Enhanced Group Sparse Regularized Nonconvex Regression for Face Recognition

Chao Zhang, Huaxiong Li, Chunlin Chen, Yuhua Qian and Xianzhong Zhou

***Abstract*—Regression analysis based methods have shown strong robustness and achieved great success in face recognition. In these methods, convex l_1-norm and nuclear norm are usually utilized to approximate the l_0-norm and rank function. However, such convex relaxations may introduce a bias and lead to a suboptimal solution. In this paper, we propose a novel Enhanced Group Sparse regularized Nonconvex Regression (EGSNR) method for robust face recognition. An upper bounded nonconvex function is introduced to replace l_1-norm for sparsity, which alleviates the bias problem and adverse effects caused by outliers. To capture the characteristics of complex errors, we propose a mixed model by combining γ-norm and matrix γ-norm induced from the nonconvex function. Furthermore, an $l_{2,\gamma}$-norm based regularizer is designed to directly seek the interclass sparsity or group sparsity instead of traditional $l_{2,1}$-norm. The locality of data, i.e., the distance between the query sample and multi-subspaces, is also taken into consideration. This enhanced group sparse regularizer enables EGSNR to learn more discriminative representation coefficients. Comprehensive experiments on several popular face datasets demonstrate that the proposed EGSNR outperforms the state-of-the-art regression based methods for robust face recognition.**

***Index Terms*—Low-rank structure, sparse representation, enhanced group sparsity, nonconvex relaxation, face recognition.**

I. Intorduction

As one of the most intensively investigated topics, face recognition (FR) has attracted much attention from the field of pattern recognition and computer vision. Numerous successful methods have been proposed and developed, including traditional [1], [2], [3], [4], [5], [6] and deep learning methods [7], [8], [9], [10]. However, sophisticated variations in face images (e.g., occlusion, illumination and expression) pose a big challenge for FR systems. Many researchers tried to develop more robust FR techniques against various noises.

Recently, regression analysis based approaches captured broad attention in the computer vision communities, which achieved great success in FR [11], [12], image analysis [13], [14], visual tracking [15], etc. Wright et al. presented a Sparse Representation Classifier (SRC) which sought a sparse representation in linear regression [12]. The l_1-norm is used to guarantee the sparsity which is the convex relaxation of intractable l_0-norm. SRC achieves some impressive performance on face recognition against pixel corruptions and occlusions [12]. Zhang et al. replaced the l_1-norm in SRC by l_2-norm and proposed a Collaborative Representation Classifier (CRC) [16]. Both the two methods are unsupervised in representation and ignore the label information of training data. In [17], the authors proposed a supervised Group Sparse Coding (GSC) method based on $l_{2,1}$-norm regularizer. GSC considers the correlations among training samples and forces the representation coefficients to be sparse at group level. Such group sparse regularizer is used in many other studies [4], [1]. In addition to the label information, sample weights learning and feature weights learning mechanisms are incorporated into regression models to improve the discrimination of representation. The former differs the roles of all training samples, such as Weighted Sparse Representation Classifier (WSRC) [18] and Locality-constrained Linear Coding (LLC) method [19], while the latter focuses on important features and alleviates the adverse influences of outlier features or pixels. Li et al. defined different mapping functions to determine the correlation between samples [20]. In [21] and [22], the authors extended LRC and CRC to Robust Linear Regression Classification (RLRC) and Robust Collaborative Representation Classification (RCRC), respectively. Yang et al. presented a Regularized Robust Coding (RRC) method by feature weights learning, which shows robustness to various outlier features [23]. Zheng et al. proposed an Iterative Re-constrained Group Sparse Classifier (IRGSC) by adaptive feature and sample weights learning [24].

It should be noted that all the methods mentioned above belong to 1D or vector-based regression models, which use l_1- or l_2-norm in loss functions. Such operations inherently assume the errors follow a Laplace or Gaussian distribution. However, in real-world scenarios, the errors are much more complicated. To deal with structural noises, Yang et al. preserved the 2D structure of error images and proposed a Nuclear norm based Matrix Regression (NMR) method [25]. NMR uses nuclear norm to measure the low-rank or approximately low-rank characteristic of errors caused by contiguous occlusions. NMR shows great potential in the presence of occlusions, shadows and reflections. Based on 1D and 2D characteristics of errors, some researchers utilize mixed norm for FR by combining vector-based norms and matrix-based norms. Luo et al. combined nuclear norm and l_1-norm in a unified model, and proposed a Nuclear-L_1

C. Zhang, H. Li, C. Chen and X. Zhou are with the Department of Control and Systems Engineering, Nanjing University, Nanjing 210093, China (e-mail: chzhang@smail.nju.edu.cn, huaxiongli@nju.edu.cn, clchen@nju.edu.cn, zhouxz@nju.edu.cn).

Y. Qian is with the Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China, and also the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006, China (e-mail: jinchengqyh@126.com).

Corresponding author: Huaxiong Li.

norm joint matrix Regression (NL_1R) method [26]. Qian et al. introduced feature weights into the mixed norm model (i.e., nuclear and l_2-norm), and proposed a Robust Nuclear norm regularized Regression (RNR) model [27]. In [28], the authors described the error by a tailored function and low-rank characteristics, which achieved robust performance in the case of complex occlusions.

These approaches generally adopt convex l_1-norm to approximate the sparsity structure, and nuclear norm for low-rank structure, which can be viewed as the extension of l_1-norm on the singular values of matrix. However, such convex relaxation is biased and leads to a suboptimal solution, since l_1-norm treats all nonzero values differently while these values contribute same in l_0-norm. Such phenomenon also exists in nuclear norm and rank function. To address this problem, nonconvex relaxations are exploited in sparse and rank minimization problems such as l_p-norm ($0 < p < 1$) [29], schatten p-norm [30], weighted nuclear norm [31], truncated nuclear norm [32], etc. Nie et al. designed a logarithmic function which has better l_0 approximation than l_1-norm, and extended it to rank minimization [33]. Xie et al. presented a Robust nuclear norm-based Matrix Regression (RMR) model via weighted nuclear norm [34], in which different singular values are assigned with different weights. Zheng et al. built a Weighted Mixed-Norm Regression (WMNR) model, which combines weighted nuclear norm and l_2-norm to cope with image corruptions [35]. Dong et al. adopted a Laplacian-uniform mixed function to describe the error distribution, and proposed a mixed model combining robust sparsity and low-rank constraints [36]. In paper [32], the authors used truncated nuclear norm to replace nuclear norm. Numerical studies have demonstrated that the nonconvex surrogates usually perform better than their convex counterparts [37], [38], [39], [40].

Although those various nonconvex relaxations (e.g., l_p-norm, schatten p-norm, weighted and truncated nuclear norm, logarithmic function induced norm) alleviate the bias problem in l_0 and rank approximation to some extent, these functions have no definite upper bound and still produce large losses caused by outliers, which may lead to the poor performance. In this paper, we propose a new nonconvex FR model called Enhanced Group Sparse Regularized Nonconvex Regression (EGSNR). We first introduce a nonconvex Minimax Concave Penalty (MCP) function with definite upper bound to approximate l_0-norm, and apply it on matrix rank problem. The boundness of MCP function alleviates the influence of outliers and improves the model robustness. Different with some other methods which only impose nonconvex relaxations on regression errors, the nonconvex function is used in all terms of our proposed model, including regularizer. The label information and class weights mechanism are also incorporated into EGSNR model to improve the representation discrimination. The main contributions are summarized as follows:

1) EGSNR utilizes mixed norm to deal with complex noises in face images. Instead of nuclear norm or l_1-norm, the nonconvex MCP function induced norm is applied to describe the low-rank and sparsity structure of representation errors. The boundness of MCP function makes EGSNR more robust to outliers. Both loss function and regularizer of EGSNR model are constrained by nonconvex functions.

2) Base on MCP function, an $l_{2,\gamma}$-norm regularizer is designed to directly seek the interclass sparsity of representation coefficients, instead of using traditional $l_{2,1}$-norm. In addition, locality constraint, i.e., class weights learning mechanism, is also introduced to improve the discrimination of representation.

3) An iterative optimization algorithm based on alternating direction method of multipliers (ADMM) framework is presented to solve EGSNR model efficiently. Comprehensive experiments on several face databases are performed to demonstrate the robustness of the proposed method to various noises, compared with the state-of-the-art regression based approaches.

The remainder of this paper is organized as follows. In Section II we review the related regression models for FR. Then, we illustrate the formulation, optimization and analysis of proposed EGNSR in Section III. Section IV reports the experiment results and analysis. Finally, Section V concludes this paper.

Notations: In this paper, matrices and vectors are written in boldface uppercase and boldface lowercase, respectively. Denote its i-th row and j-th column of matrix $\mathbf{A} = (a_{ij})$ as $\mathbf{a}^i$ and $\mathbf{a}_j$, respectively.

The $l_{2,1}$-norm of matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$ is defined as [17]:

$$\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{p} \sqrt{\sum_{j=1}^{q} a_{ij}^2} = \sum_{i=1}^{p} \|\mathbf{a}^i\|_2. \tag{1}$$

The nuclear norm of matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$ is defined as [25]:

$$\|\mathbf{A}\|_* = \sum_i \sigma_i(\mathbf{A}), \tag{2}$$

where $\sigma_i(\mathbf{A})$ is the i-th singular value of $\mathbf{A}$.

TABLE I
NOTATIONS AND DESCRIPTIONS

Notations	Descriptions
c	Number of classes
n_i	Number of training samples of the i-th class
$\mathbf{D} \in \mathbb{R}^{m\times n}$	Training matrix
$\mathbf{D}_i \in \mathbb{R}^{m\times n_i}$	Training matrix of the i-th class
$\mathbf{d}_{ij} \in \mathbb{R}^m$	The j-th training sample of the i-th class
$\mathbf{y} \in \mathbb{R}^m$	A query sample
$\mathbf{x} \in \mathbb{R}^n$	Target coefficients
$\mathbf{x}_i \in \mathbb{R}^{n_i}$	Target coefficients of the i-th class
$\mathbf{1}$	A vector with all entries being 1
$\sigma_i(\mathbf{A})$	The i-th singular value of matrix $\mathbf{A}$
$\|\cdot\|_{\gamma,*}$	Matrix γ-norm
$\|\cdot\|_\gamma$	γ-norm
$z_{i,j}$	The j-th element of vector $\mathbf{z}_i$
$\mathbf{z}_{i,k}$	The elements of the k-th class of vector $\mathbf{z}_i$
$\mathrm{T_m}(\cdot)$	The operator that converts a vector to matrix
$\mathrm{T_v}(\cdot)$	The inverse operator of $\mathrm{T_m}(\cdot)$
$\odot$	Elementwise product

The ∞-norm of vector $\mathbf{x} \in \mathbb{R}^m$ is defined as [25]:

$$\|\mathbf{x}\|_\infty = \max_i |x_i|, \ i = 1, ..., m, \tag{3}$$

where $\mathbf{x} = [x_1, x_2, ..., x_m]^T$. TABLE I lists the main notations in this paper.

II. Related Works

As analyzed in Section I, many studies attempt to find a suitable loss function to describe the error $\mathbf{e}$, e.g., vector based, matrix based or mixed losses. Their common goal is to obtain appropriate regression coefficients for classification. Specifically, for a query $\mathbf{y} \in \mathbb{R}^m$, the basic idea is

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{e}, \tag{4}$$

where the error $\mathbf{e}$ may be complicated in real world. In most existing works, $\mathbf{e}$ is normally characterized by a certain distribution like Gaussian and Laplace distribution, or the error image $\mathbf{E} = \mathrm{T_m}(\mathbf{e})$ is considered to be low-rank due to the contiguous occlusions. These methods can be unified into the following model:

$$\min_{\mathbf{x}} \phi(\mathbf{e}) + \lambda\psi(\mathbf{x}), \tag{5}$$

where $\phi(\mathbf{e})$ is the loss function, and $\psi(\mathbf{x})$ is regularization term. In the following, we give an overview on ϕ and ψ of some existing methods.

