Weighted Mixed-Norm Regularized Regression for Robust Face Identification

Jianwei Zheng, Kechen Lou, Xi Yang, Cong Bai, and Jinhui Tang, *Senior Member, IEEE*

***Abstract*— Face identification (FI) via regression-based classification has been extensively studied during the recent years. Most vector-based methods achieve appealing performance in handing the noncontiguous pixelwise noises, while some matrix-based regression methods show great potential in dealing with contiguous imagewise noises. However, there is a lack of consideration of the mixture noises case, where both contiguous and noncontiguous noises are jointly contained. In this paper, we propose a weighted mixed-norm regression (WMNR) method to cope with the mixture image corruption. WMNR reveals certain essential characteristics of FI problems and bridges the vector- and matrix-based methods. Particularly, WMNR provides two advantages for both theoretical analysis and practical implementation. First, it generalizes possible distributions of the residuals into a unified feature weighted loss function. Second, it constrains the residual image as low-rank structure that can be quantified with general nonconvex functions and a weight factor. Moreover, a new reweighted alternating direction method of multipliers algorithm is derived for the proposed WMNR model. The algorithm exhibits great computational efficiency since it divides the original optimization problem into certain subproblems with analytical solution or can be implemented in a parallel manner. Extensive experiments on several public face databases demonstrate the advantages of WMNR over the state-of-the-art regression-based approaches. More specifically, the WMNR achieves an appealing tradeoff between identification accuracy and computational efficiency. Compared with the pure vector-based methods, our approach achieves more than 10% performance improvement and saves more than 70% of runtime, especially in severe corruption scenarios. Compared with the pure matrix-based methods, although it requires slightly more computation time, the performance benefits are even larger; up to 20% improvement can be obtained.**

***Index Terms*— Alternating direction method of multipliers (ADMM), face identification (FI), matrix regression, sparse representation, weighted nonconvex norm.**

Nomenclature

m	Feature dimension
n	Data size
o	Image height
q	Image width
ν	Rank of a given matrix
$\boldsymbol{X} \in R^{m\times n}$	Dictionary matrix
$\boldsymbol{Y} \in R^{o\times q}$	Query image
$\boldsymbol{B} = \boldsymbol{U}\Delta\boldsymbol{V}^{\mathrm{T}}$	Singular value decomposition of $\boldsymbol{B}$
σ	Singular value
$\mathrm{Vec}(\cdot)$	Vectorization by column concatenation
$\mathrm{Mat}(\cdot)$	Inverse operator of $\mathrm{Vec}(\cdot)$
$\odot$	Elementwise multiplication

I. Introduction

FACE identification (FI) is one of the most attractive problems in computer vision and multimedia community, and it has been extensively studied during the past two decades. Various methods and their variants have been widely investigated and applied, such as convolutional neural networks (CNNs) [1], metric learning [2], regression analysis [3], [4], and so on. The CNN-based approach has been confirmed to be particularly successful due to its strong capability of learning extremely powerful hierarchical nonlinear representation of the image. However, the implementation of the CNN-based approach normally requires higher computation power and larger amount of training data due to the massive tuning parameters of the CNN model [5]. On the other hand, metric learning-based approaches are capable of learning the discriminative semantic information for measuring the similarities among all images under the circumstance of high dimensional and small size data. However, they are incompetent to tackle the image corruptions [6].

This paper focuses on the regression-based approaches, which have aroused broad attentions due to its interpretability of intuitive principle and robustness to specific noises [7]. The most pioneered work of regression-based methods for FI is the linear representation classifier [8], which seeks for a suitable representation of any probe sample, and identifies the sample by examining which class can lead to the minimal reconstruction error. Along this line, many studies have been evolved on the characterizations of the representation coefficients and the error term in regression problems with respect to the regression model [9].

A. Related Work

In the regression-based approaches, it is well recognized that the regularization constraints, the representation residual, and the noise/corruption are crucial issues to be considered. The regularization constraints are usually imposed upon the regression models to avoid overfitting of the representation coefficients. Wright *et al.* [10] presented a sparse representation

Manuscript received January 16, 2018; revised October 15, 2018; accepted February 9, 2019. Date of publication March 20, 2019; date of current version December 10, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001001, in part by the National Natural Science Foundation of China under Grant 61602413, Grant 61502424, Grant 61732007, Grant U1611461, and Grant 61772275, and in part by the Natural Science Foundation of Zhejiang Province, China, under Grant LY19F030016 and Grant LY18F020032. *(Corresponding authors: Xi Yang; Cong Bai.)*

J. Zheng, K. Lou, X. Yang, and C. Bai are with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: zjw@zjut.edu.cn; lkc@zjut.edu.cn; xyang@zjut.edu.cn; congbai@zjut.edu.cn).

J. Tang is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: jinhuitang@njust.edu.cn).

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

Digital Object Identifier 10.1109/TNNLS.2019.2899073

classifier (SRC) for FI by employing the l_1-norm minimization to approximate the l_0-norm constraint. Zhang *et al.* [11] asserted that it is the collaborative mechanism of l_1-norm, rather than the sparsity of the l_0-norm that renders SRC resultful, and their collaborative representation classifier (CRC) can achieve similar results as SRC but significantly save the runtime. Huang *et al.* [12] further proposed a new sparse coding method by utilizing label information and $l_{2,1}$-norm, which achieves both flat and structured sparsity for the vector representations. The corresponding model is more discriminative, and the method is more efficient and effective. In correntropy-based sparse representation (CESR) [13] and structured sparse error coding (SSEC) [14], regularization constraint is selected to be the indicator function of the nonnegative orthant, such that a nonnegative coefficient vector is enforced.

The representation residuals of aforementioned works are all measured by l_2-norm of the error vector. Such treatment inherently assumes that the residuals follow a Gaussian distribution. However, in practical FI problems, the distribution of residuals is more complicated [15], [16]. To deal with the sparse pixel corruption, Wright *et al.* [10] further assumed that the noises follow a Laplacian distribution, and they presented the robust SRC (RSRC) model. Similarly, Naseem *et al.* [17] and Zhang *et al.* [18] extended their models to the robust version using the Huber and Laplacian estimator to handle extreme pixel noise and illumination variations, respectively. However, the effectiveness of these methods is based on the correct knowledge of error distribution, which is, in fact, difficult to obtain in prior. To overcome such restriction, Yang *et al.* [19] and Zheng [20] borrowed the idea of robust regression [21] and presented the regularized robust classifier (RRC) and iterative reconstrained group sparse classifier (IRGSC), respectively. He *et al.* [13] made use of the correntropy-induced robust error metric and provided the CESR method. Although RRC, IRGSC, and CESR are proposed independently, they are all essentially a robust regression model sharing the idea that correntrogy can be considered as an M-estimator. To unify the additive model, such as SRC and the multiplicative models such as RRC, IRGSC and CESR, He *et al.* [22] proposed a framework toward generalizing the multiple half-quadratic functions in light of the maximum correntropy criterion. All of these robust-regression-correlated methods have been applied to the real-world FI problems and yielded promising results.

It is worth noting that the aforementioned regression methods all use the vector-based error model, under which the occurrence of pixel errors is assumed to be independent. This assumption does not hold when contiguous corruptions, such as occlusion, disguise, or block shadow, present. In these cases, errors are spatially correlated and contain rich structural information [23]. Moreover, the vector-based methods are time-consuming due to the high-dimensional image resolution. To overcome these limitations, Yang *et al.* [24] presented the nuclear norm-based matrix representation (NMR) model for FI. NMR not only alleviates the inherent correlations caused by contiguous noises via the involved singular value decomposition (SVD) but also directly characterizes the holistic structure of error image with efficient matrix computation. However, it has been indicated in [25] and [26] that the reconstruction performance of the convex nuclear norm will lead to a suboptimal solution. For this issue, Luo *et al.* [27] enforced low rank regularization by using Schatten p-norm to guarantee a more accurate recovery of the query sample. Similarly, Xie *et al.* [28] substituted the nuclear norm with nonconvex function for characterizing the low rank structure of the error image to achieve better identification performance. However, both nuclear norm and Schatten p-norm treat all singular values equally, which are not flexible in specific scenarios, where different rank components have different contributions.

B. Contribution and Organization

In view of the merits and demerits of the existing methods, it is quite evident that the pure vector- or matrix-based methods can only handle single type of noises, i.e., noncontiguous or contiguous, and it is reasonable to expect a more flexible way dealing with various kinds of noises. In this paper, we propose a new low-rank regularizer, named weighted nonconvex norm minimization (WN^{2}M), which provides a better approximation to the original rank minimization problem. Furthermore, assuming that the real corruption is a combination of contiguous and noncontiguous noises, we combine a tailored loss function and WN^{2}M into a unified formula for more robust FI application. Comparatively, notice that some other matrix-based methods, such as the nuclear norm regularized regression (NR) [29] for Gaussian error distribution, the nuclear-l_1 norm joint matrix regression (NL$_1$R) [30], and the robust matrix regression (RMR) [28] for the Laplacian error distribution, all adopt certain types of simple rank approximation constraints and are developed specially for some usual pixel noises, which are not flexible for the practical FI application. In summary, the contributions of this paper include the following aspects.

1) A new low-rank regularizer, named WN^{2}M, is presented to unify the nonconvex constraint and component weights into a general term by taking the genuine error structure into consideration. The general analytical solution of WN^{2}M can be derived when the weights are ranged in ascending order, and the optimum can be efficiently computed by decoupling the relevant objective function into several independent nonconvex subproblems so as to be solved in a parallel manner.
2) A general loss function is proposed to estimate the representation residuals, where a tailored error weight is learned to distinguish inliers from outliers. This weight can be obtained efficiently according to an analytical solution that contains a single parameter of clear physical meaning. Particularly, the determined weight assigns small values to noise pixels and large values to the active pixels. Such mechanism reveals the level of contributions of different pixels.
3) The novel weighted mixed-norm regression (WMNR) model, which benefits from the proposed tailored loss function as well as WN^{2}M, is provided to deal with mixture noises in FI problems. In computational implementations, the relevant minimization problem is explored by a new iterative reweighted alternating direction method of multipliers (ADMM) approach, where each

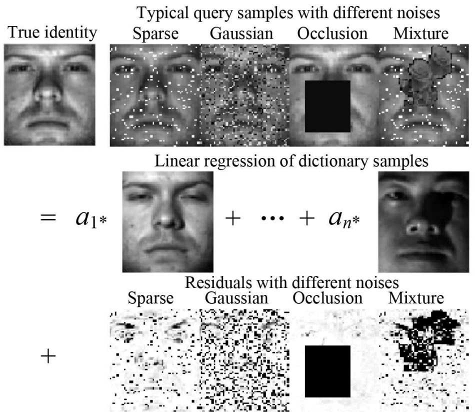

Typical query samples with different noises
True identity
Sparse
Gaussian
Occlusion
Mixture
Linear regression of dictionary samples
= a_{1*}
+ ··· + a_{n*}
Residuals with different noises
Sparse
Gaussian
Occlusion
Mixture
+

Fig. 1. Regression-based model. The query samples with various types of noises can be regressed as the linear combination of dictionary samples plus the error $\boldsymbol{e}$.

subproblem can be solved analytically and/or efficiently. Moreover, the proposed approach is flexible to cope with different types of coefficients regularization problems.

The remainder of this paper is organized as follows. In Section II, we present a unified formula for the regression-based classification. In Section III, we give a tailored loss function and the robust low-rank constraint with general nonconvex functions and then provide the solutions to these subproblems. The proposed model of WMNR and its optimization scheme are given in Section IV. Experimental results and detailed discussions on the performance of the proposed algorithms are given in Section V, and conclusions are drawn in Section VI. For convenience, some notations and theoretical proofs are provided in the Appendixes.

II. Unified Formula for Regression-Based Classification

As mentioned in Section I-A, various regression-based methods have been proposed for different kinds of FI problems, yet their main principles are quite the same, which is to derive a series of appropriate regression coefficients that may greatly facilitate the subsequent classification scheme. Having fully aware of their principal similarity and technical differences, in this section, we present a unified formulation that combines all the essential factors in the regression-based classification.

