On Reliable Multi-View Affinity Learning for Subspace Clustering

Xiaolin Xiao, *Member, IEEE*, Yue-Jiao Gong, *Senior Member, IEEE*, Zhongyun Hua, *Member, IEEE*, and Wei-Neng Chen, *Senior Member, IEEE*

***Abstract*—In multi-view subspace clustering, the low-rankness of the stacked self-representation tensor is widely accepted to capture the high-order cross-view correlation. However, using the nuclear norm as a convex surrogate of the rank function, the self-representation tensor exhibits strong connectivity with dense coefficients. When noise exists in the data, the generated affinity matrix may be unreliable for subspace clustering as it retains the connections across inter-cluster samples due to the lack of sparsity. Since both the connectivity and sparsity of the self-representation coefficients are curial for subspace clustering, we propose a Reliable Multi-View Affinity Learning (RMVAL) method so as to optimize both properties in a single model. Specifically, RMVAL employs the low-rank tensor constraint to yield a well-connected yet dense solution, and purifies the densely connected self-representation tensor by preserving only the connections in local neighborhoods using the l_1-norm regularization. This way, the strong connections on the self-representation tensor are retained and the trivial coefficients corresponding to the inter-cluster connections are suppressed, leading to a "clean" self-representation tensor and also a reliable affinity matrix. We propose an efficient algorithm to solve RMVAL using the alternating direction method of multipliers. Extensive experiments on benchmark databases have demonstrated the superiority of RMVAL.**

***Index Terms*—Affinity learning, connectivity and sparsity, low-rank tensor, multi-view subspace clustering, self-representation.**

I. Introduction

MULTIMEDIA data naturally exhibit high dimensionality and complexity in the ambient space. Usually, the high-dimensional data approximately lie in low-dimensional subspaces [1], [2], and it is possible to reconstruct the data using sparse coefficients over proper dictionaries. Considering this fact, the sparse coding and dictionary learning methods were developed with theoretical analysis and various applications [3], [4]. Later, the work in [5] shows that the database itself can be used as the dictionary as long as it "linearly spans the data space". In addition to the sparsity of the encoding coefficients, advanced properties of the coefficients, e.g., low-rankness [5], block-diagonality [6], were exploited to produce the affinity matrices, falling into the category of self-representation learning. Among the applications on self-representation learning, subspace clustering uses the representation coefficients to form the affinity matrix [5], [6], to which the spectral clustering is applied for inferring the clusters [7]. The good clustering results require reliable affinity matrices as inputs. However, due to the existence of noise and corruptions, it is non-trivial to learn accurate affinity matrices directly from the raw data [8], [9]. Moreover, the multimedia data usually endure the ambiguity, which further challenges the construction of reliable affinities since a single type of feature may not be able to portray the data relationship [10].

To resolve the ambiguity, features from heterogeneous sources or different descriptors are extracted to depict the multimedia data, and multi-view subspace clustering thus takes advantage of the complementary information residing in different views [11], [12]. Among the extensive studies, the low-rank-representation-based methods [10], [13]–[21] have become the mainstream owning to its robustness. The fundament beneath is to impose the low-rank constraints on the self-representation matrices/tensors so as to exploit the global data structure for noise removal. Continuing along this vein, many multi-view subspace clustering algorithms were developed considering the consistency and/or diversity across views. A pioneer work [13] constructed a large feature matrix by vertically concatenating all feature matrices to ensure the cross-view consistency. Nie *et al.* enforced the affinities to be consistent with the local manifolds [22]. Later, Wang *et al.* utilized the complementarity of multi-view features via the exclusivity-consistency regularization [23]. Considering the fact that previous works essentially ignore the high-order correlation across views, many methods [10], [14]–[21] were proposed with different high-order correlation measures as well as various regularizers.

Although promising performance is witnessed, the aforementioned methods cannot guarantee the reliability of the affinities since they overlooked the properties, i.e, connectivity and

Manuscript received July 15, 2020; revised October 22, 2020; accepted November 29, 2020. Date of publication December 18, 2020; date of current version December 9, 2021. This work was supported in part by the Key Project of Science and Technology Innovation 2030 supported by the Ministry of Science and Technology of China under Grant 2018AAA0101300, in part by the National Natural Science Foundation of China under Grants 62 006 080, 61 873 095, and 62 071 142, in part by China Postdoctoral Science Foundation under Grant 2019M662913, and in part by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xin Geng. *(Corresponding authors: Yue-Jiao Gong; Zhongyun Hua.)*

Xiaolin Xiao, Yue-Jiao Gong, and Wei-Neng Chen are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510 006 China (e-mail: shellyxiaolin@gmail.com; gongyuejiao@gmail.com; cwnraul634@aliyun.com).

Zhongyun Hua is with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Shenzhen 518 055, China (e-mail: huazyum@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2020.3045259.

Digital Object Identifier 10.1109/TMM.2020.3045259

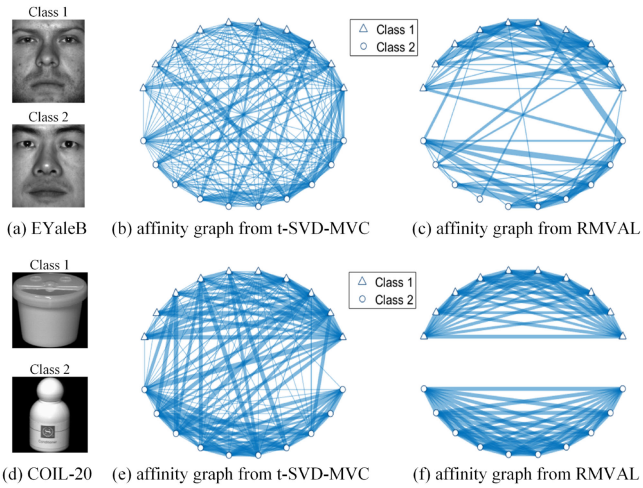

Fig. 1. Affinity graphs of two classes from (b), (c) Extended YaleB and (e), (f) COIL-20 (ten samples per class for good visualization).

sparsity, of the affinity matrix [8], [24], [25]. However, both the connectivity and sparsity of the affinity matrix play important roles in spectral clustering. Specifically, by promoting the intra-cluster connectivity, the affinity matrix shows robustness to the over-segmentation problem [25]. On the other hand, the sparsity of the affinity matrix encourages a subspace-preserving structure, i.e., samples are connected only if they come from the same subspace.[1] It has been validated that the nuclear or l_2 norm regularization leads to a dense solution [24] and the l_1 or l_0 norm results in a sparse representation [25]. In light of these facts, several methods have been developed to make a tradeoff between the connectivity and sparsity. They can be roughly grouped into three categories. First, the models in [26]–[28] bridged the gap between the connectivity and sparsity using mixed norms. Also, techniques belonging to the second category were developed to enhance the sparsity from densely connected affinity matrices by post-processing [9], [24] or iterative thresholding [8], [29]. In contrast, the works in [25], [30] were proposed to seek denser solutions from the initially sparse affinity matrices. Nevertheless, these methods consider only the single-view feature, which are not robust to the potential ambiguity residing in the multimedia data.

Considering these facts, we propose a novel Reliable Multi-View Affinity Learning (RMVAL) model for subspace clustering. The reliability of the affinity matrix is ensured by simultaneously optimizing the connectivity and sparsity properties. To illustrate the joint connectivity and sparsity, the affinity graphs learned from the proposed RMVAL and those from a typical work t-SVD-MVC [16] are given in Fig. 1, where the width of connections indicates the strength of edges. It is clear that, compared with the densely-connected affinity graphs produced from t-SVD-MVC, the intra-cluster connections are well preserved while the inter-cluster edges are suppressed using RMVAL. The general framework of the proposed RMVAL model is presented in Fig. 2. Specifically, given multi-view features, the view-specific relationship of samples is captured by the self-representation matrices (Step (a)). By stacking the self-representation matrices into a third-order tensor and then rotating, the high-order cross-view correlation is exploited with the low-rank tensor constraint, yielding a densely connected solution (Step (b)). Meanwhile, the self-representation tensor is encouraged to be sparse by preserving only the strongly connected neighbors via the l_1-norm regularization (Step (c)). Adopting an iterative optimization scheme, the "clean" self-representation tensor, in turn, promotes the learning of view-specific self-representation matrices (Step (d)). Finally, the average of the optimal self-representation tensor along the third dimension is used to generate the affinity matrix, to which the standard spectral clustering is applied for producing the clusters.

To sum up, the novelty and contributions of this work lie in the following aspects:

1) We propose a Reliable Multi-View Affinity Learning (RMVAL) model that enhances the reliability of the affinity matrix by simultaneously optimizing the connectivity and sparsity properties. By doing so, the intra-cluster samples exhibit strong connections and the inter-cluster samples are disconnected, which simultaneously prunes erroneous connections and avoids over-segmentation. To the best of our knowledge, this is the first model that guarantees both properties within a single model in the setting of multi-view affinity learning.

2) RMVAL well exploits the global and local data structures to optimize the affinity matrix. Specifically, the global structural constraint is adopted to ensure the high-order cross-view correlation via the low-rank tensor constraint. Considering the local data structure, in contrast to existing methods that learn smooth self-representation coefficients with local manifolds, RMVAL proposes to pursue the sparsity of the self-representation tensor by preserving only the strong connections in local neighborhoods. Therefore, it provides a simple yet effective way to purify the densely connected self-representation tensor, and makes it feasible to improve the robustness of the learned affinity.

3) We devise an efficient optimization algorithm to solve RMVAL using the alternating direction method of multipliers. Extensive experiments have demonstrated the effectiveness of the proposed RMVAL model.

In the rest of this paper, Section II reviews the related works and preliminaries; Section III introduces the RMVAL model; the experimental results and model analysis are presented in Section IV; finally, conclusions are drawn in Section V.

II. Related Work and Preliminaries

In this section, we first review the state-of-the-art multi-view affinity learning models and the pioneer works for affinity purification. Afterward, the notations and tensor representations are introduced to facilitate the formulation of RMVAL.