A. Loss Function

According to the characteristics of errors, the loss function ϕ can be generally described by three types of norms: vector based, matrix based and mixed norms.

(1) Vector Based Norm: The most widely adopted vector based norm is l_2-norm, i.e., $\phi(\mathbf{e}) = \|\mathbf{e}\|_2^2$, which usually performs well in most conventional tasks. However, l_2-norm is proved sensitive to outliers [35]. Therefore, l_1-norm is used in loss function which shows more robustness to sparse outliers [4]. RRC incorporates the feature weights learning into regression model to alleviate the influence of noisy features or pixels (i.e., $\phi(\mathbf{e}) = \|\mathbf{w} \odot \mathbf{e}\|_2^2$). The feature weights vector $\mathbf{w}$ is adaptively learned in the optimization iterations. Liu et al. directly use a non-squared loss rather than squared loss, i.e., $\phi(\mathbf{e}) = \|\mathbf{e}\|_2$, to improve the robustness [41]. In [42], the authors introduce a nonconvex Welsch function to estimate the errors which is more robust than traditional l_2- and l_1-norm based methods.

(2) Matrix Based Norm: Matrix based norms are usually used to describe the low-rank structure in images, and nuclear norm seems to be the most popular one which is the tightest convex envelope of rank function. For example, NMR uses nuclear norm to characterize the contiguous noises and achieves impressive performance on FR with occlusions [25]. Such strategy is also used in some other methods [26], [43]. However, as mentioned before, nuclear norm introduces a bias in which large singular values are more penalized. Thus, nonconvex matrix based norms are adopted for rank minimization. In [31], [30], [35], [36], weighted nuclear norm (i.e., $\phi(\mathbf{E}) = \sum_i w_i\delta_i(\mathbf{E})$) is used for better rank approximation, in which the larger singular values are adaptively assigned with smaller weights. In [32] and [44], the authors respectively adopt a truncated nuclear norm (i.e., $\phi(\mathbf{E}) = \|\mathbf{E}\|_t = \sum_{i=t+1}^{r} \delta_i(\mathbf{E})$ with $r = rank(\mathbf{E})$) and schatten p-norm (i.e., $\phi(\mathbf{E}) = \|\mathbf{E}\|_{Sp} = (\sum_i \delta_i^p(\mathbf{E}))^{1/p}$) to replace the nuclear norm. Actually, the schatten p-norm is equivalent to the l_p-norm of matrix singular values, and it is extended to weighted schatten p-norm [30]. In [33], Nie et al. design a logarithmic function and apply it on rank minimization (i.e., $\phi(\mathbf{E}) = (\sum_i log(\delta_i(\mathbf{E})+1))$ with convergence guarantee. Though these various models have different forms, their core idea is to more precisely approximate the rank function.

(3) Mixed Norm: Due to the complicated noises in real-world, the loss function with single norm is insufficient to describe the errors. Thus, some researches adopt mixed norms by combining vector based norms and matrix based norms in loss function. Compared with single norm based approaches, mixed norm based methods have ability to handle more complex variations in face images like illumination, noises and occlusions [36]. TABLE II summarizes some robust regression FR models with single or mixed norms in loss function. As can be observed, the weighted nuclear norm is popular since it is more generalized than others like schatten p-norm and truncated nuclear norm. However, they have no upper bound and still produce large losses caused by outliers, which are the same with l_1- and l_2-norm.

TABLE II
The formulation of different combinations of loss function ϕ and regularizer ψ

$\phi(\mathbf{e})$	$\psi(\mathbf{x})$	Formulation
$\|\mathbf{e}\|_2^2$	$\|\mathbf{x}\|_1$	SRC[12]
$\|\mathbf{e}\|_2^2$	$\|\mathbf{x}\|_2^2$	CRC[16]
$\|\mathbf{e}\|_2^2$	$\|\mathbf{x}\|_{2,1}$	GSC[17]
$\|\mathbf{e}\|_1$	$\|\|\mathbf{x}\|\|_2^2$	RCRC[22]
$\|\mathbf{w} \odot \mathbf{e}\|_2^2$	$\|\mathbf{x}\|_1$ / $\|\mathbf{x}\|_2^2$	RRC_L1 / RRC_L2 [23]
$\sum \delta_i(\mathbf{E})$	$\|\mathbf{x}\|_2^2$ / $\|\mathbf{x}\|_1$	NMR / NMR_L1[43]
$\sum w_i\delta_i(\mathbf{E})$	$\|\mathbf{x}\|_1$ / $\|\mathbf{x}\|_2^2$	RMR_L1 / RMR_L2[34]
$\sum \delta_i(\mathbf{W} \odot \mathbf{E}) + \|\mathbf{w} \odot \mathbf{e}\|_2^2$	$\|\mathbf{x}\|_2^2$	RNR[27]
$\sum \delta_i(\mathbf{E}) + \|\mathbf{e}\|_1$	$\|\mathbf{x}\|_1$	SNL$_1$R[26]
$\sum w_i\delta_i(\mathbf{E}) + \|\mathbf{s} \odot \mathbf{e}\|_2^2$	$\|\mathbf{x}\|_2^2$	WMNR[35]
$\sum w_i\delta_i(\mathbf{E}) + \|\mathbf{s} \odot \mathbf{e}\|_2^2$	$\|\mathbf{x}\|_1$	LR-LUM[36]

B. Regularization Term

Different regularizers $\psi(\mathbf{x})$ enforce the representation coefficients to have different properties. l_2-norm is usually used to constrain the magnitude of coefficients to avoid overfitting. CRC [16], RCRC [22], NMR [25], RRC_L2 [23] all adopt it for regularization. In SRC [12] and RRC_L1 [23], l_1-norm is utilized to force most of the coefficients to be zero. Considering the label information of training samples, GSC adopts $l_{2,1}$-norm to enforce the coefficients to be sparse in group level (i.e., l_2-norm on intraclass level and l_1-norm

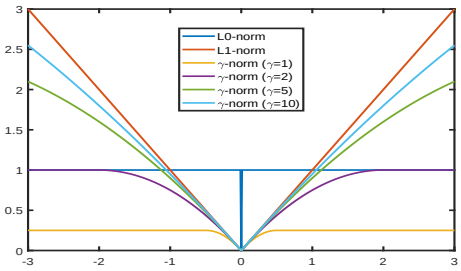


Fig. 1. The comparison of l_0, l_1 and γ-norm.

on interclass level) [17]. By incorporating the locality structure of samples, weighted l_1-, weighted l_2- and weighted $l_{2,1}$-norm are developed [18], [19]. Furthermore, some other constraints are also exploited to improve the discrimination of representation coefficients such as nonnegative constraint and k-sparse constraint [45], [46].

III. The Proposed Method

In this section, we first introduce the upper bounded MCP function induced γ-norm and the formulation of proposed model (EGSNR). Then an iterative optimization algorithm based on ADMM framework is presented to solve EGSNR model. Finally, we will make further analysis on EGSNR method.

A. MCP Function Induced Norm

Nonconvex MCP function is used in this work to approximate the l_0-norm, which is nearly unbiased and has definite upper bound. Many researchers use it for robust matrix recovery [39], matrix completion [47] and variable selection [48]. The MCP function $\rho(x;\lambda,\gamma)$ is defined as [48]:

$$\begin{aligned}\rho(x;\lambda,\gamma) &= \lambda\int_0^x(1-\frac{x}{\lambda\gamma})_+dx \\ &= (\lambda|x|-\frac{x^2}{2\gamma})I(|x|<\gamma\lambda) \\ &+\frac{\gamma\lambda^2}{2}I(|x|\geq\gamma\lambda),\end{aligned} \tag{6}$$

where $(x)_+=\max(0,x)$ and $I(\cdot)$ is the indicator function. According to the direction of [39], let $\lambda=1$ and we define the γ-norm as:

$$\|\mathbf{x}\|_\gamma=\sum_{i=1}^m\rho(\mathbf{x}_i;\gamma), \tag{7}$$

where $\mathbf{x}\in\mathbb{R}^m$ is a vector. Furthermore, γ-norm can be extended to the matrix γ-norm as follows:

$$\|\mathbf{M}\|_{\gamma,*}=\sum_{i=1}\rho(\sigma_i(\mathbf{M});\gamma), \tag{8}$$

where $\mathbf{M}\in\mathbb{R}^{p\times q}$ is a matrix. It should be noted that the MCP induced γ-norm is not a valid norm due to its violation of triangle inequality of a norm. Fig. 1 shows

Fig. 2. The effectiveness of EGSNR on removing noises and recovering face image.

the relationships among l_0-, l_1- and γ-norm. MCP function induced norms are characterized by the properties in Prop.1.

Proposition 1. *Let $\rho(x;\lambda,\gamma)$, $\|\mathbf{x}\|_\gamma$ and $\|\mathbf{M}\|_{\gamma,*}$ be defined in (6), (7) and (8) respectively, the following properties are satisfied:*

(1) $0\leq\rho(x;\lambda,\gamma)\leq\gamma\lambda^2/2$ *with left equality iff* $x=0$ *and right equality iff* $|x|\geq\gamma\lambda$;

(2) $\|\mathbf{x}\|_\gamma$ *and* $\|\mathbf{M}\|_{\gamma,*}$ *are increasing in* γ;

(3) $\lim_{\gamma\to\infty}\|\mathbf{x}\|_\gamma=\|\mathbf{x}\|_1$, $\lim_{\gamma\to\infty}\|\mathbf{M}\|_{\gamma,*}=\|\mathbf{M}\|_*$;

From Prop. 1, the upper bound of MCP loss is $\gamma\lambda^2/2$, which means it has resistance to the disturbances of outliers. On the other hand, γ-norm and matrix γ-norm have better approximation than convex l_1-norm and nuclear norm.

B. Robust Nonconvex Regression

For robust FR, we aim to develop a regression based model which is capable to accurately recognize a face image contaminated by various noises. Although the true noises in real scenarios may be complex and dense, they can be generally decomposed to two parts, i.e., low-rank structure such as contiguous occlusion and sparse structure such as pixel corruption. We use the γ-norm and matrix γ-norm defined above to build the following robust nonconvex regression model:

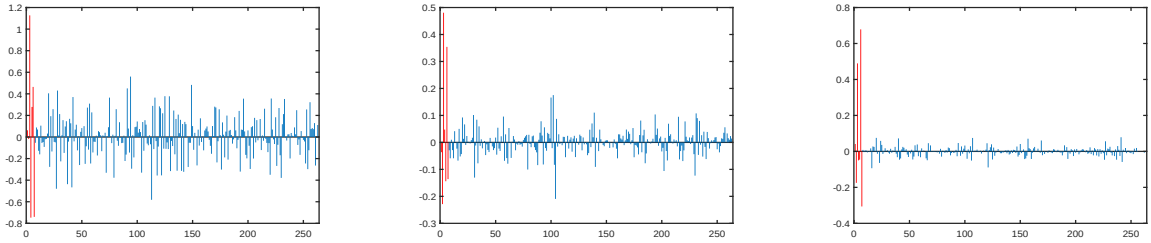
$$\begin{aligned}\min_{\mathbf{x},\mathbf{e}_1,\mathbf{e}_2}&\|\mathrm{T_m}(\mathbf{e}_1)\|_{\gamma_1,*}+\alpha\|\mathbf{e}_2\|_{\gamma_2}+\beta\psi(\mathbf{x}) \\ &\text{s.t. }\mathbf{e}_1+\mathbf{e}_2=\mathbf{y}-\mathbf{D}\mathbf{x},\end{aligned} \tag{9}$$

where γ_1 and γ_2 are tunable parameters, α and β are balance parameters. $\mathbf{e}_1$ and $\mathbf{e}_2$ denote the low-rank part and sparse part of error, respectively.

