Received August 17, 2018, accepted September 11, 2018, date of publication September 17, 2018,
date of current version October 8, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2870371

Iterative Reconstrained Low-Rank Representation via Weighted Nonconvex Regularizer

JIANWEI ZHENG[1], CHENG LU[1], HONGCHUAN YU[2], WANLIANG WANG[1], AND SHENGYONG CHEN[1], (Senior Member, IEEE)

[1]School of Computer Science and Engineering, Zhejiang University of Technology, Hangzhou 310023, China
[2]National Centre for Computer Animation, Bournemouth University, Poole BH12 5BB, U.K.

Corresponding author: Wanliang Wang (wwl@zjut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61602413, Grant 61873240, and Grant U1509207, in part by the Natural Science Foundation of Zhejiang Province under Grant Y19F030047, and in part by the Royal Society-Newton Mobility under Grant IE151018.

ABSTRACT Benefiting from the joint consideration of geometric structures and low-rank constraint, graph low-rank representation (GLRR) method has led to the state-of-the-art results in many applications. However, it faces the limitations that the structure of errors should be known a prior, the isolated construction of graph Laplacian matrix, and the over shrinkage of the leading rank components. To improve GLRR in these regards, this paper proposes a new LRR model, namely iterative reconstrained LRR via weighted nonconvex regularization, using three distinguished properties on the concerned representation matrix. The first characterizes various distributions of the errors into an adaptively learned weight factor for more flexibility of noise suppression. The second generates an accurate graph matrix from weighted observations for less afflicted by noisy features. The third employs a parameterized rational function to reveal the importance of different rank components for better approximation to the intrinsic subspace structure. Following a deep exploration of automatic thresholding, parallel update, and partial SVD operation, we derive a computationally efficient low-rank representation algorithm using an iterative reconstrained framework and accelerated proximal gradient method. Comprehensive experiments are conducted on synthetic data, image clustering, and background subtraction to achieve several quantitative benchmarks as clustering accuracy, normalized mutual information, and execution time. Results demonstrate the robustness and efficiency of IRWNR compared with other state-of-the-art models.

INDEX TERMS Low-rank representation (LRR), weighted nonconvex constraint, accelerated proximal gradient, singular value thresholding, power method.

I. INTRODUCTION

Low-rank representation (LRR) [1], as a promising approach to capture the latent subspace structure of data, has attracted broad interests and been successfully applied to extensive applications in signal processing and computer vision community, such as scene classification [2], action proposal [3], image clustering [4], [5], and transfer hashing [6], to name a few. Generally speaking, the success of LRR mainly originates from three merits: a natural hypothesis of underlying multiple low-rank subspaces in observed data, a self-expressive representation with specific noise suppression constraint, and a convex approximation of rank regularization using nuclear norm. However, these characteristics also restrict LRR's applicability for the reasons that the structure of errors should be known a prior and the intrinsic rank of the observed data might be loosely approximated.

In order to tackle heterogeneous noise sources and obtain better approximation to the original low-rank assumption, a great variety of clustering methods have been proposed recently based on the framework of LRR, i.e., to represent each sample by a linear combination of dictionary data and pursue an appropriate representation matrix via different choices of regularization and constraint terms, which can be uniformly formulated as follows [7]–[9]:

$$\min_{\boldsymbol{Z}} \gamma \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z}\|_{\mu} + \Omega(\boldsymbol{X}, \boldsymbol{Z}), \quad s.t.\, \boldsymbol{Z} \in C, \tag{1}$$

where $\boldsymbol{X} \in R^{m\times n}$ is the observed data containing n samples as its columns, $\boldsymbol{A} \in R^{m\times n}$ denotes a dictionary matrix, $\boldsymbol{Z} \in R^{n\times n}$ denotes the representation matrix (or representation matrix), $\|\cdot\|_{\mu}$ denotes a specific norm, Ω and C are some regularizers

TABLE 1. The cost functions of some related subspace clustering algorithms with respect to fidelity term $||.||_\mu$, regularizer Ω, and constraint set C.

algorithms	fidelity term	regularizer	constraint set
LRR [1]	$\gamma\|\|\boldsymbol{E}\|\|_F^2$	$\|\|\boldsymbol{Z}\|\|_*$	$\varnothing$
SMR [7]	$\gamma\|\|\boldsymbol{E}\|\|_F^2$	$\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^T)$	$\varnothing$
L2Graph [14]	$\gamma\|\|\boldsymbol{x}_i - \boldsymbol{X}\boldsymbol{z}_i\|\|_2^2$	$\|\|\boldsymbol{z}_i\|\|_2^2$	$\boldsymbol{z}_i^T\mathbf{1} = 0$
FSCNN [15]	$\gamma\|\|diag(\boldsymbol{w})\boldsymbol{E}\|\|_F^2$	$\|\|\boldsymbol{Z}\|\|_*$	$\|\|\boldsymbol{w}\|\|_0 = l$
IRIALM [16]	$\gamma\|\|\boldsymbol{W}\odot\boldsymbol{E}\|\|_F^2$	$\|\|\boldsymbol{Z}\|\|_*$	$\varnothing$
LRS [19]	none	$\sum_i(\|\|\boldsymbol{X}\boldsymbol{Z}_i\|\|_{S_p}^{2p})$	$\forall i\boldsymbol{Z}_i \subseteq \{0,1\}, \sum_i \boldsymbol{Z}_i = \boldsymbol{I}$
NSGLRR [20]	$\gamma\|\|\boldsymbol{E}\|\|_1$	$\|\|\boldsymbol{Z}\|\|_* + \beta tr(\boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^T) + \lambda\|\|\boldsymbol{Z}\|\|_1$	$\boldsymbol{Z} \geq 0$
RSS [30]	$\gamma\|\|\boldsymbol{E}\|\|_F^2$	$\|\|\boldsymbol{Z}\|\|_F^2 + \beta tr(\boldsymbol{Z}\boldsymbol{L}_{\boldsymbol{B}}\boldsymbol{Z}^T) + \lambda\|\|\boldsymbol{B}\|\|_F^2$	$\forall ij\boldsymbol{b}_i^T\mathbf{1} = 1, b_{ij} \geq 0$

and constraint set, respectively, and $\gamma > 0$ is a balance parameter.

To deal with different types of noise, many norms are exploited to measure the residual $\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z}$. For example, Frobenius norm (i.e., $\|\cdot\|_F$) is used for modeling the Gaussian noise [10], l_1-norm is employed for characterizing the Laplacian noise [9], [11], [12], and $l_{2,1}$-norm is employed for removing sample-specific outliers [1], [13]. However, these constraints work well only with a correct prior knowledge on error distribution, which is always difficult to obtain. Peng *et al*. [14] presented another error-removing method based on the property of intrasubspace projection dominance (IPD), but the IPD property itself may be disturbed by gross corruption. Kang *et al*. [15] integrated the feature selection into the residual term for revealing more accurate data relationships on the premise of corruption location fixed, which is a strong assumption and cannot be satisfied in most real-world problems. Chen and Yang [16] employed the maximum likelihood estimation principle to estimate the noise distribution, which brings about a robust framework to deal with complex errors. However, the proposed algorithm, namely iteratively reweighted inexact augmented Lagrange multiplier (IRIALM), suffers from two limitations. One is that an independent logistic function is suboptimal for the overall optimization. The other is that the model loses the reweighting essence by integrating the weights update into the ADMM framework [17], [18].

For representation matrix $\boldsymbol{Z}$, the regularizer has been extensively studied using l_1-norm [11], Frobenius norm [10], [14], nuclear norm ($\|\cdot\|_*$) [1], [15], [18], [19], and mixture (e.g., $\|\cdot\|_* + F$ [9], $l_1 + \|\cdot\|_*$ [20]) or replacement [21] of some of them. Despite the success of these convex surrogate functions, recently there have been numerous attempts on employing nonconvex ones to approximate the intrinsic structure of data. In terms of the singular values, the key idea is that the larger ones are more informative and should be less penalized. Such attempts include different nonconvex surrogates (e.g., truncated operator [22], capped-l_1 penalty [23], logarithm constraint [24], and the well-known Schatten p quasi-norm [25]), weighted nuclear norm [26], and their mixture such as weighted Schatten p-Norm [27]. Empirically, these attempts achieve better performance than the convex counterparts. Moreover, theoretical results have also been established [28]. However, the resultant optimization problem is much more challenging. Most existing optimization algorithms that work with the nuclear norm cannot be applied. In addition, most existing nonconvex optimization methods take at least $O(mn^2)$ time complexity at each iteration for a $m \times n$ matrix (assuming $m \geq n$), which is expensive for large matrices.

Some recent works focus on geometric structures to yield a better representation matrix $\boldsymbol{Z}$ with the assumption that if two samples are close in the intrinsic manifold of the data, then the representations of these two points in the subspace are also close to each other [29]. This idea inspires Hu *et al*. [7] to propose a smooth representation (SMR) clustering that explicitly takes into account the local structure of input data. Similarly, Yin *et al*. [20] proposed a nonnegative sparse graph LRR (NSGLRR) method by incorporating the Laplacian matrix into the cost function. To overcome the drawback of suboptimal results, some joint optimization schemes [30], [31] have been proposed, that is, to simultaneously learn the representations and the affinity matrix. Peng *et al*. [32] further imposed a rank constraint on Laplacian matrix for more explicit block-diagonal results. However, the deterministic rank may be unknown for many practical problems. Furthermore, all the involved similarity metric may deteriorate by noises, which causes severe performance degradation.

Table 1 shows the cost functions of some related methods following framework (1), where vector $\boldsymbol{w}$ is the feature weights, $\boldsymbol{L}$ denotes the Laplacian matrix and $\boldsymbol{B}$ means the similarity matrix. Detailed explanations can be found in the original paper and is omitted here due to space limit. Motivated by these works and based on the above analysis, our goal is to overcome the limitations in the properties of noise resistant constraint, weighted nonconvex regularization, and joint affinity matrix learning, as well as to integrate them into a unified formulation. To this end, we propose iterative reconstrained low-rank representation model via weighted nonconvex regularization (IRWNR), which renders clearer block-diagonal representation matrix and facilitates subspace clustering in noisy scenarios. Our main contributions include:

1) Introducing a factor matrix $\boldsymbol{W}$ to the error term in Eq. (1), which can adaptively penalize the residual entries. Particularly, the factor matrix can not only distinguish outliers from data but also reweight the contributions of the active features individually.

2) Integrating the learned $\boldsymbol{W}$ with the update of Laplacian matrix $\boldsymbol{L}$ to improve the robustness. As a result, the underlying structure can be genuinely exploited to ensure a well-behaved representation matrix $\boldsymbol{Z}$.
3) Presenting a nonconvex weighted Rational function, i.e., $\Omega = \sum s_i\sigma_i(\boldsymbol{Z})/(1 + a\sigma_i(\boldsymbol{Z})/2)$, to approximate the singular values effectively, while keeping convexity of the whole cost function via a referenced span of a.
4) Deriving an efficient proximal gradient algorithm by the observation that the obtained singular values can be automatically thresholded. Specifically, a block approximation of the leading singular vectors and a terminal comparison of the singular values to the threshold values are employed for fast implementation.

Notations 1: In the sequel, vectors and matrices are denoted by lowercase boldface and uppercase boldface, respectively. For any matrix $\boldsymbol{A} \in R^{m\times n}$, $\mathrm{vec}(\boldsymbol{A})$, $\mathrm{tr}(\boldsymbol{A})$ and $\boldsymbol{A}^T$ are its vector form [33], trace and transpose respectively, $\boldsymbol{a}_i$ is the ith column of $\boldsymbol{A}$, and a_{ij} is the jth element in $\boldsymbol{a}_i$. $\|\boldsymbol{A}\|_F = (\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{A}))^{1/2}$ denotes the Frobenius norm, $\|\boldsymbol{A}\|_* = \sum_i \sigma_i(\boldsymbol{A})$ denotes the nuclear norm, and $\|\boldsymbol{A}\|_2 = \max_{|\boldsymbol{x}|_2\neq 0} \frac{|\boldsymbol{A}\boldsymbol{x}|_2}{|\boldsymbol{x}|_2}$ is the spectral norm. $\boldsymbol{I}$ and $\boldsymbol{1}$ denote identity matrix and all 1 vector, respectively, with their dimensions following the context.