As illustrated in Fig. 1, the fundamental idea behind the regression-based model is that any query samples with various types of noises can be represented as the superposition of dictionary samples and the residual $\boldsymbol{e}$. To be more specific, given any query face image $\boldsymbol{Y} \in R^{o\times q}$, it is possible to show

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{y} = \mathrm{Vec}(\boldsymbol{Y}) \in R^m$ with $m = o \times q$, $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_c] \in R^{m\times n}$ is the dictionary matrix with the set of samples from c individuals, and $\boldsymbol{a}$ is the coefficient vector used for computing the minimum class residual and determining the face identity [8]. From Fig. 1, we can see that the residual image $\boldsymbol{e}$ normally possesses the following two characteristics.

1) It follows certain distribution that can be faithfully reflected by a tailored loss function, e.g., the Gaussian function [10], the Laplacian function [18], or the Logistic function [19].
2) The residual image $\boldsymbol{E} = \mathrm{Mat}(\boldsymbol{e})$ is always considered to be low-rank since many of its entries are useless due to the corruption of contiguous noises, e.g., certain occlusion images.

Considering the nature of the regression-based model and the residual image, we present a unified criterion to obtain the regression coefficients

$$\min \sum_{i=1}^{m} \phi(e_i) + \sum_{i=1}^{\nu} g(\sigma_i(\boldsymbol{E})) + \lambda\vartheta(\boldsymbol{a}) \tag{2}$$

where ϕ is a loss function that is widely used and normally formulated by the M-estimators in various forms [31], e_i is the ith residual entry, g is a surrogate function with low-rank constraint, and $\vartheta(\boldsymbol{a})$ is the regularization of $\boldsymbol{a}$ with λ being a balance parameter. It is evident that the formulation of (2) generalizes all essential components of the regression-based analysis, and it is the footstone of our WMNR method and may be beneficial to derive new approaches. In addition, different choices of ϕ, g, and ϑ lead to various regression-based methods. In the following, we give an overview of the existed methods regarding loss function ϕ, low-rank constraint g, and coefficient regularization ϑ.

A. Loss Function

The minimization of loss functions with different types of constraints lead to the core technology for the noncontiguous noise suppression. The most widely adopted one is the Gaussian function that employs l_2-norm to characterize the reconstruction residual [8], [10], i.e., $\sum \phi(e_i) = \|\boldsymbol{e}\|_2^2$. Although l_2-norm behaves well in most of the routine classification tasks, it has been theoretically proved to be sensitive to sparse outliers [19], [32]. Comparatively, RSRC [10] resorts to the Laplacian function for l_1-sparsity constrained maximum likelihood estimation solution. Likewise, RCRC [18] characterizes the fidelity with l_1-norm, i.e., $\sum \phi(e_i) = \|\boldsymbol{e}\|_1$, for robustness to sparse corruption. However, the l_1-sparsity constraint normally makes the computational complexity of both RSRC and RCRC high. Inspired by the robust regression theory, RRC [19] models the representations as a weighted regression problem and adopts an independent mapping function to characterize the noises. Specifically, RRC uses $\sum \phi(e_i) = \|\boldsymbol{w} \odot \boldsymbol{e}\|_2^2$ as the loss function, and the weight $\boldsymbol{w}$ is set to be the Logistic function as follows:

$$w_i = \frac{\exp\left(-\beta e_i^2 + \beta\theta\right)}{1 + \exp\left(-\beta e_i^2 + \beta\theta\right)} \tag{3}$$

whose parameters β and θ are both positive scalars for desirable restraint to noises. With the help of $\boldsymbol{w}$, RRC is capable of distinguishing the effective features from the invalid ones.

B. Low-Rank Constraint

The aforementioned loss functions characterize the image error in a pixel-by-pixel manner so as to tackle the noncontiguous corruption problems. However, they neglect the structural information contained in the residual image $\boldsymbol{E}$. To remedy this issue, the rank minimization constraint, i.e., rank($\boldsymbol{E}$), is widely used to determine the regression coefficients. In practice, rank($\boldsymbol{E}$) is generally converted into nuclear norm, i.e., $\|\boldsymbol{E}\|_* = \sum \sigma_i(\boldsymbol{E})$, for optimization tractability [33]. A variety of recently proposed methods, including NMR [24], NL$_1$R [30], NR [29], and so on, all adopt $\sum g(\sigma_i(\boldsymbol{E})) = \|\boldsymbol{E}\|_*$ to approximate the rank constraint.

It is known that the nuclear norm minimization is equivalent to low-rank constraint under necessary incoherence conditions [28]. However, the corresponding solution is always suboptimal to the original rank minimization since it is a loose approximation. This fact motivates us to pursue the nonconvex surrogate function to approximate the rank constraint. S$_p$L$_q$ [27] adopts the Schatten p-norm, i.e., $\sum g(\sigma_i(\boldsymbol{E})) = \|\boldsymbol{E}\|_{S_p}^p = \sum \sigma_i(\boldsymbol{E})^p$, $p \in (0, 1]$, to regress the query samples, which possesses two merits. First, it follows the classical Abel theorem that the Schatten p-norm constraint has algebraic roots when $p = 1/2, 2/3$, which results in the analytical solution of subproblems in S$_p$L$_q$ optimization. Second, S$_p$L$_q$ can alleviate the correlations among features in residual matrix $\boldsymbol{E}$ and make the distribution approximately be Gaussian. Another way to improve the performance of nuclear norm is to treat different rank components unequally as in RMR [28], which presents a weighted nuclear norm constrained matrix regression for FI, i.e., $\sum g(\sigma_i(\boldsymbol{E})) = \|\boldsymbol{E}\|_{s,*} = \sum s_i \sigma_i(\boldsymbol{E})$, where the weight s_i ensures a better approximation to the original rank minimization problem.

C. Coefficients Regularization

Different regularizers lead to different properties of representation coefficients $\boldsymbol{a}$. A carefully chosen $\vartheta(\boldsymbol{a})$ will force $\boldsymbol{a}$ to be concentrated in reasonable areas so that it can be regressed elaborately by samples in $\boldsymbol{X}$. The most pioneered confine of $\vartheta(\boldsymbol{a})$ is the l_1-norm, which forces most of the coefficients related to other subjects to be zero. SRC [10], RRC [19], and NR [29] all adopt this constraint into their cost functions for sparsity. Another popular choice is the l_2-norm that makes the unconcerned part of $\boldsymbol{a}$ to be small in magnitude but not absolutely zero. Due to its high efficiency, l_2-norm has been widely used in CRC [18], NMR [24], RMR [28], and so on. By jointly considering the sparsity and collaboration, GSC [12] adopts l_2-norm for intraclass coefficients and l_1-norm for interclass coefficients. This group property makes the samples from the same class prefer to hold the flat values, i.e., it is expected that the samples from the correct subject can dominate the coding coefficients via $\vartheta(\boldsymbol{a}) = \|\boldsymbol{a}\|_{2,1} = \sum_{i=1}^{c} \|a_i\|_2$. Besides, in CESR [13] and SSEC [14], $\vartheta(\boldsymbol{a})$ is selected to be the indicator function of the nonnegative orthant as $I(a_i) = a_i$, when $a_i > 0$, or $I(a_i) = 0$, when $a_i \le 0$, such that a nonnegative regularization term $\boldsymbol{a} \ge 0$ is enforced.

III. Robust Loss Approximations for Mixture Corruption

In Section II, we generalize most of the existing methods regarding the regression coefficient into a unified formula consisting of ϕ, g, and ϑ, by which different regression-based classification methods can be implemented. In this section, we focus on the feature learned loss function and the weighted nonconvex low-rank constraint so as to handle both the contiguous and noncontiguous image corruptions simultaneously in the robust FI problem.

A. Feature Learned Loss Function

In general, it is a challenging problem to predefine ϕ for the regression residual due to the diversity of noise variations. A natural assumption is that the unknown function ϕ is symmetric around zero and differentiable [4], [19]. Furthermore, the function $\phi(r^{1/2})$ should be concave and increasing when $r > 0$ [36]. These conditions lead to the fact that $\phi(r)'/r$ is decreasing for $r > 0$. Under these assumptions, $\phi(r)$ can be represented as follows (see [37]):

$$\phi(r) = \inf_{\zeta > 0} \frac{1}{2}\zeta r^2 - \phi^*\left(\frac{1}{2}\zeta\right) \tag{4}$$

where $\phi^*(\zeta)$ is the conjugate function of $\phi(r^{1/2})$ with variational parameter ζ. Substituting (4) into the loss function (2), we can write the subproblem with respect to ϕ as follows:

$$\min_{w} \frac{1}{2}\|\sqrt{\boldsymbol{W}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})\|_2^2 + \varphi(\boldsymbol{w}) \tag{5}$$

where $\boldsymbol{W} = diag(\boldsymbol{w})$ with $\boldsymbol{w} = (\zeta_i, \ldots, \zeta_m)$, $\varphi(\boldsymbol{w}) = \sum_{i=1}^{m} \phi^*(\zeta_i/2)$, and $\zeta = \phi(r)'/r$ is the optimal value of the dual function $\phi^*((1/2)\zeta) = \inf_r (1/2)\zeta r^2 - \phi(r)$.

For the residual term $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}$ in subproblem (5), a desirable $\boldsymbol{w}$ should be able to assign large weights to pixels with small error, whereas small weights to pixels with large error. This property distinguishes effective features from the invalid ones when the dictionary with clear images and query with corrupted ones are given. To achieve such property, any nonincreasing function of the form $\phi(r)'/r$ is optional. Specifically, if we employ $w_i = 2$ or $w_i = 1/|e_i|$, $i = 1, \ldots, m$, then the loss function (5) reduces to the Gaussian [10] or Laplacian [18] distribution, respectively. However, from Fig. 1, we can see that the real-world corruption may be more complex, e.g., the mixture of sparse noises and block occlusion. With regard to various types of noises, we introduce the adaptive weight estimation technique into optimization procedure to improve the robustness and flexibility of the method.

For the regularization term $\varphi(\boldsymbol{w})$ of subproblem (5), the usual selection can be l_2-norm, l_1-norm, $l_{2,1}$-norm, as well as nonnegative orthant. Considering that all the weights should be nonnegative, we further add a probability constraint onto $\boldsymbol{w}$ for feature balance and numerical stability, i.e., $\boldsymbol{w}^T\boldsymbol{1} = 1$ and $\boldsymbol{w} \ge 0$. Under these conditions, the l_1-norm and nonnegative orthant are constant or naturally met. Besides, without any prior knowledge, it is unable to partition the weights into different groups. Thus, we adopt l_2-norm here to obtain nontrivial and closed-form solution.

By the choice of residual term and regularization term, our loss function turns out to be

$$\min_{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{1}=1, \boldsymbol{w} \ge 0} \frac{1}{2}\|\sqrt{\boldsymbol{W}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})\|_2^2 + \gamma\,\|\boldsymbol{w}\|_2^2 \tag{6}$$

where γ is a tunable parameter. Let $\boldsymbol{X} = [\boldsymbol{f}_1; \boldsymbol{f}_2; , \ldots, ; \boldsymbol{f}_m]$ with $\boldsymbol{f}_i^{\mathrm{T}} \in R^n$ being the ith row of $\boldsymbol{X}$, and $e_i = y_i - \boldsymbol{f}_i\boldsymbol{a}$ denotes the component error with respect to vector $\boldsymbol{a}$. Problem (6) can be rewritten as follows:

$$\min_{\boldsymbol{w}} \sum_{i=1}^{m}\left\{w_i e_i^2+\gamma w_i^2\right\} = \min_{\boldsymbol{w}}\left\|\boldsymbol{w}+\frac{\boldsymbol{d}}{2\gamma}\right\|_2^2$$
$$\text{s.t.} \quad \boldsymbol{w}^{\mathrm{T}}\boldsymbol{1}=1, \quad w_i \geq 0, \ i=1\sim m \tag{7}$$

where we denote $d_i = e_i^2$ for notation simplicity. The Lagrangian function of (7) is

$$L(\boldsymbol{w},\alpha,\boldsymbol{v})=\frac{1}{2}\left\|\boldsymbol{w}+\frac{\boldsymbol{d}}{2\gamma}\right\|_2^2-\alpha(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{1}-1)-\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w} \tag{8}$$

with two Lagrangian multipliers $\alpha \geq 0$ and $\boldsymbol{v} \geq 0$. According to the Karush–Kuhn–Tucher (KKT) condition [37], the optimal solution $\boldsymbol{w}$ is

$$\boldsymbol{w}=\left(-\frac{\boldsymbol{d}}{2\gamma}+\alpha\right)_+ \tag{9}$$

where $(\cdot)_+$ keeps the positive elements unaltered and sets the rest to be zero.