The label information of training samples is important for robust FR. In GSC, the label information is utilized to improve the sparsity of coefficients at group level with $l_{2,1}$-norm. The $l_{2,1}$-norm can improve the group sparsity but depress the values of coefficients within each class. To address this problem, we directly seek the sparsity of the l_2-norm of coefficients for each class. Based on the γ-norm clarified in Section III-A, we define the following $l_{2,\gamma}$-norm for group sparsity:

$$\|\mathbf{x}\|_{2,\gamma}=\|\mathbf{u}\|_\gamma,\mathbf{u}=[\mathbf{u}_1,...,\mathbf{u}_c]^T, \tag{10}$$

where $\mathbf{u}_i=\|\mathbf{x}_i\|_2$. It can be seen that $l_{2,\gamma}$-norm directly forces the coefficients of some classes to be zeros via γ-norm. In addition, locality structure helps to learn more discriminative representation since different classes have



(a) (b) (c)

Fig. 3. The coefficients of (a) SRC, (b) GSC, and (c) EGSNR of a face image from ExYaleB dataset. The coefficients in red correspond to the correct class.

different contributions in representation. Thus, the enhanced group sparse regularizer $\psi(\mathbf{x})$ in (9) is defined as follows:

$$\psi(\mathbf{x}) = \|\mathbf{w} \odot \mathbf{u}\|_{\gamma_3}, \tag{11}$$

where $\mathbf{w} = [w_1, w_2, ..., w_c]^T \in \mathbb{R}^c$ is the weight vector. $\mathbf{w}$ imposes penalty on all classes with different weights. Obviously, w_i should be large if the query sample $\mathbf{y}$ is far from the i-th class. In other words, w_i is positively correlated with the distance between $\mathbf{y}$ and the subspace spanned by the training samples of the i-th class. Inspired by LRC [11], we first solve the following least squares problem:

$$\mathbf{x}_i = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}_i\mathbf{x}\|^2 = (\mathbf{D}_i^T\mathbf{D}_i)^{-1}\mathbf{D}_i^T\mathbf{y}. \tag{12}$$

Problem (12) can be solved by $(\mathbf{D}_i^T\mathbf{D}_i + \lambda\mathbf{I})^{-1}\mathbf{D}_i^T\mathbf{y}$ if $\mathbf{D}_i^T\mathbf{D}_i$ is singular. We use the class-specific residual to characterize the distance between $\mathbf{y}$ and each subspace. The residual of the i-th class can be computed by $r_i = \|\mathbf{y} - \mathbf{D}_i\mathbf{x}_i\|_2$. Then the weight w_i of the i-th class can be defined as

$$w_i = \frac{r_i - r_{\min}}{r_{\max} - r_{\min}}, \tag{13}$$

where $r_{\min}$ and $r_{\max}$ are the minimum and maximum of residuals $\{r_i\}_{i=1}^c$, respectively. Obviously, the value of w_i locates in the range of [0,1]. The enhanced group sparsity regularizer $\psi(\mathbf{x})$ combines the nonconvex relaxation, group sparsity and locality structure of data, which helps learn more discriminative coefficients for classification.

Substituting (11) into (9), we obtain the final Enhanced Group Sparse regularized Nonconvex Regression (EGSNR) model as follows:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u},\mathbf{e}_1,\mathbf{e}_2} & \|\mathrm{T_m}(\mathbf{e}_1)\|_{\gamma_1,*} + \alpha\|\mathbf{e}_2\|_{\gamma_2} + \beta\|\mathbf{w} \odot \mathbf{u}\|_{\gamma_3} \\ \text{s.t. } & \mathbf{e}_1 + \mathbf{e}_2 = \mathbf{y} - \mathbf{D}\mathbf{x},\ u_i = \|\mathbf{x}_i\|_2, \\ & \mathbf{u} = [u_1, ..., u_c]^T. \end{aligned} \tag{14}$$

From (14), we can observe the relationships between EGSNR and other robust regression models. Mixed losses, low-rank and sparse, are utilized to describe the error matrix. Nonconvex relaxation with upper bound is adopted to seek a better and more robust sparsity or low-rank approximation. EGSNR imposes nonconvex constraints on loss and regularization term simultaneously. In addition, the group sparse constraint is enhanced by class-wise sparsity and locality structure of data. Fig. 2 shows the effectiveness of EGSNR on removing noises, in which a face image with mixed noises is decomposed to a clean image (i.e., $\mathbf{Dx}$ in EGSNR), low-rank noises and sparse noises (i.e., $\mathbf{e}_1$ and $\mathbf{e}_2$ respectively). Besides the reconstructed image, Fig. 3 shows the representation coefficients of EGSNR as well as SRC and GSC, which adopt traditional l_1-, $l_{2,1}$-norm for sparsity respectively. It can be clearly seen that EGSNR obtains more sparse and discriminative coefficients than SRC and GSC, which is beneficial for classification.

C. Optimization

In this section, we solve the EGSNR model (14) via ADMM algorithm, which has been widely applied in convex and nonconvex minimization problems [49], [50], [35], [51].

To solve problem (14), we first introduce two auxiliary variables $\mathbf{g} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^c$, and the original problem is converted to the following equivalent problem:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u},\mathbf{e}_1,\mathbf{e}_2,\mathbf{v},\mathbf{g}} & \|\mathrm{T_m}(\mathbf{e}_1)\|_{\gamma_1,*} + \alpha\|\mathbf{e}_2\|_{\gamma_2} + \beta\|\mathbf{v}\|_{\gamma_3} \\ \text{s.t. } & \mathbf{e}_1 + \mathbf{e}_2 = \mathbf{y} - \mathbf{D}\mathbf{x},\ \mathbf{x} = \mathbf{g}, \\ & \tilde{\mathbf{g}} = \mathbf{u},\ \mathbf{v} = \mathbf{w} \odot \mathbf{u}, \end{aligned} \tag{15}$$

where $\tilde{\mathbf{g}} = [\|\mathbf{g}_1\|_2, ..., \|\mathbf{g}_c\|_2]^T \in \mathbb{R}^c$.

To solve (15) is equivalent to minimize the augmented Lagrange function $\mathcal{L}_\mu$ defined as:

$$\begin{aligned} \mathcal{L}_\mu = & \|\mathrm{T_m}(\mathbf{e}_1)\|_{\gamma_1,*} + \alpha\|\mathbf{e}_2\|_{\gamma_2} + \beta\|\mathbf{v}\|_{\gamma_3} \\ & + \mathbf{z}_1^T(\mathbf{y} - \mathbf{D}\mathbf{x} - \mathbf{e}_1 - \mathbf{e}_2) + \mathbf{z}_2^T(\mathbf{x} - \mathbf{g}) \\ & + \mathbf{z}_3^T(\tilde{\mathbf{g}} - \mathbf{u}) + \mathbf{z}_4^T(\mathbf{v} - \mathbf{w} \odot \mathbf{u}) \\ & + \frac{\mu}{2}(\|\mathbf{y} - \mathbf{D}\mathbf{x} - \mathbf{e}_1 - \mathbf{e}_2\|_2^2 + \|\mathbf{x} - \mathbf{g}\|_2^2 \\ & + \|\tilde{\mathbf{g}} - \mathbf{u}\|_2^2 + \|\mathbf{v} - \mathbf{w} \odot \mathbf{u}\|_2^2), \end{aligned} \tag{16}$$

where $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4$ are the Lagrange multiplier vectors, and $\mu > 0$ is a penalty factor. ADMM is an iterative algorithm and the augmented Lagrange function is minimized by solving the subproblems w.r.t. each unknown variable iteratively, in which each subproblem can be solved efficiently. In the k-th iteration, it contains following seven steps to update all variables.

Step 1. Update $\mathbf{e}_1$: fix other variables, and update $\mathbf{e}_1$ by solving the following optimization problem:

$$\min_{\mathbf{e}_1} \ \|\mathrm{T_m}(\mathbf{e}_1)\|_{\gamma_1,*} + \mathbf{z}_1^T(\mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}_1-\mathbf{e}_2) + \frac{\mu}{2}\|\mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}_1-\mathbf{e}_2\|_2^2. \tag{17}$$

By simple manipulation, problem (17) is equivalent to the following problem:

$$\min_{\mathbf{e}_1} \frac{1}{\mu}\|\mathrm{T_m}(\mathbf{e}_1)\|_{\gamma_1,*} + \frac{1}{2}\|\mathbf{e}_1-\mathbf{h}_1\|_2^2, \tag{18}$$

where $\mathbf{h}_1 = \mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}_2+\frac{1}{\mu}\mathbf{z}_1$. The problem (18) can be solved by following theorem:

Theorem 1. [47] *Given the SVD* $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ *of matrix* $\mathbf{Y}$*, the optimal solution of*

$$\min_{\mathbf{Y}} \ \lambda\|\mathbf{Y}\|_{\gamma,*} + \frac{1}{2}\|\mathbf{Y}-\mathbf{E}\|_F^2,$$

with $\gamma > \lambda$ *can be obtained by*

$$S_{\lambda,\gamma}(\mathbf{Y}) = \mathbf{U}\mathbf{\Sigma}_{\lambda,\gamma}\mathbf{V}^T,$$

where $\mathbf{\Sigma}_{\lambda,\gamma}$=diag$(S_{\lambda,\gamma}(\sigma_1),...,S_{\lambda,\gamma}(\sigma_r))$ *is a matrix with the diagonal element*

$$S_{\lambda,\gamma}(\sigma_i) = \begin{cases} \sigma_i, & \sigma_i \geq \gamma, \\ \dfrac{\sigma_i-\lambda}{1-\frac{\lambda}{\gamma}}, & \lambda \leq \sigma_i < \gamma, \\ 0, & \sigma_i < \lambda. \end{cases}$$

According to Theorem 1, problem (18) has a closed-form solution which can be written as:

$$\mathrm{T_m}(\hat{\mathbf{e}}_1) = S_{\frac{1}{\mu},\gamma_1}(\mathrm{T_m}(\mathbf{h}_1)), \Leftrightarrow \quad \hat{\mathbf{e}}_1 = \mathrm{T_v}\left(S_{\frac{1}{\mu},\gamma_1}(\mathrm{T_m}(\mathbf{h}_1))\right), \tag{19}$$

where $\mathrm{T_v}(\cdot)$ is the inverse operator of $\mathrm{T_m}(\cdot)$ that reshapes a vector to matrix.

Step 2. Update $\mathbf{e}_2$: fix other variables, and update $\mathbf{e}_2$ by solving the following problem:

$$\min_{\mathbf{e}_2} \ \frac{\alpha}{\mu}\|\mathbf{e}_2\|_{\gamma_2} + \frac{1}{2}\|\mathbf{e}_2-\mathbf{h}_2\|_2^2, \tag{20}$$

where $\mathbf{h}_2 = \mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}_1+\frac{1}{\mu}\mathbf{z}_1$. Problem (20) can be decomposed to a series of independent minimization problems w.r.t. $\{e_{2,i}\}_{i=1}^m$:

$$\min_{e_{2,i}} \ \frac{\alpha}{\mu}\rho(e_{2,i};\gamma_2) + \frac{1}{2}(e_{2,i}-h_{2,i})^2, \tag{21}$$

where $e_{2,i}$ and $h_{2,i}$ are the i-th component of $\mathbf{e}_2$ and $\mathbf{h}_2$, respectively. Although $\rho(\cdot)$ is a nonconvex function, problem (21) has a closed-form solution according to [29]:

$$\hat{e}_{2,i} = \begin{cases} t_1, & \text{if} \ \ \rho(t_1;\gamma_2) \leq \rho(t_2;\gamma_2), \\ t_2, & \text{otherwise}, \end{cases} \tag{22}$$

where t_1 and t_2 are obtained by solving:

$$t_1 = \arg\min_t \frac{1}{2}(t-h_{2,i})^2 + \frac{\alpha}{\mu}(|t| - \frac{t^2}{2\gamma_2}), \ \text{s.t.}|t| \leq \gamma_2, \tag{23}$$

Algorithm 1 ADMM Algorithm for EGSNR

Input: The training matrix $\mathbf{D} \in \mathbb{R}^{m\times n}$, query sample $\mathbf{y} \in \mathbb{R}^m$, the model parameters $\gamma_1, \gamma_2, \gamma_3, \alpha, \beta, \mu, \delta, \mu_{\max}$, the convergence criteria parameter ϵ.