II. PROPOSED IRWNR MODEL

To begin with, we first briefly introduce the basic graph LRR model [3], [8], [20] as an instance of (1). Our changes will then be addressed accordingly. For clarity, we assume that the observed data $\boldsymbol{X}$ are from either a single subspace or a union of multiple subspaces. The prototype of GLRR is formulated as

$$\min_{\boldsymbol{Z}} \gamma\|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z}\|_\mu + \|\boldsymbol{Z}\|_* + \beta\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^T), \tag{2}$$

where μ can be determined by the error priori information, the regularization $\Omega(\boldsymbol{X}, \boldsymbol{Z})$ of Eq. (1) is expressed as the mixture of $\|\boldsymbol{Z}\|_*$ and $\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^T)$ with balancing parameter β, in which Laplacian matrix $\boldsymbol{L}$ is constructed either directly from the raw data or from their learned representations.

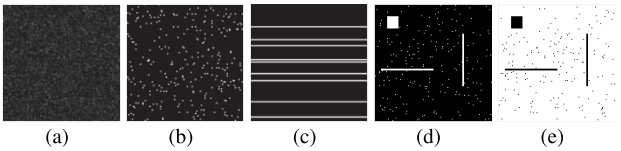


FIGURE 1. Illustrating four types of errors as well as our weight factor: (a) Gaussian noise [10], (b) Laplacian noise [11], (c) Feature outliers [15], (d) Mixed corruption, (e) A well-estimated weight matrix W.

A. WEIGHTED FEATURE LEARNING FOR ERROR PENALIZING

The fidelity term in (2), i.e., $||\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z}||_\mu$, generally refers to the deviation between reconstructed matrix $\boldsymbol{A}\boldsymbol{Z}$ and input data $\boldsymbol{X}$. It could exhibit as noise, missed entries, corruptions and outliers in practice. Fig. 1 illustrates several common types of error $\boldsymbol{E}$ under the context of subspace clustering. In real scenarios, the distribution of errors may be more complex than just a single status shown in Fig. 1(a), (b) and (c). For example, Fig. 1(d) shows a specific corruption mixed by Laplacian noise and certain structural occlusions, which cannot be correctly removed either by fixed distribution assumptions [1], [10], [11] or a simple feature selection scheme [15]. As shown in Fig. 1(e), we attempt to suppress the noise with an intuitive idea, i.e., introducing a weight variable $\boldsymbol{W}$ to adaptively penalize the individual elements of the error matrix. Specifically, we try to assign smaller weights (black points in Fig. 1(e)) to the noisy features and assign greater weights (white regions in Fig. 1(e)) to the clean features. With this criterion, the fidelity term shall have the ability to identify the important features and reinforce the effect of them during optimization so as to adaptively learn a more robust representation matrix.

In our implementation, we consider the weight matrix $\boldsymbol{W}$ in probability form, i.e., $w_{ij} > 0$ and $\boldsymbol{1}^T\boldsymbol{W}\boldsymbol{1} = 1$, and use the Frobenius norm for simplicity, i.e., $||\cdot||_\mu = ||\cdot||_F$. Following the aforementioned criterion, a smaller e_{ij} probably corresponds to the clean features, thus should be assigned a larger probability w_{ij}. Therefore, a natural method to determine the weight matrix is solving the following problem:

$$\min_{\boldsymbol{1}^T\boldsymbol{W}\boldsymbol{1}=1, w_{ij}\geq 0} \gamma\|\boldsymbol{W}^{1/2} \odot (\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z})\|_F^2. \tag{3}$$

However, problem (3) has a trivial solution, only the minimal residual can be assigned with weight 1 and all the other residuals cannot be effectively weighted (with weight 0). Accordingly, we add regularization $\|\boldsymbol{W}\|_F^2$ to avoid the trivial solution, i.e.,

$$\min_{\boldsymbol{1}^T\boldsymbol{W}\boldsymbol{1}=1, w_{ij}\geq 0} \gamma\|\boldsymbol{W}^{1/2} \odot (\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z})\|_F^2 + \lambda\|\boldsymbol{W}\|_F^2, \tag{4}$$

where λ is a balance parameter. We will see in Subsection III.A that problem (4) can be solved with a closed form solution.

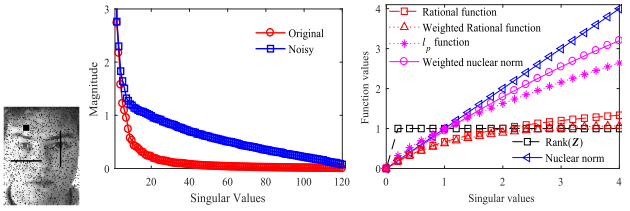


FIGURE 2. An instance of (a) corrupted sample and its (b) rank components, as well as some (c) penalty functions.

B. WEIGHTED RATIONAL FUNCTION FOR RANK APPROXIMATION

Although the nuclear norm used in model (2) is the tightest convex approximation to the rank constraint, the obtained solution may seriously deviate from the original one particularly in the presence of noises. Fig. 2 illustrates a noisy

face image, its rank components, and some penalty functions. It can be noted that the weighted nuclear norm, i.e., $\sum_{i=1}^{n} s_i\sigma_i$, the nonconvex l_p constraint, i.e., $\sum_{i=1}^{n} \sigma_i^p$, and the Rational function, i.e., $r(\boldsymbol{\sigma}) = \sum_{i=1}^{n} \sigma_i/(1 + a\sigma_i/2)$, are all tighter approximations than the nuclear norm. Moreover, by further adding a weight factor $\boldsymbol{s}$ onto Rational function, it approximates the natural rank constraint more closely, especially for the large singular values. Fig. 2 (b) shows the larger rank components of the noisy image are closely coherent with the original ones, while the smaller singular values deviate far away from the original ones. Based on these observations, we introduce the weighted Rational function to penalize larger singular values less than smaller ones, i.e.,

$$r(\boldsymbol{s}, \boldsymbol{\sigma}) = \frac{\boldsymbol{s}\boldsymbol{\sigma}}{1 + a\boldsymbol{\sigma}/2}, \tag{5}$$

where $\boldsymbol{s}$ is a given weight, $\boldsymbol{\sigma}$ is some singular value, and a is the parameter to be determined.

C. THE OVERALL COST FUNCTION

We propose to marry the above two merits with the GLRR model (2), such that the advantages of feature reweighting scheme and intrinsic rank approximation can be taken simultaneously in a single model as follows:

$$\begin{aligned} J(\boldsymbol{W}, \boldsymbol{L}, \boldsymbol{Z}) = & \min_{\mathbf{1}^T\boldsymbol{W}\mathbf{1}=1, w_{ij}\geq 0} \gamma\|\boldsymbol{W}^{1/2} \odot (\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z})\|_F^2 \\ & + \lambda\|\boldsymbol{W}\|_F^2 + \beta\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^T) + \sum_{i=1}^{n} r(s_i, \sigma_i). \end{aligned} \tag{6}$$

While model (6) seeks linear relationships of the data with reweighted features, it considers Laplacian matrix $\boldsymbol{L}$ using raw features. In this case, the noisy features may be included to measure pair-wise similarities of the data and the resulting $\boldsymbol{L}$ would be potentially suboptimal. To remedy this issue, we further require that the model should respect the intrinsic geometrical structure in the subspace of clean features. That is, we construct the Laplacian matrix $\boldsymbol{L}$ based on $\boldsymbol{W} \odot \boldsymbol{X}$ instead of $\boldsymbol{X}$ and update it iteratively along with the aforementioned reweighting scheme. Therefore, model (6) is further formulated as follows:

$$\begin{aligned} J(\boldsymbol{W}, \boldsymbol{L}, \boldsymbol{Z}) = & \min_{\mathbf{1}^T\boldsymbol{W}\mathbf{1}=1, w_{ij}\geq 0} \gamma\|\boldsymbol{W}^{1/2} \odot (\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z})\|_F^2 \\ & + \lambda\|\boldsymbol{W}\|_F^2 + \beta\mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}_{\boldsymbol{W}}\boldsymbol{Z}^T) + \sum_{i=1}^{n} r(s_i, \sigma_i). \end{aligned} \tag{7}$$

Model (7) is the final cost function of our method, namely iterative reconstrained LRR with weighted nonconvex regularization (IRWNR).

III. OPTIMIZATION ALGORITHM

In recent years, there has been a big literature to address multivariable optimization problems, such as the alternating direction method (ADM) [33]–[35] and the iterative reweighted method (IRM) [18], [25], [36]. Although ADM has drawn considerable attention, the convergence is not guaranteed for problems containing more than two variables. Accordingly, we apply IRM to our optimization problem (7) with three variables $\boldsymbol{W}$, $\boldsymbol{L}$, and $\boldsymbol{Z}$. Since we not only reweight $\boldsymbol{W}$ but also update $\boldsymbol{L}$ to constrain $\boldsymbol{Z}$, the term "iterative reweighted" is transformed into "iterative reconstrained" in this paper.

Following the IRM framework [18], the sketch of our optimization scheme is described in Algorithm 1, where the detailed selection of dictionary $\boldsymbol{A}$ hinges on the concerned applications. When the pending data contains multiple subspaces, the observed data $\boldsymbol{X}$ is a natural selection as the dictionary (i.e., $\boldsymbol{A} = \boldsymbol{X}$). If there is only a single subspace, the identity matrix $\boldsymbol{I}$ can be chosen as $\boldsymbol{A}$. In the following, the update rules for $\boldsymbol{W}$, $\boldsymbol{L}$, and $\boldsymbol{Z}$ are sequentially described. Moreover, an efficient singular value thresholding (SVT) operator is presented to speed up each iteration. Finally, the complexity and convergence analysis are given.

Algorithm 1 Our Optimization Approach Following IRM Framework

Input: Data matrix $\boldsymbol{X} \in R^{m\times n}$, dictionary matrix $\boldsymbol{A} \in R^{m\times n}$, and parameter γ, β, l; Set $k = 0$ and initialize $\boldsymbol{Z}_0 \in R^{n\times n}$.
Output: $\boldsymbol{Z}_k$.

1: **While** not converged **do**
2: Estimate weights matrix $\boldsymbol{W}_{k+1}$ as Eq. (10), where $\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z}_k$;
3: Update Laplacian matrix $\boldsymbol{L}_{k+1}$ as Eq. (11);
4: Inner loop for $\boldsymbol{Z}_{k+1}$ with $\boldsymbol{W}_{k+1}$ and $\boldsymbol{L}_{k+1}$.
5: $k = k + 1$;
6: **end while**

A. THE CLOSED FORM SOLUTION FOR W AND L

By keeping all other variables fixed, the subproblme of updating $\boldsymbol{W}$ turns out to be Eq. (4). Following KKT condition, the optimal $\boldsymbol{W}$ can be directly computed as

$$\boldsymbol{W} = \left(\kappa - \frac{\boldsymbol{E}^2}{2\lambda}\right)_+, \tag{8}$$

where $\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z}$, $\boldsymbol{E}^2$ denotes a matrix whose elements are e_{ij}^2 with a slight abuse of symbols, κ is the Lagrangian multipliers of constraint $\mathbf{1}^T\boldsymbol{W}\mathbf{1} = 1$, and $(\cdot)_+$ denotes a nonnegative operator. Without loss of generality, suppose the elements of $\mathrm{vec}(\boldsymbol{E}^2)$ are in non-descending order, then $\mathrm{vec}(\boldsymbol{W})$ will be in non-ascending order. Let the optimal $\mathrm{vec}(\boldsymbol{W})$ has l zero elements related to noises, i.e., the $(mn-l+1)$th element equals 0, where $mn = m\times n$. This together with the constraint $\mathbf{1}^T\boldsymbol{W}\mathbf{1} = 1$ leads to

$$\begin{cases} \kappa = \dfrac{1}{mn-l} + \sum_{j=1}^{mn-l} \dfrac{\boldsymbol{e}_j^2}{2\lambda(mn-l)}, \\ \lambda = (mn-l)\dfrac{\boldsymbol{e}_{mn-l+1}^2}{2} - \dfrac{1}{2}\sum_{j=1}^{mn-l} \boldsymbol{e}_j^2. \end{cases} \tag{9}$$

With derived κ and λ, $\boldsymbol{W}$ can be analytically expressed by

$$\boldsymbol{W} = (e^2_{mn-l+1} - \boldsymbol{E}^2)/[(mn-l)e^2_{mn-l+1} - \sum_{j=1}^{mn-l} e_j^2]. \quad (10)$$

Note that in Eq. (10), the number of zero elements l is much easier to tune than the regularization parameter λ since the value of λ could be from zero to infinite and l is an integer having explicit meaning. In our experiments, we let $l = \lfloor \rho m \rfloor$, where $\rho \in \{0.6, 0.7, 0.8, 0.9\}$, and $\lfloor \rho m \rfloor$ outputs the largest integer smaller than ρm.