In a practical FI problem, a sparse $\boldsymbol{w}$ is preferable due to the impact of various noises. Another advantage of sparsity is that the computational cost would be alleviated to some extent. Without loss of generality, we assume $d_1, \ldots, d_m$ are arranged in the ascending order, and the optimal $\boldsymbol{w}$ has $k > 0$ nonzero weights, i.e., $w_k > 0$ and $w_{k+1} = 0$. These assumptions imply that

$$-\frac{d_{k+1}}{2\gamma}+\alpha=0. \tag{10}$$

Besides, from constraint $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{1}=1$, we get

$$\sum_{j=1}^{k}\left(-\frac{d_j}{2\gamma}+\alpha\right)=1 \Rightarrow \alpha=\frac{1}{k}+\sum_{j=1}^{k}\frac{d_j}{2\gamma k}. \tag{11}$$

Combining (10) and (11), we get

$$\gamma=\left(kd_{k+1}-\sum_{j=1}^{k}d_j\right)/2. \tag{12}$$

With the derived α and γ, the optimal $\boldsymbol{w}$ can be determined as

$$\boldsymbol{w}=(d_{k+1}-\boldsymbol{d})/\left(kd_{k+1}-\sum_{j=1}^{k}d_j\right). \tag{13}$$

Fig. 2 exhibits some typical weights function, including Gaussian, Laplacian, Logistic, and the one proposed by (13). It is clear that the Gaussian distribution treats all features equally, no matter whether it is inlier or not. On the other hand, the Laplacian distribution assigns higher values to features with smaller residuals. However, the weight tends to infinity when the residual is close to zero, which causes numerical instability. While Logistic fidelity assigns larger weights to inliers and smaller weights to outliers within the bound [0, 1], it has two undetermined parameters β and θ that require exhausting fine-tuning procedure. Moreover, it is unreasonable to assign equivalent weights to the uncorrupted features since different active pixels may contribute differently to the final classification results. Our proposed $\boldsymbol{w}$ also assigns small weights to large residuals for noises suppression, while it assigns significant weights to small ones for features ranking.

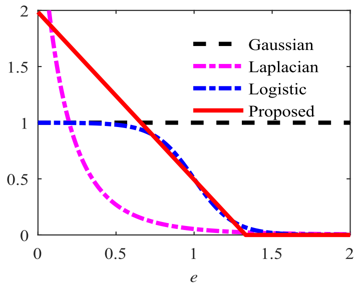


Fig. 2. Typical weight $\boldsymbol{w}$ for loss fidelity.

Furthermore, the analytical solution of (13) has only one tunable parameter k that has specific physical meaning, i.e., the number of nonzero weights, and it is easier to set than the two parameters in Logistic fidelity.

B. Weighted Nonconvex Low-Rank Constraint

Inspired by $\mathrm{S}_p\mathrm{L}_q$ [27] and RMR [28], in this section, we present a new low-rank regularizer, named WN^{2}M, for structure the mining subproblem with respect to the second term in criterion (2). The proposed weighted nonconvex norm of residual matrix $\boldsymbol{E} \in R^{o\times q}$ is defined as

$$\|\boldsymbol{E}\|_{s,g}=\sum_{i=1}^{\nu}s_i g(\sigma_i) \tag{14}$$

where $\boldsymbol{s}=[s_1, \ldots, s_\nu]^{\mathrm{T}}$ is the vector of nonnegative weights, and $g(\cdot)$ is a continuous, concave, and nondecreasing function for better approximation of the rank constraint. Table I shows several well-known nonconvex surrogate functions g, including l_p quasi norm $(0 < p < 1)$ [27], Logarithm [38], minimax concave penalty [39], exponential type penalty [40], Geman [41], Laplace [42], and LogExp [43], as well as their first supergradients. Numerical studies [33], [41] have demonstrated that the nonconvex surrogates usually outperform their convex counterparts in the field of error correction and image recovery.

Under (14), our low-rank approximation aims to find an appropriate matrix $\boldsymbol{E}$, which provides the closest approximation of a given intermediate matrix $\boldsymbol{G}$ in the sense of the Frobenius norm fidelity, that is

$$\boldsymbol{E}=\arg\min_{\boldsymbol{E}}\frac{1}{2}\|\boldsymbol{E}-\boldsymbol{G}\|_F^2+\|\boldsymbol{E}\|_{s,g}. \tag{15}$$

In (15), the integration of weighted scheme and nonconvex norm makes the problem more challenging than that only one of them is considered. As discussed in [44], the problem of nonconvex Schatten p-norm relaxation can be directly decomposed into several independent subproblems. However, the resulting solution may not fit (15) due to the general nonconvex function and the newly added weights. To achieve feasible problem decomposition, Lemma 1 is introduced first (all the proofs of the lemmas and theorems are provided in the Appendixes).

Lemma 1: Given the SVD of $\boldsymbol{G}$ as $\boldsymbol{G}=\boldsymbol{U\Delta V}^{\mathrm{T}}$, where the singular values are $\{\sigma_i, i=1,2,\ldots,\nu\}$, then the optimal solution of (15) will be $\boldsymbol{E}=\boldsymbol{U\Sigma V}^{\mathrm{T}}$ with $\boldsymbol{\Sigma} = diag\{\delta_i,$

TABLE I

WELL-KNOWN NONCONVEX SURROGATE FUNCTIONS AND THEIR FIRST SUPERGRADIENTS

Function	$g(\sigma), \mu > 0$	First Supergradient
l_p [27]	$\mu\sigma^p$	$\begin{cases} +\infty, & \text{if} \quad \sigma = 0, \\ \mu p \sigma^{p-1}, & \text{if} \quad \sigma > 0. \end{cases}$
Logarithm [38]	$\frac{\mu}{\log(p+1)} \log(p\sigma + 1)$	$\frac{p\mu}{(p\sigma+1)\log(p+1)}$
MCP [39]	$\begin{cases} \mu\sigma - \frac{\sigma^2}{2p}, & \text{if} \quad \sigma < p\mu, \\ \frac{1}{2}p\mu^2, & \text{if} \quad \sigma \geq p\mu. \end{cases}$	$\begin{cases} \mu - \frac{\sigma}{p}, & \text{if} \quad \sigma < p\mu, \\ 0, & \text{if} \quad \sigma \geq p\mu. \end{cases}$
ETP [40]	$\frac{\mu}{1-\exp(-p)}(1 - \exp(-p\sigma))$	$\frac{\mu p}{1-\exp(-p)} \exp(-p\sigma)$
Geman [41]	$\frac{\mu\sigma}{\sigma+p}$	$\frac{\mu p}{(\sigma+p)^2}$
Laplace [42]	$\mu(1 - \exp(-\frac{\sigma}{p}))$	$\frac{\mu}{p} \exp(-\frac{\sigma}{p})$
LogExp [43]	$\mu \log(2/(1 + \exp(-\sigma/p)))$	$\frac{\mu}{p(1+\exp(\sigma/p))}$

$i = 1, 2, \ldots, \nu\}$, where δ_i is solved by the following problem:

$$\begin{cases} \min\limits_{\delta} \sum\limits_{i=1}^{\nu} \left(\frac{1}{2}(\delta_i - \sigma_i)^2 + s_i g(\delta_i) \right) \\ \text{s.t.} \quad \delta_i \geq 0 \quad \text{and } \delta_i \geq \delta_j, \text{ for } i \leq j. \end{cases} \tag{16}$$

Lemma 1 can be considered as the intermediate step of problem conversion. However, it is still very challenging to solve problem (16) due to the nonnegative ($\delta_i \geq 0$) and order ($\delta_i > \delta_j$, $i < j$) constraints. Intuitively, if these two conditions can be discarded, then problem (16) would be solved in a parallel manner with respect to

$$\min f_i(\delta) = \frac{1}{2}(\delta_i - \sigma_i)^2 + s_i g(\delta_i), \quad i = 1, \ldots, \nu. \tag{17}$$

Taking l_p, Logarithm, and Geman functions listed in Table I for example, Fig. 3 illustrates the function $f_i(\delta)$ with varying $\boldsymbol{\sigma}$ and $\boldsymbol{s}$. The weights $\boldsymbol{s}$ are set in nondescending order as $\{0.8, 1.5, 2.5, 2.5, 2.5\}$ corresponding to $\{\sigma_i, i = 1, \ldots, 5\}$. Since the solution to (17) is in the range of $[0, \sigma]$ for $\sigma > 0$ and $[\sigma, 0]$ for $\sigma < 0$ [44], without loss of generality, we only take the case $\sigma > 0$ for consideration. As shown in Fig. 3, for all the nonconvex functions in Table I with fixed μ and p, there exists a certain threshold sequence $\boldsymbol{\tau}$. When $\sigma_i < \tau_i$, the minimum of function $f_i(\delta)$ is located at $\delta_i = 0$. Otherwise, a specific positive δ_i would be optimal for the minimal $f_i(\delta)$. From the blue lines with two square marks and according to [44], a correct thresholding τ_i and the corresponding δ_i^* can be determined by letting $f_i(\delta)$ equal to $f_i(0)$, that is

$$\frac{1}{2}\left(\delta_i^* - \tau_i\right)^2 + s_i g\left(\delta_i^*\right) = \frac{1}{2}\tau_i^2 \tag{18}$$

from which we can get

$$s_i g\left(\delta_i^*\right) = \frac{1}{2}\delta_i^{*2} + s_i \delta_i^* g'\left(\delta_i^*\right) \tag{19}$$

$$\tau_i = \delta_i^* + s_i g'\left(\delta_i^*\right) \tag{20}$$

where g' is the first supergradient of the detailed selection of nonconvex function g in Table I. With the determined τ_i and σ_i, $f_i(\delta)$ has the unique minimum that satisfies

$$\delta_i - \sigma_i + s_i g'(\delta_i) = 0 \tag{21}$$

Algorithm 1 WN^{2}M Complexity

```
Input: $\boldsymbol{G}, \boldsymbol{s}, t_m$, where $t_m$ denotes the terminal index.
Output: the optimal $\boldsymbol{E}$
1. $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{V}^T$ with $\boldsymbol{\Delta} = diag\{\sigma_i, i = 1, 2, \ldots, \nu\}$;
                                                        $O(oq\min(o, q))$
2. Solve (20) and (21) to obtain $\boldsymbol{\tau}$.                 $O(\nu)$
3. Let $\delta_i = 0$, for all the $|\sigma_i| < \tau_i$;
4. Let $\delta_i^0 = |\sigma_i|$ all the $|\sigma_i| \geq \tau_i$, then conduct step 6 in parallel;
5. for $k = 1, \ldots, t_m$ do
6.    $\delta_i^k = |\sigma_i| - s_i g'(\delta_i^{k-1})$                 $O(\nu)$
7.    $k = k + 1$;
8. end
9. $\delta_i = \text{sgn}(\sigma_i)\delta_i^{t_m}$;
10. $\boldsymbol{\Sigma} = diag\{\delta_i, i = 1, 2, \ldots, \nu\}$;
11. Return $\boldsymbol{E} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$
```

in the range of (δ^*, ∞) for $\sigma_i \in (\tau_i, \infty)$.

To sum up, we present WN^{2}M in Algorithm 1 for solving subproblem (15), which involves two main issues: the achievement of threshold τ and the fast searching of the optimal solution $\boldsymbol{E}$.