A. Multi-View Affinity Learning

By far, the self-representation-based multi-view affinity learning has attracted great attention [10]. Existing methods apply different principles to integrate the view-specific self-representation matrices to obtain the final affinity matrix. A

[1]The sparsity/subspace-preserving property is also called the block-diagonal property in [6] where nonzero entries exist when samples come from the same subspace.

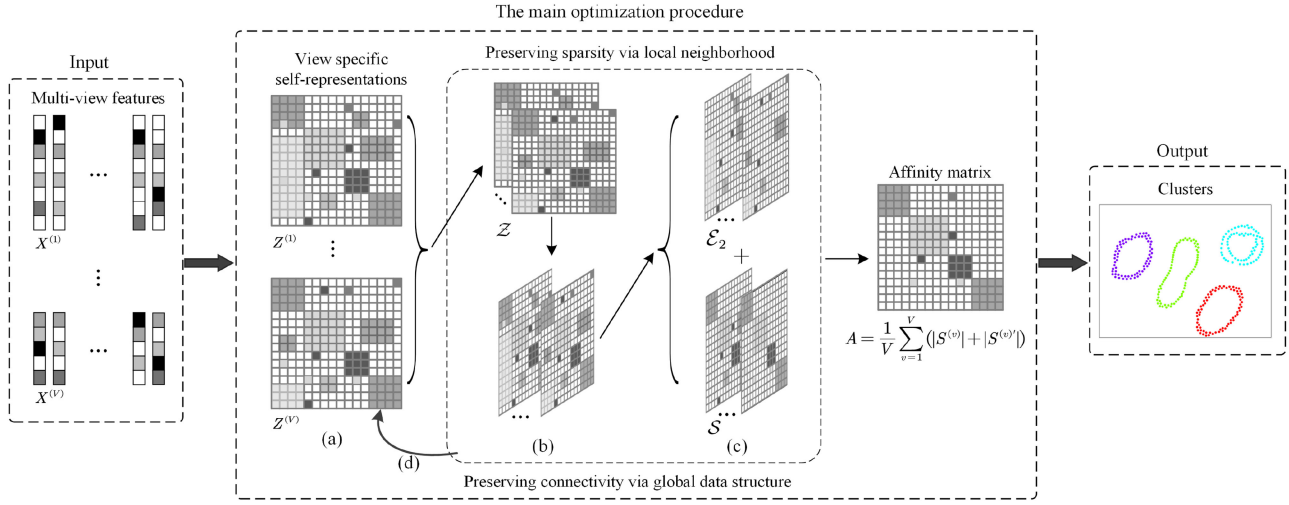

Fig. 2. Framework of the proposed RMVAL model.

common strategy is to explore the multi-view consistency. Zhang *et al.* [14] stacked the representation matrices into a third-order tensor and imposed the unfolding-based low-rank tensor constraint to preserve the consistency within all tensor modes. To overcome the representation deficiency of the unfolding operation, the tensor low-rankness derived from tensor-product is exploited to holistically capture the global structure of the self-representation tensor [15], [16]. The t-SVD-MSC model proposed in [16] reached a milestone for multi-view learning by jointly exploring the complicated cross-sample and cross-view relationship. Since t-SVD-MSC considers only the global data structure, many followers [17]–[19], [22] further improve the performance by employing the local manifolds and/or the nonlinear data structures. Additionally, the weights of different views are evaluated to emphasize the informative views [10], [18], [31]. Besides, several works devised different strategies to relieve the high computation burdens [19], [20], [32].

B. Purifying the Densely Connected Affinity Matrix

The self-representation-based multi-view affinity models represent samples as linear combinations of other samples in the same database and enforce the specific properties, e.g., sparse, low-rank, block-diagonal, to produce the affinity matrices. Recently, the low-rank constraint is more popular since it preserves the global data structure with a low complexity. However, the generated affinity matrix tends to be densely connected since the nuclear norm is used to approximate the rank function. As pointed out in [8], [24], [25], an optimal affinity matrix should not only exhibit strong intra-cluster connectivity but also be sparse to produce a subspace-preserving solution. To this end, different methods were developed to purify the dense affinity matrix.

Note that, owing to the property of the "intra-subspace projection dominance," the intra-cluster self-representation coefficients tend to be relatively larger than the inter-cluster coefficients in many cases [8]. Thus, a straightforward idea is to simply preserve several largest coefficients on the affinity matrix. However, this may lead to a connectivity issue that produces over-segmented subspaces since samples are not well-connected [33]. Peng *et al.* [8] filtered out the trivial values on the representation coefficients using iterative thresholding ridge regression, which partially resolves this problem by the iterative scheme. Yet, this method lacks the global structural constraint, leading to unsatisfactory clustering results. The work in [9] devised a post-processing technique to purify the densely connected affinity graph via selecting "good neighbors," which are defined as the neighbors that maintain shared neighbors. By keeping only the good neighbors, the affinity matrix is endowed with both connectivity and sparsity properties. Nevertheless, we experimentally found that the effectiveness of good neighbors relies heavily on the number of selected good neighbors, requiring the prior knowledge on the size of the clusters. However, it is impractical to estimate the number of samples per cluster in real applications. Moreover, on the unbalanced databases, such a number may not even exist. This observation can be confirmed by the experiments in Section IV-C.

C. Tensor Representation

In this paper, the uppercase and calligraphy letters denote the matrices and the third-order tensors respectively. Given $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we can split $\mathcal{X}$ into submodules $\{X^{(v)}\}_{v=1}^{n_3}$ where $X^{(v)}$ denotes the v-th frontal slice. The Frobenius norm (F-norm) and l_1-norm of $\mathcal{X}$ are defined as $\|\mathcal{X}\|_F := (\sum_{i,j,k} |\mathcal{X}(i,j,k)|^2)^{\frac{1}{2}}$ and $\|\mathcal{X}\|_1 := \sum_{i,j,k} |\mathcal{X}(i,j,k)|$ respectively. $\mathcal{X}_f := fft(\mathcal{X}, [\,], 3)$ applies Fast Fourier Transform (FFT).

The tensor-product (t-product) generalizes the matrix multiplication for multi-way data [34]. Based on t-product, the tensor low-rankness is well defined to model the global structure of the multi-way data. Here, we briefly review the main operators and the readers can refer to [34]–[37] for a comprehensive study.

Definition 1: [34] As shown in Fig. 3, for $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, the Singular Value Decomposition (**t-SVD**) is defined as $\mathcal{X} := \mathcal{W} * \mathcal{A} * \mathcal{B}'$, where $*$ denotes t-product, $\mathcal{W} \in \mathbb{R}^{n_1 \times n_1 \times n_3}$ and $\mathcal{B} \in \mathbb{R}^{n_2 \times n_2 \times n_3}$ are orthogonal tensors, and $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is f-diagonal (i.e., all frontal slices are diagonal matrices).

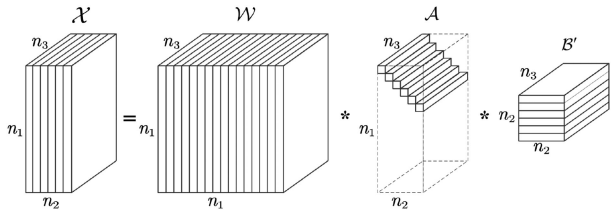

Fig. 3. T-SVD of an $n_1 \times n_2 \times n_3$ tensor $\mathcal{X}$.

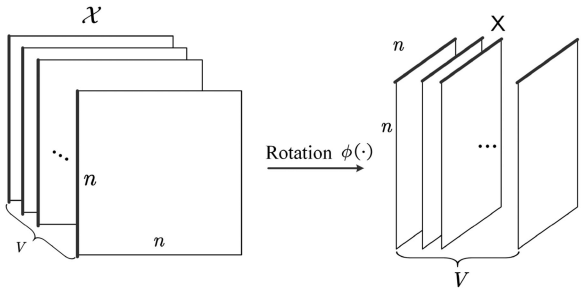

Fig. 4. Rotation of the representation tensor.

Definition 2: [16] The t-SVD-based Tensor Nuclear Norm (**t-SVD-TNN**) of $\mathcal{X}$ is defined as the sum of the singular values of all frontal slices of $\mathcal{X}_f$, i.e., $\|\mathcal{X}\|_{\circledast} := \sum_{k=1}^{n_3} \|X_f^{(k)}\|_* := \sum_{i=1}^{min\{n_1,n_2\}} \sum_{k=1}^{n_3} |A_f^{(k)}(i,i)|$, where $A_f^{(k)}$ comes from the complex-valued matrix SVD where $X_f^{(k)} = W_f^{(k)} A_f^{(k)} B_f^{(k)'}$.

The t-SVD-TNN is proved to be the tightest convex relaxation to the l_1-norm of the tensor multi-rank (Theorem 2.4.1, [38]). Nevertheless, for capturing the high-order relationship in multi-view learning, directly applying the t-SVD-TNN on the representation tensor is sub-optimal since the t-SVD-TNN is orientation-dependent [16], and the following rotation operator is adopted to solve this problem.

Definition 3: [16] As shown in Fig. 4, a rotation operator is defined to transform $\mathcal{X} \in \mathbb{R}^{n\times n\times V}$ to $\mathsf{X} \in \mathbb{R}^{n\times V\times n}$ as $\mathsf{X} = \phi(\mathcal{X}) = shiftdim(\mathcal{X}, 1)$.

In this work, we use $\circledast$ to represent the t-SVD-TNN imposed on the rotated representation tensor. Finally, the basic notations are summarized in Table I.