With $\boldsymbol{W}$ given and following the basic computation steps as [7] and [20], we update our graph term as

$$\begin{cases} \mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}_{\boldsymbol{W}}\boldsymbol{Z}^T) = \sum_{i=1}^{n}\sum_{j=1}^{n} \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2 d_{ij}, \\ d_{ij} = \exp(\dfrac{-\|\boldsymbol{w}_i \odot \boldsymbol{x}_i - \boldsymbol{w}_j \odot \boldsymbol{x}_j\|^2}{\theta^2}), \end{cases} \quad (11)$$

where $\boldsymbol{L}_{\boldsymbol{W}} = \boldsymbol{D}^d - \boldsymbol{D}$ with $\boldsymbol{D}^d$ being a diagonal matrix and its jth diagonal element $D^d_{jj} = \sum_i d_{ij}$, and d_{ij} is the (i,j)th entry of $\boldsymbol{D}$ that denotes the similarity of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ with a given parameter θ (We set it as the standard variance of $\boldsymbol{X}$). The main difference between (11) and the existing methods such as [7] and [20] lies in the incorporation of $\boldsymbol{W}$ into consideration, which ensures the learned graph be close to the intrinsic geometrical structure formed by the clean feature subspace.

B. THE SOLVING SCHEME FOR Z

The suboptimization of $\boldsymbol{Z}$ in Algorithm 1 (step 4) can be formulated as

$$F(\boldsymbol{Z}) = \overbrace{\gamma\|\boldsymbol{W}^{1/2} \odot (\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z})\|_F^2 + \beta \mathrm{tr}(\boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^T)}^{f(\boldsymbol{Z})} + \overbrace{\sum_{i=1}^{n} r(s_i, \sigma_i)}^{r(\boldsymbol{Z})}, \quad (12)$$

which also can be solved via ADM[33]–[35] framework by updating the residual $\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z}$ and $\boldsymbol{Z}$ alternatively in a Gauss-Seidel manner. Unfortunately, applying ADM for (12) results in a tricky problem as

$$\min_{\boldsymbol{Z}} \|\boldsymbol{A}\boldsymbol{Z} - \boldsymbol{C}\|_F^2 + \sum_{i=1}^{n} r(s_i, \sigma_i), \quad (13)$$

where $\boldsymbol{C}$ is an intermediate constant matrix during the update steps. The accelerated proximal gradient (APG) method [38] can be introduced to solve subproblem (13) as [20]. However, the resulting double loops involving ADM and APG make it run unbearably slow. To deal with this challenge, we divide (12) into two terms, i.e., $F(\boldsymbol{Z}) = f(\boldsymbol{Z}) + r(\boldsymbol{Z})$. Due to the fact that f is ν-Lipschitz smooth, i.e., $\|\nabla f(\boldsymbol{Z}_1) - \nabla f(\boldsymbol{Z}_2)\|_2 \leq \nu\|\boldsymbol{Z}_1 - \boldsymbol{Z}_2\|_2$ and $r(\boldsymbol{Z})$ is smooth and nonconvex, APG can be directly applied to (12), instead of any sub-optimization in ADM framework. Particularly, the APG can update $\boldsymbol{Z}$ by

$$\begin{aligned} \boldsymbol{Z}_{k+1} &= \arg\min_{\boldsymbol{Z}} \frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{Z}_k + \eta\nabla f(\boldsymbol{Z}_k)\|_F^2 + \eta r(\boldsymbol{Z}) \\ &= \mathrm{prox}_{\eta r}(\boldsymbol{Z}_k - \eta\nabla f(\boldsymbol{Z}_k)) \end{aligned} \quad (14)$$

at iteration k, where $0 < \eta < 1/\nu$ is the stepsize, and

$$\nabla f(\boldsymbol{Z}_k) = 2\gamma(\boldsymbol{W} \odot \boldsymbol{A})^T(\boldsymbol{W} \odot (\boldsymbol{A}\boldsymbol{Z}_k - \boldsymbol{X})) + 2\beta(\boldsymbol{Z}_k\boldsymbol{L}) \quad (15)$$

is the gradient of $f(\boldsymbol{Z})$.

Recently, APG methods have been extensively studied [39]–[41] and extended to problems of matrix completion and robust principal component analysis [42], [43]. The state-of-the-art is the efficient inexact proximal gradient (EIPG) algorithm [41], within whose framework our subproblem of $\boldsymbol{Z}$ can be solved as in Algorithm 2. Each iteration requires only one proximal step (step 12). Acceleration is performed in step 4 and the objective is then checked in step 6 to determine whether $\boldsymbol{Y}_{t+1}$ is accepted (steps 5-7).

Algorithm 2 EIPG for Subproblem (12)

Input: Estimated $\boldsymbol{W}$ and $\boldsymbol{L}$, parameter $\eta \in (0, 1/\nu)$, $k = 0$.
Output: $\boldsymbol{Z}$.

1: $\boldsymbol{Z}_0 = 0, \boldsymbol{Z}_1 \in R^{n\times n}$ follows $N(0,1)$;
2: **While** not converged **do**
3: $k = k + 1$;
4: $\boldsymbol{Y}_k = \boldsymbol{Z}_k + \frac{k-1}{k+2}(\boldsymbol{Z}_k - \boldsymbol{Z}_{k-1})$
5: $\bigtriangleup_k = \max_{t=\max(1,k-3),\ldots,k} F(\boldsymbol{Z}_t)$;
6: **if** $F(\boldsymbol{Y}_k) \leq \bigtriangleup_k$ **then**
7: $\boldsymbol{G}_k = \boldsymbol{Y}_k$
8: **else**
9: $\boldsymbol{G}_k = \boldsymbol{Z}_k$
10: **end if**
11: $\boldsymbol{\Theta}_k = G_k - \eta\nabla f(\boldsymbol{G}_k)$;
12: $\boldsymbol{Z}_{k+1} = \mathrm{prox}_{\eta r}(\boldsymbol{\Theta}_k)$.
13: **end while**

The remaining problem of Algorithm 2 lies in the proximal operator (step 12). Due to the nonconvexity of $r(\boldsymbol{Z})$ and the existence of weights $\boldsymbol{s}$, traditional algorithms such as singular value thresholding (SVT) [44], iteratively reweighted norm minimization (IRNN) [25], and generalized proximal gradient (GPG) [45] cannot be directly employed here. In the following, we show how the solution of proximal step in Algorithm 2 can be achieved.

To preserve the major components of the input data, $\boldsymbol{s} = \nabla r(\sigma(\boldsymbol{Z}_k))$ is employed to penalize the smaller singular values as much as possible. Therefore, the weights s_i, $i = 1, \ldots, n$, are in ascending order with the premise that the singular values σ_i, $i = 1, \ldots, n$, are descending. Under these conditions, subproblem (16) can be illustrated in Fig. 3, where $a = 4.9$, $\boldsymbol{\sigma} = [2.7, 2.2, 1.5, 0.9, 0.2]^T$, and $\boldsymbol{s} = [0.8, 1.5, 2.5, 2.7, 3.0]^T$. It can be noted that the three red lines are convex and their minimal points have the property of $\delta_i^* \geq \delta_j^*$ for $s_i \leq s_j$, $i < j$, whereas the blue and pink lines are nonconvex and their minimal points all lie in $\delta^* = 0$; From this observation, we conjecture that, under certain conditions, the proximal operator $\mathrm{prox}_{\eta r}(\boldsymbol{\Theta}_k)$ may be strictly convex and solvable in a parallel way. Lemma 1 and 2 are presented to validate our hypotheses.

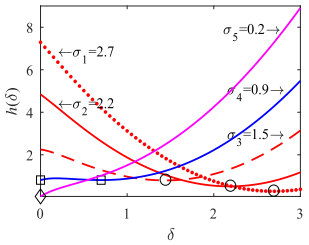


FIGURE 3. Illustration of function $h(\delta_i)$ with s_i in nondescending order. The marked points denote the global optimums of $h(\delta_i)$.

Lemma 1: Given the weights satisfying $0 \le s_1 \le s_2 \le \cdots \le s_n$, $\text{prox}_{\eta r}(\boldsymbol{\Theta}_k)$ can be decoupled into independent subproblems as

$$\min_{\delta_i \ge 0} h(\delta_i) = \frac{1}{2}(\delta_i - \sigma_i(\boldsymbol{\Theta}_k))^2 + \frac{\eta s_i \delta_i}{1 + a\delta_i/2}, \tag{16}$$

and their optimal solutions satisfy the order constraint $\delta_1 \ge \delta_2 \ge \cdots \ge \delta_n$.

Proof: By replacing the l_p constraint with the rational function, we obtain Lemma 1 similarly as [27, Th. 2]. □

Lemma 2: Despite the nonconvexity of rational penalty function, the proximal operator $\text{prox}_{\eta r}(\cdot)$ in (14) is strictly convex if $0 < a < 1/(\eta \max(\boldsymbol{s}))$.

Proof: All omitted proofs can be found in the appendix. □

Theorem 1: Let $\boldsymbol{\Theta}_k = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be the SVD of $\boldsymbol{\Theta}_k$. If $0 < a < 1/(\eta \max(\boldsymbol{s}))$, then the global minimizer of step 12 in Algorithm 2 is

$$\boldsymbol{Z}_{k+1} = \boldsymbol{U}\boldsymbol{\Xi}\boldsymbol{V}^T, \tag{17}$$

where $\boldsymbol{\Xi}$ is the threshold function outputting a diagonal matrix, whose subproblems are defined in (16) with solutions generated by Eq. (18) and (19).

Moreover, from the blue line in Fig. 3, we can see that there exists a specific τ acting as the threshold. Especially when $\delta = \tau$, two different points share the same minimal function value, i.e., $h(\delta)^* = h(0)$. By applying this equality to (16), we have

$$\tau_i = \frac{2\sqrt{s_i a \eta} - 1}{2}, \quad i = 1, \cdots, n. \tag{18}$$

$$\delta_i - \sigma_i + \frac{s_i \eta}{(1 + a\delta_i/2)^2} = 0, \quad i = 1, \cdots, n. \tag{19}$$

By combining all these results together, the proximal operator can be computed through Theorem 1. Specifically, when $\sigma_i < \tau$, we get $\delta^* = 0$ directly; When $\sigma_i \ge \tau$, then δ^* can be achieved from Eq. (19), i.e., $\delta_i^{k+1} = \sigma_i - \frac{s_i \eta}{(1 + a\delta_i^k/2)^2}$, $i = 1, \cdots, n$. Note that this is also an iterative procedure. However, empirically we found that satisfactory results can be obtained within 2 iterations.