Output: Optimal coefficients vector $\mathbf{x}^k$.

Initialization: $\mathbf{x}^0 = \mathbf{0}, \mathbf{e}_1^0 = \mathbf{y}, \mathbf{e}_2^0 = \mathbf{0}, \mathbf{z}_1 = \mathbf{0}, \mathbf{z}_2 = \mathbf{0}, \mathbf{z}_3 = \mathbf{z}_4 = \mathbf{0}$.
Compute class weights $\mathbf{w}$ by Eqs. (12), (13).
Compute $\mathbf{M} = (\mathbf{D}^T\mathbf{D}+\mathbf{I})^{-1}$, $\mathbf{H} = (\mathbf{W}^T\mathbf{W}+\mathbf{I})^{-1}$.
while not converged **do**
  Update $\mathbf{e}_1$: Let $\mathbf{h}_1^{k+1} = \mathbf{y}-\mathbf{D}\mathbf{x}^k-\mathbf{e}_2^k+\frac{1}{\mu^k}\mathbf{z}_1^k$, $\mathbf{e}_1^{k+1} = \mathrm{T_v}\left(S_{\frac{1}{\mu^k},\gamma_1}(\mathrm{T_m}(\mathbf{h}_1^k))\right)$;
  Update $\mathbf{e}_2$: Compute $\{e_{2,i}^{k+1}\}_{i=1}^m$ by Eqs. (22), (23), (24). $\mathbf{e}_2^{k+1} = [e_{2,1}^{k+1},...,e_{2,m}^{k+1}]^T$;
  Update $\mathbf{x}$: Let $\mathbf{h}_3^{k+1} = \mathbf{y}-\mathbf{e}_1^{k+1}-\mathbf{e}_2^{k+1}+\frac{1}{\mu^k}\mathbf{z}_1^k$, $\mathbf{x}^{k+1} = \mathbf{M}(\mathbf{D}^T\mathbf{h}_3^{k+1}+\mathbf{g}^k-\frac{1}{\mu^k}\mathbf{z}_2^k)$;
  Update $\mathbf{g}$: Compute $\{\mathbf{g}_i^{k+1}\}_{i=1}^n$ by Eq. (30). $\mathbf{g}^{k+1} = [\mathbf{g}_1^{k+1},...,\mathbf{g}_n^{k+1}]^T$;
  Update $\mathbf{u}$: Let $\tilde{\mathbf{g}}^{k+1} = [||\mathbf{g}_1^{k+1}||_2,...,||\mathbf{g}_c^{k+1}||_2]$, $\mathbf{u}^{k+1} = \mathbf{H}(\mathbf{g}^{k+1}+\mathbf{W}^T\mathbf{v}^k+(\mathbf{z}_3^k+\mathbf{W}^T\mathbf{z}_4^k)/\mu^k)$;
  Update $\mathbf{v}$: Compute $\{v_i^{k+1}\}_{i=1}^c$ by Eq. (35). $\mathbf{v}^{k+1} = [v_1^{k+1},...,v_c^{k+1}]^T$;
  Update the Lagrange multiplier vectors $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4$ and penalty factor μ by Eqs. (36);
  $k := k+1$;
  Check the convergence criteria (37).
end while
return $\mathbf{x}^k$.

$$t_2 = \arg\min_t \frac{1}{2}(t-h_{2,i})^2 + \frac{\alpha\gamma_2}{2\mu}, \ \text{s.t.}|t| \geq \gamma_2. \tag{24}$$

Problems (23) and (24) are quadratic functions and can be easily solved.

Step 3. Update $\mathbf{x}$: fix other variables, we can obtain $\mathbf{x}$ by solving the following problem:

$$\min_{\mathbf{x}} \ \mathbf{z}_1^T(\mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}_1-\mathbf{e}_2) + \mathbf{z}_2^T(\mathbf{x}-\mathbf{g}) + \frac{\mu}{2}(\|\mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}_1-\mathbf{e}_2\|_2^2 + \|\mathbf{x}-\mathbf{g}\|_2^2). \tag{25}$$

To minimize problem (25) is equivalent to solve the following problem:

$$\min_{\mathbf{x}} \ \|\mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}_1-\mathbf{e}_2+\frac{1}{\mu}\mathbf{z}_1\|_2^2 + \|\mathbf{x}-\mathbf{g}+\frac{1}{\mu}\mathbf{z}_2\|_2^2. \tag{26}$$

This is a typical least squares problem which leads to a closed-form solution. By setting the derivative of (26) w.r.t. $\mathbf{x}$ to zero, we can obtain its optimal solution:

$$\hat{\mathbf{x}} = (\mathbf{D}^T\mathbf{D}+\mathbf{I})^{-1}(\mathbf{D}^T\mathbf{h}_3+\mathbf{g}-\frac{1}{\mu}\mathbf{z}_2), \tag{27}$$

where $\mathbf{h}_3 = \mathbf{y}-\mathbf{e}_1-\mathbf{e}_2+\frac{1}{\mu}\mathbf{z}_1$ and $\mathbf{I}$ is the identity matrix.

Step 4. Update $\mathbf{g}$: with other variables fixed, $\mathbf{g}$ can be

calculated by solving the following minimization problem:

$$\begin{aligned}\min_{\mathbf{g}}\ & \mathbf{z}_2^T(\mathbf{x}-\mathbf{g})+\mathbf{z}_3^T(\tilde{\mathbf{g}}-\mathbf{u}) \\ & +\frac{\mu}{2}(\|\mathbf{x}-\mathbf{g}\|_2^2+\|\tilde{\mathbf{g}}-\mathbf{u}\|_2^2).\end{aligned} \tag{28}$$

It should be noted that $\mathbf{g}\in\mathbb{R}^n$ and $\tilde{\mathbf{g}}\in\mathbb{R}^c$ are in different dimensions. We rewrite (28) as the following formulation:

$$\min_{\mathbf{g}}\sum_{i=1}^{c}\big[(z_{3,i}-\mu u_i)\|\mathbf{g}_i\|_2+\mu\|\mathbf{g}_i-\frac{\mu\mathbf{x}_i+\mathbf{z}_{2,i}}{2\mu}\|_2^2\big],\Leftrightarrow$$

$$\sum_{i=1}^{c}\min_{\mathbf{g}_i}\big[\frac{z_{3,i}-\mu u_i}{2\mu}\|\mathbf{g}_i\|_2+\frac{1}{2}\|\mathbf{g}_i-\frac{\mu\mathbf{x}_i+\mathbf{z}_{2,i}}{2\mu}\|_2^2\big], \tag{29}$$

where $z_{3,i}$ is the i-th element of vector $\mathbf{z}_3$, and $\mathbf{z}_{2,i}\in\mathbb{R}^{n_i}$ is the components associated with the i-th class of vector $\mathbf{z}_2$. From (29), it can be easily obtained that solving $\mathbf{g}$ is equivalent to solve each $\mathbf{g}_i$ independently. Problem (29) can be solved by the following theorem:

Theorem 2. [52] *Given $\mathbf{t}\in\mathbb{R}^m$ and $\lambda>0$, the optimal solution $\tilde{\mathbf{s}}$ of*

$$\min_{\mathbf{s}\in\mathbb{R}^m}\ \lambda\|\mathbf{s}\|_2+\frac{1}{2}\|\mathbf{s}-\mathbf{t}\|_2^2,$$

is given by

$$\tilde{\mathbf{s}}=\max(1-\frac{\lambda}{\|\mathbf{t}\|_2})\mathbf{t}.$$

According to Theorem 2, problem (29) has a closed-form solution for each $\mathbf{g}_i$, i.e.,

$$\hat{\mathbf{g}}_i=\max\big(1-\frac{(z_{3,i}-\mu u_i)}{2\mu\|\mathbf{r}_i\|_2},0\big)\frac{\mu\mathbf{x}_i+\mathbf{z}_{2,i}}{2\mu}, \tag{30}$$

The optimal $\hat{\mathbf{g}}$ is the concatenation of $\{\hat{\mathbf{g}}_i\}_{i=1}^c$:

$$\hat{\mathbf{g}}=[\hat{\mathbf{g}}_1,...,\hat{\mathbf{g}}_c]^T. \tag{31}$$

Step 5. Update $\mathbf{u}$: with other variables fixed, $\mathbf{u}$ can be calculated by solving the following problem:

$$\begin{aligned}\min_{\mathbf{u}}\ & \mathbf{z}_3^T(\tilde{\mathbf{g}}-\mathbf{u})+\mathbf{z}_4^T(\mathbf{v}-\mathbf{w}\odot\mathbf{u}) \\ & +\frac{\mu}{2}(\|\tilde{\mathbf{g}}-\mathbf{u}\|_2^2+\|\mathbf{v}-\mathbf{w}\odot\mathbf{u}\|_2^2).\end{aligned} \tag{32}$$

Similar to the optimization strategy of $\mathbf{x}$, we can get the closed-form solution of (32), which can be written as:

$$\hat{\mathbf{u}}=(\mathbf{W}^T\mathbf{W}+\mathbf{I})^{-1}(\tilde{\mathbf{g}}+\mathbf{W}^T\mathbf{v}+\frac{\mathbf{z}_3+\mathbf{W}^T\mathbf{z}_4}{\mu}), \tag{33}$$

where $\mathbf{W}$ is a diagonal matrix with $\mathbf{W}_{ii}=\mathbf{w}_i$.

Step 6. Update $\mathbf{v}$: fix other variables, we can calculate $\mathbf{v}$ by solving the following problem:

$$\min_{\mathbf{v}}\ \frac{\beta}{\mu}\|\mathbf{v}\|_{\gamma_3}+\frac{1}{2}\|\mathbf{v}-\mathbf{w}\odot\mathbf{u}+\frac{1}{\mu}\mathbf{z}_4\|_2^2. \tag{34}$$

Same as the optimization of $\mathbf{e}_2$, we can solve a sequence subproblems w.r.t. $\{v_i\}_{i=1}^c$ and the final solution of (34) is:

$$\hat{\mathbf{v}}=[\hat{v}_1,...,\hat{v}_c]^T, \tag{35}$$

where $\{\hat{v}_i\}_{i=1}^c$ is the optimum of each subproblem w.r.t. $\{v_i\}_{i=1}^c$, respectively.

Algorithm 2 EGSNR based Classification

Input: Training matrix $\mathbf{D}\in\mathbb{R}^{m\times n}$, a query $\mathbf{y}\in\mathbb{R}^m$.
Output: The predicted label of $\mathbf{y}$.

1: Compute the optimal coefficients $\tilde{\mathbf{x}}$ of $\mathbf{y}$ by performing Algorithm 1.
2: Compute the residuals:

$$s_i(\mathbf{y})=\|\mathrm{T_m}(\mathbf{D}\tilde{\mathbf{x}}-\mathbf{D}_i\tilde{\mathbf{x}}_i)\|_{\gamma_1,*},\ i=1,2,...,c.$$

3: Predict the label of $\mathbf{y}$: $\mathrm{Label}(\mathbf{y})=\arg\min_i s_i(\mathbf{y})$.