III. Reliable Multi-View Affinity Learning

A. Motivation

The t-SVD-based Multi-view Subspace Clustering (t-SVD-MSC) is a typical model for multi-view learning that imposes the t-SVD-TNN on the rotated self-representation tensor, so as to simultaneously and thoroughly utilize the high-order correlation across both views and samples. The t-SVD-MSC model is formulated as

$$\begin{aligned} \min_{\mathcal{Z},E}\ & \|\mathcal{Z}\|_{\circledast} + \lambda\|E\|_{2,1} \\ \text{s.t. } & \{X^{(v)} = X^{(v)}Z^{(v)} + E^{(v)}\}_{v=1}^{V}, \\ & \mathcal{Z} = (Z^{(1)}, Z^{(2)}, \ldots, Z^{(V)}), \\ & E = [E^{(1)}; E^{(2)}; \cdots; E^{(V)}], \end{aligned} \tag{1}$$

TABLE I
Summary of Commonly Used Notations

Notation	Description
X	matrix
$\mathcal{X}$	third-order tensor
$X^{(v)} = \mathcal{X}(:,:,v)$	v-th frontal slice of $\mathcal{X}$ (i.e, matrix)
$\{X^{(v)}\}_{v=1}^{n_3}$	collection of frontal slices of $\mathcal{X}$
$\mathcal{X}_f$	$\mathcal{X}$ in the Fourier space (3-rd direction)
$\mathsf{X} = \phi(\mathcal{X})$	rotated $\mathcal{X}$ (shown in Fig. 2)
$\mathcal{X} * \mathcal{Y}$	t-product of $\mathcal{X}$ and $\mathcal{Y}$
$'$	(conjugate) transpose
$\lvert\cdot\rvert$	absolute value
$\|\cdot\|_*$	matrix nuclear norm
$\|\cdot\|_1$	matrix/tensor l_1-norm
$\|\cdot\|_{2,1}$	matrix $l_{2,1}$-norm
$\|\cdot\|_{\bar{\circledast}}$	t-SVD-based Tensor Nuclear Norm
$\|\cdot\|_{\circledast}$	t-SVD-TNN on rotated tensor
$(\cdot)_+$	positive part of $(\cdot)$

$\{X^{(v)}\}$ is interchangeable with $\{X^{(v)}\}_{v=1}^{n_3}$ for concise.

where $X^{(v)}$ is the feature matrix from the v-th view; $\mathcal{Z} \in \mathbb{R}^{n\times n\times V}$ is obtained by stacking the representation matrices $\{Z^{(v)}\}$ along the third dimension, and $\|\mathcal{Z}\|_{\circledast}$ imposes the t-SVD-TNN on the rotated $\mathcal{Z}$; E is constructed by vertically concatenating the view-specific error matrices, and $\|E\|_{2,1} = \sum_j (\sum_i |E(i,j)|^2)^{\frac{1}{2}}$. The $l_{2,1}$-norm is commonly adopted to deal with sample-specific corruptions and outliers [5]. Afterward, the optimal affinity matrix is obtained from $A = \frac{1}{V}\sum_{v=1}^{V}(|Z^{(v)}| + |Z^{(v)'}|)$.

Although t-SVD-MVC significantly outperforms previous methods, the t-SVD-TNN leads to a well-connected solution with dense coefficients. As such, the intra-cluster connectivity of the affinity matrix is guaranteed whereas the sparsity is sacrificed. This way, the affinity matrix learned from t-SVD-MSC may not be sufficiently reliable, as a dense solution is likely to contain erroneous connections when noise exists.

B. Proposed Model

Considering the limitation of t-SVD-MSC, we propose a Reliable Multi-View Affinity Learning (RMVAL) model to jointly optimize the connectivity and sparsity of the affinity matrix. RMVAL achieves the reliability by simultaneously preserving well-connected samples and pruning erroneous connections, resulting in effective and accurate clustering results. Our RMVAL model is formulated as

$$\begin{aligned} \min_{\substack{\{Z^{(v)}\},\mathcal{S},\\ E_1,\mathcal{E}_2}}\ & \sum_{v=1}^{V} \|Z^{(v)}\|_* + \lambda_1\|\mathcal{S}\|_{\circledast} + \lambda_2\|E_1\|_{2,1} + \lambda_3\|\mathcal{E}_2\|_1 \\ \text{s.t. } & \{X^{(v)} = X^{(v)}Z^{(v)} + E_1^{(v)}\}_{v=1}^{V}, \\ & E_1 = [E_1^{(1)}; E_1^{(2)}; \cdots; E_1^{(V)}], \\ & \mathcal{Z} = (Z^{(1)}, Z^{(2)}, \ldots, Z^{(V)}), \\ & \mathcal{Z} = \mathcal{S} + \mathcal{E}_2, \end{aligned} \tag{2}$$

where $\mathcal{Z} = (Z^{(1)}, Z^{(2)}, \ldots, Z^{(V)})$ coincides with that in Eq. (1) and E_1 equals to E in Eq. (1). We decompose $\mathcal{Z}$ into $\mathcal{S}$ and $\mathcal{E}_2$, corresponding to the "clean" self-representation tensor and trivial inter-cluster connections respectively. Accordingly, the rotated t-SVD-TNN $\|\cdot\|_{\circledast}$ and the l_1-norm are used to preserve the global data structure and the local neighborhoods respectively. The affinity matrix is then set to $A = \frac{1}{V}\sum_{v=1}^{V}(|S^{(v)}| + |S^{(v)'}|)$.

Please note that, it is non-trivial to construct a reliable affinity matrix from multi-view features mainly because (1) the noise and corruptions naturally exist in the raw features over all views, and (2) to deal with the high-order cross-view correlation, the global data structure should be considered, which induces dense solutions and tends to destroy the subspace-preserving property. In view of improving the reliability of the affinity matrix, RMVAL advances in the following aspects:

- To solve (1), for each type of feature, we use the low-rank induced self-representation matrix $Z^{(v)}$ to capture the global data structure. Going beyond the local similarities, the low-rank constraint shows robustness since it can approximately recover the subspace structures when the feature matrix is contaminated. Moreover, we concatenate the error matrices from different views and use the $l_{2,1}$-norm to encourage the columns of $\{E_1^{(v)}\}_{v=1}^{V}$ to have consistent magnitudes. Thus, it can effectively handle the sample-specific corruptions. This way, both the intra-view information and the consistency across all views are well considered for eliminating the noise and corruptions.
- For (2), by representing $\mathcal{Z} = \mathcal{S} + \mathcal{E}_2$ via the tube-wise notation, $\mathcal{Z}(:,j,k) = \mathcal{S}(:,j,k) + \mathcal{E}_2(:,j,k)$ essentially decomposes $\mathcal{Z}(:,j,k)$ into two parts: $\mathcal{S}(:,j,k)$ represents the compact representation coefficients for the j-th sample from the k-th view and $\mathcal{E}_2(:,j,k)$ denotes the trivial values corresponding to the unwelcome inter-cluster connections. Thus, we apply the t-SVD-TNN on $\mathcal{S}$ to find a densely connected solution. On the other hand, the l_1-norm encourages a sparse solution since it preserves only the large coefficients, i.e, strong connections in local neighborhoods.

C. Optimization

It is intractable to solve Eq. (2) directly since variables are coupled. We thus devise an efficient optimization algorithm within the framework of the Alternating Direction Method Of Multipliers (ADMM) [39]. By introducing an auxiliary set $\{U^{(v)} = Z^{(v)}\}$, the augmented Lagrange function of Eq. (2) is formulated as

$$\begin{aligned}\mathcal{L}(\{Z^{(v)}\},\{U^{(v)}\},E_1,\mathcal{E}_2,\mathcal{S}) &= \sum_{v=1}^{V}\|U^{(v)}\|_* + \lambda_1\|\mathcal{S}\|_{\circledast} \\ &+ \lambda_2\|E_1\|_{2,1} + \lambda_3\|\mathcal{E}_2\|_1 + \frac{\rho}{2}\left(\sum_{v=1}^{V}\left(\|U^{(v)} - Z^{(v)} + \frac{Y_1^{(v)}}{\rho}\|_F^2\right.\right. \\ &\left.\left. + \|X^{(v)} - X^{(v)}Z^{(v)} - E_1^{(v)} + \frac{Y_2^{(v)}}{\rho}\|_F^2\right) + \|\mathcal{Z} - \mathcal{S} - \mathcal{E}_2 + \frac{\mathcal{Y}_3}{\rho}\|_F^2\right),\end{aligned} \tag{3}$$

where $\{Y_1^{(v)}\}$, $\{Y_2^{(v)}\}$, $\mathcal{Y}_3$ are the Lagrange multipliers, and $\rho > 0$ is the penalty parameter. The variables $\{Z^{(v)}\}, \{U^{(v)}\}, E_1, \mathcal{E}_2, \mathcal{S}$ can be alternately optimized by minimizing Eq. (3) when other variables are fixed.

1) $\{Z^{(v)}\}$–subproblem: The optimization with respect to $\{Z^{(v)}\}$ is view-independent, and we take the v-th view as an example. Fixing other variables except $Z^{(v)}$, the problem reduces to

$$\begin{aligned}\min_{Z^{(v)}} \|U^{(v)} - Z^{(v)} + \frac{Y_1^{(v)}}{\rho}\|_F^2 + \|X^{(v)} - X^{(v)}Z^{(v)} \\ - E_1^{(v)} + \frac{Y_2^{(v)}}{\rho}\|_F^2 + \|Z^{(v)} - S^{(v)} - E_2^{(v)} + \frac{Y_3^{(v)}}{\rho}\|_F^2,\end{aligned} \tag{4}$$

where $S^{(v)}$, $E_2^{(v)}$, $Y_3^{(v)}$ are the v-th frontal slices of $\mathcal{S}$, $\mathcal{E}_2$, $\mathcal{Y}_3$ respectively. The closed-form solution to Eq. (4) is obtained by setting its derivation to zero. Therefore,

$$Z^{(v)*} = (X^{(v)}X^{(v)'} + 2I)^{-1}(M_1 + X^{(v)'}M_2 + M_3), \tag{5}$$

where the temporal variables are set to $M_1 = U^{(v)} + \frac{Y_1^{(v)}}{\rho}$, $M_2 = X^{(v)} - E_1^{(v)} + \frac{Y_2^{(v)}}{\rho}$, and $M_3 = S^{(v)} + E_2^{(v)} - \frac{Y_3^{(v)}}{\rho}$.