C. THE ACCELERATION FOR PROXIMAL OPERATOR

The SVD operation is the main burden in Theorem 1, which needs to be conducted at each iteration. Given $\boldsymbol{U}_q$ formed by the first q left singular vectors of $\boldsymbol{\Theta}_k$, Proposition 1 shows that $\text{prox}_{\eta r}(\boldsymbol{\Theta}_k)$ can be obtained based on a smaller matrix $\boldsymbol{Q}$. To obtain such a $\boldsymbol{Q}$, reference [43] resorts to the power method [46], [47] and successfully approximats the SVT in nuclear norm constrained problems. Nevertheless, the algorithm in [43] is designed to tackle a fixed rank problem, i.e., the rank of the objective matrix should be given in advance, which may not be promised in real-world applications. Recent works [48], [49] turn to a fixed precision problem by minimizing the rank of a SVD approximation given some desired error tolerance. On that basis, we attempt to conduct proximal operator through solving an adaptive thresholding problem, which does not require any rank parameter or error threshold given in advance. To clarify it, a rank shrinkage SVD algorithm is presented, in which the required singular values are gradually estimated by incrementally building up the blocked SVD approximation.

Proposition 1 [43]: Assume that $\boldsymbol{Q} \in R^{n \times q}$, where $q \ge \text{rank}(\boldsymbol{\Theta}_k)$, is orthogonal and $\text{span}(\boldsymbol{U}_q) \subseteq \text{span}(\boldsymbol{Q})$. Then, $\text{prox}_{\eta r}(\boldsymbol{\Theta}_k) = \boldsymbol{Q}\text{prox}_{\eta r}(\boldsymbol{Q}^T\boldsymbol{\Theta}_k)$.

Algorithm 3 $\text{Prox}_{\eta r}(\boldsymbol{\Theta}_k)$ With Efficient SVD and Automatic SVT

Input: $\boldsymbol{\Theta}_k$, block size b, set $i = 0$.

Output: Estimated left singular vectors $\boldsymbol{U}_{\boldsymbol{Q}}$, right singular vectors $\boldsymbol{V}_{\boldsymbol{Q}}$, and thresholded singular values $\boldsymbol{\Sigma}_\delta$.

1: **While** not converged **do**
2: $i = i + 1$;
3: $\boldsymbol{\Omega}_i = \text{randn}(n, b)$;
4: $\boldsymbol{Q}_i = \text{PowerScheme}(\boldsymbol{V}_k; \boldsymbol{\Omega}_i)$ [43][46];
5: $\boldsymbol{Q}_i = \text{orth}(\boldsymbol{Q}_i - \sum_{j=1}^{i-1} \boldsymbol{Q}_j\boldsymbol{Q}_j^T\boldsymbol{Q}_i)$, $\boldsymbol{Q} = [\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_i]$;
6: $\boldsymbol{B}_i = \boldsymbol{Q}_i^T\boldsymbol{V}_k$, $\boldsymbol{B} = [\boldsymbol{B}_1, \ldots, \boldsymbol{B}_i]$;
7: $\boldsymbol{V}_k = \boldsymbol{V}_k - \boldsymbol{Q}_i\boldsymbol{B}_i$;
8: $[\boldsymbol{Q}_t, \boldsymbol{R}_t] = \text{qr}(\boldsymbol{B}^T, 0)$, $[\boldsymbol{U}_t, \boldsymbol{\Sigma}_t, \boldsymbol{V}_t] = \text{svd}(\boldsymbol{R}_t)$;
9: Obtain $\boldsymbol{\tau}$ by (18);
10: **if** $\max(\boldsymbol{\tau}) > \min(\boldsymbol{\Sigma})$, then **end while**, **otherwise** move to step 2;
11: Given $\boldsymbol{\tau}$, update $\boldsymbol{\delta}$ by (19);
12: $\boldsymbol{U}_{\boldsymbol{Q}} = \boldsymbol{Q}\boldsymbol{V}_t$, $\boldsymbol{V}_{\boldsymbol{Q}} = \boldsymbol{Q}_t\boldsymbol{U}_t$.

The entire procedure for our proximal operator is shown in Algorithm 3. Steps 3-7 use the power method and block-based SVD approximation to efficiently build up an orthogonal matrix $\boldsymbol{Q}$ that approximates $\text{span}(\boldsymbol{U}_{\boldsymbol{Q}})$. Steps 8-10 perform a small SVD and check the stop criterion. Though SVD operation is still needed, $\boldsymbol{R}_t$ in step 8 is much smaller than the original matrix $\boldsymbol{\Theta}_k$. In step 11, the singular values are thresholded using Theorem 1.

So far, we have presented a new approach to achieve a more appropriate LRR representation matrix $\boldsymbol{Z}$. The learned $\boldsymbol{Z}$ can be used to construct an affinity matrix as $(|\boldsymbol{Z}| + |\boldsymbol{Z}^T|)/2$, which can be further fed into the spectral clustering method, e.g., normalized cut, for data segmentation. Additionally, it can also be applied to the low-rank recovery of the contaminated data $\boldsymbol{X}$.

TABLE 2. Several options of weight functions.

Constant	Inverse [18]	Gaussian [37]	Truncated [32]	Logistic [16]
1	$1/\|e_{ij}\|$	$\exp(-e_{ij}^2/2\sigma^2)$	$\begin{cases} 0, & e_{ij} > \varepsilon \\ 1, & \text{otherwise} \end{cases}$	$\frac{\exp(\alpha(\beta-e_{ij}^2))}{1+(\exp(\alpha(\beta-e_{ij}^2))}$

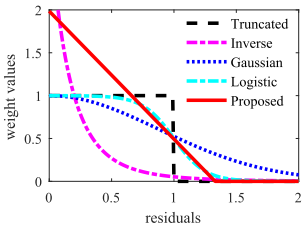


FIGURE 4. Different weighting functions.

D. ALGORITHM ANALYSIS

The flexibility of our model mainly derives from the reweighting matrix $\boldsymbol{W}$ and the penalty function $r(\boldsymbol{s}, \boldsymbol{\sigma})$. Note that different implementations of $\boldsymbol{W}$ will lead to different constraints for the error term in (2). When all entries of $\boldsymbol{W}$ are constantly assigned to 1, it acts as the Frobenius norm. When we set the weights to be inversely proportional to the magnitudes of residuals, i.e., $w_{ij} = 1/|e_{ij}|$, then the constraint turns into l_1-norm. More applicable functions are illustrated in Table 2. Fig. 4 further illustrates the weighting function of (10) and some others from Table 2. It can be noted that the proposed reweighting matrix $\boldsymbol{W}$ is more reasonable than others. The Gaussian, Logistic, and Truncated functions all assign smaller weights to the corrupted features, but they ignore the different activity of the useful features. The Inverse function assigns different weights to the useful features, but its value tends to infinity when the residual is close to 0, which causes numerical instability.

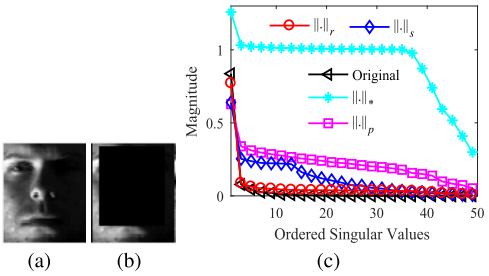


FIGURE 5. Comparison of various rank constraints. (a) Original image (b) 80% occluded image (c) Estimated singular values.

To demonstrate the role of penalty function $r(\boldsymbol{s}, \boldsymbol{\sigma})$, we take a face image with 80% of its region occluded by a rectangular block as an example. With all the parameters set as described in the experimental section, Fig. 5 shows that the estimated singular values by $r(\boldsymbol{s}, \boldsymbol{\sigma})$ ($\|\cdot\|_r$) is much closer to the ground truth compared with nuclear norm ($\|\cdot\|_*$), Schatten p-norm ($\|\cdot\|_p$), and weighted treatment ($\|\cdot\|_s$). Moreover, the new constraint has more solid theoretical support for the global minimum and model selection (see Subsection III.B).

The main computational complexity of IRWNR lies in the computation of proximal operator (step 12 of Algorithm 2). With the learned $\boldsymbol{U_Q}$,$\boldsymbol{V_Q}$ and $\boldsymbol{\Sigma_\delta}$, the dot product by $\boldsymbol{W}$, the matrix multiplication, and the SVD operation cost $O(mn)$, $O(mnq)$, and $O(mq^2)$, respectively, where q is the estimated rank of the current $\boldsymbol{Z}_k$. In contrast, exact proximal operator takes $O(mn^2)$ time, and is much slower as $n \gg q$ in most real-world problems.

For the convergence analysis, we present Theorem 2 to demonstrate the convergence of subproblem $F(\boldsymbol{Z})$, which combing with the closed form solutions for $\boldsymbol{W}$ and $\boldsymbol{L}$ leads to $J(\boldsymbol{W}_k, \boldsymbol{L}_k, \boldsymbol{Z}_k) \geq J(\boldsymbol{W}_{k+1}, \boldsymbol{L}_{k+1}, \boldsymbol{Z}_k) \geq J(\boldsymbol{W}_{k+1}, \boldsymbol{L}_{k+1}, \boldsymbol{Z}_{k+1})$. Note that the sequence $\{J_k\}_{i=0}^{\infty}$ is bounded from below by zero, hence convergent.

Theorem 2: Given $\eta \in (0, 1/\nu)$, the sequence $\{\boldsymbol{Z}_k\}$ generated by problem (12) satisfies the following properties:

(1) $F(\boldsymbol{Z}_k) - F(\boldsymbol{Z}_{k+1}) \geq \frac{1}{2}(\frac{1}{\eta} - \nu)\|\boldsymbol{Z}_k - \boldsymbol{Z}_{k+1}\|_F^2 \geq 0$;
(2) $\lim_{k\to\infty} ||\boldsymbol{Z}_k - \boldsymbol{Z}_{k+1}||_F^2 = 0$.

IV. EXPERIMENTAL RESULTS

We investigate the performance of our proposed subspace clustering method, IRWNR, by conducting comprehensive experiments on synthetic data and real-world problems, such as image clustering [20] and background subtraction [51]. We compare IRWNR to several recently developed low-rank recovery methods, including LRR [1], SMR [7], RSS [30], FSCNN [15], IRIALM [16], NSGLRR [20], LRS [19], L2Graph [14], ROSL [51] and ncRPCA [12]. Competing methods in different experiments are chosen according to their mathematical conditions as well as the applicability for specific tasks. For a fair comparison, the balance parameters of all the competing methods, e.g., γ and β, are traversed in{1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2} to report the best result. The remaining parameters of all compared algorithms are searched from a candidate set as suggested in their papers and tuned to achieve the best performance. Unless specified otherwise, for IRWNR, the parameters η and a are set as $0.6\max(\gamma\|\boldsymbol{X}^T\boldsymbol{X}\|_2, \beta\|\boldsymbol{L}^0\|_2)$ and $0.9/(\eta\max(s_i))$ in our experiments. To quantitatively and effectively evaluate the clustering performance, we utilized two metrics, accuracy (AC) and normalized mutual information (NMI) [20], as well as execution time (in second) in our experiments. Note that execution time is obtained by the average of running the corresponding tests 10 times. All experiments were

TABLE 3. The clustering metrics of NSGLRR, RSS, FSCNN, L2Graph, WNR, and IRWNR for synthetic data.

