Waltham, MA, USA. His current research interests include data mining and machine learning, especially ensemble learning.

Dr. Liu has served as a reviewer for many IEEE Transactions and journals including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), and the IEEE TRANSACTIONS ON BIG DATA (TBD). He has also served on the program committee for the conferences including AAAI, IJCAI, and NIPS. He is an Associate Editor of *IEEE Computational Intelligence Magazine*.

Sheng Li (S'11–M'17–SM'19) received the B.Eng. degree in computer science and engineering and the M.Eng. degree in information security from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2010 and 2012, respectively, and the Ph.D. degree in computer engineering from Northeastern University, Boston, MA, USA, in 2017.

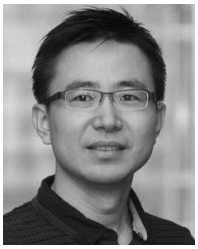
From 2017 to 2018, he was a Research Scientist with Adobe Research. Since 2018, he has been a Tenure-Track Assistant Professor with the Department of Computer Science, University of Georgia. He has authored or co-authored more than 70 papers at leading conferences and journals. His current research interests include robust machine learning, deep learning, visual intelligence, and behavior modeling.

Dr. Li was a recipient of the Adobe Data Science Research Award in 2019 and the best paper awards (or nominations) at SDM 2014, IEEE ICME 2014, and IEEE FG 2013. He serves as an Associate Editor of *IEEE Computational Intelligence Magazine*, *Neurocomputing*, *IET Image Processing*, and *SPIE Journal of Electronic Imaging*. He serves on the Editorial Board of *Neural Computing and Applications*. He has also served as a reviewer for several IEEE Transactions and a Program Committee Member for NIPS, ICML, IJCAI, AAAI, CVPR, ICCV, and KDD.

Zhengming Ding (S'14–M'18) received the B.Eng. degree in information security and the M.Eng. degree in computer software and theory from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2010 and 2013, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA, in 2018.

Since 2018, he has been a Faculty Member with the Department of Computer, Information and Technology, Indiana University–Purdue University Indianapolis, Indianapolis, IN, USA. His current research interests include transfer learning, multiview learning, and deep learning.

Dr. Ding was a recipient of the National Institute of Justice Fellowship from 2016 to 2018, the Best Paper Award (SPIE 2016), and the Best Paper Candidate (ACM MM 2017). He is currently an Associate Editor of the *Journal of Electronic Imaging* (JEI).



Yun Fu (S'07–M'08–SM'11–F'19) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, Xi'an, China, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign.

Since 2012, he has been an Interdisciplinary Faculty Member with the College of Engineering and Khoury College of Computer and Information Sciences, Northeastern University, Boston, MA, USA. He has authored or co-authored leading journals, books/book chapters, and international conferences/workshops. His current research interests include machine learning, computational intelligence, big data mining, computer vision, pattern recognition, and cyber-physical systems.

Dr. Fu is a fellow of IAPR, OSA, and SPIE, a Lifetime Distinguished Member of ACM, a Lifetime Member of AAAI and Institute of Mathematical Statistics, a member of ACM Future of Computing Academy, Global Young Academy, AAAS, INNS, and a Beckman Graduate Fellow from 2007 to 2008. He was a recipient of seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, Grainger Foundation; nine Best Paper Awards from IEEE, IAPR, SPIE, SIAM; and many major Industrial Research Awards from Google, Samsung, and Adobe, etc. He serves as an Associate Editor, the Chair, a PC Member, and a reviewer for many top journals and international conferences/workshops. He is currently an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEANING SYSTEMS (TNNLS) and the IEEE TRANSACTIONS OF IMAGE PROCESSING (TIP).