We now return to cope with the nonnegative and order constraints. Borrowing the idea of [44], the nonnegative condition is naturally satisfied when no weights are encountered. Unfortunately, the solutions of the decoupled $f_i(\delta)$ may not satisfy the order constraint when weights occur. From the red lines and the minimum circles in Fig. 3, we further hypothesize that the order constraint can be naturally satisfied when $\boldsymbol{s}$ being a nondescending sequence. To confirm this hypothesis, we present Theorem 1 as follows.

Theorem 1: Given the weights $\boldsymbol{s}$ satisfying $0 \leq s_1 \leq s_2, \ldots, \leq s_\nu$, the optimal solutions $\boldsymbol{\delta}$ of the decoupled subproblems in (17) have the same nonascending order as $\boldsymbol{\sigma}$.

According to Theorem 1, we implement Algorithm 1 with $\boldsymbol{s}$ in the nondescending order. Generally, the singular values of an image matrix are always sorted in the descending order, and the larger singular values usually correspond to the subspaces

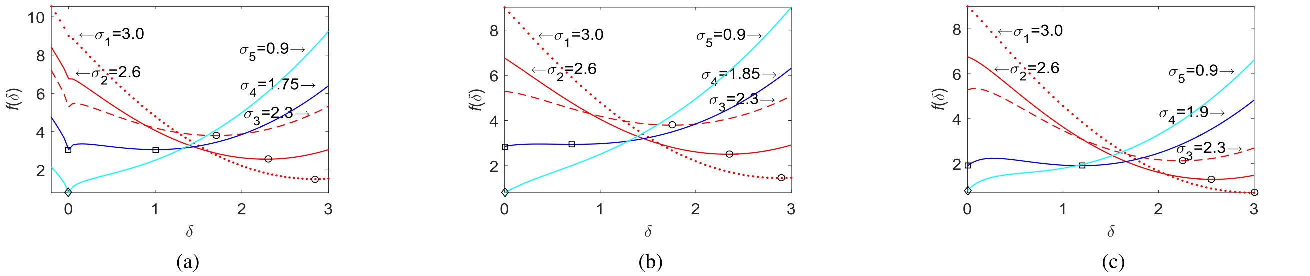


Fig. 3. Typical nonconvex functions of $f_i(\delta)$ with different $\boldsymbol{s}$ and σ. (a) l_p function with $\mu = 1$ and $p = 0.6$. (b) Logarithm function with $\mu = 1$ and $\gamma = 1.6$. (c) Geman function with $\mu = 1$ and $\gamma = 0.4$.

of more important components. The nondescending order is significant for the practical FI problem since the pixels with larger rank components should be penalized less than the remainder pixels, and therefore, the maintenance of dominant image information would be guaranteed.

We update weight vector as $\boldsymbol{s} = g'(\boldsymbol{\sigma})$ in this paper. Taking one face image from Extended Yale B (ExYaleB) occluded with 80% black block as the example, Fig. 4 illustrates the reconstructed faces and learned singular values by different methods. Fig. 4(a) and (b) are the original and occluded images, while Fig. 4(c)–(f) are the reconstructed images from NMR with nuclear norm, S$_p$L$_q$ with Schatten p-norm, RMR with weighted nuclear norm, and WN^{2}M with the mixed norm, respectively. It can be seen that WN^{2}M obtains much more faithful face images with the aid of both nonconvex constraint and ascending weights. In Fig. 4(g), we use the occluded image directly to calculate its original singular values. The learned singular values of NMR deviate severely from the genuine ones. S$_p$L$_q$ performs better than NMR with much more genuine rank components. RMR achieves even better performance with more zero singular values produced by the massive black coverage. However, our method still outperforms all others, and its curve is nearly the same as the original one.

IV. Proposed Method

In this section, the WMNR method with the relevant optimization procedure is given; some special cases are also discussed. To cope with contiguous and noncontiguous noises in a more robust manner, we substitute the tailored loss function (6) and the weighted nonconvex low-rank constraint (14) into the unified formula (2), which leads to the cost function of the proposed WMNR model

$$\begin{aligned} J(\boldsymbol{a}, \boldsymbol{w}) &= \|\sqrt{\boldsymbol{W}}\boldsymbol{e}\|_2^2 + \gamma\,\|\boldsymbol{w}\|_2^2 + \|\boldsymbol{E}\|_{s,g} + \lambda\vartheta(\boldsymbol{a}) \\ \text{s.t.}\ \ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{1} &= 1, \quad w_i \geq 0, \quad i = 1 \sim m. \end{aligned} \tag{22}$$

A minimizer $J(\boldsymbol{a}, \boldsymbol{w})$ can be obtained by alternately updating the weights $\boldsymbol{w}$ with $\boldsymbol{a}$ being fixed and updating the representation coefficients $\boldsymbol{a}$ with $\boldsymbol{w}$ fixed. Especially, the resulting subproblem of $\boldsymbol{w}$ can be solved by (13). The remaining problem is to find an efficient iterative algorithm for updating $\boldsymbol{a}$.

A. Optimization

In this section, we adopt the well-known ADMM [45] method to efficiently solve the subproblem with respect to $\boldsymbol{a}$.

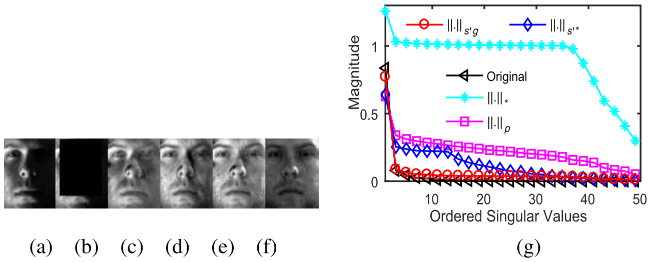


Fig. 4. Comparison of various low-rank constraints. (a) Original face, (b) 80% occluded face, as well as the reconstructed images from (c) $\|\cdot\|_*$, (d) $\|\cdot\|_p$, (e) $\|\cdot\|_{S,*}$, and (f) $\|\cdot\|_{S,g}$. (g) Approximation results with x- and y-axes being the ordered singular values and their magnitudes, respectively.

Based on our model, the cost function (22) can be reformulated as follows:

$$\begin{aligned} &\min\ \ \|\sqrt{\boldsymbol{W}}\boldsymbol{e}\|_2^2 + \|\boldsymbol{E}\|_{s,g} + \lambda\vartheta(\boldsymbol{a}) \\ &\text{s.t.}\ \boldsymbol{y} - \boldsymbol{X}\boldsymbol{a} = \boldsymbol{e}, \quad \boldsymbol{a} = \boldsymbol{u} \end{aligned} \tag{23}$$

where $\vartheta(\boldsymbol{a})$ can be any regularization term, as described in Section II-C. According to (23), the augmented Lagrange function L_ρ is written as follows:

$$\begin{aligned} L_\rho = \|\sqrt{\boldsymbol{W}}\boldsymbol{e}\|_2^2 &+ \|\boldsymbol{E}\|_{s,g} + \lambda\vartheta(\boldsymbol{u}) + \boldsymbol{z}_1^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a} - \boldsymbol{e}) \\ &+ \boldsymbol{z}_2^{\mathrm{T}}(\boldsymbol{a} - \boldsymbol{u}) + \frac{\rho}{2}\left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a} - \boldsymbol{e}\|_2^2 + \|\boldsymbol{a} - \boldsymbol{u}\|_2^2\right) \end{aligned} \tag{24}$$

where ρ is the penalty parameter, and $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are the Lagrange multipliers. Denote l as the iteration index, and $\boldsymbol{w}^{l+1}$ being fixed, the update of ADMM variables goes as follows:

$$\boldsymbol{e}_{l+1} = \arg\min_{\boldsymbol{e}} L(\boldsymbol{e}, \boldsymbol{a}_l, \boldsymbol{z}_{1,l}, \boldsymbol{w}^{l+1}) \tag{25a}$$

$$\boldsymbol{u}_{l+1} = \arg\min_{\boldsymbol{u}} L(\boldsymbol{u}, \boldsymbol{a}_l, \boldsymbol{z}_{2,l}) \tag{25b}$$

$$\boldsymbol{a}_{l+1} = \arg\min_{\boldsymbol{a}} L(\boldsymbol{a}, \boldsymbol{e}_{l+1}, \boldsymbol{u}_{l+1}, \boldsymbol{z}_{1,l}, \boldsymbol{z}_{2,l}) \tag{25c}$$

$$\boldsymbol{z}_{1,l+1} = \boldsymbol{z}_{1,l} + \rho(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}_{l+1} - \boldsymbol{e}_{l+1}) \tag{25d}$$

$$\boldsymbol{z}_{2,l+1} = \boldsymbol{z}_{2,l} + \rho(\boldsymbol{a}_{l+1} - \boldsymbol{u}_{l+1}). \tag{25e}$$

Fixing $\boldsymbol{a}_l$ and $\boldsymbol{z}_{1,l}$, (25a) can be expressed as

$$\begin{aligned} \boldsymbol{e}_{l+1} = \arg\min_{\boldsymbol{e}} \|\sqrt{\boldsymbol{W}}\boldsymbol{e}\|_2^2 &+ \|\boldsymbol{E}\|_{s,g} \\ &+ \frac{\rho}{2}\left(\|\boldsymbol{e} - (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}_l + \boldsymbol{z}_{1,l}/\rho\|_2^2)\right). \end{aligned} \tag{26}$$

We employ a two-step optimization procedure to compute $\boldsymbol{e}_{l+1}$. In the first step, we solve the weighted loss function with a penalty regularization term

$$\boldsymbol{e}_{l+1}^1 = \arg\min_{\boldsymbol{e}} \|\sqrt{\boldsymbol{W}}\boldsymbol{e}\|_2^2 + \frac{\rho}{2}\left(\|\boldsymbol{e} - (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}_l + \boldsymbol{z}_{1,l}/\rho)\|_2^2\right) \tag{27}$$

which has an analytical solution as

$$\boldsymbol{e}_{l+1}^1 = (\boldsymbol{I} + 2\boldsymbol{W}/\rho)^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}_l + \boldsymbol{z}_{1,l}/\rho) \tag{28}$$

where the first term is a diagonal matrix with respect to the fixed $\boldsymbol{W}$. Thus, we only need to conduct an elementwise multiplication between the two terms in (28), which is computationally efficient. In the second step, we apply the weighted nonconvex constraint to $\boldsymbol{e}^1_{l+1}$ for a desirable low-rank space as

$$\boldsymbol{e}_{l+1} = \arg\min_{\boldsymbol{e}} \|\boldsymbol{E}\|_{s,g} + \frac{\rho}{2}\left\|\boldsymbol{E} - \boldsymbol{E}^1_{l+1}\right\|_F^2 \qquad (29)$$

where $\boldsymbol{E}^1_{l+1} = \mathrm{Mat}(\boldsymbol{e}^1_{l+1})$, and it can be solved efficiently by Algorithm 1.

Fixing $\boldsymbol{a}_l$ and $z_{2,l}$, the subproblem of $\boldsymbol{u}_{l+1}$, (25b), becomes

$$\boldsymbol{u}_{l+1} = \arg\min_{\boldsymbol{u}} \lambda\vartheta(\boldsymbol{u}) + \frac{\rho}{2}\left(\|\boldsymbol{u} - \boldsymbol{a} - z_2/\rho\|_2^2\right) \qquad (30)$$

whose solution with a different $\vartheta(\boldsymbol{u})$ is given by

$$\boldsymbol{u}_{l+1} = \begin{cases} D_{\lambda/\rho}(\boldsymbol{a}_l + z_{2,l}/\rho), & l_1\text{-norm [27]} \\ (\boldsymbol{a}_l + z_{2,l}/\rho)/(1 + 2\lambda/\rho), & l_2\text{-norm} \\ \mathrm{Vec}(D_{i,\lambda/\rho}(\boldsymbol{a}_l + z_{2,l}/\rho), i = 1, \ldots, c), & l_{2,1}\text{-norm [45]} \\ \lambda(\boldsymbol{a}_l + z_{2,l}/\rho)_+, & \text{nonnegative orthant [4].} \end{cases} \qquad (31)$$

Fixing $\boldsymbol{e}_{l+1}$, $\boldsymbol{u}_{l+1}$, $z_{1,l}$, and $z_{2,l}$, the update of $\boldsymbol{a}_{l+1}$ as given in (25c) can be expressed as

$$\boldsymbol{a}_{l+1} = \arg\min_{\boldsymbol{a}} \left(\|\boldsymbol{X}\boldsymbol{a} - \boldsymbol{g}_a\|_2^2 + \|\boldsymbol{a} - \boldsymbol{g}_u\|_2^2\right) \qquad (32)$$

where $\boldsymbol{g}_a = \boldsymbol{y} - \boldsymbol{e}_{l+1} + z_{1,l}/\rho$ and $\boldsymbol{g}_u = \boldsymbol{u} - z_{2,l}/\rho$ are the two auxiliary variables. Equation (32) has a closed-form solution

$$\boldsymbol{a}_{l+1} = \boldsymbol{C}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{g}_a + \boldsymbol{g}_u) \qquad (33)$$

where $\boldsymbol{C} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{I})^{-1}$ can be computed in advance and cached offline.