Noise ratios	NSGLRR		RSS		FSCNN		L2Graph		WNR		IRWNR	
	AC	NMI	AC	NMI	AC	NMI	AC	NMI	AC	NMI	AC	NMI
0.0	100	100	100	100	100	100	100	100	100	100	100	100
0.1	100	100	25.8	4.7	85.8	65.9	90.8	75.2	91.3	76.9	100	100
0.2	21.5	1.5	21.0	1.2	22.1	3.0	21.8	3.0	22.8	8.7	100	100
0.3	20.3	1.0	20.3	1.1	20.0	1.1	20.2	2.4	20.1	3.3	100	100
0.4	20.2	0.9	20.2	0.8	20.0	0.8	20.1	0.8	20.1	3.2	97.6	97.5
0.5	20.0	0.6	20.1	0.6	20.0	0.6	20.1	0.7	20.0	1.8	67.2	43.6

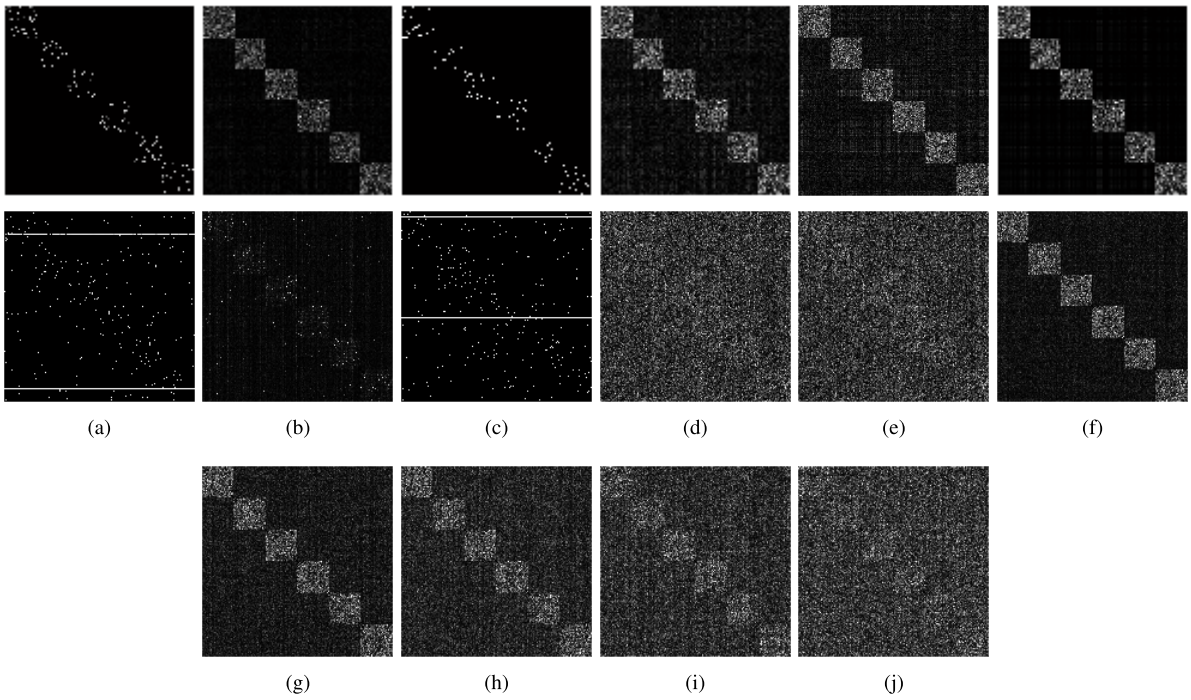


FIGURE 6. The learned affinity matrix from (a) RSS, (b) NSGLRR, (c) L2Graph, (d) FSCNN, (e) WNR and (f) IRWNR. The first and second rows show the results from 0% and 10% noise corruption, respectively. The third row shows the results of IRWNR under noise ratios (g) 20%, (h) 30%, (i) 40%, and (j) 50%.

implemented in MATLAB, and are run under a laptop with Intel(R) Core(TM) 2.4-GHz i7 CPU and 8.0-GB RAM.[1]

A. SYNTHETIC DATA

We first verify the robustness of our method to different levels of mixture noise. Six independent subspaces with 25 intrinsic dimensions and 500 ambient dimensions are generated. There are altogether 1200 data points collected by randomly generated 200 samples from each subspace. Moreover, different proportions of samples are selected to be corrupted by Gaussian noise with variance 0.5 and sparse noise with variance 5. In this experiment, we compare IRWNR with several state-of-the-art subspace clustering algorithms, i.e., NSGLRR [20], RSS [30], FSCNN [15], and L2Graph [14]. Besides, by removing the weight factor $\boldsymbol{W}$ from IRWNR, the reduced version WNR is also used as a competing method to verify the effectiveness of our reweighting strategy. Table 3 and Fig. 6 show the clustering metrics and part of the learned affinity matrices with corruption proportions range in $\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

As can be seen from Table 3, all the six competing methods achieve 100% clustering results in the noiseless scenario. The learned affinity matrices in the first row of Fig. 6 also reveal clear block-diagonal structure. However, the affinity matrices from NSGLRR, RSS, FSCNN, L2Graph, and WNR are still less perfect than our methods. For NSGLRR, FSCNN, and WNR, there are more inter-cluster relationships in the affinity matrices. For RSS and L2Graph, the intra-cluster similarities are much too sparse. Both of these two properties will deteriorate their practical application. Furthermore, the performance of NSGLRR, RSS, FSCNN, L2Graph, and WNR drops sharply as the noise level increases. Among them, NSGLRR,

[1]The demo code has been provided as a supplemental material, and will be released after the review phase at http://www.escience.cn/people/zhengjianwei/index.html

FSCNN, L2Graph, and WNR perform poorly when the noise ration is 20%. RSS even fails with only 10% noise corruption. This indicates that a simple prior distribution approximation, e.g., Gaussian (Frobenius norm) or Laplacian (l_1 norm), cannot correctly uncover the intrinsic similarity information of data when data contain mixture noise. With the aid of feature weight $\boldsymbol{W}$, IRWNR is more robust than other competing methods and obtains 100% accuracy even under 30% noise level. Fig. 6 (g)-(j) show that the learned affinity matrices of IRWNR become increasingly blur along with higher noise ratio. However, the diagonal structure achieved by IRWNR under 40% noise level is still clearer than the results obtained by other methods under 10% noise level. The direct comparison between IRWNR and WNR also shows that the weight factor $\boldsymbol{W}$ in the fidelity term plays a key role for noise suppression. Without $\boldsymbol{W}$, WNR performs similarly as the conventional methods, which loses the ability to discriminate inlier features from outlier ones.

B. IMAGE CLUSTERING

Six image datasets summarized in Table 4 are used for evaluating the performance of the methods. For computational efficiency, we downsize the images from their original form to the smaller one for reducing the dimensionality of the data. For example, all the AR images are downsized and normalized from 165×120 to 83×60.

TABLE 4. Uesd image datasets. *n*, *m*, and *c* denote the sample size, the feature dimension, and the number of subjects.

Datasets	n	m	c
USPS	2913	256	3
JAFFE	213	676	10
MNIST	2000	784	10
COIL	1440	1024	20
ORL	400	1024	40
AR	1200	4980	100

1) CLUSTERING ON UNOCCLUDED IMAGES

The first five datasets in Table 4 are used in this experiment. Table 5 shows the performance of nine competing methods, where the last two columns are with the average experimental results of all the evaluated datasets. Our first observation is that the proposed method outstandingly outperforms the state-of-the-art methods on average. For all the datasets, IRWNR achieves the best results in the tests except with ORL, where it is second best. Even though IRIALM and FSCNN share the highest AC and NMI with IRWNR in JAFFE, their performance are unsatisfactory in MNIST and COIL. The overall performance of SMR ranks second benefiting from the locality preservation property of Laplacian constraint and the stability of closed-form solution. Although NSGLRR integrates the Laplacian regularization as well as rank constraint together and performs better than other competing methods except IRWNR in USPS, its results fluctuate sharply in different datasets due to the poor convergence of inexact ADMM algorithm. The other methods all fail to obtain even one best result over the five testing datasets. Compared to SMR, L2Graph, FSCNN, NSGLRR, and IRIALM, which ranks second to fifth with the overall performance, our IRWNR has an improvement of 2.32%, 5.70%, 6.44%, 7.12%, and 11.34% in accuracy, respectively.

(a) (b)

FIGURE 7. Some disguised images from the AR dataset. (a) with glass, (b) with scarf.

2) CLUSTERING ON OCCLUDED IMAGES

AR database contains over 4000 images corresponding to 126 individuals. There are 26 images available for each subject, among which 6 images are with disguise of glass and 6 images are with disguise of scarf, as it is shown in Fig. 7. The disguised images from 100 people (50 for male and 50 for female) are selected in this experiment to examine the robustness of the competing methods. We compare our proposal with SMR, L2Graph, FSCNN, NSGLRR, and IRIALM since they perform better than LRR, LRS and RSS on unoccluded images.

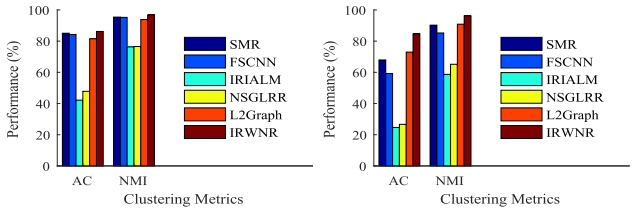


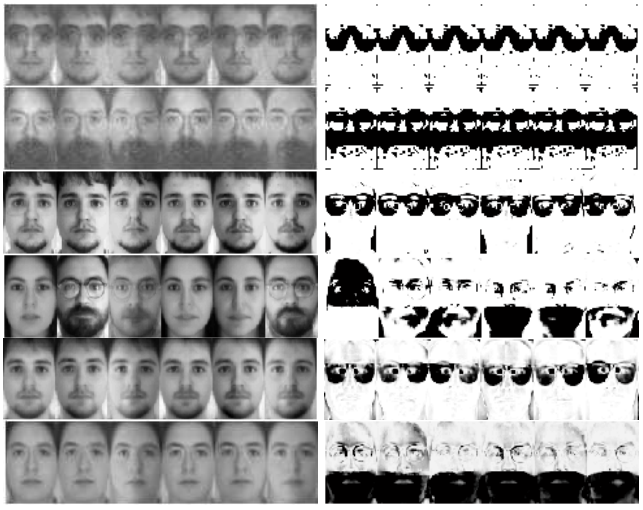
FIGURE 8. Clustering performance (%) on the AR dataset with real disguise of (a) glass and (b) scarf.

Fig. 8 shows the clustering AC and NMI of the compared methods for the disguised AR dataset. Similar to the previous problem, IRWNR gives the best performance outperforming other competing methods. NSGLRR, another joint optimization method integrating both the properties of locality preservation and low-rank constraint, performs poorer than FSCNN, L2Graph, SMR, and the proposed method. Even if IRIALM can also handle outliers due to its weighted residual mechanism, it does not show satisfactory results compared with others. By the aid of feature selection mechanism, FSCNN outperforms L2Graph, NSGLRR, and IRIALM under the glass disguised scenario. Interestingly, the performance of SMR, which does not include any noise resistant property, again ranks second in this experiment. On average, our method achieves an improvement of 1.27%, 1.75%, 32.26%, 29.31%, and 3.82% over SMR, FSCNN, IRIALM, NSGLRR, and L2Graph, respectively. This further

TABLE 5. Clustering performance (%) on five different image data sets.

Method	USPS		JAFFE		MNIST		COIL		ORL		Average	
	AC	NMI	AC	NMI	AC	NMI	AC	NMI	AC	NMI	AC	NMI
LRR	93.2	81.5	94.2	94.4	50.9	51.1	59.0	69.6	51.8	76.4	69.8	74.6
LRS	78.7	69.7	71.8	78.3	35.4	27.6	49.7	56.6	53.2	75.1	57.8	61.5
SMR	94.0	81.2	97.3	97.2	62.3	61.8	70.2	79.5	72.5	84.7	79.3	80.9
RSS	94.5	81.6	32.3	28.5	51.7	55.8	76.4	87.9	21.2	41.7	55.2	59.1
FSCNN	93.5	81.8	99.5	99.1	55.0	54.2	62.0	75.0	65.7	81.8	75.1	78.4
IRIALM	94.0	82.0	99.5	99.1	46.0	44.2	45.3	59.1	66.4	83.2	70.2	73.5
NSGLRR	94.6	81.9	98.5	97.8	51.9	52.9	62.0	72.4	65.3	79.8	74.5	77.0
L2Graph	93.2	81.2	97.6	96.5	59.1	54.6	59.7	73.7	69.8	83.8	75.9	78.0
IRWNR	94.6	82.0	99.5	99.1	66.9	64.8	82.6	89.0	70.2	84.1	81.6	83.9

verifies the superiority of locality preservation, noises elimination, and weighted nonconvex rank sparsity.