Step 7. Update the Lagrange multiplier vectors and the penalty factor by the following equations with other variables fixed:

$$\begin{aligned}\mathbf{z}_1&=\mathbf{z}_1+\mu(\mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}_1-\mathbf{e}_2),\\ \mathbf{z}_2&=\mathbf{z}_2+\mu(\mathbf{x}-\mathbf{g}),\\ \mathbf{z}_3&=\mathbf{z}_3+\mu(\tilde{\mathbf{g}}-\mathbf{u}),\\ \mathbf{z}_4&=\mathbf{z}_4+\mu(\mathbf{v}-\mathbf{w}\odot\mathbf{u}),\\ \mu&=\min(\mu_{\max},\delta\mu),\end{aligned} \tag{36}$$

where the parameters $\mu_{\max}$ and $\delta>1$ are manually set.

Convergence criteria. ADMM algorithm solves the original objective function via a sequence of subproblems w.r.t. each unknown variable iteratively. To achieve an optimal solution, it is important to adopt suitable stopping criteria. Following the suggestions in [49], the stopping criteria for EGSNR are defined as:

$$\begin{cases}\|\mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}_1-\mathbf{e}_2\|_\infty<\epsilon,\\ \|\mathbf{x}-\mathbf{g}\|_\infty<\epsilon,\\ \|\tilde{\mathbf{g}}-\mathbf{u}\|_\infty<\epsilon,\\ \|\mathbf{v}-\mathbf{w}\odot\mathbf{u}\|_\infty<\epsilon,\end{cases} \tag{37}$$

where $\epsilon>0$ is a small tolerance error.

Now, we can efficiently solve the EGSNR by Eqs. (19), (22), (27), (30), (33), (35) and (36) iteratively. Algorithm 1 summarizes the entire ADMM for EGSNR in detail. After obtaining the optimal representation coefficients for a given query sample, we use the Algorithm 2 to classify it. The convergence analysis for ADMM algorithm has been widely studied in [49]. Fig. 4 plots the convergence curves of Algorithm 1 on five face datasets used in our experiments. To show the convergence curves clearly, the objective function values are normalized by dividing the maximum value. It is clear that objective function loss drops to a stable value eventually, which indicates that the proposed optimization algorithm for solving EGSNR has good convergence properties.

D. Computational Analysis

Computational complexity is an important issue when estimating the performance of an algorithm [52], [49], [53]. In this section, we make a discussion on the computational cost of Algorithm 1. In EGSNR, the class weights $\mathbf{w}$ is computed once in advance, thus the major computational cost is spent on the iterations. Given the image size $m=p\times q$ and the

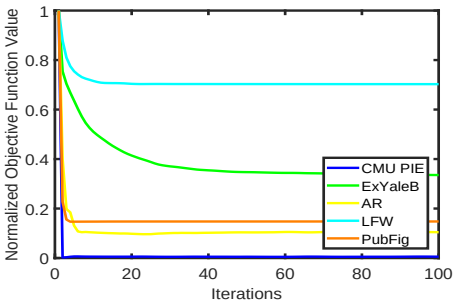


Fig. 4. The normalized convergence curves of Algorithm 1 on five datasets.

TABLE III
RECOGNITION RATES (%) COMPARISON ON EXYALEB WITH DIFFERENT TRAINING SETS (S1 DENOTES SUBSET 1 FOR TRAINING).

Method	s1	s2	s3	s4	s5
SRC	69.08	69.41	83.54	91.88	75.82
CRC	68.92	69.51	83.27	92.85	75.94
CSC	73.41	71.04	87.32	96.17	78.36
SLRC	70.31	70.82	85.62	93.42	75.33
RRC_L1	72.36	73.47	87.32	94.21	84.62
RRC_L2	71.45	72.84	86.24	93.66	85.41
NMR	78.20	78.50	88.41	95.30	87.76
F-IRNNLS	73.71	78.65	87.61	96.68	83.30
F-LR-IRNNLS	80.29	82.07	88.83	96.78	89.24
WMNR	83.68	84.27	93.01	98.37	89.47
GF	88.80	87.84	93.86	96.73	70.59
EGSNR	**90.42**	**92.70**	**95.35**	**98.97**	**91.33**

number of training samples n, the complexity of SVD operation for updating $\mathbf{e}_1$ is $O(pq^2)$ assuming $p > q$. The cost of computing $\mathbf{e}_2$ is $O(m)$. The optimization of $\mathbf{x}$ involves matrix inverse and multiplication computation. Noting that $(\mathbf{D}^T\mathbf{D}+\mathbf{I})^{-1}$ is fixed in iterations, we can compute and store it by pseudo-inverse in advance. Thus, the cost of $\mathbf{x}$ is $O(n^2)$. For updating $\mathbf{g}$, the computational complexity is $O(ct^2)$, assuming each class has t training samples. Similar to $\mathbf{x}$, the computational consumption of $\mathbf{w}$ is $O(c^2)$. The time cost of computing $\mathbf{v}$ is $O(c)$. Thus, the total computational cost of Algorithm 1 is $O(k(pq^2 + m + n^2 + ct^2 + c^2))$ if there are k iterations.

IV. EXPERIMENTS

In this section, we conduct experiments on several public available face datasets, including Extended Yale B (ExYaleB) [54], CMU PIE [55], AR [56], LFW [57] and PubFig [58], to validate the robustness and effectiveness of EGSNR. Several most-related regression based FR methods are tested for comparison, such as SRC [12], CRC [16], CSC [4], SLRC [5], RRC_L1 [23], RRC_L2 [23], NMR [25], F-IRNNLS [28], F-LR-IRNNLS [28] and WMNR [35]. The l_1-norm minimization problem in SRC is solved by Homotopy algorithm [59], and the balance parameter of CRC is fine-tuned to report their best results. The parameters of other approaches are set following the authors' suggestions. In EGSNR, $\gamma_1 = \gamma_2 = \gamma_3 = 3$, $\mu = 1$, $\delta = 1.01$ and $\epsilon = 0.001$ are adopted in our experiments.

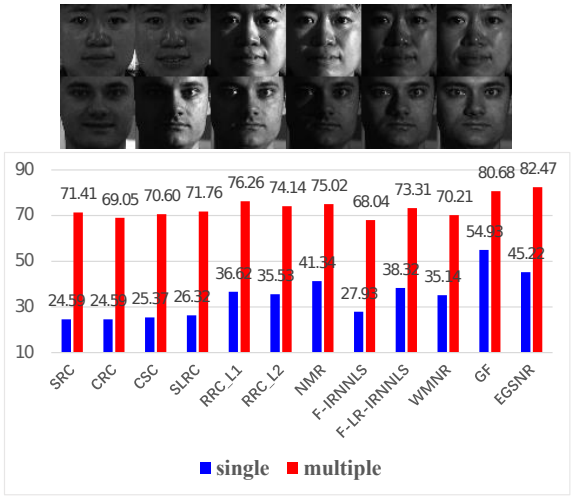


Fig. 5. Some typical images of CMU PIE dataset and recognition rates (%) comparison of different methods under single and multiple training samples protocols.

A. FR With Illumination Changes

We first investigate the robustness of EGSNR to various illumination changes on ExYaleB dataset, which consists of 2414 frontal face images over 38 individuals. The whole dataset is divided into five subsets in accordance with the illumination conditions in images [54]. From subset 1 to 5, the face images characterize slight-moderate-severe illumination changes. All images are resized to 48×42 pixels. For the five subsets, we adopt the cross-validation strategy, in which each subset is used for training and the rest for testing respectively. Specially, GradientFace (GF) is used for comparison, which is a typical method for face recognition under varying illumination [60]. TABLE III lists the experimental results of different methods on ExYaleB with different training sets. It can be observed that EGSNR achieves the best performance in all cases. F-LR-IRNNLS, WMNR and GF also obtain competitive results, and GF outperforms other methods like WMNR, F-LR-IRNNLS when s1-s3 are used for training. However, the performance of GF is relatively poor when s5 is used as training set, which indicates that it may not well deal with extreme illumination changes in s5. Differently, EGSNR still achieves over 90% recognition accuracy when the images contain extreme illumination changes.

In the second experiment, we conduct tests on CMU PIE dataset, which contains 68 individuals with total 41,368 face images. 1629 face images of the 68 individuals are chosen for tests. All images are resized to 32×32 pixels. $M(= 1, 5)$ images per subject are randomly selected for training and the rest for testing, corresponding to single and multiple training samples protocols respectively. Fig. 5 shows some face images of CMU PIE dataset and the performance of different methods under two protocols. Since the illumination changes in PIE is smaller than those in ExYaleB, GF achieves best performance under single sample protocol and EGSNR ranks the second. However, under multiple samples protocol,

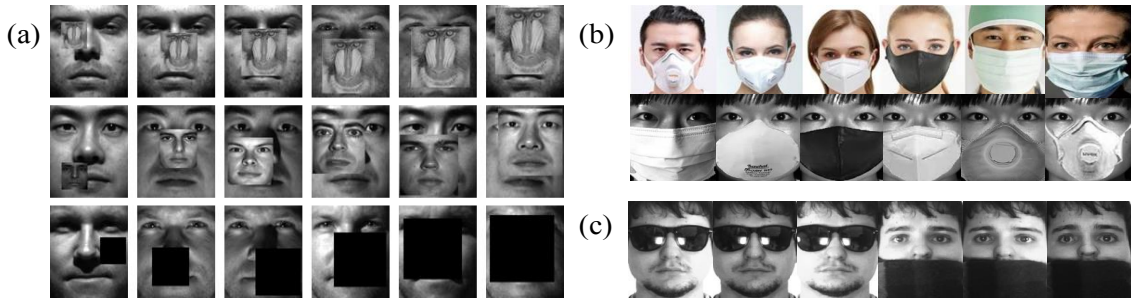


Fig. 6. Some test images used in our experiments. (a) face images from ExYaleB with six levels (i.e., 10% $\sim$ 60%) and three types (i.e., baboon, human face and black block) of occlusions. (b) The top row shows some face images with mask occlusions from the Internet and the second row shows some face images from ExYaleB with manually set mask occlusions. (c) Some face images with sunglasses and scarves occlusions from AR dataset.

EGSNR outperforms all other methods including GF. In addition, it should be noted that GF is specially designed for varying illumination conditions, while EGSNR can deal with not only illumination but also facial occlusions which will be proved later. The experimental results on ExYaleB and CMU PIE demonstrate the effectiveness and robustness of EGSNR to illumination changes.

B. FR With Contiguous Occlusion

In this section, we design experiments on ExYaleB and AR datasets to validate the robustness of EGSNR to contiguous occlusions. We utilize subset 1 of ExYaleB as training set, and the images of subset 3 are imposed various facial occlusions as test set [25]. AR contains over 4000 face images of 126 individuals with different illumination, expression and occlusion conditions. Total 2600 face images (1400 non-occluded images, 600 images with sunglasses and 600 images with scarves) of 100 individuals are used in our experiment. The images from AR are resized to 50×40.