The implementation of WMNR is presented in Algorithm 2. The convergence properties of ADMM algorithm and the weights term have been comprehensively studied in [20] and [45]. With regard to Algorithm 2, we only need to enforce the terminal criterion for ADMM and the weights sequence as $\max\{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{a}-\boldsymbol{e}\|_2, \|\boldsymbol{a}-\boldsymbol{u}\|_2\} < \varepsilon$ and $\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|_2/\|\boldsymbol{w}_{t-1}\|_2 < \varepsilon$, respectively.

From the right columns of Algorithms 1 and 2, it is obvious that the computational complexity of WMNR is mainly determined by performing SVD and the matrix multiplications. Let t and l be the overall iterations of outer loop and inner loops, the whole computational complexity of WMNR is $O(tl(o,q\min(o,q) + n^2 + mn))$.

B. Two Special Cases

Although WMNR is focusing on the compound corruption involving both contiguous and noncontiguous noises, it has close relation to the vector- and matrix-based approaches. Two special cases can be derived with certain simplifications of WMNR. One special case is to eliminate step 6 of Algorithm 2; then, WMNR reduces to a pure vector-based method. It is worth noting that the time complexity of the resulting method is $(O(n^2m + nm))$, and it is lower than other vector-based methods, such as RRC [19] and IRGSC [20].

Another special case is to reform WMNR into a robust low-rank method by considering the contiguous noises only. In this case, we discard the tailored loss function in formula (22) and rewrite our model as

$$\begin{aligned} &\min\ \|\mathrm{Mat}(\boldsymbol{e})\|_{s,g} + \lambda\vartheta(\boldsymbol{a}) \\ &\text{s.t. } \boldsymbol{y} - \boldsymbol{X}\boldsymbol{a} = \boldsymbol{e},\ \boldsymbol{a} = \boldsymbol{u}. \end{aligned} \qquad (34)$$

Algorithm 2 WMNR Complexity

```
Input: $y, A, \mu, \lambda, \rho$, and $\varepsilon$.
Output: the optimal $a$ and $w$.
Initialize $t = 0$, $a^t = 1/n$;
Repeat
  1. $t = t + 1$;
  2. Estimate the feature weights by Eq. (13);                    $O(k)$
  Repeat
    3. Initialize $l = 0$, $a_l = a^t$, $z_{1,l} = 0$, $z_{2,l} = 0$;
    4. $l = l + 1$;
    5. estimate $e^1_l$ by Eq. (28) for contiguous errors; $O(mn)$
    6. estimate $e_l$ by using Algorithm 1 for noncontiguous
       errors;                                                   $O(oq\min(o,q))$
    7. find $u_l$ with respect to different regularization term using
       Eq. (31);
                                                                   $O(n)$
    8. Update $a_l$ by Eq. (33);                                  $O(n^2 + mn)$
    9. Update $z_{1,l}$ and $z_{2,l}$ by Eq. (25d) and Eq. (25e),
       respectively.                                               $O(m)$
  Until converge
  10. $a^t = a_l$;
Until converge
```

With the elimination of steps regarding $\boldsymbol{w}$ in Algorithm 2, problem (34) has similar optimization steps as Algorithm 1. The only required revision lies in solving (15) with a new intermediate matrix $\boldsymbol{G} = \mathrm{Mat}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}_l + z_{1,l}/\rho)$. We term this version of our model as the weighted nonconvex norm regression (WN^{2}R) with time complexity $O(l(oq\min(o,q) + n^2 + mn))$, which is comparable with the reported computational complexity of NMR [24], S$_p$L$_q$ [27], and RMR [28].

C. Identification Scheme

Borrowing the ideas from both of RRC and S$_p$L$_q$, we jointly use the weighted Frobenius and nonconvex low-rank norm as a robust metric for the practical FI problem. Given the optimal $\boldsymbol{a}$ and the dictionary samples $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, …, $\boldsymbol{X}_c$ from different subjects, the approximated image $\boldsymbol{y}'$ can be represented as $\boldsymbol{y}' = \boldsymbol{X}_1\boldsymbol{a}_1 + \boldsymbol{X}_2\boldsymbol{a}_2 +, \ldots, +\boldsymbol{X}_c\boldsymbol{a}_c$. Let Θ_i be the characteristic function that selects the coefficients affiliated to the ith class. One can get the corresponding class reconstruction error as

$$\boldsymbol{e}_i(\boldsymbol{y}) = \|\sqrt{\boldsymbol{W}}(\boldsymbol{y}' - \boldsymbol{X}(\Theta_i(\boldsymbol{a})))\|_2 + \|\mathrm{Mat}(\boldsymbol{y}' - \boldsymbol{X}(\Theta_i(\boldsymbol{a})))\|_{s,g} \qquad (35)$$

where $\boldsymbol{W}$, $\boldsymbol{a}$, and $\boldsymbol{s}$ are all the learned solution from Algorithm 2. The final decision rule is

$$\mathrm{identity}(\boldsymbol{y}) = \arg\min_i \boldsymbol{e}_i(\boldsymbol{y}). \qquad (36)$$

TABLE II

SETUP OF THE TRAVERSED PARAMETERS IN THE EXPERIMENTS

Parameters	Specification	Models
balance parameter λ	{1e-5,1e-4,1e-2,1e-1,1e0}	all
penalty parameter ρ	{1e-2,1e-1,1e0,1e1}	all
nonzero weight percent	{0.8,0.7,0.6,0.5,0.4}	RRC, IRSGC, WMNR
nonconvex parameter p	0.9	RMR, WMNR
convergence tolerance ε	1e-3	all
maximum iterations t_m	200	all

V. EXPERIMENTAL AND DISCUSSION

Five benchmark face image databases, namely, the CMU PIE face database,[1] the AR database,[2] the ExYaleB database,[3] the LFW database,[4] and the PubFig database [46], are selected to evaluate the effectiveness and robustness of our proposed methods. Several recently proposed regression-based approaches, including RRC, IRGSC, NMR, NR, and RMR, are used for comparisons. RRC and IRGSC are the vector-based methods, which preprocess each face sample as a column vector by connecting corresponding gray intensities in series. For RRC, the l_1-norm regularization is used since it performs relatively better than the l_2-norm. The other three methods are matrix-based classifiers, which directly take the gray faces as input samples. To generate the best identification rate, Table II shows the parameter setups for all competing methods. All experiments are implemented in MATLAB R2014a on a PC with 3.0-GHz CPU and 12-GB RAM.

A. FI With Pose Variations

We first employ the CMU PIE face database to validate the performance of WMNR in FI with pose variations (PVs). The whole PIE database contains 68 subjects with 41 368 face images. All the images are captured under varying poses, illuminations, and expressions. In our experiment, the five different poses (pose05, pose07, pose09, pose27, and pose29) under different illuminations and expressions are used. The images are all manually aligned and cropped to be 64 × 64 with 256 gray levels. In this subset, the 3329 near frontal images from pose27 are used as the training set, and the remaining samples are used for testing. Some face images of one subject with different illuminations and expressions are shown in Fig. 5. Table III lists the identification accuracy of seven competing methods under different PVs. It can be observed that the vector-based methods achieve higher accuracy than the matrix-based methods in general. Our WMNR algorithm outperforms most of the other compared algorithms except for the pose07 subset, in which RRC achieves the best accuracy. Moreover, our WN^{2}R algorithm achieves the best identification rate among all the matrix-based methods.

B. FI With Random Corruptions

In this section, we design three experiments to investigate the robustness of our proposed methods in dealing with different levels of contiguous and noncontiguous noises, namely, the pixel corruption, the square block corruption, and the nonsquare block corruption. The ExYaleB database, which contains about 2414 frontal face images of 38 subjects, is adopted here for evaluation. We randomly select 30 images per subject to form the dictionary, and the rest samples are used as the query set. All samples are equally normalized and cropped to be with the size 64 × 50.

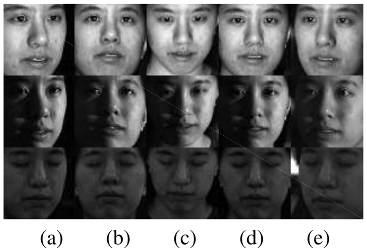


Fig. 5. Some samples in PIE database. (a) pose05. (b) pose07. (c) pose09. (d) pose27. (e) pose29.

TABLE III

PERFORMANCE COMPARISONS (%) UNDER PVS ON PIE DATABASE (BOLD AND ITALIC FONTS DENOTE THE BEST ACCURACY AND OUR PROPOSED METHODS, RESPECTIVELY)

Methods	Different pose variations			
	pose05	pose07	pose09	pose29
RRC	73.95	**88.26**	95.12	67.68
IRGSC	73.95	85.05	93.68	67.68
NMR	72.75	78.53	82.87	63.30
NR	72.16	78.22	84.10	63.30
RMR	73.23	82.52	87.46	65.75
WN^{2}R	74.85	84.05	87.77	66.97
WMNR	**76.65**	87.73	**96.54**	**73.20**

1) FI With Pixel Corruption: In this experiment, certain percentages of pixels for each test image are randomly replaced by the uniformly distributed noises. Since all the competing algorithms have been proved to be with desirable performance under mild pixel corruption, we increase the corruption ratio from 50% to 80%. The experimental results for all the compared methods are listed in Table IV. The first observation is that the matrix-based methods, i.e., NMR, NR, RMR, and WN^{2}R, all perform poorly in this scenario. Among them, our WN^{2}R achieves the highest identification rate but still lags at least 24%, 32%, 33%, and 36%, behind other vector-based methods under 50%, 60%, 70%, and 80% pixel corruptions, respectively. The reason is that the noncontiguous noise lacks of structural characteristics, which obstructs the matrix-based methods to distinguish the noises from essential features. For vector-based methods, IRGSC outperforms RRC under all levels of pixel corruption, which demonstrate the superiority of the underlying group constraint in the scenario of the pose-fixed variations. Comparatively, WMNR lags 0.05% and 0.95% behind RRC and IRGSC, respectively, under 50% pixel corruption but outperforms both RRC and IRGSC when the corruption level reaches more than 50%. Specifically, the average improvement of WMNR over IRGSC is accurately 4.06%.