2) $\{U^{(v)}\}$–subproblem: The auxiliary variables $\{U^{(v)}\}$ are introduced to separate the view-specific low-rank constraint and self-representation learning, and the optimal $U^{(v)}$ is obtained by solving

$$\min_{U^{(v)}} \|U^{(v)}\|_* + \|U^{(v)} - Z^{(v)} + \frac{\rho}{2}\frac{Y_1^{(v)}}{\rho}\|_F^2. \tag{6}$$

Eq. (6) can be optimized via the singular value shrinkage introduced in Theorem 1 (Theorem 2.1, [40]).

Theorem 1: Given a matrix M_4 whose SVD is denoted by $M_4 = ABC'$ and a constant $\sigma_1 > 0$, the optimal solution to $\min_U \frac{1}{2}\|U - M_4\|_F^2 + \sigma_1\|U\|_*$ is obtained at $U^* = A(B - diag(\sigma_1))_+C'$, where $diag(\sigma_1)$ constructs a diagonal matrix whose sizes equal to those of B and the diagonal elements are σ_1; $(\cdot)_+$ denotes the positive part of $(\cdot)$, namely, $(B - diag(\sigma_1))_+ = \max(B - diag(\sigma_1), 0)$.

3) E_1–subproblem: Fixing other variables except E_1 and introducing a temporary variable M_5 by concatenating $\{X^{(v)} - X^{(v)}Z^{(v)} + \frac{Y_1^{(v)}}{\rho}\}_{v=1}^{V}$ along the column direction, the E_1–subproblem can be optimized by solving a group Lasso problem

$$\min_{E_1} \frac{1}{2}\|E_1 - M_5\|_F^2 + \frac{\lambda_2}{\rho}\|E_1\|_{2,1}. \tag{7}$$

We introduce the following Theorem 2 (Lemma 3.1, [41]) to solve Eq. (7).

Theorem 2: Given a matrix M_5 and a constant $\sigma_2 > 0$, the optimal solution to $\min_{E_1} \frac{1}{2}\|E_1 - M_5\|_F^2 + \sigma_2\|E_1\|_{2,1}$ is obtained at $E_1^*(:,j) = (1 - \frac{\sigma_2}{\|M_5(:,j)\|_2})_+ M_5(:,j)$, where j is the column index.

4) $\mathcal{E}_2$–subproblem: The optimization with respect to $\mathcal{E}_2$ reduces to a tensor lasso problem:

$$\min_{\mathcal{E}_2} \frac{1}{2}\|\mathcal{E}_2 - \mathcal{Z} + \mathcal{S} - \frac{\mathcal{Y}_3}{\rho}\|_F^2 + \frac{\lambda_3}{\rho}\|\mathcal{E}_2\|_1. \tag{8}$$

Let $\mathcal{M}_6 = \mathcal{Z} - \mathcal{S} + \frac{\mathcal{Y}_3}{\rho}$, $\mathbf{e}_2$ and $\mathbf{m}_6$ be the vectorization of $\mathcal{E}_2$ and $\mathcal{M}_6$ respectively. Eq. (8) is equivalent to minimize $\frac{1}{2}\|\mathbf{e}_2 - \mathbf{m}_6\|_2^2 + \frac{\lambda_3}{\rho}\|\mathbf{e}_2\|_1$, and the solution is obtained at

$$\mathbf{e}_2^* = (1 - \frac{\lambda_3/\rho}{|\mathbf{m}_6|})_+ \mathbf{m}_6. \tag{9}$$

The optimal solution to Eq. (8) is then obtained by reshaping $\mathbf{e}_2$ into a tensor form.

5) $\mathcal{S}$–subproblem: Fixing other variables except $\mathcal{S}$, the problem equals to solve

$$\min_{\mathcal{S}} \frac{1}{2}\|\mathcal{S} - \mathcal{Z} + \mathcal{E}_2 - \frac{\mathcal{Y}_3}{\rho}\|_F^2 + \frac{\lambda_1}{\rho}\|\mathcal{S}\|_{\circledast}. \tag{10}$$

Eq. (10) is the t-SVD-TNN minimization problem that can be solved via the tensor tubal-shrinkage presented in Theorem 3 (Theorem 2, [16]).

Theorem 3: Given a tensor $\mathcal{M}_7$ and a constant $\sigma_3 > 0$, let the t-SVD of $\mathcal{M}_7$ be $\mathcal{ABC}'$, the optimal solution to $\min_{\mathcal{S}} \frac{1}{2}\|\mathcal{S} - \mathcal{M}_7\|_F^2 + \sigma_3\|\mathcal{S}\|_{\circledast}$ is obtained at $\mathcal{S}^* = \mathcal{A} * \theta_{n_3\sigma_3}(\mathcal{B}) * \mathcal{C}'$, where $\theta_{n_3\sigma_3}(\mathcal{B}) = \mathcal{B} * \mathcal{J}$ and $\mathcal{J}$ is an f-diagonal tensor whose diagonal element in the Fourier domain is $\mathcal{J}(i,i,j) = (1 - \frac{n_3\sigma_3}{\mathcal{B}(i,i,j)})_+$.

6) Multipliers and penalty parameter: In each iteration, the multipliers and penalty parameter are updated as

$$\begin{aligned}
\{Y_1^{(v)*}\} &= \{\Theta_1^{(v)} + \rho(U^{(v)} - Z^{(v)})\};\\
\{Y_2^{(v)*}\} &= \{\Theta_2^{(v)} + \rho(X^{(v)} - X^{(v)}Z^{(v)} - E_1^{(v)})\};\\
\mathcal{Y}_3^* &= \Theta_3 + \rho(\mathcal{Z} - \mathcal{S} - \mathcal{E}_2);\\
\rho^* &= \min\{\beta * \rho,\ \rho_{\max}\}.
\end{aligned} \tag{11}$$

The parameter β is introduced to adopt the varying penalty parameter scheme until a maximum value $\rho_{\max}$ is achieved [39]. By doing so, the empirical convergence speed is accelerated and the performance is less dependent to the initialization of ρ.

Since Eq. (3) consists three constraints, we define the corresponding residuals as

$$\begin{aligned}
r1 &= \max\{\|U^{(v)} - Z^{(v)}\|_\infty\};\\
r2 &= \max\{\|X^{(v)} - X^{(v)}Z^{(v)} - E_1^{(v)}\|_\infty\};\\
r3 &= \|\mathcal{Z} - \mathcal{S} - \mathcal{E}_2\|_\infty,
\end{aligned} \tag{12}$$

where $\|\cdot\|_\infty$ equals to the maximum number of the matrix/tensor.

The stopping criteria is met when all residuals are small enough as the algorithm proceeds. Finally, the affinity matrix is calculated from the optimal self-representation tensor and then used to yield clusters. The whole procedure of the RMVAL model is summarized in Algorithm 1.

Algorithm 1: RMVAL for Subspace Clustering

Input : Multi-view features: $\{X^{(v)}\}$; parameters: λ_1, λ_2, λ_3.
Output: Optimal clusters.
Initialize $\{Y_1^{(v)}\}, \{Y_2^{(v)}\}, \mathcal{Y}_3$ to $\mathbf{0}$; initialize ρ to 10^{-3}; $\rho^{max} = 10^{10}$, $\beta = 2$, $\epsilon = 10^{-7}$.
repeat
  Update $\{Z^{(v)}\}$ by Eq. (5).
  Update $\{U^{(v)}\}$ according to Theorem 1.
  Update E_1 according to Theorem 2.
  Update $\mathcal{E}_2$ by Eq. (9).
  Update $\mathcal{S}$ according to Theorem 3.
  Update multipliers and penalty parameter by Eq. (12).
until $r_1 < \epsilon$ & $r_2 < \epsilon$ & $r_3 < \epsilon$;
$A = \frac{1}{V}\sum_{v=1}^{V}\left(|S^{(v)}| + |S^{(v)'}|\right)$.
Applying the spectral clustering on A to find the clusters.

D. Discussion

In the following, we analyze the convergence property and the complexity of RMVAL.

1) Convergence Analysis: RMVAL is optimized within the ADMM framework, consisting five subproblems (blocks) with respect to $\{Z^{(v)}\}$, $\{U^{(v)}\}$, E_1, $\mathcal{E}_2$, and $\mathcal{S}$. Unfortunately, the convergence of ADMM for multi-block optimization cannot be proved in a theoretical manner [42]. Following previous works [16]–[18], [21], in Section IV-C5, we will investigate the empirical convergency of RMVAL. In brief, RMVAL exhibits good convergence behaviors in real scenarios.

2) Computation Complexity: To solve RMVAL, an iterative optimization algorithm is designed. Let T be the number of iterations, the computation costs within one iteration are calculated as:

- $\{Z^{(v)}\}$–subproblem: the cost of this subproblem is insignificant since the matrix inverse operation in Eq. (5) can be pre-computed and is used across all iterations and all views;
- $\{U^{(v)}\}$–subproblem: to solve Eq. (6), the singular value shrinkage operation necessitates the computation of matrix SVD, and thus, it needs $\mathcal{O}(Vn^3)$ for all views;
- the E_1 and $\mathcal{E}_2$ subproblems consist column-wise and element-wise thresholding respectively. Their costs are negligible;
- $\mathcal{S}$–subproblem: the t-SVD-TNN optimization consists of calculating 3D FFT, inverse FFT, matrix SVD and multiplication. Since the tensor tubal-shrinkage is applied on the rotated representation tensor and $n \gg V$, this subproblem is dominated by the FFT operations with the complexity of $\mathcal{O}(Vn^2\log(n))$.

As illustrated above, the total computation complexity of RMVAL is $\mathcal{O}(TVn^3)$.

IV. Experiments

In this section, RMVAL is compared with the state-of-the-arts to examine its effectiveness, and we also provide an in-depth analysis on the properties of RMVAL.