(a) (b)

FIGURE 9. Recovery of AR face images with real disguise. (a) Recovered images. (b) Estimated weight maps.

FSCNN, IRIALM, and IRWNR all adopt the idea of weights learning mechanism to eliminate part of the useless features in input data. FSCNN and IRIALM resort to feature selection strategy and an independent logistic function, respectively, for this purpose, whereas IRWNR utilizes an Frobenius norm constrained problem to reveal different contributions of input features. Fig. 9 illustrates the reconstructed images and the learned weights maps corresponding to the listed face images of Fig. 7, where the top, middle, and bottom two rows are the results from FSCNN, IRIALM, and, IRWNR respectively. From the top two rows in Fig. 9, we see that the estimated weights from FSCNN are consistent for all the samples, which coincide with the essence of feature selection, but violate the randomness of noises distribution. Consequently, FSCNN fails in regaining both a clear human face and an accurate weights map. From the third row in Fig. 9, IRIALM successfully recovers the face images from glass occlusion. The learned weights also accurately draw out the location of glass. However, there are major mistakes in the result of IRIALM (the fourth row of Fig. 9) from scarf occlusion, such as wrong individuals and bad elimination. The reason that IRIALM works reliably in small rang occlusion (glass) but poorly in relatively large area occlusion (scarf) comes from the fact that it involves the update of weights variable into the inner loop of ADMM framework, which leads to local optimization and poor convergence. In the bottom two rows, our method clearly learned the accurate face image and weights maps not only from the glass occlusion but also from the scarf occlusion. Moreover, compare our results to the ones from IRIALM, one can see that IRIALM assigns 0 (black region) to the deemed occlusion pixels and assigns 1 (white region) to the deemed non-occlusion pixels. However, our method assigns 0 to the occlusion pixels but assigns meaningful values (grey region) to the non-occlusion pixels, which exhibits different contributions of useful features and leads to a better clustering results.

(a) (b) (c) (d)

FIGURE 10. Example frames from the testing videos of (a) escalator, (b) highway, (c) office, and (d) sroom.

C. BACKGROUND SUBTRACTION

Background subtraction from video sequences is an important step in many applications, including traffic monitoring, and abnormal behavior detection. Surveillance videos from a fixed camera can be naturally modeled by our algorithm with $\boldsymbol{A} = \boldsymbol{I}$ and $\beta = 0$. We test the proposed IRWNR and other state-of-the-art background-foreground separation methods, including LRR [1], ncRPCA [12], and ROSL [51], on video sequences of escalator [51], highway [51], office [52], and sroom [52]. Fig. 10 illustrates some frames of these videos. Since the frame size of original videos is very high dimensional, e.g. 480 × 640 for office, we rescaled the frames

TABLE 6. Rescaled video size of four involved datasets.

features	escalator	highway	office	sroom
pixels per frame	130×160	240×320	130×160	180×240
stacked size	20800	76800	19200	43200
total frames	197	1698	181	1521

from different videos to smaller size as shown in Table 6 for computational tractability. To measure the performance quantitatively, we use NMSE $= \|\boldsymbol{Z} - \boldsymbol{O}\|_F / \|\boldsymbol{O}\|_F$ and PSNR$(\boldsymbol{Z}, \boldsymbol{O})$ as our experimental results, where $\boldsymbol{O}$ is the ground truth and PSNR is a built-in function of MATLAB.

The quantitative results by competing methods are illustrated in Table 7, where *na* denotes *not applicable* since it costs more than 24 hours for LRR to learn an optimal $\boldsymbol{Z}$. The proposed method shows higher PSNR and lower NMSE compared to other algorithms. ROSL outperforms LRR and ncRPCA in all four datasets, but lags behind IRWNR by 0.024 for NMSE and 10.69 for PSNR on average of the used videos. Notice that the subspace rank, which is unknown for the other three competing methods, should be designated in advance for ROSL, which makes it easier for low-rank approximation.

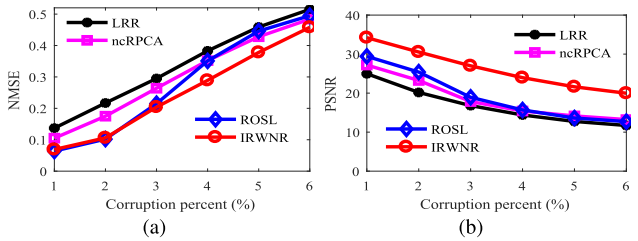


FIGURE 11. (a) NMSE and (b) PSNR of LRR, ncRPCA, ROSL, and IRWNR with the mixture corruption percentage ranging from 10% to 60%.

To test the robustness of IRWNR on background subtraction problem, we made further evaluation under the influence of mixed types of corruption. Different portions of input frames, from 10 to 60 percent, are simultaneously occluded by composite noise corruption (a randomly generated block together with Gaussian pixel corruption) at random locations. The experimental results are shown in Fig. 11. The figure shows that the performance of all competing methods degrades along with the increasing corruption portions. Encouragingly, IRWNR again outperforms other methods in all corruption levels especially for PSNR. Although LRR adopts group constraint for noise resistance, it does not fit in the mixed types of corruption. ROSL performs well on the lower level of corruption portion, but whose performance degrades more sharply than ncRPCA under heavier corruption.

Fig. 12 presents a visual comparison of background modeling results using different algorithms on two representative frames from office and sroom, where the binary foreground map (Fig. 12(c), (f)) is generated by running a median filter on the estimated sparse error [50]. As shown in Fig. 12, all the competing methods can extract clear background and separate foreground region with certain accuracy. However, various degrees of blur and ghost shadow exist in LRR, ncRPCA, and ROSL recovery images, leading to incomplete foreground segmentation. In contrast, our method obtains much cleaner background, especially for the pictures of filtered error, which demonstrates that our approach is more practically applicable.

D. EXECUTION TIME

In order to compare the computational complexity of different low-rank representation methods, we measure the execution time of competing algorithms. LRR, RSS, and LRS are omitted in this subsection due to their poor performance in our previous experiments. Since FSCNN, IRIALM, NSGLRR, and our IRWNR are all iterative approaches, many factors such as initialization, step sizes, maximum number of iterations and choice of balancing parameters can affect their running time. Thus, we report the objective values versus execution time of these approaches under their optimal tuned parameters. Fig. 13 illustrates the experimental results on MNIST and COIL datasets. The same stopping criterion is used, namely, the methods keep running until the difference of objective values between consecutive iterations is smaller than $tol =$ 1e-4. Besides, it costs SMR and L2Graph 90.24s and 25.81s respectively in MNIST dataset as well as 27.63s and 11.05s respectively in COIL dataset.

As can be seen from the figure, the proposed approach is computationally more efficient compared to other iterative based clustering methods. In MNIST dataset, NSGLRR converges faster than IRWNR at the initial stage of iterations. However, the decreasing objective values turn to increase at certain point due to the inexactly linearizing approximation to the cost function in [20]. An intuitive interpretation for the relatively slower convergence of IRWNR at the initial stage is that the initial variable $\boldsymbol{Z}$ has higher rank (may be full rank), which weakens the function of efficient SVD approximation in Algorithm 3. The continuous update of variable $\boldsymbol{Z}$ makes its rank closer to the authentic lower one, which impels Algorithm 3 to execute and converge faster and faster. SMR and L2Graph show faster execution time than the iterative methods due to their closed-form solution. However, the complexity for the Sylvester equation of SMR is $O(n^3)$, which prevents it from applying to even larger data size. Although L2Graph shows the fastest running time, its clustering accuracy is much lower than that of SMR and ours. Table 7 also shows the execution time of all the competing approaches in background subtraction. Again, IRWNR runs faster than other state-of-the-art methods except for the sroom video, where our method lags behind ROSL 1.58s. Encouragingly, our method is efficient even when the matrix size is large. For highway dataset, IRWNR is almost four times faster than ROSL.

E. PARAMETER ANALYSIS

Recall that the effectiveness of the proposed approach mainly comes from three properties, i.e. (i) the adaptively learned

TABLE 7. NMSE, PSNR and Execution time (in seconds) on the video background subtraction experiments.

Method	escalator			highway			office			sroom		
	NMSE	PSNR	Time	NMSE	PSNR	Time	NMSE	PSNR	Time	NMSE	PSNR	Time
LRR	0.081	27.45	20.77	*na*	*na*	*na*	0.054	33.29	10.75	0.089	25.04	2626
ncRPCA	0.067	29.04	10.14	0.145	23.17	2746	0.040	35.85	4.91	0.111	23.16	211.7
ROSL	0.066	29.17	3.06	0.075	29.56	8390	0.033	37.54	1.35	0.081	25.88	21.10
IRWNR	0.061	37.45	0.80	0.069	42.03	2335	0.014	45.02	0.51	0.015	40.40	22.68

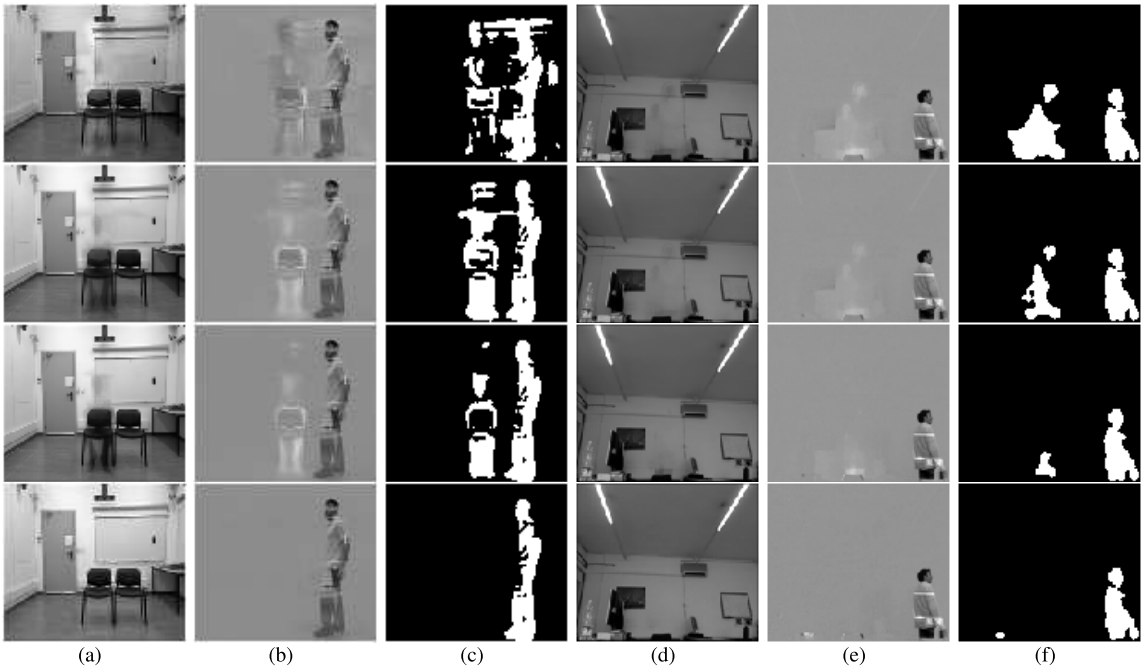


FIGURE 12. Example results of (a)(d) background, (b)(e) foreground, and (c)(f) filtered $\boldsymbol{E}$ on *sroom* (left) and *office* (right) datasets, as recovered by LRR, ncRPCA, ROSL and IRWNR (Sequentially from top to bottom).