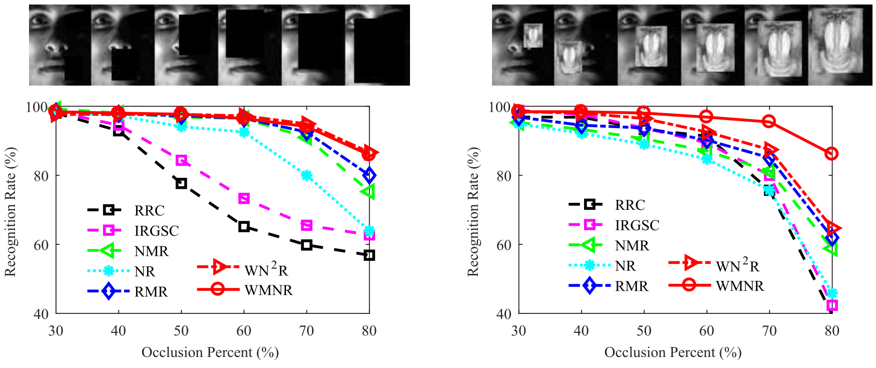
2) FI With Square Block Corruption: In this experiment, each query sample is corrupted by a randomly located square block of a pure black or baboon image with varying occlusion levels. The experimental results under compared methods are

[1] https://www.ri.cmu.edu/publications/the-cmu-pose-illumination-and-expression-pie-database-of-human-faces

[2] http://www2.ece.ohio-state.edu/∼aleix/ARdatabase.html

[3] http://vision.ucsd.edu/∼leekc/ExtYaleDatabase/ExtYaleB.html

[4] http://vis-www.cs.umass.edu/lfw



(a) (b)

Fig. 6. Identification rates (%) of RRC, IRGSC, NMR, NR, RMR, WN^{2}R, and WMNR, under different levels of (a) black block and (b) baboon block occlusions.

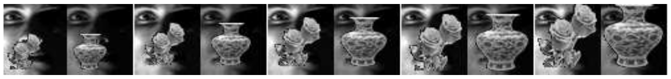
TABLE IV

Performance Comparisons (%) Under Different Levels of Pixel Corruption on ExYaleB Database (Bold and Italic Fonts Denote the Best Accuracy and Our Proposed Methods, Respectively)

Methods	Levels of pixel corruption			
	50%	60%	70%	80%
RRC	96.52	91.00	82.14	60.39
IRGSC	**97.42**	95.70	85.35	64.29
NMR	67.84	52.94	31.37	15.29
NR	68.24	53.73	29.80	15.29
RMR	67.06	50.98	26.67	12.55
WN^{2}R	72.16	58.82	48.78	23.29
WMNR	96.47	**95.90**	**89.80**	**76.84**

summarized in Fig. 6. The images on top of Fig. 6 illustrate the occlusion levels varying from 30% to 80% percentages. From Fig. 6, we can see that RRC and IRGSC underperform NMR, NR, RMR, and WN^{2}R in these two scenarios. It is reasonable from theoretical analysis that the low-rank constraint excels at mining the structural information. In Fig. 6(a), WN^{2}R ranks first under the black block occlusion due to the clear low-rank characteristics. From the results in Fig. 6(b), we find out that all methods, except WMNR, perform poorer when the baboon is used. We attribute this to the fact that the baboon object exhibits more similar features as the face than the pure black block does. Thus, it is much more challenging for the compared methods to distinguish the inliers from outliers.

3) FI With Nonsquare Block Corruption: In the third experiment, we consider two different nonsquare objects to occlude the query images, as shown in Fig. 7. Similar to the previous experiment, block occlusion is evaluated by placing the nonsquare images (rose and vase) on each test image. The location of the occlusion is randomly positioned and is unknown during testing. We consider different percentages that the images being covered by the occluded object from 50% to 90%. Average identification rates of ten runs for the different levels of rose and vase occlusions (VOs) are shown in Tables V and VI, respectively.



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

Fig. 7. Two images, rose and vase, used for (a) 50%, (b) 60%, (c) 70%, (d) 80%, and (e) 90% nonsquare block occlusion.

TABLE V

Performance Comparisons (%) Under Different Levels of RO on ExYaleB Database (Bold and Italic Fonts Denote the Best Accuracy and Our Proposed Methods, Respectively)

Methods	Levels of rose corruption				
	50%	60%	70%	80%	90%
RRC	**98.82**	98.03	94.12	69.80	16.47
IRGSC	98.43	97.25	92.16	64.31	11.76
NMR	93.73	91.76	90.20	87.06	79.61
NR	95.29	92.94	90.98	85.88	77.65
RMR	96.86	93.33	92.55	88.63	82.35
WN^{2}R	96.86	94.12	92.94	88.63	82.35
WMNR	98.65	**98.12**	**96.47**	**94.50**	**90.41**

TABLE VI

Performance Comparisons (%) Under Different Levels of VO on ExYaleB Database (Bold and Italic Fonts Denote the Best Accuracy and Our Proposed Methods, Respectively)

Methods	Levels of vase corruption				
	50%	60%	70%	80%	90%
RRC	97.65	**97.65**	93.33	79.61	19.61
IRGSC	97.25	95.69	89.80	76.08	17.25
NMR	90.59	85.88	82.75	76.47	70.98
NR	91.76	89.41	85.10	78.43	63.92
RMR	93.33	91.76	89.02	86.67	80.00
WN^{2}R	94.12	91.76	89.02	86.67	82.35
WMNR	**98.00**	**97.65**	**96.47**	**92.94**	**89.88**

From Fig. 7, it is clear that the actual coverage area for both cases with nonsquare images is smaller than that in the square block scenarios. Consequently, the identification rates of all the competing methods are relatively higher than those of square images under the same occlusion percentage. Furthermore, from Tables V and VI, we can see that all methods perform better when the rose image occurs. We attribute this to the fact that the vase object exhibits more textures than the rose does and confuses with the real facial features, which makes FI more difficult. RRC and IRGSC achieve desirable accuracy when the occlusion percentage is lower than 70%. However, their identification rates drop sharply when the coverage percent is up to 80%. It can be seen from Table V that the accuracy of RRC is 98.82% and 16.47% when the occlusion rates are 50% and 90%, respectively. This is because a larger coverage rate leads to more structural noises, while the vector-

TABLE VII

PERFORMANCE COMPARISONS (%) ON AR DATABASE WITH TWO DIFFERENT DISGUISES (BOLD AND ITALIC FONTS DENOTE THE BEST ACCURACY AND OUR PROPOSED METHODS, RESPECTIVELY)

Methods	Sunglass		Scarf	
	Session1	Session2	Session1	Session2
RRC	99.00	89.33	93.33	76.33
IRGSC	99.00	90.00	**95.33**	80.33
NMR	74.00	40.00	72.67	45.33
NR	73.67	37.33	71.33	44.33
RMR	96.76	68.36	88.67	65.00
WN^{2}R	98.00	70.67	89.67	65.33
WMNR	**99.33**	**91.00**	**95.33**	**82.67**

based methods are not adept at this type of corruption. Comparatively, the performances of NMR, NR, RMR, and WN^{2}R are good when the occlusion level becomes higher, but these matrix-based methods show no advantages when the occlusion level is relative low. Our WMNR occupies the best identification rate in most experimental results, as shown in Tables V and VI, and the gap gets larger as the occlusion rate gets higher. This demonstrates that WMNR is more robust than the others for FI under various types of contiguous occlusions.

C. FI With Real Disguise

In this section, we evaluate the robustness of our methods with real disguise in two scenarios: 1) faces with sunglasses and 2) faces with scarves. In addition, we would like to test the proposed methods in cases, where few training samples are available per subject and the query samples are with variations of illumination and longer data acquisition interval. Thus, 400 neutral images with nonoccluded frontal views in session 1 from the AR database are used as the dictionary, while the disguised images from sessions 1 and 2 are used for testing. Table VII lists the results by competing methods under the image resolution of 42 × 30. Interestingly, the two famous matrix-based methods, NMR and NR, which claim to be robust to continuous noise, perform poorly in this experiment due to the limited training samples and severe illumination changes. RMR and WN^{2}R achieve better identification rate, which demonstrate the superiority by unequally treatment of singular values and nonconvex constraint. RRC and IRGSC outperform RMR and WN^{2}R, benefiting from their reweighting mechanism. WMNR takes all the preferable properties mentioned earlier into consideration and obtains the best identification rate.

D. FI With Mixed Corruption

In this experiment, we test the performance of our methods for the case of mixture noise. In this case, both pixel corruption and block occlusion degrade the query images from the ExYaleB and PIE databases. The basic experimental settings are similar as Sections V-A and V-B. Each test image from ExYaleB database is corrupted by noises following the uniform distribution, and the percentages of those randomly chosen noise pixels are from 10% to 60%. Then, we place the occlusion image on each corrupted test image. With the PIE data set, experiments are conducted with 40% pixel corruption and VO. Some example query images with this degradation from ExYaleB and PIE are shown in Figs. 8 and 9, respectively.

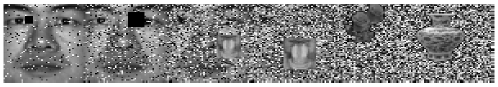


Fig. 8. Some samples with mixture noise (from 10% to 60%) on ExYaleB database.

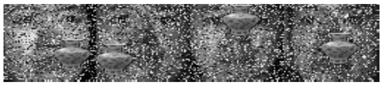


Fig. 9. Some samples with 40% mixture noise and PVs on PIE database.

Fig. 10 and Table VIII list the experimental results of all competing methods on ExYaleB and PIE database, respectively. It can be seen that WN^{2}R consistently achieves better accuracy compared to RMR, NMR, and NR, which demonstrates that the joint reweighting and nonconvex constraint generally results in a closer approximation of the intrinsic rank function than just using one of these two tricks. Nevertheless, all the matrix-based methods clearly perform poorer than the vector-based ones. This is because, the advantages of NMR, NR, RMR, and WN^{2}R are structure mining and low-rank approximation, which indicates that the matrix-based methods are more effective when dealing with the contiguous noises. However, it can be seen from Figs. 8 and 9 that the noncontiguous noise is dominant in the query images. On the contrary, the vector-based methods performs better with respect to the noncontiguous noises, and among those methods, our WMNR effectively combines the merits of matrix operation and feature learning, which yields better identification rate compared to the IRGSC and RRC.

E. FI With Uncontrolled Setting

The face images used in the aforementioned experiments are all captured in a controlled environment. In this section, we further test our methods in two uncontrolled databases: the LFW database and the PubFig database. LFW contains the images of 5749 different subjects. We gather the subjects that contain more than ten samples and then get a data set with 158 subjects from LFW-a, a revised version of LFW [47]. For each subject, five samples are randomly selected for training and another five samples for testing. The images are all cropped and resized to 32 × 32. On the PubFig data set, we follow the same experiment setting as in [20] and [28]; 20 images for each individual are randomly selected, and in total, 100 subjects are chosen for our experiments, each image is resized to 64 × 64 pixels. Ten images for each subject are used as dictionary images, and the rest are used as test set.

Table IX exhibits the identification results of all competing methods on these two databases. Our first observation is that the accuracy of all these regression-based methods is not comparable with the human-level performance. However, WMNR still ranks first in these challenging settings. Although IRGSC obtains identical identification rate in LFW, it lags behind our WMNR by 3.1% in PubFig. Moreover, WN^{2}R occupies the first place among the matrix-based methods, which again verifies the superiority of our proposed reweighting scheme.

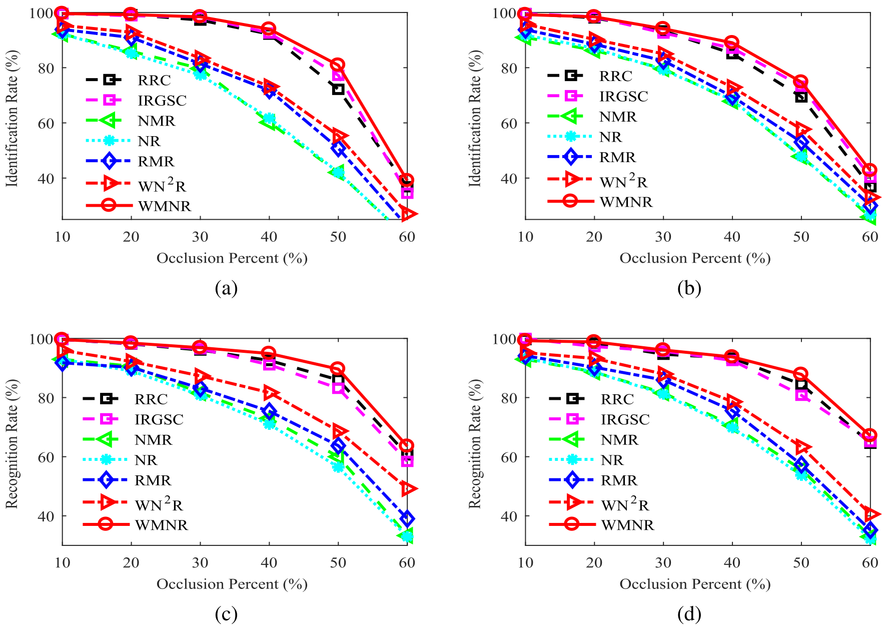


Fig. 10. Identification rates (%) under different levels of mixture corruption on ExYaleB database. (a) Black block. (b) Baboon. (c) Rose. (d) Vase.