A. Experimental Settings

Databases. Six widely used databases are chosen for experiments with the contents varying on faces (EYaleB,[2] Notting-Hill [43]), scenes (Scene-15 [44], MITIndoor-67 [45]),

[2][Online]. Available: https://cvc.yale.edu/projects/yalefacesB/yalefacesB.html

TABLE II
STATISTICS OF THE DATASETS

Content	Dataset	Samples	Clusters	Samples per cluster
Face	EYaleB	640	10	64
	Notting-Hill	4460	5	595-1402
Scene	Sene-15	4485	15	210-410
	MITIndoor-67	5360	67	77-83
Object	COIL-20	1440	20	72
	Caltech-101	8677	101	31-800

and objects (COIL-20,[3] Caltech-101 [46]). The statistics of databases are reported in Table II. We follow the settings in [17] to generate the multi-view features. Specifically, the intensity, Local binary patterns, and Gabor features are extracted for EYaleB, Notting-Hill, and COIL-20. As to Scene-15, MITIndoor-67, and Caltech-101, we also extract three handcrafted features, i.e., pyramid histograms of visual words, pairwise rotation invariant co-occurrence local binary pattern, and census transform histogram. Since MITIndoor-67 and Caltech-101 are relatively challenging datasets for clustering, we further adopt the last layers from VGG19 and Inception-V3 networks[4] as the forth views respectively to take advantage of the powerful deep features.

Competitors. Eight state-of-the-art multi-view learning algorithms are chosen for comparison, including: (1) Low-rank Tensor constrained Multi-view Subspace Clustering (LT-MSC, in *ICCV 2015*) [14], (2) Multi-view Learning with Adaptive Neighbors (MLAN, in *AAAI 2017*) [22], (3) Exclusivity-Consistency regularized Multi-view Subspace Clustering (ECMSC, in *CVPR 2017*) [23], (4) t-SVD-based Multi-view Subspace Clustering (t-SVD-MSC, in *IJCV 2018*) [16], (5) Hyper-Laplacian Regularized Multi-linear Multi-View Self-representation (HLR-M^2 VS, in *TCybern 2018*) [17], (6) Essential Tensor Learning for Multi-view Spectral Clustering (ETLMSC in *TIP, 2019*) [20], (7) Jointly Learning kernel representation tensor and affinity matrix for Multi-View Clustering (JLMVC, in *TMM 2019*) [18], and (8) Unified Graph and Low-rank Tensor Learning (UGLTL, in *AAAI 2020*) [19]. In addition, the standard SPectral Clustering (SPC) [7] and the Low-Rank Representation (LRR) based clustering [5] are applied on all views, and the best clustering results are reported as performance baselines.

Among them, LT-MSC, ECMSC, t-SVD-MSC, HLR-M^2 VS, and JLMVC are the subspace-learning-based models since they use the self-representation matrices to capture the relationship of samples in each view. Our RMVAL model also belongs to this category. MLAN, ETLMSC, and UGLTL are graph-based models by capturing the view-specific relationship of samples using local graphs.

Evaluation Metrics. Following the literature [16]–[18], [21], we use six popular metrics for performance evaluation, i.e., ACCuracy (ACC), Normalized Mutual Information (NMI), Adjusted Rand index (AR), F-score, precision and recall. For all measures, higher values denote better performance. These metrics are correlated but each metric favors a specific characteristic of the clustering results. Usually, a comprehensive comparison is provided by jointly considering all metrics [16].

[3][Online]. Available: https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php

[4][Online]. Available: https://keras.io/zh/applications/

B. Performance Comparison

To avoid the interference of randomness, all experiments are repeated ten times and the mean results are recorded in Tables III–V. Overall, RMVAL obtains the best or comparable performance on databases with different contents. Specifically,

1) Face Image Clustering: It is challenging to cluster the images in EYaleB since this database is collected under well controlled positions and scales with extreme illumination conditions. For example, two images belonging to a same person but with large illumination variations may have a larger distance than two images from different persons taken in similar lighting conditions (shown in Fig. 3 in [47]). That is, the intra-cluster distances are prone to be larger than the inter-cluster distances when there exists severe illumination changes. As shown in the left part of Table III, RMVAL is the best-performing method on EYaleB and it outperforms the second best competitor JLMVC by the margins of 5.0%, 3.7%, 8.9%, 7.9%, 9.2%, 6.6% on the six evaluation metrics respectively. Meanwhile, ETLMSC and UGLTL produce unsatisfactory results probably because both methods directly adopt the Euclidean distance to calculate the local graphs, and thus may be easily affected by large illumination changes. This indicates that compared with the subspace-learning-based methods, the graph-based models, i.e., MLAN, ETLMSC, and UGLTL, suffer problems when the initial data graphs are inadequate to capture the relationship of samples.

On Notting-Hill, RMVAL can correctly recover all clusters, and HLR-M^2 VS ranks in the second place with near-optimal results. While HLR-M^2 VS explores the local manifolds using the hyper-Laplacian regularizer, RMVAL works in a simple yet effective way that emphasizes the connectivity and sparsity of the affinity matrix. Generally, compared to the results on EYaleB, all algorithms obtain much better performance on Notting-Hill. This is because the Notting-Hill database exhibits relatively small intra-cluster discrepancies since no large illumination variations are included [48] and thus the clustering task is relatively easy.

2) Scene Image Clustering: The Scene-15 database contains scene images from 10 outdoor classes (forest, highway, etc.) and 5 indoor categories (office, bedroom, etc.). The performance of different algorithms on Scene-15 are reported on the left part of Table IV. Specifically, RMVAL, JLMVC, and UGLTL obtain comparable results and show noticeable improvements over other methods. The competitive advantages of these three models primarily stem from the t-SVD-TNN imposed on the rotated representation tensor. Furthermore, they perform enhanced affinity learning with different concerns: JLMVC adopts the kernel trick to handle the nonlinear data structures; UGLTL learns the affinity matrix based on projected graph learning; meanwhile, RMVAL carefully balances the connectivity and sparsity of the affinity matrix via seeking the global and local data structures.

TABLE III
CLUSTERING RESULTS ON EYALEB AND NOTTING-HILL

	EYaleB ($\lambda_1 = 6$, $\lambda_2 = 0$, $\lambda_3 = 0$)						Notting-Hill ($\lambda_1 = 2$, $\lambda_2 = 0.2$, $\lambda_3 = 0.02$)					
	ACC	NMI	AR	F-score	Precision	Recall	ACC	NMI	AR	F-score	Precision	Recall
SPC$best$	0.366	0.360	0.255	0.308	0.296	0.310	0.816	0.732	0.712	0.775	0.780	0.776
LRR$best$	0.615	0.627	0.451	0.508	0.481	0.539	0.794	0.579	0.558	0.653	0.672	0.636
LT-MSC	0.626	0.637	0.459	0.521	0.485	0.539	0.868	0.779	0.777	0.825	0.830	0.814
MLAN	0.346	0.352	0.093	0.213	0.159	0.321	0.584	0.476	0.301	0.584	0.380	0.748
ECMSC	0.783	0.759	0.544	0.597	0.513	0.718	0.767	0.817	0.678	0.764	0.637	0.954
t-SVD-MSC	0.652	0.667	0.500	0.550	0.514	0.590	0.957	0.900	0.900	0.922	0.937	0.907
HLR-M^{2}VS	0.670	0.703	0.529	0.577	0.560	0.595	0.996	0.982	0.990	0.986	0.989	0.984
ETLMSC	0.325	0.307	0.179	0.262	0.257	0.267	0.951	0.911	0.898	0.924	0.940	0.908
JLMVC	0.910	0.897	0.832	0.849	0.837	0.860	-	-	-	-	-	-
UGLTL	0.338	0.344	0.152	0.242	0.224	0.264	0.950	0.921	0.903	0.924	0.939	0.910
RMVAL	**0.956**	**0.930**	**0.906**	**0.916**	**0.914**	**0.917**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

JLMVC runs out of memory when being applied to Notting-Hill, MITIndoor-67, and Caltech-101.

TABLE IV
CLUSTERING RESULTS ON SCENE-15 AND MITINDOOR-67

	Scene-15 ($\lambda_1 = 10$, $\lambda_2 = 1$, $\lambda_3 = 0.05$)						MITIndoor-67 ($\lambda_1 = 5$, $\lambda_2 = 1$, $\lambda_3 = 0.05$)					
	ACC	NMI	AR	F-score	Precision	Recall	ACC	NMI	AR	F-score	Precision	Recall
SPC$best$	0.437	0.421	0.270	0.321	0.314	0.329	0.443	0.559	0.304	0.315	0.294	0.340
LRR$best$	0.445	0.426	0.272	0.324	0.316	0.333	0.120	0.226	0.031	0.045	0.044	0.047
LT-MSC	0.574	0.571	0.424	0.465	0.452	0.479	0.431	0.546	0.280	0.290	0.279	0.306
MLAN	0.331	0.475	0.151	0.248	0.150	0.731	0.232	0.408	0.012	0.041	0.021	0.662
ECMSC	0.457	0.463	0.303	0.357	0.318	0.408	0.469	0.590	0.323	0.333	0.314	0.355
t-SVD-MSC	0.812	0.858	0.771	0.788	0.743	0.839	0.684	0.750	0.555	0.562	0.543	0.582
HLR-M^{2}VS	0.878	0.895	0.850	0.861	0.850	0.871	0.802	0.866	0.730	0.734	0.713	0.757
ETLMSC	0.878	0.902	0.851	0.862	0.848	0.877	0.775	0.899	0.729	0.733	0.709	0.758
JLMVC	**0.988**	0.975	0.975	0.977	0.979	0.975	-	-	-	-	-	-
UGLTL	0.976	0.960	0.952	0.955	0.961	0.950	**0.948**	**0.979**	**0.940**	**0.940**	**0.930**	0.951
RMVAL	**0.988**	**0.982**	**0.979**	**0.980**	**0.982**	**0.978**	0.936	0.978	0.931	0.932	0.905	**0.961**