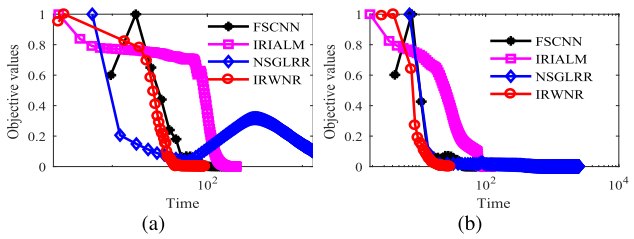


FIGURE 13. Objective values versus execution time (in seconds) on the (a) MNIST and (b) COIL datasets.

feature factor $\boldsymbol{W}$; (ii) the iteratively updated Laplacian matrix $\boldsymbol{L}$; and the nonconvex constraint r (iii) with or (iv) without the weight $\boldsymbol{s}$. Their individual contributions are controlled by parameters γ, β, and a respectively. To demonstrate the role of these properties, we simply set the corresponding parameter to 0 and leave the remaining model applied in different datasets. Table 8 shows the experimental results from different combinations of these properties on MNIST, COIL, and AR datasets. The baseline comes from Eq. (8) with $l = n$, $\beta = 0$, and the constraint being nuclear norm, which adopts none of these properties. On the other side, the proposed IRWNR adopts all. As can be seen, although the contributions are different, all these properties generate

TABLE 8. Effectiveness of three properties on AC and NMI (%) in MNIST, COIL, and AR datasets.

Properties	MNIST		COIL		AR	
	AC	NMI	AC	NMI	AC	NMI
baseline	51.8	53.2	64.4	75.3	80.0	93.9
with i	51.9	53.5	68.7	77.7	85.3	96.1
with ii	54.1	56.6	77.6	85.4	82.4	94.3
with iii	63.7	62.9	66.8	77.2	80.0	93.9
with iv	62.2	61.0	65.2	76.7	80.0	93.9
with i and ii	55.1	56.6	78.5	87.6	86.2	96.8
with i and iii	64.2	63.8	69.1	78.3	85.3	96.1
with i and iv	63.7	63.2	68.7	77.8	85.3	96.1
with ii and iii	65.2	63.2	81.8	88.1	82.4	94.3
with ii and iv	64.0	62.8	81.5	87.9	82.4	94.3
with all	66.9	64.8	82.6	89.1	86.2	96.8

positive impact on the performance. Specifically, property (i) contributes little on non-occluded MNIST dataset, but performs well on occluded AR dataset (under glass corruption). Property (ii) is useful on all these three datasets, especially for COIL. Besides, while property (iii) does not work in AR, its performance improvement in MNIST is noticeable. These results demonstrate that our approach is capable of dealing

with different types of data distribution in practical scenarios. Moreover, the clustering results generated with property (iii) is consistently better (at least equal) than those generated with property (iv), which verifies the superiority of the newly added weight factor $\boldsymbol{s}$.

V. CONCLUSION

In this paper, we propose a new low-rank representation method, IRWNR, which marries the merits of feature learning, manifold update, and weighted nonconvex constraint. Among these merits, the first reveals different contributions of input features in the learning process, the second guarantees IRWNR to construct a better Laplacian matrix for more accurately capturing the intrinsic structure of the data, and the third ensures a closer approximation to the latent low-rank representation matrix. A reconstrained inexact APG framework is presented to solve our IRWNR model. Furthermore, based on a key observation that the singular values can be automatically thresholded, we approximate the proximal operator by a smaller matrix and the power method. Experiments on synthetic data, six image segmentation datasets, and four video sequences demonstrate that IRWNR is not only robust to different types of noises and heterogeneous data distribution, but also more efficient than other state-of-the-art iterative methods.

There are several ways to further improve or extend the proposed approach: 1) Although the theoretical analysis and experimental studies gave a span for setting of some parameters such as η and a, it is challenging to determine the optimal value of the remaining parameters. Thus, a deep exploration on model selection is crucial to more general applications. 2) The proposed method works slowly in the initial stage of SVT approximation due to the higher rank of representation matrix. One way to remedy this problem is to investigate a useful warm-start strategy. The other is to augment the proposed model by imposing explicit rank constraint on Laplacian matrix or representation matrix. 3) Despite the promising results, IRWNR is restricted to data clustering with limited sample size, i.e., $n \leq 5000$, since that it runs in batch mode. We will dedicate to implement our method in an incremental manner, so as to make it workable in large-scale applications or dynamic data streams.

APPENDIX A
PROOF OF LEMMA 2

Proof: For simplicity of notations, we denote $\sigma_i(\boldsymbol{Z})$ as σ_i. With a given matrix $\boldsymbol{\Theta}_k \in R^{n\times n}$, problem (14) can be rewritten as

$$\begin{aligned}\text{prox}_{\eta r}(\boldsymbol{\Theta}_k) &= \min_{\boldsymbol{Z}} \frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\Theta}_k\|_F^2 + \eta\sum_{i=1}^{n} r(s_i,\sigma_i)\\ &= \min_{\boldsymbol{Z}} \frac{1}{2}\text{tr}(\boldsymbol{Z}^T\boldsymbol{Z}) - \text{tr}(\boldsymbol{Z}\boldsymbol{\Theta}_k^T) + \eta\sum_{i=1}^{n} r(s_i,\sigma_i)\\ &= \min_{\boldsymbol{Z}} -\text{tr}(\boldsymbol{Z}\boldsymbol{\Theta}_k^T) + \sum_{i=1}^{n}\frac{1}{2}\sigma_i^2 + \eta r(s_i,\sigma_i). \quad (20)\end{aligned}$$

Note that $\text{tr}(\boldsymbol{Z}\boldsymbol{\Theta}_k^T)$ is linear in $\boldsymbol{Z}$ and the summation operation preserves convexity. Hence, problem (20) is strictly convex if $q_i = \frac{1}{2}\sigma_i^2 + \eta r(s_i,\sigma_i)$ is strictly convex. To this end, it suffices to reveal that the second derivative of all q_i, $i = 1,\cdots,n$, are positive, i.e. $q_i''(\sigma_i) > 0$ for all $\sigma_i \geq 0$, which generates

$$1 - \eta r''(s_i,\sigma_i) > 0. \quad (21)$$

With the premise that $r''(s_i,\sigma_i) = s_i r''(\sigma_i)$, and $0 > r''(\sigma_i) \geq r''(0) = -a$, (21) leads to $0 < a < 1/(\eta\max(\boldsymbol{s}))$. □

APPENDIX B
PROOF OF THEOREM 1

Proof: Let $\boldsymbol{\Theta}_k = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be the SVD of $\boldsymbol{\Theta}_k$. Since $r(\boldsymbol{Z})$ and Frobenius norm are all unitary invariant, we have

$$\begin{aligned}\boldsymbol{Z}_{k+1} &= \arg\min_{\boldsymbol{Z}} \frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\Theta}_k\|_f^2 + \eta r(\boldsymbol{Z})\\ &= \arg\min_{\boldsymbol{Z}} \frac{1}{2}\|\boldsymbol{U}\boldsymbol{Z}\boldsymbol{V}^T - \boldsymbol{\Sigma}\|_F^2 + \eta r(\boldsymbol{U}\boldsymbol{Z}\boldsymbol{V}^T)\\ &= \boldsymbol{U}\arg\min_{\boldsymbol{Z}}\{\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\Sigma}\|_F^2 + \eta r(\boldsymbol{Z})\}\boldsymbol{V}^T. \quad (22)\end{aligned}$$

Thus, we need to prove that

$$\boldsymbol{\Xi}(\boldsymbol{\Sigma},a,\eta) = \min_{\boldsymbol{Z}}\{\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\Sigma}\|_F^2 + \eta r(\boldsymbol{Z})\} \quad (23)$$

is the optimal solution. Notice that with Lemma 2, (23) is strictly convex since $0 < a < 1/(\eta\max(\boldsymbol{s}))$, and hence ensures a unique minimum. Let $\boldsymbol{Z} = \boldsymbol{U}_{\boldsymbol{Z}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{V}_{\boldsymbol{Z}}^T$ be the SVD of $\boldsymbol{Z}$, we further have

$$\begin{aligned}\|\boldsymbol{Z}-\boldsymbol{\Sigma}\|_F^2 &= \|\boldsymbol{Z}\|_F^2 + \|\boldsymbol{\Sigma}\|_F^2 - 2\text{tr}(\boldsymbol{Z}^T\boldsymbol{\Sigma})\\ &\geq \|\boldsymbol{\Sigma}_{\boldsymbol{Z}}\|_F^2 + \|\boldsymbol{\Sigma}\|_F^2 - 2\text{tr}(\boldsymbol{\Sigma}_{\boldsymbol{Z}}^T\boldsymbol{\Sigma})\\ &= \|\boldsymbol{\Sigma}_{\boldsymbol{Z}}-\boldsymbol{\Sigma}\|_F^2, \quad (24)\end{aligned}$$

where the inequality comes from von NeumannÂąÂŕs theory. Inequality (24) implies that

$$\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\Sigma}\|_F^2 + \eta r(\boldsymbol{Z}) \geq \frac{1}{2}\|\boldsymbol{\Sigma}_{\boldsymbol{Z}}-\boldsymbol{\Sigma}\|_F^2 + \eta r(\boldsymbol{\Sigma}_{\boldsymbol{Z}}). \quad (25)$$

Note that the equality holds if $\boldsymbol{Z} = \boldsymbol{\Sigma}_{\boldsymbol{Z}}$. Therefore, problem (23) can be reduced to minimize the right side of problem (25), which is separable from Lemma 1. Hence, the solution can be achieved by applying $\boldsymbol{\Xi}$ to the entries of $\boldsymbol{\Sigma}$. □

APPENDIX C
PROOF OF THEOREM 2

Proof: First, from Theorem 1, we have

$$\begin{aligned}&\langle\nabla f(\boldsymbol{Z}_k), \boldsymbol{Z}_{k+1}-\boldsymbol{Z}_k\rangle + \frac{1}{2\eta}\|\boldsymbol{Z}_{k+1}-\boldsymbol{Z}_k\|_F^2\\ &\quad + \sum_{i=1}^{n} r(s_{k+1,i},\sigma_i(\boldsymbol{Z}_{k+1})) \leq \langle\nabla f(\boldsymbol{Z}_k), \boldsymbol{Z}_k-\boldsymbol{Z}_k\rangle\\ &\quad + \frac{1}{2\eta}\|\boldsymbol{Z}_k-\boldsymbol{Z}_k\|_F^2 + \sum_{i=1}^{n} r(s_{k,i},\sigma_i(\boldsymbol{Z}_k))\end{aligned}$$

which can be reformulated as

$$\begin{aligned}&\langle \nabla f(\mathbf{Z}_k), \mathbf{Z}_k - \mathbf{Z}_{k+1}\rangle \\ &\geq -\sum_{i=1}^{n} r(s_{k,i}, \sigma_i(\mathbf{Z}_k)) \\ &\quad + \sum_{i=1}^{n} r(s_{k+1,i}, \sigma_i(\mathbf{Z}_{k+1})) + \frac{1}{2\eta}\|\mathbf{Z}_{k+1} - \mathbf{Z}_k\|_F^2 \end{aligned} \quad (26)$$