TABLE VIII

PERFORMANCE COMPARISONS (%) UNDER 40% PERCENTAGE OF MC ON PIE DATABASE (BOLD AND ITALIC FONTS DENOTE THE BEST ACCURACY AND OUR PROPOSED METHODS, RESPECTIVELY)

Methods	Different pose variations			
	pose05	pose07	pose09	pose29
RRC	61.38	**82.51**	**85.37**	55.96
IRGSC	63.47	82.20	84.10	55.35
NMR	39.22	35.89	40.67	23.24
NR	36.83	34.66	38.53	22.32
RMR	42.43	36.36	41.31	25.24
WN^{2}R	43.43	37.67	42.68	28.90
WMNR	**65.87**	81.98	84.93	**56.71**

TABLE IX

PERFORMANCE COMPARISONS (%) ON THE LFW AND PUBFIG DATABASES (BOLD AND ITALIC FONTS DENOTE THE BEST ACCURACY AND OUR PROPOSED METHODS, RESPECTIVELY)

Databases	RRC	IRGSC	NMR	NR	RMR	*WN^{2}R*	*WMNR*
LFW	53.26	**54.43**	44.05	46.96	44.18	49.62	**54.43**
PubFig	43.23	43.00	44.50	43.80	44.60	44.60	**46.10**

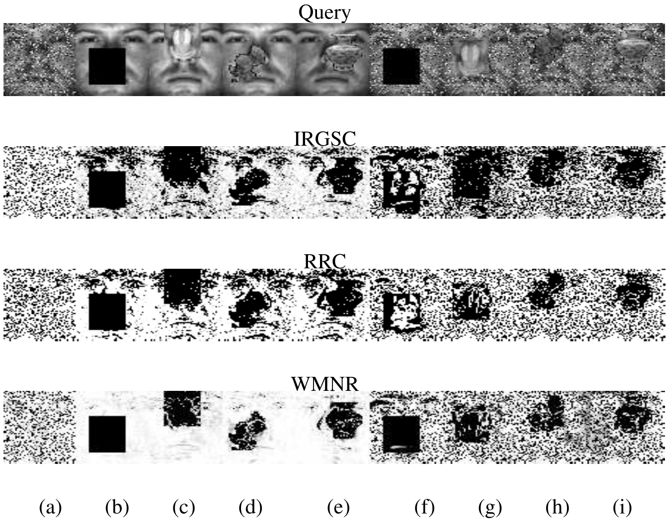


Fig. 11. Estimated weight maps under (a) pixel corruption, block occlusion with (b) black block, (c) baboon, (d) rose, or (e) vase, and pixel corruption mixed with (f) black, (g) baboon, (h) rose, or (i) VO.

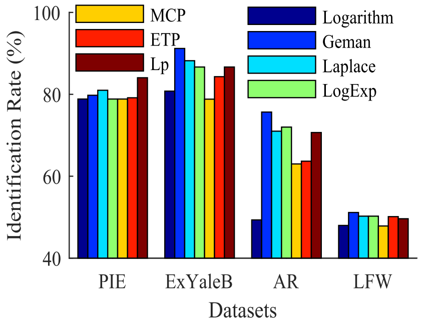


Fig. 12. Recognition rates with different nonconvex functions on a different database.

F. Behaviors of Feature Weights and Nonconvex Constraint

We discuss the impact of feature weights and nonconvex constraint on the identification performance in this section. First of all, it is necessary to verify whether the learned feature weights distinguish the inliers and outliers as we expected. Fig. 11 shows the behavior of feature weights for IRGSC, RRC, and WMNR under different types of corruption. The first row lists the query samples. The second to the fourth rows are the estimated weight maps for IRGSC, RRC, and WMNR, respectively, where black values (near to zero) represent detected outliers by the competing methods. We observe that WMNR works better than IRGSC and RRC. For block occlusion, WMNR detects the outlier objects more accurately. Most of the black regions in the weight maps are concentrated on the occluded area. IRGSC and RRC also detect the right occlusion area, but they identify a number of inlier pixels as outliers. Similarly for mixture corruption, IRGSC and RRC assign too many small weights to the unoccluded area.

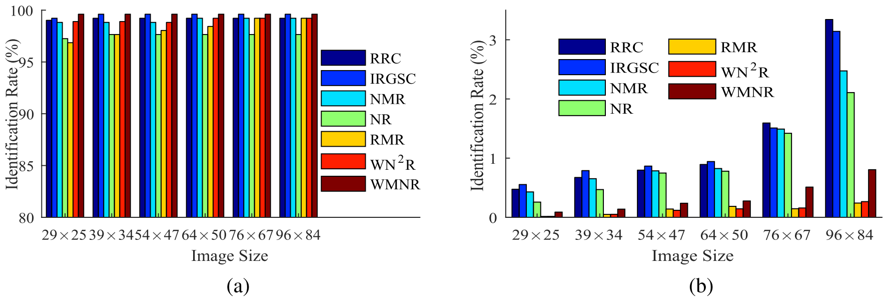


Fig. 13. Performance of all the competing methods under different image sizes on the ExYaleB database. (a) Identification rate. (b) Average running time per sample.

TABLE X

RUNTIME (IN SECONDS) OF COMPETING METHODS ON DIFFERENT DATABASES AND EXPERIMENTAL SCENARIOS

Methods	Average runtime per sample							
	PIE		ExYaleB			AR	LFW	PubFig
	(PV)	(MC)	(RC)	(RO)	(VO)			
RRC	55.45	78.68	6.18	8.67	8.37	1.63	3.56	6.33
IRGSC	24.18	33.04	3.29	5.45	4.95	1.46	2.34	7.94
NMR	2.15	5.98	0.43	0.28	0.31	0.28	0.49	0.61
NR	2.77	5.83	0.65	0.25	0.27	0.27	0.15	0.63
RMR	3.36	0.46	0.27	0.16	0.21	0.17	0.06	0.27
WN^{2}R	2.94	0.45	0.19	0.14	0.17	0.16	0.05	0.26
WMNR	11.54	8.63	0.71	1.96	1.90	0.33	0.77	0.84

The reason for this is that there is no spatial correlation constraint among the weights within IRGSC and RRC.

As discussed in Section II-B, we mainly focus on the nonconvex functions listed in Table I. For highlighting the importance of different nonconvex functions to our WN^{2}R methods, in this section, we conduct four experiments using the images from PIE pose07, ExYaleB with 80% black block occlusion, AR sunglass in session 2, and uncontrolled LFW data set, respectively. Fig. 12 lists the experimental results with different nonconvex functions on these four experiments. From Fig. 12, one can observe that different nonconvex constraints lead to comparable, but not so close, identification rates. The l_p nonconvex function obtains the best results in PIE database. However, the Geman constraint ranks first in other three experiments. These results prove that we can further improve our proposed methods by fine-tuning the surrogate function. However, the theoretical explanation of which constraint may lead to the best performance is still under investigation.

G. Comparison Analysis of Runtime

Apart from accuracy, computational cost is another important performance indicator for the proposed classifiers. In this section, we compare our methods, i.e., WN^{2}R and WMNR, with other competing ones under different applications. Table X lists the average running time of identifying one query sample on five databases and several practical scenarios, including the PIE database with PV and mixed corruption (MC), the ExYaleB database with random corruption, rose occlusion (RO), VO, as well as the AR, LFW, and PubFig databases. All the experimental settings follow those given in Sections V-A–V-E.

From Table X, our first observation is that all the matrix-based methods clearly achieve better efficiency than the vector-based methods. Combining with the accuracy results listed in the previous experiments, our WN^{2}R algorithm not only achieves the best identification rate among all the matrix-based methods but also occupies the first place in terms of computational efficiency. Although the efficiency of RMR is very close to WN^{2}R, it is behind our method clearly in terms of accuracy. Recall that the accuracy of RRC and IRGSC outperforms the results generated by the matrix-based methods in most tests expect for the block occlusion scenario. One can say that the vector-based methods sacrifice more runtime for better effectiveness. With regard to this viewpoint, our WMNR method takes the merits from both the vector- and matrix-based methods and achieves an appealing tradeoff between the runtime and accuracy. We can see that its accuracy ranks first in most experiments and also consumes much less runtime than RRC and IRGSC.

To further evaluate the efficiency of WN^{2}R and WMNR, we test all the methods on the ExYaleB database with different image sizes and without any manual corruption. Fig. 13 illustrates the experimental results under image resolutions 29 × 25, 39 × 34, 54 × 47, 64 × 50, 76 × 67, and 96 × 84. From this Fig. 13, we can see that all the competing methods achieve desirable and similar accuracy in the clear environment. However, NR and RMR still underperform other methods due to their coarse approximation to the essential rank function. In terms of efficiency, all the compared methods cost more runtime along with the increasing of the image sizes. Specifically, WMNR is much more efficient than the pure vector-based methods and slightly slower than the pure matrix-based methods. Moreover, WN^{2}R and RMR share the first place among all the matrix-based methods.

VI. CONCLUSION

In this paper, we present a general formulation to deal with the mixture image corruption, i.e., the concurrence of noncontiguous and contiguous noises. A novel mixed-norm constrained regression model, named WMNR, is proposed, which provides two merits. The first tackles the noncontiguous noises with uncertain distributions into a feature weighted loss function. The second characterizes the contiguous residual image as a low-rank problem under the constraints of general the nonconvex functions and ordered rank weights. A computationally efficient optimization scheme with respect to the WMNR model is derived. In such scheme, by proper problem decomposition, certain analytical solutions to some of the resulting subproblems can be obtained, and a new reweighted ADMM algorithm is provided to handle the rest in a parallel computation manner. Moreover, two special cases of WMNR are naturally deduced to manage the scenarios, where single kind of noise occurs. One is designed for noncontiguous noises only. It has the similar cost function as two well-known vector-based approaches, i.e., RRC and IRGSC, but is with much less computational cost from the theoretical point of view. The other one, named WN^{2}R, is designed especially for contiguous noises and can be considered as an extension to the existing matrix-based methods. Extensive experimental results verify that the presented methods are more robust to PV, pixel corruption, block occlusion, real disguise, and mixture noises compared to the state-of-the-art regression-based approaches for FI problem. We also empirically show that WMNR achieves well balance of the performance and the computational complexity, i.e., WMNR obtains much higher accuracy and compatible efficiency with respect to the pure matrix-based methods, whereas it requires much less runtime to achieve better or similar performance with respect to the pure vector-based methods.

There are still some issues that deserve our further investigation. On one hand, although a direct application of CNN or its variants into small data problem may not ensure favorable results, we can borrow the convolutional idea to replace our linear representation operation for richer feature information [47]. On the other hand, the full SVD step in WN^{2}M takes $O(oq\min(o, q))$ time complexity at each iteration, which is expensive on large matrices. An online or partial SVD computation will greatly enrich our model's applicability to high-resolution image.

Appendix A
Notations

Unless specified otherwise, throughout this paper, the capital bold and lowercase bold symbols are used to represent matrices and vectors, respectively. For any matrix $\boldsymbol{B}$, $\boldsymbol{b}_i$ is the ith column of $\boldsymbol{B}$ and b_{ij} is the jth element in $\boldsymbol{b}_i$. $\boldsymbol{B}^T$ and tr($\boldsymbol{B}$) denote the transpose and trace of the matrix $\boldsymbol{B}$, respectively. $\boldsymbol{1}$ denotes the vector with all entries being 1. Some notations and abbreviations frequently used throughout this paper are given in the Nomenclature section.