TABLE V
CLUSTERING RESULTS ON COIL-20 AND CALTECH-101

	COIL-20 ($\lambda_1 = 10$, $\lambda_2 = 0.5$, $\lambda_3 = 0.05$)						Caltech-101 ($\lambda_1 = 10$, $\lambda_2 = 1$, $\lambda_3 = 0.1$)					
	ACC	NMI	AR	F-score	Precision	Recall	ACC	NMI	AR	F-score	Precision	Recall
SPC$best$	0.627	0.806	0.619	0.640	0.596	0.692	0.484	0.723	0.319	0.34	0.597	0.235
LRR$best$	0.761	0.829	0.720	0.734	0.717	0.751	0.510	0.728	0.304	0.339	0.627	0.231
LT-MSC	0.804	0.860	0.748	0.760	0.741	0.776	0.559	0.788	0.393	0.403	0.670	0.288
MLAN	0.862	0.961	0.835	0.844	0.758	0.953	0.579	0.748	0.222	0.265	0.173	0.560
ECMSC	0.782	0.942	0.781	0.794	0.695	0.925	0.359	0.606	0.273	0.286	0.433	0.214
t-SVD-MSC	0.830	0.884	0.786	0.800	0.785	0.808	0.607	0.858	0.430	0.440	0.742	0.323
HLR-M^{2}VS	0.852	0.960	0.833	0.842	0.757	0.949	0.650	0.872	0.463	0.472	0.760	0.343
ETLMSC	0.877	0.947	0.862	0.869	0.830	0.914	0.639	0.899	0.456	0.456	0.825	0.324
JLMVC	0.945	0.970	0.937	0.940	0.940	0.941	-	-	-	-	-	-
UGLTL	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.669	0.902	0.504	0.513	**0.960**	0.365
RMVAL	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.851**	**0.934**	**0.911**	**0.913**	0.917	**0.908**

Compared to Scene-15, MITIndoor-67 is more challenging since it contains 67 indoor scenes such that the between-cluster distances are relatively small. By comparing the clustering results on the right part of Table IV, we find that RMVAL and UGLTL obtain promising performance with large margins over their competitors. Although relying on different strategies, both methods well explore the local neighborhoods and the global low-rankness to capture the relationship of data. In contrast, the LRR model makes use of only the global data structure while MLAN focuses on exploiting local neighbors. The unsatisfactory performance of LRR and MLAN shows the necessity of jointly seeking the global and local data structures when processing complex datasets.

3) Generic Object Clustering: COIL-20 consists objects of simple shapes (toys, cups, etc.) with different geometric characteristics. Images of each objects are taken at pose intervals of five degrees, corresponding to 72 images per class. As such, the clusters in the COIL-20 dataset are endowed with clear manifold structures. This can be validated from the results on the left part of Table V, where most methods achieve promising performance. Although good performance can be expected generally owning to the simple intrinsic data structures, RMVAL and UGLTL consistently show advantages over their competitors as they can correctly recover all subspace structures.

Compared with COIL-20, Caltech-101 is a much larger and more complicated dataset. The challenges of processing Caltech-101 mainly come from the large number of clusters, uncontrolled conditions, and unbalanced clusters. As recorded in the right part of Table V, although several algorithms achieve good precision rates, their clustering results are still far from

satisfactory, particularly for the recall rate and the related F-score, AR, ACC. Since the recall rate is evaluated by taking the clustering as a series of decisions over all sample pairs [16], the low recall value indicates that, given a sample, previous methods cannot fully identify the samples within the same subspace. As a comparison, RMVAL outperforms the second best method by the margins of 48.8%, 78.0%, 80.8%, and 27.2% in terms of the recall, F-score, AR, and ACC values. RMVAL improves the NMI score by around 3.6% over the best competitor, and the precision rate of RMVAL ranks in the second place, slightly lower than that of UGLTL. Overall, the proposed RMVAL model is able to handle the relatively large and challenging dataset.

C. In-Depth Analysis of RMVAL

In this section, examples are provided to visualize the subspace structures and the properties of the affinity matrices learned from RMVAL, and then, we compare RMVAL with a newly proposed post-processing technique. Afterward, the sensitivity of parameter and the empirical convergence behaviors are examined.

1) Subspace Structures Uncovered From the Affinity Matrix: To provide an intuitive illustration on the effectiveness of the proposed RMVAL model, we visualize the discovered subspace structures by setting the affinity matrix as $A = \frac{1}{V}\sum_{v=1}^{V}(|Z^{(v)}| + |Z^{(v)'}|)$ and $A = \frac{1}{V}\sum_{v=1}^{V}(|S^{(v)}| + |S^{(v)'}|)$ respectively. Following the methods in [21], [49], we use the t-Distributed Stochastic Neighbor Embedding (t-SNE) [50] for visualization since it is an effective technique for revealing the structures of the high-dimensional data that lie on the union of several subspaces. The visual comparison is shown in Fig. 5, where different colors represent different clusters. By simultaneously optimizing the connectivity and sparsity properties, the global and local data structures are jointly considered to enhance the reliability of the affinity matrix. Thus, as shown in the second column of Fig. 5, the affinity matrices learned from RMVAL reveal relatively clear subspace structures. As a comparison, by setting $A = \frac{1}{V}\sum_{v=1}^{V}(|Z^{(v)}| + |Z^{(v)'}|)$, only the global low-rankness of the data structures is considered. It is obvious that the discovered subspace structures are worse since more clusters are mixed as shown in the first column of Fig. 5. Please note that we omit the visualization results on the MITIndoor-67 and Caltech-101 databases since it is hard to distinguish the subspace structures when the cluster numbers are large.

2) Visualizing the Connectivity and Sparsity of the Affinity Matrix: As shown in Fig. 6, compared with the affinity matrices learned from t-SVD-MVC [16], those from RMVAL exhibit much clear connections by jointly maintaining the the connectivity and sparsity. Conceptually, owing to the property of "intra-subspace projection dominance" [8], the intra-cluster coefficients tend to be relatively larger than the inter-cluster coefficients in many cases. Thus, by jointly considering the global low-rankness and preserving the strong connections in local neighborhoods, the inter-cluster connections are alleviated while the intra-cluster connections are relatively enhanced.

3) Comparison With the Post-Processing Technique in [24]: Given a densely connected affinity matrix, the work in [24]

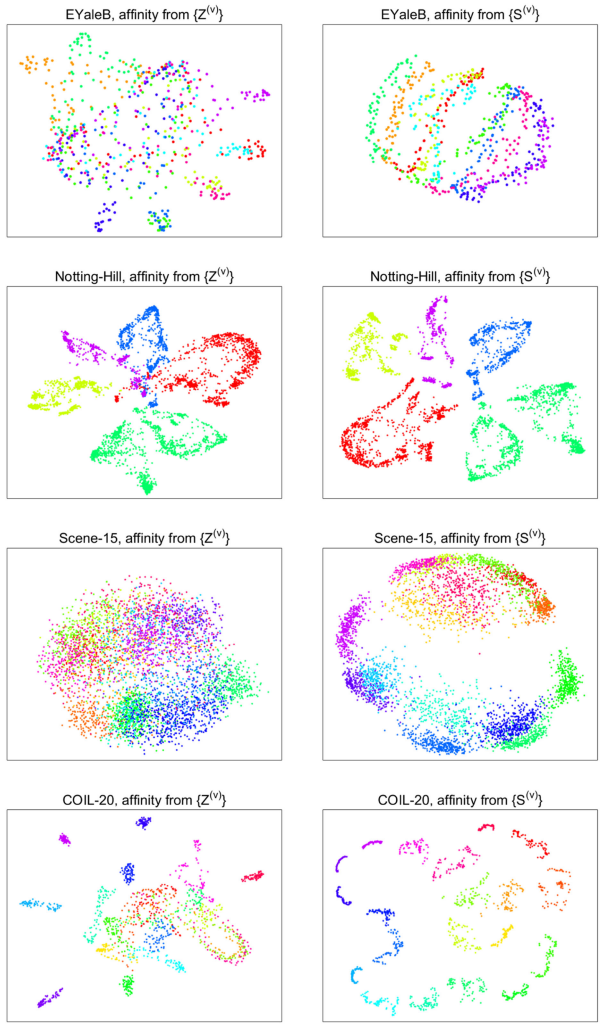

Fig. 5. Visualization of the subspace structures uncovered from different affinity matrices via t-SNE.

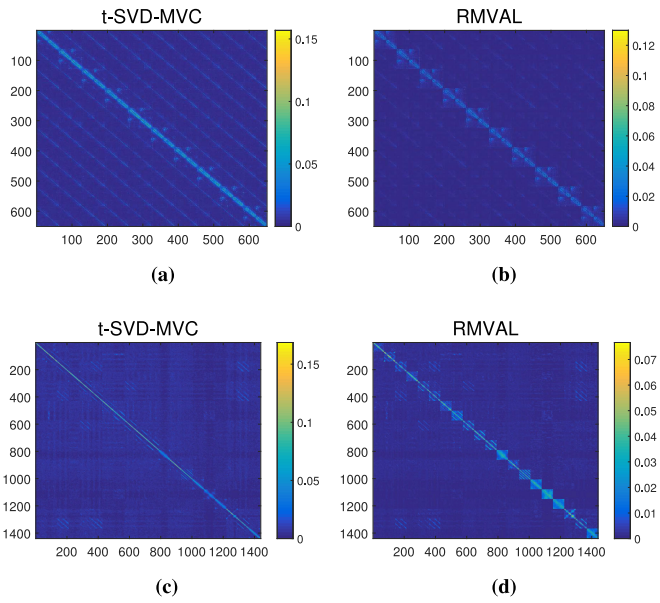

Fig. 6. Illustration of affinity matrices from t-SVD-MVC and RMVAL on (a), (b) EYaleB and (c), (d) COIL-20.