1) FR With Square Block Occlusion: In this experiment, we design three random square block occlusions on ExYaleB. We set increasing levels of square block on test images, from 10% to 60%, with an unrelated image as occlusion (e.g., baboon, human face or black block), which is used in many studies [25], [23], [35]. Fig. 6(a) shows some test images with different levels and types of occlusions. The experimental results are illustrated in Fig. 7. The first row in Fig. 7 shows the experimental results under multiple training samples protocol, while the second row shows those under single training sample protocol. In the mode of multiple training samples, we can observe that EGSNR is more robust other methods with the increase of occlusion area. From Fig. 7(a) and Fig. 7(b), when the occlusion level is less than 40%, RRC_L1, RRC_L2, NMR, F-IR-IRNNLS and WMNR can obtain competitive performance. However, the recognition rates drop fast of these method with 60% occlusion, while EGSNR still achieves impressive 94.86% and 92.95% accuracy. This difference becomes evident in the case of black block occlusion, which is more extreme than baboon and human face occlusion. From Fig. 7(c), the performance of F-IRNNLS, F-LR-IRNNLS and WMNR degrades quickly when the occlusion ratio increases over 40%. EGSNR remains relatively stable and achieves the best performance. In particular, under single sample training protocol, the advantage of EGSNR over other methods is more significant, which is clearly presented in Fig. 7(d), (e) and (f). This encouraging performance indicates that EGSNR can be applied in difficult scenarios with only few training samples, although it is not specifically designed for few training samples condition. These results demonstrate that EGSNR is more robust and powerful to complex occlusion compared with other methods.

TABLE IV
RECOGNITION RATES (%) COMPARISON OF DIFFERENT METHODS ON EXYALEB AND AR FACE DATASETS UNDER MASKS, SUNGLASSES AND SCARVES OCCLUSION SCENARIOS.

Method	ExYaleB	AR	
	masks	sunglasses	scarves
SRC	28.57	43.17	50.50
CRC	27.62	43.17	55.17
CSC	32.19	55.83	55.83
SLRC	30.68	46.67	53.17
RRC_L1	75.62	80.83	59.17
RRC_L2	73.43	77.33	57.83
NMR	76.38	85.91	61.67
F-IRNNLS	67.24	84.50	65.33
F-LR-IRNNLS	75.81	73.33	69.50
WMNR	95.47	89.00	83.67
EGSNR	**97.44**	**91.00**	**88.67**

2) FR With Real-world Disguise: In this experiment, we evaluate the performance of EGSNR against three common real-world disguises: masks, sunglasses and scarves. For ExYaleB, the testing images are occluded by different kinds of masks. For AR, 800 non-occluded images are used for training and 1200 images with sunglasses and scarves for testing. Fig. 6(b) and (c) show some test images with various disguises. TABLE IV reports the recognition rates of different methods under the three occlusion scenarios. We can observe that EGSNR outperforms other methods with three types of disguises. When the occluded area gets larger (e.g., wearing scarves), the performance of NMR, F-IRNNLS and WMNR drops significantly, while EGSNR keeps relatively stable and achieves average 5.00%, 19.17%, 27.00% higher accuracy than WMNR, F-LR-IRNNLS and

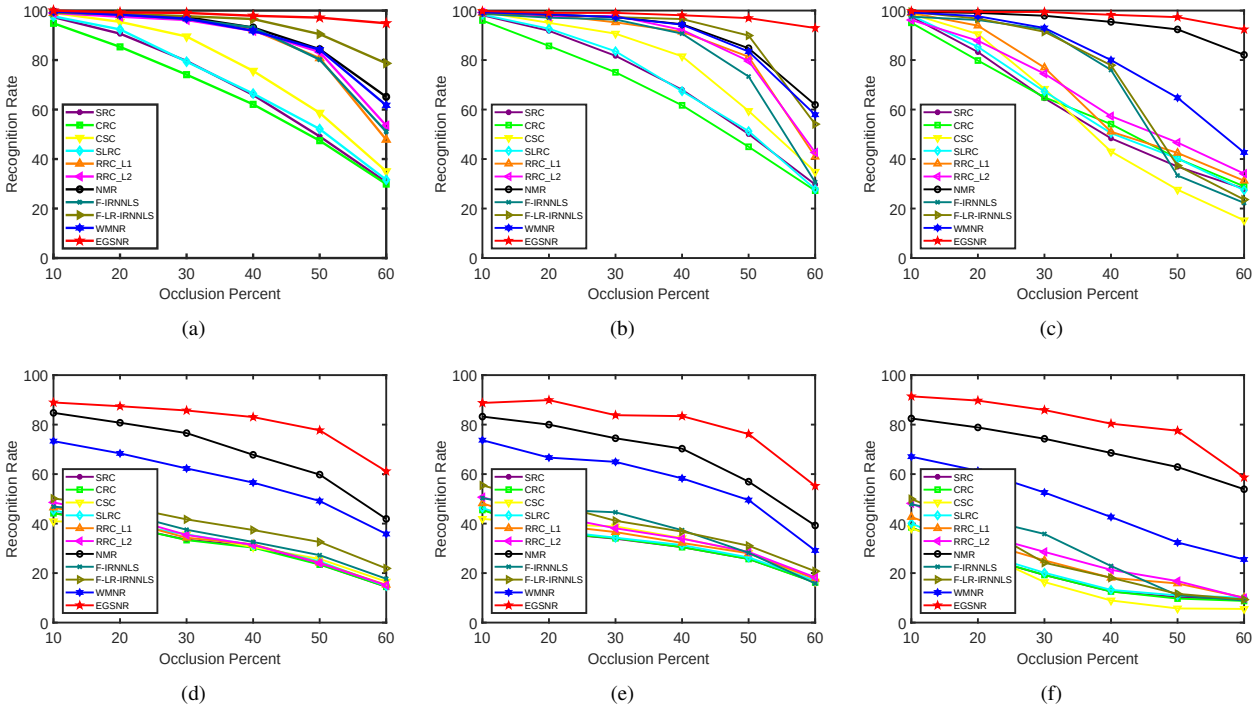


Fig. 7. Recognition rates (%) of different methods on ExYaleB under different experiment settings. (a), (b) and (c): with multiple training samples under 10%∼60% baboon, human face and black block occlusion, respectively. (d), (e) and (f): with single training sample under 10%∼60% baboon, human face and black block occlusion, respectively.

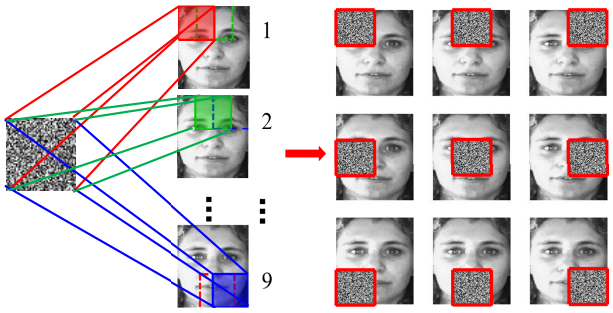


Fig. 8. The test images from AR with different facial regions occluded.

NMR, respectively. These experimental results demonstrate that EGSNR is more robust to recognize faces with real-world disguises like masks, sunglasses and scarves.

3) FR with Different Facial Regions Occluded: In this experiment, we investigate the performance of EGSNR to contiguous occlusions in different facial regions on AR dataset. 600 non-occluded face images with lighting changes per person are selected as test images. For each one, we impose a square block whose elements are random numbers between 0 and 255 as occlusion. The length of this block is denoted as l and step size is dx and dy in horizontal and vertical direction respectively. We set $l = 20$, $dx = 10$, $dy = 15$ in our experiments. Thus, there are total nine experiment settings, denoted as 1, 2, ... , 9, as shown in Fig. 8. The nine regions of occlusion cover the whole face. The recognition rates of all the competing methods in the nine cases are exhibited in TABLE V. As can be clearly seen, EGSNR outperforms other methods in all cases. F-LR-IRNNLS and WMNR also show robustness and achieve comparable results. However, the recognition rates of other methods vary significantly with different regions occluded, although the occlusion area is the same. For example, the ranges of F-LR-IRNNLS and WMNR are 11.00% and 16.00% respectively, while that of EGSNR is only 6.16%. It implies that EGSNR is less sensitive to the occluded facial regions than other compared methods. Besides, the average accuracy improvements of EGSNR over NMR, F-LR-IRNNLS and WMNR are 18.54%, 11.12% and 7.43% respectively, verifying the superiority of EGSNR.

C. FR With Mixed Noises

In this experiment, we evaluate the robustness of EGSNR to mixed noises. Both pixel corruptions and contiguous occlusions are imposed on test images. Specially, except baboon occlusion, different levels (i.e., 10%–50%) of random pixel noises are used to contaminate test images. The basic settings on training samples are the same as those in Section IV-B. Fig. 9 shows some face images with different levels of mixed noises. The 30% mixed noises mean 30% baboon occlusion plus 30% pixel corruptions, as marked in Fig. 9. TABLE VI lists the recognition rates of different methods.

We can see that EGSNR achieves the best recognition rates in all cases. The performance of NMR is not desirable, since it only considers the low-rank representation error. F-LR-IRNNLS and WMNR achieve competitive recognition

TABLE V
RECOGNITION RATES (%) COMPARISON OF DIFFERENT METHODS ON AR DATASET WITH OCCLUSION ON DIFFERENT FACIAL REGIONS.

Method	1	2	3	4	5	6	7	8	9
SRC	57.17	49.67	43.00	70.83	48.33	50.67	70.00	68.66	70.50
CRC	55.67	49.17	44.50	67.33	48.00	52.50	68.83	70.33	69.83
CSC	75.00	65.83	68.17	86.50	77.00	75.50	86.17	88.50	83.67
SLRC	57.67	50.00	43.67	71.17	48.17	51.33	70.50	69.17	70.83
RRC_L1	78.16	68.00	71.33	93.00	81.83	78.00	90.50	90.17	86.83
RRC_L2	77.50	75.83	74.67	90.67	83.33	81.83	88.17	86.00	85.33
NMR	72.50	71.50	61.67	84.17	76.33	76.83	88.16	90.50	86.83
F-IRNNLS	74.00	73.17	72.33	81.67	80.17	78.50	88.50	87.67	78.50
F-LR-IRNNLS	85.17	82.50	80.33	90.83	84.17	83.33	91.33	91.17	84.67
WMNR	89.33	80.17	81.50	95.50	90.33	90.33	96.17	94.83	90.33
EGSNR	**96.33**	**96.50**	**92.67**	**98.33**	**97.33**	**97.67**	**98.83**	**98.83**	**98.83**

Fig. 9. The face images with mixed noises (from 0% to 50%).

TABLE VI
RECOGNITION RATES (%) COMPARISON OF DIFFERENT METHODS ON EXYALEB WITH INCREASING LEVEL OF MIXED NOISES.

Method	10%	20%	30%	40%	50%
SRC	88.00	68.76	47.24	28.57	16.38
CRC	88.38	69.52	47.05	28.38	16.00
CSC	96.57	93.52	77.33	49.90	14.29
SLRC	88.95	70.10	48.57	28.57	16.76
RRC_L1	98.29	95.24	84.19	52.95	25.71
RRC_L2	94.67	91.62	78.10	43.05	16.00
NMR	96.57	82.47	59.05	30.10	14.29
F-IRNNLS	96.68	90.29	83.43	54.67	20.19
F-LR-IRNNLS	98.10	94.29	87.24	60.38	24.38
WMNR	97.90	96.95	87.62	63.24	26.10
EGSNR	**99.05**	**96.57**	**92.76**	**81.71**	**57.33**

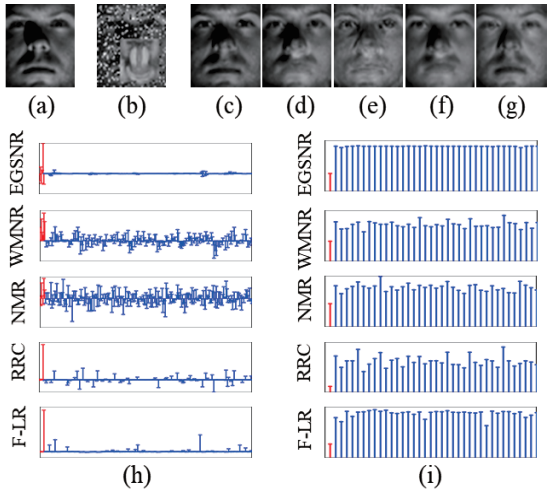


Fig. 10. Recognition with mixed noises. (a) Original image from ExYaleB. (b) Test image with mixed noises. The reconstructed image of (c) EGSNR, (d) WMNR, (e) NMR, (f) RRC_L1 and (g) F-LR-IRNNLS. (h) Representation coefficients. (i) Residuals of each class (F-LR denotes the F-LR-IRNNLS method, and the correct class is marked in red).

accuracies when the noises are mild. However, when the level of mixed noises reaches 50%, their performance is poor. SRC, CRC, CSC and SLRC are not robust to mixed noises. Fig. 10 illustrates the recognition processes of EGSNR and several robust methods on a test sample. Fig. 10(a) and (b) show the original face image and contaminated image, (c)-(g) show the reconstructed images of EGSNR, WMNR, NMR, RRC_L1, and F-LR-IRNNLS, (h) and (i) show the representation coefficients and residuals of each class with correct class marked in red. We can observe that EGSNR obtains a clear face image which is similar to original image, while the reconstructed images of other methods have different degrees of distortion. Besides, EGSNR achieves the sparsest representation coefficients, in which only the correct class has large values and the coefficients of other irrelevant classes are almost depressed to zero. The differences of representation residuals between correct class and all other classes are also significant in EGSNR. This implies that compared with other robust methods, EGSNR has more resistance against mixed noises.