Appendix B
Proof of Lemma 1

Assume $\boldsymbol{E} = \boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{B}^{\mathrm{T}}$ is the optimal solution of (15), whose singular values arrange in the descending order as given by $\boldsymbol{\Delta}$. According to the trace inequality of John von Neumann [48] $\mathrm{tr}(\boldsymbol{EG}) \le \sum_{i=1}^{\nu} \delta_i\sigma_i$, we have

$$\begin{aligned}\|\boldsymbol{E}-\boldsymbol{G}\|_F^2 &= \mathrm{tr}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}) + \mathrm{tr}(\boldsymbol{\Delta}^T\boldsymbol{\Delta}) - 2\mathrm{tr}(\boldsymbol{E}^{\mathrm{T}}\boldsymbol{G})\\ &\ge \mathrm{tr}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}) + \mathrm{tr}(\boldsymbol{\Delta}^T\boldsymbol{\Delta} - 2\mathrm{tr}(\boldsymbol{\Sigma}^T\boldsymbol{\Delta}) = \|\boldsymbol{\Sigma}-\boldsymbol{\Delta}\|_F^2.\end{aligned} \tag{37}$$

This leads to

$$\frac{1}{2}\|\boldsymbol{E}-\boldsymbol{G}\|_F^2 + \mathrm{tr}(\boldsymbol{S}g(\boldsymbol{\Delta})) \ge \frac{1}{2}\|\boldsymbol{\Sigma}-\boldsymbol{\Delta}\|_F^2 + \mathrm{tr}(\boldsymbol{S}g(\boldsymbol{\Delta})). \tag{38}$$

Note that when $\boldsymbol{B} = \boldsymbol{V}$ and $\boldsymbol{Q} = \boldsymbol{U}$, the equality of (37) holds according to the von Neumann theorem. Thus minimizing (15) can be reduced to problem (16). □

Appendix C
Proof of Theorem 1

From (20), we can see that τ_i is a monotonically increasing function with regard to s_i. Given a fixed σ and any two weights with $s_i \le s_j$, we describe our analysis in three different situations.

When $\sigma \le \tau(s_i)$ and $\sigma \le \tau(s_j)$, we have $\delta_i(s_i) = \delta_j(s_j) = 0$; when $\sigma > \tau(s_i)$ and $\sigma \le \tau(s_j)$, we have $\delta_j(s_j) = 0$ and $\delta_i(s_i) > 0$. Furthermore, when $\sigma > \tau(s_i)$ and $\sigma > \tau(s_j)$, we refer to Algorithm 1 for revealing the sequence of $\delta_j(s_j)$ and $\delta_i(s_i)$. Initially, we have $\delta_i(s_i) = \delta_j(s_j) = |\sigma|$, and then, they are iteratively updated by $\delta^k = |\sigma| - sg'(\delta^{k-1})$. Since g is a concave nondecreasing function and $s_i \le s_j$, we can obtain that $\delta_i^{t_m}(s_i) \ge \delta_j^{t_m}(s_j)$. In conclusion, considering $\delta(\sigma, s)$ as an implicit function with respect to σ and s, we have $\delta_i(s_i) \ge \delta_j(s_j)$, and $s_i \le s_j$ holds for a fixed σ. On the other hand, for a fixed $\boldsymbol{s}$, it has been proven in [49] that $\delta_i(\sigma_i) \ge \delta_j(\sigma_j)$, $\sigma_i \ge \sigma_j$. These two inequalities lead to $\delta_i(\sigma_i,\, s_i) \ge \delta_j(\sigma_j, s_j)$ for $\sigma_i \ge \sigma_j$, $s_i \le s_j$, and $i \le j$. □

References

[1] J. Tang *et al.*, "Discriminative deep quantization hashing for face image retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6154–6162, Dec. 2018.

[2] L. Lin *et al.*, "Cross-domain visual matching via generalized similarity measure and feature learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1089–1102, Jun. 2017.

[3] J. Zheng *et al.*, "Kernel group sparse representation classifier via structural and non-convex constraints," *Neurocomputing*, vol. 296, pp. 1–11, Jun. 2018.

[4] M. Liadis *et al.*, "Robust and low-rank representation for fast face identification with occlusions," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2203–2218, May 2017.

[5] H. Li *et al.*, "Multimodal 2D+3D facial expression recognition with deep fusion convolutional neural network," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2816–2831, Dec. 2017.

[6] H. Wang *et al.*, "Semantic discriminative metric learning for image similarity measurement," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1579–1589, Aug. 2016.

[7] Y. Xu *et al.*, "A new discriminative sparse representation method for robust face recognition via l_2 regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2233–2242, Oct. 2017.

[8] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106–2112, Nov. 2010.

[9] Z. Zhang *et al.*, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, no. 1, pp. 490–530, May 2015.

[10] J. Wright *et al.*, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[11] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 471–478.

[12] J. Huang *et al.*, "Supervised and projected sparse coding for image classification," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 438–444.

[13] 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.

[14] X.-X. Li *et al.*, "Structured sparse error coding for face recognition with occlusion," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1889–1900, May 2013.

[15] J. Chen *et al.*, "Matrix variate distribution-induced sparse representation for robust image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2291–2300, Oct. 2015.

[16] X. Jiang and J. Lai, "Sparse and dense hybrid representation via dictionary decomposition for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1067–1079, May 2015.

[17] I. Naseem, R. Togneri, and M. Bennamoun, "Robust regression for face recognition," *Pattern Recognit.*, vol. 45, no. 1, pp. 104–118, Jan. 2012.

[18] L. Zhang *et al.* (2012). "Collaborative representation based classification for face recognition." [Online]. Available: https://arxiv.org/abs/1204.2358

[19] M. Yang *et al.*, "Regularized robust coding for face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1753–1766, May 2013.

[20] J. Zheng *et al.*, "Iterative re-constrained group sparse face recognition with adaptive weights learning," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2408–2423, May 2017.

[21] Z. Zhang, "Parameter estimation techniques: A tutorial with application to conic fitting," *Image Vis. Comput.*, vol. 15, no. 1, pp. 59–76, 1997.

[22] R. He *et al.*, "Half-quadratic based iterative minimization for robust sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 261–275, Feb. 2014.

[23] J. Tang *et al.*, "Generalized deep transfer networks for knowledge propagation in heterogeneous domains," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 12, no. 4s, p. 68, Nov. 2016.

[24] 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.

[25] Y. Xie *et al.*, "Weighted schatten p-norm minimization for image denoising and background subtraction," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4842–4857, Oct. 2016.

[26] S. Gu *et al.*, "Weighted nuclear norm minimization with application to image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2862–2869.

[27] L. Luo *et al.*, "Robust image regression based on the extended matrix variate power exponential distribution of dependent noise," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 9, pp. 2168–2182, Sep. 2017.

[28] J. Xie *et al.*, "Robust nuclear norm-based matrix regression with applications to robust face recognition," *IEEE Trans. Image Process*, vol. 26, no. 5, pp. 2286–2295, May 2017.

[29] J. Qian *et al.*, "Robust nuclear norm regularized regression for face recognition with occlusion," *Pattern Recognit.*, vol. 48, no. 10, pp. 3145–3159, Oct. 2015.

[30] L. Luo *et al.*, "Nuclear-L_1 norm joint regression for face reconstruction and recognition with mixed noise," *Pattern Recognit.*, vol. 48, no. 12, pp. 3811–3824, Dec. 2015.

[31] M. Yang, L. Zhang, and D. Zhang, "Efficient misalignment-robust representation for real-time face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 850–863.

[32] U. Srinivas *et al.*, "Structured sparse priors for image classification," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1763–1776, Jun. 2015.

[33] C. Lu, Z. Lin, and S. Yan, "Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization," *IEEE Trans. Image Process.*, vol. 24, no. 2, pp. 646–654, Feb. 2015.

[34] J. Jiang *et al.*, "Noise robust face hallucination via locality-constrained representation," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1268–1281, Aug. 2014.

[35] W. Liu *et al.*, "KCRC-LCD: Discriminative kernel collaborative representation with locality constrained dictionary for visual categorization," *Pattern Recognit.*, vol. 48, no. 10, pp. 3076–3092, Oct. 2015.

[36] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig, "Strong sub- and super-Gaussianity," in *Proc. Int. Conf. Latent Variable Anal. Signal Separat.*, 2010, pp. 303–310.

[37] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1997.

[38] J. H. Friedman, "Fast sparse regression and classification," *Int. J. Forecasting*, vol. 28, no. 3, pp. 722–738, Sep. 2012.

[39] Z.-F. Jin *et al.*, "An alternating direction method with continuation for nonconvex low rank minimization," *J. Sci. Comput.*, vol. 66, no. 2, pp. 849–869, Feb. 2016.

[40] C. Gao *et al.*, "A feasible nonconvex relaxation approach to feature selection," in *Proc. 25th AAAI Conf. Artif. Intell.*, 2011, pp. 356–361.

[41] C. Lu *et al.*, "Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 829–839, Feb. 2016.

[42] J. Trzasko and A. Manduca, "Highly undersampled magnetic resonance image reconstruction via homotopic ℓ_0-minimization," *IEEE Trans. Med. Imag.*, vol. 28, no. 1, pp. 106–121, Jan. 2009.

[43] L. B. Montefusco, D. Lazzaro, and S. Papi, "A fast algorithm for nonconvex approaches to sparse recovery problems," *Signal Process.*, vol. 93, no. 9, pp. 2636–2647, Sep. 2013.

[44] W. Zuo *et al.*, "A generalized iterated shrinkage algorithm for nonconvex sparse coding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 217–224.

[45] S. Boyd *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[46] N. Kumar *et al.*, "Attribute and simile classifiers for face verification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 365–372.

[47] H. Chang *et al.*, "Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1182–1194, May 2018.

[48] M. Nikolova, "Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares," *SIAM Multiscale Modeling Simulation*, vol. 4, no. 3, pp. 960–991, 2005.

[49] S. Chrétien and T. Wei, "Von Neumann's trace inequality for tensors," *Linear Algebra Appl.*, vol. 482, pp. 149–157, Oct. 2015.

Jianwei Zheng received the M.S. degree in electrical and information engineering and the Ph.D. degree in control theory and control engineering from the Zhejiang University of Technology, Hangzhou, China, in 2005 and 2010, respectively.

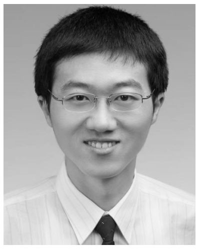
Since 2010, he has been with the College of Computer Science, Zhejiang University of Technology. He is currently a Visiting Fellow with the National Center for Computer Animation, Bournemouth University, Poole, U.K. His current research interests include sparse coding, low-rank decomposition, and nonconvex optimization.

Kechen Lou received the M.S. degree from the School of Computer Science and Engineering, Zhejiang University of Technology, Hangzhou, China, in 2019, where he is currently pursuing the Ph.D. degree.

His current research interests include image and video enhancement, pattern recognition, and machine learning.

Xi Yang received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2004, the M.S. degree from Zhejiang University, Hangzhou, China, in 2007, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2011.

Since 2012, he has been with the College of Computer Science and Engineering, Zhejiang University of Technology, Hangzhou. His current research interests include nonlinear control and optimization with applications in nonlinear output regulation, multiagent systems, intelligent transportation systems, and image processing.



Cong Bai received the B.E. degree from Shandong University, Jinan, China, in 2003, the M.E. degree from Shanghai University, Shanghai, China, in 2009, and the Ph.D. degree from the National Institute of Applied Sciences of Rennes, Rennes, France, in 2013.

From 2003 to 2006, he was with the School of Information Science and Engineering, Shandong Agricultural University, Tai'an, China. Since 2013, he has been the Faculty of the College of Computer Science, Zhejiang University of Technology, Hangzhou, China. His current research interests include computer vision and multimedia retrieval.

Jinhui Tang (M'08–SM'14) received the B.Eng. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively.

From 2008 to 2010, he was a Research Fellow with the School of Computing, National University of Singapore, Singapore. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. He has authored over 150 papers in top-tier journals and conferences. His current research interests include multimedia analysis and search, computer vision, and machine learning.