TABLE VI
CLUSTERING PERFORMANCE OF T-SVD-MVC, T-SVD-MVC WITH POST-PROCESSING, AND RMVAL

	ACC	NMI	AR	F-score	Precision	Recall	ACC	NMI	AR	F-score	Precision	Recall
	EYaleB						Notting-Hill					
t-SVD-MSC	0.652	0.667	0.500	0.550	0.514	0.590	0.957	0.900	0.900	0.922	0.937	0.907
t-SVD-MSC-post	0.849	0.862	0.778	0.801	0.777	0.825	0.939 ↓	0.878 ↓	0.872 ↓	0.900 ↓	0.892 ↓	0.908
RMVAL	0.956	0.930	0.906	0.916	0.914	0.917	1.000	1.000	1.000	1.000	1.000	1.000
	Scene-15						MITIndoor-67					
t-SVD-MSC	0.812	0.858	0.771	0.788	0.743	0.839	0.684	0.750	0.555	0.562	0.543	0.582
t-SVD-MSC-post	0.637 ↓	0.666 ↓	0.524 ↓	0.560 ↓	0.511 ↓	0.619 ↓	0.725	0.772	0.588	0.625	0.614	0.636
RMVAL	0.988	0.982	0.979	0.980	0.982	0.978	0.936	0.978	0.931	0.932	0.905	0.961
	COIL-20						Caltech-101					
t-SVD-MSC	0.830	0.884	0.786	0.800	0.785	0.808	0.607	0.858	0.430	0.440	0.742	0.323
t-SVD-MSC-post	1.000	1.000	1.000	1.000	1.000	1.000	0.531 ↓	0.753 ↓	0.355 ↓	0.366 ↓	0.618 ↓	0.261 ↓
RMVAL	1.000	1.000	1.000	1.000	1.000	1.000	0.851	0.934	0.911	0.913	0.917	0.908

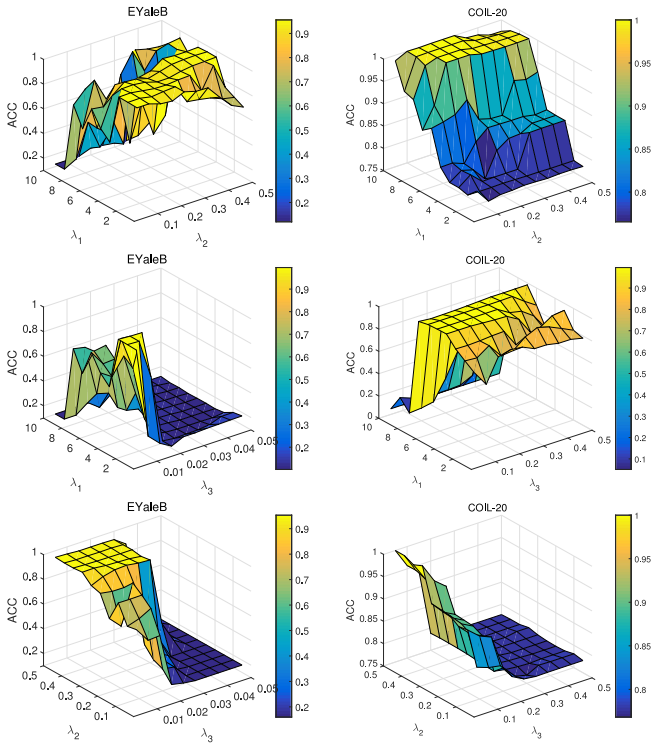

Fig. 7. ACC of RMVAL with different parameter settings.

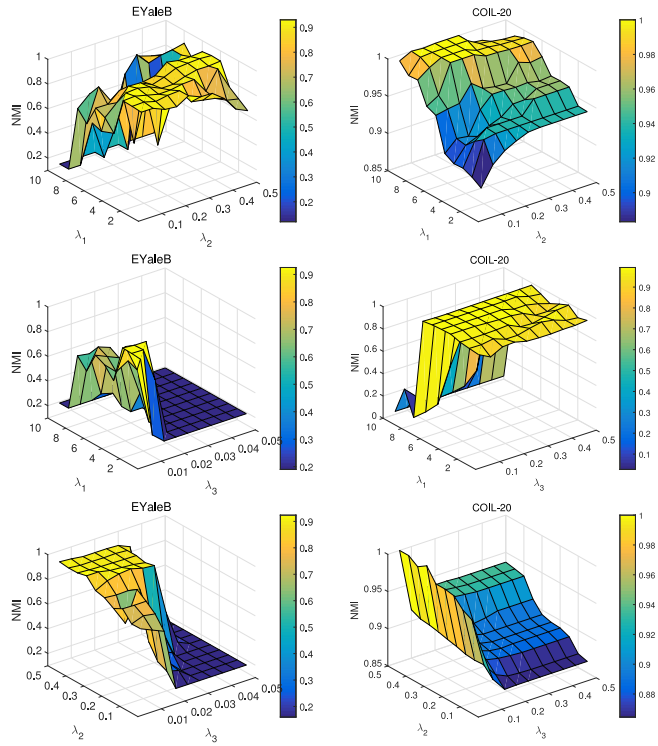

Fig. 8. NMI of RMVAL with different parameter settings.

developed a post-processing technique to seek sparse solutions from the densely connected samples by keeping only the connections belonging to "good neighbors". However, as mentioned in Section II-B, this technique may suffer problem in practical applications since it requires the prior knowledge on the number of good neighbors. We now compare the performance of t-SVD-MVC, t-SVD-MVC followed by the post-processing technique, and RMVAL to discuss this issue, and the result are recorded in Table VI. Note that we tune the parameters in [24] in large ranges according to the statistics of the datasets so as to always obtain its best performance. We observe that, (1) RMVAL obtains the best performance in all cases; (2) adopting the post-processing technique, t-SVD-MVC exhibits noticeable performance gains on well balanced databases, i.e., EYaleB, MITIndoor-67, and COIL-20; (3) when being applied to the unbalanced datasets, i.e., Notting-Hill, Scene-15, and Caltech-101, unfortunately, the performance degrades with post-processing. This may come from the fact that it is hard to set an appropriate number of "good neighbors" for the unbalanced datasets.

4) Parameter Sensitivity: There are three tradeoff parameters λ_1, λ_2, and λ_3 in RMVAL, corresponding to the terms of global structural constraint, view-specific self-representation, and sparsity preservation, respectively. We first coarsely locate the parameters within the range of $\{0.01, 0.1, 1, 10, 100\}$, and then narrow the ranges experimentally to select λ_1 from $\{1, 2, \ldots, 10\}$, tune λ_2 from $\{0.1, 0.2, \ldots, 1\}$, and choose λ_3 from $[0.01, 1]$. In the following, we plot the ACC and NMI scores on the EYaleB and COIL-20 databases to show the performance sensitivity over different parameter settings. Generally, the performance of RMVAL is less sensitive to λ_2, and is significantly affected by the values of λ_1 and λ_3. From Figs. 7 and 8, we find that the optimal values of λ_3 on EYaleB are much larger than that on COIL-20. This is because the EYaleB dataset may exhibit large inter-cluster similarities due to the severe

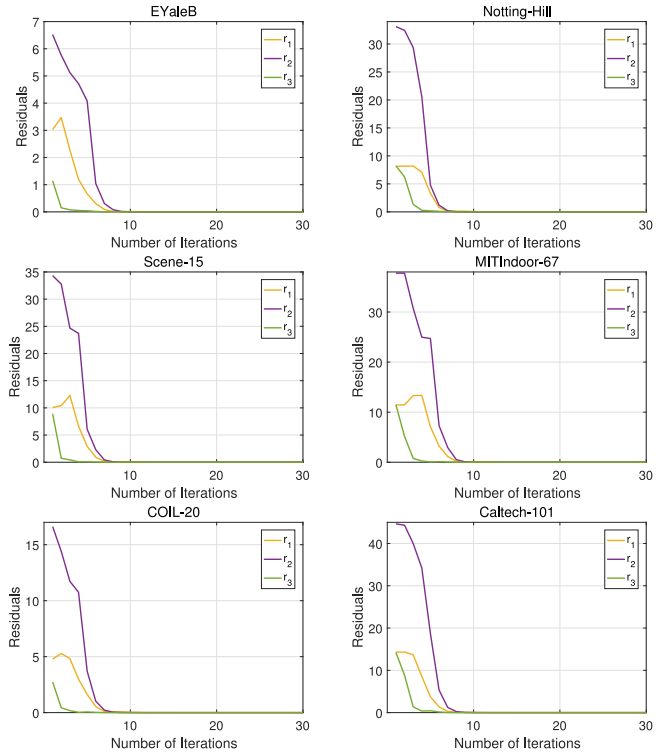

Fig. 9. Empirical convergence curves of RMVAL.

illumination changes, and thus, the property of intra-subspace projection dominance may broken. That is, there exists large coefficients among inter-cluster samples and directly setting large λ_3 values may filter out too many intra-cluster connections. This observation coincides the results in Fig. 6(a), where the sub-diagonal elements on the affinity matrices are relatively large, corresponding to images belong to different persons but with the same illumination conditions. In the contrary, COIL-20 consists relatively simple data structures and thus λ_3 can be well tuned from a large range.

5) Empirical Convergence: Within the ADMM framework, a fast convergence rate can be expected by adopting the varying penalty scheme to adjust the parameter ρ [39]. In Fig. 9, we plot the residual curves when processing all databases. It can be observed that the residuals tend to be stable within 10 iterations, showing the efficiency of Algorithm 1.

V. Conclusion

In this paper, we propose a Reliable Multi-View Affinity Learning (RMVAL) model that simultaneously optimizes the connectivity and sparsity of the affinity. By doing so, the intra-subspace samples are well-connected on the affinity matrix and the inter-subspace samples are disconnected, leading to accurate clustering results. RMVAL well exploits the global and local data structures: it encourages the intra-subspace samples to be densely connected via global low-rankness and purifies the dense solution by preserving only the strong connections in neighborhoods. We devise an efficient algorithm to solve the RMVAL model within the ADMM framework. Extensive experiments on six benchmark datasets validated the effectiveness of RMVAL.