D. FR With Uncontrolled Setting

The face images tested in previous experiments are all captured in strictly controlled environment. In this section, we extend our experiments on two uncontrolled face datasets: LFW and PubFig. For LFW, we use a subset of LFW (i.e., LFW-a) and total 1580 images of 158 subjects are selected for experiments. For PubFig, 2000 images of 200 subjects are used. All the images are cropped and resized to 32×32 pixels. We randomly select half images of each person for training and the other half for testing. PCA with 98% energy preserved is used in SRC, CRC, CSC, RRC_L1 and RRC_L2 for computational efficiency. Some example images of LFW and PubFig are shown in Fig. 11.

The recognition rates of SRC, CRC, CSC, SLRC, RRC_L1, RRC_L2, NMR, F-IRNNLS, F-LR-IRNNLS, WMNR and EGSNR on LFW-a and PubFig are shown in TABLE VII. We can see that EGSNR is superior to other methods. F-IRNNLS, F-LR-IRNNNLS and WMNR also achieve competitive results. These experiment results further

TABLE VII
RECOGNITION RATES (%) OF DIFFERENT METHODS ON LFW AND PUBFIG DATASETS (IRNNLS1 AND IRNNLS2 REPRESENT F-IRNNLS AND F-LR-IRNNLS METHOD RESPECTIVELY).

Dataset	SRC	CRC	CSC	SRLC	RRC_L1	RRC_L2	NMR	IRNNLS1	IRNNLS2	WMNR	EGSNR
LFW	39.37	40.13	42.03	39.49	41.39	42.66	40.37	46.84	43.05	46.62	**49.49**
PubFig	38.10	37.20	39.40	38.40	42.80	41.30	40.10	39.40	43.90	44.70	**45.60**

Fig. 11. Some face images of (a) LFW and (b) PubFig dataset.

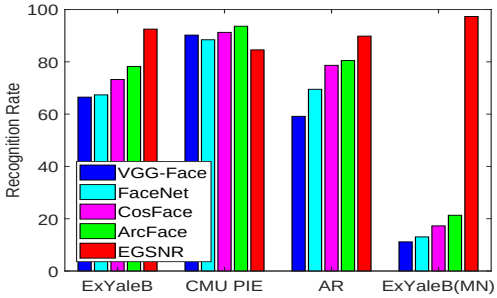


Fig. 12. Recognition rates (%) comparison of VGG-Face, FaceNet, CosFace, ArcFace and EGSNR. ExYaleB (MN) means the test images are contaminated by 10% mixed noises.

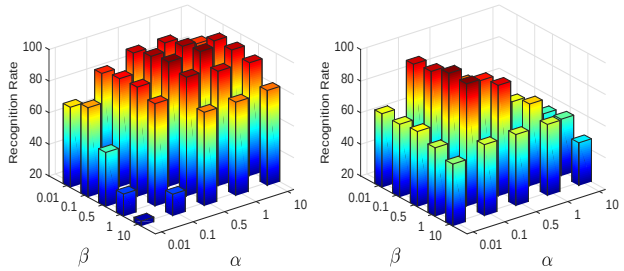


Fig. 13. Recognition rates (%) of EGSNR versus α and β on ExYaleB dataset with (a) masks occlusion and (b) 10% mixed noises.

TABLE VIII
RECOGNITION RATES (%) OF EGSNR AND ITS THREE VARIATIONS ON EXYALEB WITH MIXED NOISES AT DIFFERENT LEVELS.

Method	10%	20%	30%	40%	50%
EGSNR-s	98.43	93.67	84.76	67.43	31.05
EGSNR-v	98.29	91.62	74.29	36.00	13.52
EGSNR-t	99.05	95.24	90.10	78.38	53.14
EGSNR	**99.05**	**96.57**	**92.76**	**81.71**	**57.33**

confirm that EGSNR is more robust than other regression based methods in FR.

E. Compared With CNN Based Methods

CNN based methods have achieved great success in many computer vision and image analysis tasks in recent years [9]. In this experiment, we compare our EGSNR with some typical deep learning models to investigate their robustness to various noises (e.g., extreme illumination changes, large occlusions and complex noises). Following [61], [62], we use the pre-trained CNN models to extract features and nearest neighbor with cosine distance metric for classification. Four popular and publicly available deep learning models on VGG-Face [7], FaceNet [8], CosFace [10] and ArcFace [9], are used in our experiment, which are well-trained and evaluated on very large wild face datasets. CNN based methods and EGSNR are tested on ExYaleB, CMU PIE and AR datasets to evaluate the performance against extreme illumination changes, contiguous occlusions and mixed noises. The LFW and PubFig are not included due to the excellent performance of the deep models on these large wild datasets. For ExYaleB and PIE, the experimental settings are same as Section IV-A (multiple training samples protocol for PIE). For AR datasets, the face images with sunglasses and scarves are used for testing. ExYaleB with 10% mixed noises, i.e., ExYaleB (MN), are also used for comparison, and the experimental settings are same as IV-C. Fig. 12 shows the performance of four CNN based methods and EGSNR.

It can be seen that CNN based methods obtain slightly higher recognition rates than EGSNR on CMU PIE, since the illumination changes in CMU PIE dataset are small. However, EGSNR significant outperforms CNN based methods on ExYaleB and AR face datasets. Specifically, on ExYaleB (MN), the performance of CNN based methods is poor. The main reason is that there are complex noises (e.g., severe shadows, large occlusions and mixed noises) in test sets, while the training sets consist of clean face images and CNN models cannot acquire any prior knowledge about the noises. EGSNR adopts low-rank and sparse structures to characterize the representation error, and is capable to handle various noises. Finetuning the deep neural networks by adding some specific noises in training set may promote their robustness to the specific noises, however, they may not generalize well for other new noises and this topic is beyond the scope of this paper. In addition, it should be mentioned that EGSNR consumes much less training and computing resources than deep learning methods.

F. Ablation Study and Parameter Analysis

In proposed EGSNR model (14), matrix γ-norm and vector γ-norm are used to characterize different potential noises, i.e., low-rank and sparse noises. Besides, an $l_{2,\gamma}$-norm is used to replace traditional $l_{2,1}$-norm to promote

the group sparsity. In this section, we conduct ablation experiments to verify the effect of them separately. Three variations of EGSNR are derived, i.e., EGSNR-s, EGSNR-v and EGSNR-t. EGSNR-s and EGSNR-v discard the first term and second term in EGSNR model (14) respectively, and EGSNR-t uses $l_{2,1}$-norm in EGSNR instead of $l_{2,\gamma}$-norm. We compare EGSNR and its three variations on ExYaleB dataset with mixed noises at different levels, and the experimental results are reported in TABLE VIII. We can observe that: 1) EGSNR outperforms EGSNR-s and EGSNR-v in all cases, which demonstrates that single low-rank or sparse constraint cannot well deal with the complex noises in face images, and it is beneficial to combine two constraints. 2) EGSNR outperforms EGSNR-t which demonstrates the effectiveness and superiority of proposed $l_{2,\gamma}$-norm compared with traditional $l_{2,1}$-norm. 3) EGSNR-s outperforms EGSNR-v which implies that single matrix based norm may be more suitable than single vector based norm for handling mixed noises.

From the objective function (14), there are two important tunable parameters α and β in EGSNR model. α makes a balance between low-rank error and sparse error, and β is the regularizer parameter which controls the strength of enhanced group sparsity. To achieve the satisfactory performance of EGSNR, it is necessary to analyze its parameter sensitivity. We first define two candidate sets $\{0.01, 0.1, 0.5, 1, 10\}$ and $\{0.01, 0.1, 0.5, 1, 10\}$ for α and β, respectively. Then, with different combinations of the two parameters, EGSNR is performed on ExYaleB under contiguous masks occlusion and 10% mixed noises, respectively. Fig. 13 shows the recognition performance of EGSNR versus α and β. We can observe that both parameters impact the performance of proposed method. When the face images are occluded by masks (i.e., Fig. 13(a)), the performance is not very sensitive to α and β when the two parameters locate in $[0.1, 0.5]$. When the test images are contaminated by complex noises (i.e., Fig. 13(b)), the performance is sensitive to α since it affects the loss of sparse noises. In general, EGSNR can obtain satisfactory performance when α and β locate in $[0.1, 1]$. However, it is still difficult to find the optimal parameter α and β for different datasets. A simple way for parameter setting is to determine each one with the other fixed. In this paper, we first fix β as 0.5 and search the optimal α in the interval $[0.1, 1]$. Then α is fixed as the found value and search in $[0.1, 1]$ for optimal β.

V. Conclusion and Future Work

In this paper, an Enhanced Group Sparse regularized Nonconvex Regression (EGSNR) model is proposed for robust face recognition. EGSNR utilizes mixed norms to model the representation residuals and shows robustness to gross errors. The nonconvex MCP function is introduced to estimate the l_0-norm and extended on matrix for rank approximation. To improve the discrimination of representation, locality and group sparse structures are considered simultaneously in EGSNR. An $l_{2,\gamma}$-norm is proposed to enhance the group sparsity instead of using traditional $l_{2,1}$-norm. Based on ADMM framework, an iterative algorithm is presented to solve EGSNR model. Experimental results on several popular face datasets demonstrate the effectiveness and robustness of the proposed method in dealing with complex occlusions and noises.

There are still some issues on proposed method which deserve our further investigation. In EGSNR, the potential noises are jointly modeled by low-rank and sparse structures. However, the noises in face images may be much complicated in real-world which may be not simply described by these structures. How to extend the proposed model for general noises deserves further study. Besides, the performance of EGSNR on wild datasets (e.g., LFW) is not desirable compared with deep learning methods. Incorporating robust face alignment methods into EGSNR may boost the performance on these wild face datasets. Finally, EGSNR assumes that the training set contains all person identities of test set (i.e., closed-set face recognition), while open-set face recognition is a more challenging and important topic in practice. Extending EGSNR for open-set face recognition task is an interesting and practical subject for future work.

Acknowledgments

The authors would like to thank the editor and anonymous reviewers for their constructive and valuable comments and suggestions. This work was partially supported by the National Key Research and Development Program of China (Nos. 2018YFB1402600, 2016YFD0702100), and the National Natural Science Foundation of China (Nos. 61672332, 62073160, 61876079, 71671086).

References

[1] Y. Sun, Q. Liu, J. Tang, and D. Tao, "Learning discriminative dictionary for group sparse representation," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3816–3828, 2014.