Second, since $f(\mathbf{Z})$ is Lipchitz smooth, we have

$$\begin{aligned}&f(\mathbf{Z}_k) - f(\mathbf{Z}_{k+1}) \\ &\geq \langle \nabla f(\mathbf{Z}_k), \mathbf{Z}_k - \mathbf{Z}_{k+1}\rangle \\ &\quad -\frac{\nu}{2}\|\mathbf{Z}_{k+1} - \mathbf{Z}_k\|_F^2 \geq -\sum_{i=1}^{n} r(s_{k,i}, \sigma_i(\mathbf{Z}_k)) \\ &\quad + \sum_{i=1}^{n} r(s_{k+1,i}, \sigma_i(\mathbf{Z}_{k+1})) + \frac{1}{2}(\frac{1}{\eta} - \nu)\|\mathbf{Z}_k - \mathbf{Z}_{k+1}\|_F^2, \end{aligned} \quad (27)$$

where the first and second inequality come from [25, Definition 1] and inequality (26), respectively. Since $\eta \in (0, 1/\nu)$, inequality (27) further leads to

$$F(\mathbf{Z}_k) - F(\mathbf{Z}_{k+1}) \geq \frac{1}{2}(\frac{1}{\eta} - \nu)\|\mathbf{Z}_k - \mathbf{Z}_{k+1}\|_F^2 \geq 0, \quad (28)$$

which demonstrates that $F(\mathbf{Z}_k)$ is monotonically deceasing. By summing up all the inequality with $k = 1, 2, \cdots, \infty$, we get $\lim_{k\to\infty} ||\mathbf{Z}_k - \mathbf{Z}_{k+1}||_F^2 = 0$. □

REFERENCES

[1] 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.
[2] K. Guo, X. Xu, and D. Tao, "Discriminative GoDec+ for classification," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3414–3429, Jul. 2017.
[3] H. Zhu *et al.*, "YoTube: Searching action proposal via recurrent and static regression networks," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2609–2622, Jun. 2018.
[4] Y. Y. Chen, L. Zhang, and Z. Yi, "Subspace clustering using a low-rank constrained autoencoder," *Inf. Sci.*, vol. 424, pp. 27–38, Jan. 2018.
[5] X. Peng, S. Xiao, J. Feng, W.-Y. Yau, and Z. Yi, "Deep subspace clustering with sparsity prior," in *Proc. 25th Int. Conf. Artif. Intell.*, New York, NY, USA, Jul. 2016, pp. 1925–1931.
[6] J. T. Zhou *et al.*, "Transfer hashing: From shallow to deep," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2018.2827036.
[7] H. Hu, Z. Lin, J. Feng, and J. Zhou, "Smooth representation clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 3834–3841.
[8] C.-G. Li, C. You, and R. Vidal, "Structured sparse subspace clustering: A joint affinity learning and subspace clustering framework," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2988–3001, Jun. 2017.
[9] M. Yin, J. Gao, Z. Lin, Q. Shi, and Y. Guo, "Dual graph regularized latent low-rank representation for subspace clustering," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4918–4933, Dec. 2015.
[10] C. Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan, "Robust and efficient subspace segmentation via least squares regression," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 347–360.
[11] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.
[12] Z. Kang, C. Peng, and Q. Cheng, "Robust PCA via nonconvex rank approximation," in *Proc. IEEE Conf. Data Mining*, Atlantic, NJ, USA, Nov. 2015, pp. 211–220.
[13] P. Li, J. Yu, M. Yang, L. Zhang, D. Cai, and X. Li, "Constrained low-rank learning using least squares-based regularization," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4250–4262, Dec. 2017.
[14] 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.
[15] C. Peng, Z. Kang, M. Yang, and Q. Cheng, "Feature selection embedded subspace clustering," *IEEE Signal Process. Lett.*, vol. 23, no. 7, pp. 1018–1022, Jul. 2016.
[16] J. Chen and J. Yang, "Robust subspace segmentation via low-rank representation," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1432–1445, Aug. 2014.
[17] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized robust coding for face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1753–1766, May 2013.
[18] J. Yang, X. Yang, X. Ye, and C. Hou, "Reconstruction of structurally-incomplete matrices with reweighted low-rank and sparsity priors," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1158–1172, Mar. 2017.
[19] F. Nie and H. Huang, "Subspace clustering via new low-rank model with discrete group structure constraint," in *Proc. Int. Joint Conf. Artif. Intell.*, New York, NY, USA, Jul. 2016, pp. 1874–1880.
[20] M. Yin, J. Gao, and Z. Lin, "Laplacian regularized low-rank representation and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 504–517, Mar. 2016.
[21] X. Peng, C. Lu, Z. Yi, and H. J. Tang, "Connections between nuclear-norm and frobenius-norm-based representations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 218–224, Jan. 2018.
[22] J. Zheng, M. Qin, H. Yu, and W. Wang, "An efficient truncated nuclear norm constrained matrix completion for image inpainting," in *Proc. Comput. Graph. Int.*, Bintan Island, Indonesia, Jun. 2018, pp. 97–106.
[23] T. Zhang, "Analysis of multi-stage convex relaxation for sparse regularization," *J. Mach. Learn. Res.*, vol. 11, pp. 1081–1107, Mar. 2010.
[24] J. Zheng, H. Qiu, W. Sheng, X. Yang, and H. Yu, "Kernel group sparse representation classifier via structural and non-convex constraints," *Neurocomputing*, vol. 296, pp. 1–11, May 2018.
[25] C. Lu, J. Tang, S. Yan, and Z. Lin, "Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 829–839, Feb. 2016.
[26] S. Gu, Q. Xie, D. Meng, W. Zuo, X. Feng, and L. Zhang, "Weighted nuclear norm minimization and its applications to low level vision," *Int. J. Comput. Vis.*, vol. 121, no. 2, pp. 183–208, Jan. 2017.
[27] Y. Xie, S. H. Gu, Y. Liu, W. M. Zuo, W. S. Zhang, and L. Zhang, "Weighted Schatten p-norm minimization for image denoising and background subtraction," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4842–4857, Oct. 2016.
[28] H. Gui, J. Han, and Q. Gu, "Towards faster rates and oracle property for low-rank matrix estimation," in *Proc. Int. Conf. Mach. Learn.*, New York, NY, USA, Jun. 2016, pp. 2300–2309.
[29] H. Liu, J. Wu, T. Liu, D. Tao, and Y. Fu, "Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 5, pp. 1129–1143, May 2017.
[30] X. Guo, "Robust subspace segmentation by simultaneously learning data representations and their affinity matrix," in *Proc. Int. Conf. Artif. Intell.*, Buenos Aires, Argentina, Jul. 2015, pp. 3547–3553.
[31] M. Yin, S. Xie, Z. Wu, Y. Zhang, and J. Gao, "Subspace clustering via learning an adaptive low-rank graph," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3716–3728, Aug. 2018.
[32] C. Peng, Z. Kang, and Q. Cheng, "Integrating feature and graph learning with low-rank representation," *Neurocomputing*, vol. 249, pp. 106–116, Aug. 2017.
[33] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, "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.
[34] X. Ren and Z. Lin, "Linearized alternating direction method with adaptive penalty and warm Starts for fast solving transform invariant low-rank textures," *Int. J. Comput. Vis.*, vol. 104, no. 1, pp. 1–14, Aug. 2013.
[35] C. Lu, J. Feng, S. Yan, and Z. Lin, "A unified alternating direction method of multipliers by majorization minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 527–541, Mar. 2018.
[36] J. Zheng, P. Yang, S. Chen, G. Shen, and W. Wang, "Iterative re-constrained group sparse face recognition with adaptive weights learning," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2408–2423, May 2017.

[37] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1561–1576, Aug. 2011.
[38] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
[39] H. Li and Z. Lin, "Accelerated proximal gradient methods for nonconvex programming," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, vol. 1, Dec. 2015, pp. 379–387.
[40] S. Ghadimi and G. Lan, "Accelerated gradient methods for nonconvex nonlinear and stochastic programming," *Math. Program.*, vol. 156, no. 1, pp. 59–99, 2016.
[41] Q. Yao, J. T. Kwok, F. Gao, W. Chen, and T. Y. Liu, "Efficient inexact proximal gradient algorithm for nonconvex problems," in *Proc. Int. Conf. Artif. Intell.*, Melbourne, VIC, Australia, Aug. 2017, pp. 3308–3314.
[42] C.-J. Hsieh and P. Olsen, "Nuclear norm minimization via active subspace selection," in *Proc. Int. Conf. Mach. Learn.*, Beijing, China, 2014, pp. 575–583.
[43] Q. Yao, J. T. Kwok, and W. Zhong, "Fast low-rank matrix learning with nonconvex regularization," in *Proc. Int. Conf. Data Mining*, Nov. 2015, pp. 539–548.
[44] J.-F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.
[45] C. Lu, C. Zhu, C. Xu, S. Yan, and Z. Lin, "Generalized singular value thresholding," in *Proc. 29th AAAI Conf. Artif. Intell.*, Austin, TX, USA, Jan. 2015, pp. 1805–1811.
[46] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Rev.*, vol. 53, no. 2, pp. 217–288, 2011.
[47] P.-G. Martinsson and S. Voronin, "A randomized blocked algorithm for efficiently computing rank-revealing factorizations of matrices," *SIAM J. Sci. Comput.*, vol. 38, no. 5, pp. s485–s507, 2016.
[48] Y. Li and W. Yu. (2017). "A fast implementation of singular value thresholding algorithm using recycling rank revealing randomized singular value decomposition." [Online]. Available: https://arxiv.org/abs/1704.05528
[49] T.-H. Oh, Y. Matsushita, Y.-W. Tai, and I. S. Kweon, "Fast randomized singular value thresholding for low-rank optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 376–391, Feb. 2018.
[50] T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, and E.-H. Zahzah, "Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset," *Comput. Sci. Rev.*, vol. 23, pp. 1–71, 2017.
[51] X. B. Shu, F. Porikli, and N. Ahuja, "Robust orthonormal subspace learning: Efficient recovery of corrupted low-rank matrices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 3874–3881.
[52] J. G. Vila, "Parametric region-based foreground segmentation in planar and multi-view sequences," Ph.D. dissertation, Dept. Signal Theory Commun., Polytechn. Univ. Catalonia, Barcelona, Spain, 2013.

JIANWEI ZHENG received the B.Sc. degree in 2005 and the Ph.D. degree in control theory and control engineering from the Zhejiang University of Technology, China, in 2010. He is currently an Associate Professor with the School of Computer Science and Engineering, Zhejiang University of Technology. He has published over 40 academic articles in reputable journals and conferences, including the IEEE TIP, *Neurocomputing*, *Visual Computer*, *Applied Intelligence*, PCM, and CGI.

CHENG LU received the M.S. degree from the School of Computer Science and Engineering, Zhejiang University of Technology, China, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include image and video enhancement, pattern recognition, and machine learning.

HONGCHUAN YU received the Ph.D. degree in computer vision from the Institute of Intelligent Machine, Chinese Academy of Sciences, in 2000. He is currently a Senior Lecturer of computer graphics with the National Centre for Computer Animation, Bournemouth University. After that, he was a Research Fellow at Tsinghua University, Nanyang Technological University, Singapore, and The University of Western Australia, Perth, Australia. He has published over 70 academic articles in reputable journals and conferences, and regularly served as PC members/referees for international journals and conferences, including the IEEE TPAMI, the IEEE TIP, the IEEE TVCG, IVC, PR, CVIU, PRL, CAD, and CGI.

WANLIANG WANG received the Ph.D. degree in control theory and control engineering from Tongji University, Shanghai, China, in 2001. In 2002, he visited the Institute of Science and Technology, The University of Manchester, and also visited the Georgia Institute of Technology, and the University of Michigan, USA. He is currently a Full Professor with the Zhejiang University of Technology, China. His current research interests include artificial intelligence. He was a recipient of the National Outstanding Teacher Award in 2008 and the First National Teacher of Ten Thousand Plan Award in 2014.

SHENGYONG CHEN (M'01–SM'10) received the Ph.D. degree in computer vision from the City University of Hong Kong in 2003. He was with the University of Hamburg from 2006 to 2007. He is currently a Professor with the Tianjin University of Technology and also with the Zhejiang University of Technology, China. He has authored over 100 scientific papers in international journals. His research interests include computer vision, robotics, and image analysis. He is a fellow of the IET and a Senior Member of the CCF. He received the Fellowship from the Alexander von Humboldt Foundation of Germany. He received the National Outstanding Youth Foundation Award of China in 2013.