As RMVAL purifies the densely connected self-representation tensor via the l_1-norm regularization, it essentially assumes the equal reliability of different views and adopts the same strategy to process each view. However, this assumption does not always hold. For example, some features may exhibit large inter-class margins such that the induced self-representation matrices have clear connectivity and sparsity. In this case, the small coefficients also correspond to intra-cluster connections, and thus, directing cutting off these small values may reduce the reliability of the affinity matrix. Our future will study this problem.

References

[1] B. Wang *et al.*, "Learning adaptive neighborhood graph on grassmann manifolds for video/image-set subspace clustering," *IEEE Trans. Multimedia*, vol. 23, pp. 216–227, 2021.
[2] Z. Wang, L. Wang, J. Wan, and H. Huang, "Shared low-rank correlation embedding for multiple feature fusion," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2020.3003747.
[3] R. Rubinstein, T. Faktor, and M. Elad, "K-SVD dictionary-learning for the analysis sparse model," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2012, pp. 5405–5408.
[4] X. Li *et al.*, "Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7086–7098, Nov. 2014.
[5] G. Liu *et al.*, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
[6] C. Lu, J. Feng, Z. Lin, T. Mei, and S. Yan, "Subspace clustering by block diagonal representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 487–501, Feb. 2019.
[7] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *proc Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.
[8] X. Peng, Z. Yi, and H. Tang, "Robust subspace clustering via thresholding ridge regression, in am assoc," *Artif. Intell.*, vol. 25, 2015, pp. 3827–3833.
[9] X. Peng, Z. Yu, Z. Yi, and H. Tang, "Constructing the l2-graph for robust subspace learning and subspace clustering," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1053–1066, Apr. 2017.
[10] C. Tang *et al.*, "Learning a joint affinity graph for multiview subspace clustering," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1724–1736, Jul. 2019.
[11] H. Gao, F. Nie, X. Li, and H. Huang, "Multi-view subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 4238–4246.
[12] Y. Li, M. Yang, and Z. M. Zhang, "A survey of multi-view representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 1863–1883, Oct. 2019.
[13] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan, "Multi-task low-rank affinity pursuit for image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2439–2446.
[14] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao, "Low-rank tensor constrained multiview subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1582–1590.
[15] M. Yin, J. Gao, S. Xie, and Y. Guo, "Multiview subspace clustering via tensorial t-product representation," *IEEE Trans. Neural Netw. Learn. Syst*, no. 99, pp. 1–14, 2018.
[16] Y. Xie *et al.*, "On unifying multi-view self-representations for clustering by tensor multi-rank minimization," *Int. J. Comput. Vis.*, vol. 126, no. 11, pp. 1157–1179, 2018.
[17] Y. Xie, W. Zhang, Y. Qu, L. Dai, and D. Tao, "Hyper-laplacian regularized multilinear multiview self-representations for clustering and semisupervised learning," *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 572–586, Feb. 2020.
[18] Y. Chen, X. Xiao, and Y. Zhou, "Jointly learning kernel representation tensor and affinity matrix for multi-view clustering," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 1985–1997, Aug. 2020.
[19] J. Wu, X. Xie, L. Nie, Z. Lin, and H. Zha, "Unified graph and low-rank tensor learning for multi-view clustering, in am assoc," *Artif. Intell.*, 2020, pp. 6388–6395.

[20] J. Wu, Z. Lin, and H. Zha, "Essential tensor learning for multi-view spectral clustering," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5910–5922, Dec. 2019.
[21] X. Xiao, Y. Chen, Y.-J. Gong, and Y. Zhou, "Prior knowledge regularized multiview self-representation and its applications," *IEEE Trans. Neural Netw. Learn. Syst.*, 2020, doi: 10.1109/TNNLS.2020.2984625.
[22] F. Nie, G. Cai, and X. Li, "Multi-view clustering and semi-supervised classification with adaptive neighbours, in am assoc," *Artif. Intell.*, 2017, pp. 2408–2414.
[23] X. Wang, X. Guo, Z. Lei, C. Zhang, and S. Z. Li, "Exclusivity-consistency regularized multi-view subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 923–931.
[24] J. Yang, J. Liang, K. Wang, P. L. Rosin, and M.-H. Yang, "Subspace clustering via good neighbors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1537–1544, Jun. 2020.
[25] Y. Chen, C.-G. Li, and Y. Chong, "Stochastic sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 4155–4164.
[26] C. Lu, J. Feng, Z. Lin, and S. Yan, "Correlation adaptive subspace segmentation by trace lasso," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1345–1352.
[27] C. You, C.-G. Li, D. P. Robinson, and R. Vidal, "Oracle based active set algorithm for scalable elastic net subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 3928–3937.
[28] Y.-X. Wang, H. Xu, and C. Leng, "Provable subspace clustering: When lrr meets ssc," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5406–5432, Sep. 2019.
[29] S. Sarfraz, V. Sharma, and R. Stiefelhagen, "Efficient parameter-free clustering using first neighbor relations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8934–8943.
[30] H. Liu, X. Yang, L. J. Latecki, and S. Yan, "Dense neighborhoods on affinity graph," *Int. J. Comput. Vis.*, vol. 98, no. 1, pp. 65–82, 2012.
[31] Z. Li, F. Nie, X. Chang, Y. Yang, C. Zhang, and N. Sebe, "Dynamic affinity graph construction for spectral clustering using multiple features," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6323–6332, Dec. 2018.
[32] X. Peng, Z. Huang, J. Lv, H. Zhu, and J. T. Zhou, "COMIC: Multi-view clustering without parameter selection," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5092–5101.
[33] B. Nasihatkon and R. Hartley, "Graph connectivity in sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 2137–2144.
[34] M. E. Kilmer and C. D. Martin, "Factorization strategies for third-order tensors," *Linear Alg. Appl.*, vol. 435, no. 3, pp. 641–658, 2011.
[35] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover, "Third-order tensors as operators on matrices: A. theoretical and computational framework with applications in imaging," *SIAM J. Matrix Anal. Appl.*, vol. 34, no. 1, pp. 148–172, 2013.
[36] E. Kernfeld, S. Aeron, and M. Kilmer, "Clustering multi-way data: A novel algebraic approach," 2014, *arXiv:1412.7056v2*.
[37] P. Zhou, C. Lu, Z. Lin, and C. Zhang, "Tensor factorization for low-rank tensor completion," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1152–1163, Mar. 2018.
[38] Z. Zhang, G. Ely, S. Aeron, N. Hao, and M. Kilmer, "Novel methods for multilinear data completion and de-noising based on tensor-svd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 3842–3849.
[39] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation, in proc," *Adv. Neural Inf. Process. Syst.*, 2011, pp. 612–620.
[40] J.-F. Cai, E. J. Candés, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.
[41] J. Huang, F. Nie, H. Huang, and C. Ding, "Robust manifold nonnegative matrix factorization," *ACM Trans. Knowl. Discov. Data*, vol. 8, no. 3, p. 11, 2014.
[42] C. Chen, B. He, Y. Ye, and X. Yuan, "The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent," *Math. Prog..*, vol. 155, no. 1-2, pp. 57–79, 2016.
[43] T. Zhou, C. Zhang, C. Gong, H. Bhaskar, and J. Yang, "Multiview latent space learning with feature redundancy minimization," *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1655–1668, Apr. 2020.
[44] A. Oliva and A. Torralba, "Modeling the shape of the scene: A. holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[45] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 413–420.
[46] F.-F. Li, F. Rob, and P. Pietro, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Comput. Vis. Image Underst.*, vol. 106, no. 1, pp. 59–70, 2007.
[47] J. Yang *et al.*, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 156–171, Jan. 2017.
[48] 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.
[49] C. Zhang, Q. Hu, H. Fu, P. Zhu, and X. Cao, "Latent multi-view subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 4279–4287.
[50] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, Nov., pp. 2579–2605, 2008.

Xiaolin Xiao (Member, IEEE) received the B.E. degree from Wuhan University, China, in 2013 and the Ph.D. degree from the University of Macau, Macau, China, in 2019. She is currently a Postdoc Fellow with the School of Computer Science and Engineering, South China University of Technology, China. Her research interests include multi-view learning and color image processing and understanding.

Yue-Jiao Gong (Senior Member, IEEE) received the B.S. and Ph.D. degrees in computer science from Sun Yat-sen University, China, in 2010 and 2014, respectively. She is currently a Full Professor with the School of Computer Science and Engineering, South China University of Technology, China. Her research interests include evolutionary computation, swarm intelligence, machine learning, as well as their applications to image processing and smart city. She has published more than 80 papers, including more than 40 IEEE TRANSACTIONS papers, in her research area.

Zhongyun Hua (Member, IEEE) received the B.S. degree from Chongqing University, Chongqing, China, in 2011, and the M.S. and Ph.D. degrees from University of Macau, Macau, China, in 2013 and 2016, respectively, all in software engineering. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Shenzhen, China. His research interests include chaotic system, multimedia security and image processing.

Wei-Neng Chen (Senior Member, IEEE) received the bachelor's and Ph.D. degrees from Sun Yat-sen University, Guangzhou, China, in 2006 and 2012, respectively. He is currently a Professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou. He has coauthored more than 100 international journal and conference papers, including more than 50 papers published in the IEEE TRANSACTIONS journals. His current research interests include computational intelligence, swarm intelligence, network science, and their applications.

Dr. Chen was the recipient of the IEEE Computational Intelligence Society (CIS) Outstanding Dissertation Award in 2016, and the National Science Fund for Excellent Young Scholars in 2016. He is currently the Vice-Chair of the IEEE Guangzhou Section. He is also a Committee Member of the IEEE CIS Emerging Topics Task Force. He is an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, AND THE COMPLEX & INTELLIGENT SYSTEMS.