[2] J. Lu, Y.-P. Tan, and G. Wang, "Discriminative multimanifold analysis for face recognition from a single training sample per person." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 39–51, 2013.

[3] 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, 2015.

[4] J. Lai and X. Jiang, "Classwise sparse and collaborative patch representation for face recognition," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3261–3272, 2016.

[5] W. Deng, J. Hu, and J. Guo, "Face recognition via collaborative representation: Its discriminant nature and superposed representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2513–2521, 2018.

[6] H. Li, L. Zhang, B. Huang, and X. Zhou, "Sequential three-way decision and granulation for cost-sensitive face recognition," *Knowl. Based Syst.*, vol. 91, pp. 241–251, 2016.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[8] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4690–4699.

[10] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *CVPR*, 2018, pp. 5265–5274.

[11] 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, 2010.

[12] J. Wright, A. Y. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.

[13] L. Jing and M. K. Ng, "Sparse label-indicator optimization methods for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1002–1014, 2013.

[14] J. Xu, L. Zhang, and D. Zhang, "A trilateral weighted sparse coding scheme for real-world image denoising," in *ECCV*, 2018, pp. 20–36.

[15] Z. He, S. Yi, Y.-M. Cheung, X. You, and Y. Y. Tang, "Robust object tracking via key patch sparse representation," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 354–364, 2016.

[16] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *ICCV*, 2012, pp. 471–478.

[17] J. Huang, F. Nie, H. Huang, and C. Ding, "Supervised and projected sparse coding for image classification," in *AAAI*, 2013.

[18] C.-Y. Lu, H. Min, J. Gui, L. Zhu, and Y.-K. Lei, "Face recognition via weighted sparse representation," *J. Vis. Commun. Image R.*, vol. 24, no. 2, pp. 111–116, 2013.

[19] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010, pp. 3360–3367.

[20] F. Li, Y. Qian, J. Wang, C. Dang, and L. Jing, "Clustering ensemble based on sample's stability," *Artif. Intell.*, vol. 273, pp. 37–55, 2019.

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

[22] S. Cai, L. Zhang, W. Zuo, and X. Feng, "A probabilistic collaborative representation based approach for pattern classification," in *CVPR*, 2016, pp. 2950–2959.

[23] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized robust coding for face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1753–1766, 2013.

[24] J. Zheng, P. Yang, S. Chen, G. Shen, and W. Wang, "Iterative re-constrained group sparse face recognition with adaptive weights learning," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2408–2423, 2017.

[25] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 156–171, 2016.

[26] L. Luo, J. Yang, J. Qian, and Y. Tai, "Nuclear-l1 norm joint regression for face reconstruction and recognition with mixed noise," *Pattern Recognit.*, vol. 48, no. 12, pp. 3811–3824, 2015.

[27] J. Qian, L. Luo, J. Yang, F. Zhang, and Z. Lin, "Robust nuclear norm regularized regression for face recognition with occlusion." *Pattern Recognit.*, vol. 48, no. 10, pp. 3145–3159, 2015.

[28] M. Iliadis, H. Wang, R. Molina, and A. K. Katsaggelos, "Robust and low-rank representation for fast face identification with occlusions," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2203–2218, 2017.

[29] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye, "A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems," in *ICML*, 2013, pp. 37–45.

[30] Y. Xie, S. Gu, Y. Liu, W. Zuo, W. Zhang, and L. Zhang, "Weighted schatten p-norm minimization for image denoising and background subtraction," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4842–4857, 2016.

[31] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *CVPR*, 2014, pp. 2862–2869.

[32] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, "Fast and accurate matrix completion via truncated nuclear norm regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2117–2130, 2012.

[33] F. Nie, Z. Hu, and X. Li, "Matrix completion based on non-convex low-rank approximation," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2378–2388, 2019.

[34] J. Xie, J. Yang, J. J. Qian, Y. Tai, and H. M. Zhang, "Robust nuclear norm-based matrix regression with applications to robust face recognition," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2286–2295, 2017.

[35] J. Zheng, K. Lou, X. Yang, C. Bai, and J. Tang, "Weighted mixed-norm regularized regression for robust face identification," *IEEE Trans. Neural Netw. Learn. Syst.*, 2019.

[36] J. Dong, H. Zheng, and L. Lian, "Low-rank laplacian-uniform mixed model for robust face recognition," in *CVPR*, 2019, pp. 11 897–11 906.

[37] C. Lu, J. Tang, S. Yan, and Z. Lin, "Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 829–839, 2015.

[38] C. Lu, C. Zhu, C. Xu, S. Yan, and Z. Lin, "Generalized singular value thresholding," in *AAAI*, 2015.

[39] S. Wang, D. Liu, and Z. Zhang, "Nonconvex relaxation approaches to robust matrix recovery," in *IJCAI*, 2013.

[40] H. Zhang, J. Yang, J. Xie, J. Qian, and B. Zhang, "Weighted sparse coding regularized nonconvex matrix regression for robust face recognition," *Inf. Sci.*, vol. 394, pp. 1–17, 2017.

[41] Y. Liu, Y. Guo, H. Wang, F. Nie, and H. Huang, "Semi-supervised classifications via elastic and robust embedding," in *AAAI*, 2017.

[42] J. Ke, C. Gong, T. Liu, L. Zhao, and D. Tao, "Laplacian welsch regularization for robust semisupervised learning," *IEEE Trans. Cybern.*, pp. 1–14, 2020, 10.1109/TCYB.2019.2953337.

[43] J. Chen, J. Yang, L. Luo, J. Qian, and W. Xu, "Matrix variate distribution-induced sparse representation for robust image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2291–2300, 2015.

[44] F. Nie, H. Huang, and C. Ding, "Low-rank matrix recovery via efficient schatten p-norm minimization," in *AAAI*, 2012.

[45] H. Liu, Z. Wu, X. Li, D. Cai, and T. S. Huang, "Constrained nonnegative matrix factorization for image representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1299–1311, 2011.

[46] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, 2015.

[47] D. Liu, T. Zhou, H. Qian, C. Xu, and Z. Zhang, "A nearly unbiased matrix completion approach," in *MLKD*, 2013, pp. 210–225.

[48] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, 2010.

[49] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *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.

[50] C. Chen, B. He, Y. Ye, and X. Yuan, "The direct extension of admm for multi-block convex minimization problems is not necessarily convergent," *Math. Program.*, vol. 155, no. 1-2, pp. 57–79, 2016.

[51] Y. Wang, W. Yin, and J. Zeng, "Global convergence of admm in nonconvex nonsmooth optimization," *J. Sci. Comput.*, vol. 78, no. 1, pp. 29–63, 2019.

[52] J. Yang and Y. Zhang, "Alternating direction algorithms for ℓ_1-problems in compressive sensing," *SIAM J. Sci. Comput.*, vol. 33, no. 1, pp. 250–278, 2011.

[53] Y. Qian, J. Liang, W. Pedrycz, and C. Dang, "Positive approximation: an accelerator for attribute reduction in rough set theory," *Artif. Intell.*, vol. 174, no. 9-10, pp. 597–618, 2010.

[54] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 6, pp. 643–660, 2001.

[55] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, 2003.

[56] A. M. Martinez and R. Benavente, "The AR face database," CVC, Barcelona, Spain, Tech. Rep., 1998.

[57] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[58] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *ICCV*, 2009, pp. 365–372.

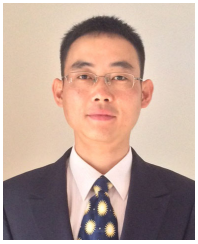
[59] M. S. Asif and J. Romberg, "Sparse recovery of streaming signals using l_1-homotopy," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4209–4223, 2014.

[60] T. Zhang, Y. Y. Tang, B. Fang, Z. Shang, and X. Liu, "Face recognition under varying illumination using gradientfaces," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2599–2606, 2009.

[61] M. Mehdipour Ghazi and H. K. Ekenel, "A comprehensive analysis of deep learning based representation for face recognition," in *CVPR Workshops*, 2016, pp. 34–41.

[62] C. Y. Wu and J. J. Ding, "Occluded face recognition using low-rank regression with generalized gradient direction," *Pattern Recognit.*, vol. 80, pp. 256–268, 2018.

Chao Zhang (S'19) received the B.E. degree in automation from Nanjing University, Nanjing, China in 2018. He is currently pursuing the M.E. degree in the Department of Control and Systems Engineering, Nanjing University, Nanjing, China, and also working as a researcher at the Research Center for Novel Technology of Intelligent Equipment, Nanjing University, Nanjing, China. His current research interests include machine learning, pattern recognition, and computer vision.



Huaxiong Li (M'11) received the M.E. degree in control theory and control engineering from Southeast University, Nanjing, China, in 2006, and Ph.D. degree from Nanjing University, Nanjing, China, in 2009. He is currently an Associate Professor with the Department of Control and Systems Engineering, Nanjing University, Nanjing, China. He was a visiting scholar at the Department of Computer Science, University of Regina, Canada, from 2007 to 2008, and a visiting scholar at the University of Hong Kong, Hong Kong, China, in 2010. His current research interests include machine learning, pattern recognition, and computer vision. He is a Committee Member of JiangSu association of Artificial Intelligence (JSAI) Pattern Recognition Committee, and a Committee Member of China Association of Artificial Intelligence (CAAI) Machine Learning Committee.

Chunlin Chen (S'05-M'06) received the B.E. degree in automatic control and Ph.D. degree in control science and engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2001 and 2006, respectively. He is currently a Professor and the Chair of the Department of Control and Systems Engineering, Nanjing University, Nanjing, China. He was a visiting scholar at Princeton University, Princeton, USA, from 2012 to 2013. He had visiting positions at the University of New South Wales Canberra at ADFA, Australia, and the City University of Hong Kong, Hong Kong, China.

His recent research interests include machine learning, pattern recognition, intelligent information processing, and quantum control. He is a Co-Chair of Technical Committee on Quantum Cybernetics, IEEE Systems, Man and Cybernetics Society. He is a Committee Member of JiangSu association of Artificial Intelligence (JSAI) Pattern Recognition Committee, and a Committee Member of China Association of Artificial Intelligence (CAAI) Machine Learning Committee.

Yuhua Qian received the M.S. and Ph.D. degrees in computers with applications from Shanxi University, Taiyuan, China, in 2005 and 2011, respectively.

He is currently a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education, Shanxi University. He is best known for multi-granulation rough sets in learning from categorical data and granular computing. He is involved in research on pattern recognition, feature selection, rough set theory, granular computing, and artificial intelligence. He has authored over 80 articles on these topics in international journals. He served on the Editorial Board of the International Journal of Knowledge-Based Organizations and Artificial Intelligence Research. He has served as the Program Chair or Special Issue Chair of the Conference on Rough Sets and Knowledge Technology, the Joint Rough Set Symposium, and the Conference on Industrial Instrumentation and Control, and a PC Member of many machine learning, data mining, and granular computing conferences.

Xianzhong Zhou (M'10) received the B.S. and M.S. degrees in System Engineering and the Ph.D. degree in Control Theory and Application from Nanjing University of Science and Technology, Nanjing, China, in 1982, 1985, and 1996, respectively. He is currently a Professor with the Department of Control and Systems Engineering, Nanjing University, Nanjing, China, and the Director of the Research Center for Novel Technology of Intelligent Equipment, Nanjing University, Nanjing, China. His current research interests include eye view vision systems, intelligent information processing, and future integrated automation systems. Prof. Zhou was among the people selected for 333 Engineering of Jiangsu Province, China in 2002 and 2007, and the Excellent Science and Technology Worker of Jiangsu Province, China, in 2000. He is the Executive Director of Systems Engineering Society of China and the Honor President of Systems Engineering Society of Jiangsu Province, China.