Efficient Image Classification via Structured Low-Rank Matrix Factorization Regression

Hengmin Zhang, *Member, IEEE*, Jian Yang, *Member, IEEE*, Jianjun Qian, *Member, IEEE*, Guangwei Gao, *Senior Member, IEEE*, Xiangyuan Lan, Zhiyuan Zha, *Senior Member, IEEE*, and Bihan Wen, *Senior Member, IEEE*

***Abstract*— In real-world applications involving sparse coding and low-rank matrix recovery problems, linear regression methods usually struggle to effectively capture the structured correlations present in data matrices. This limitation arises from representation approaches that treat images as vectors and handle testing samples individually, overlooking these correlations. To address these challenges, we propose a novel approach that leverages the low-rank property to capture the global and intrinsic structure of residual and coefficient matrices, departing from the assumption of independent and identically distributed (I.I.D) data. Our method introduces nonconvex and nonsmooth low-rank matrix regression models guided by the extended matrix variate power exponential distribution (M.P.E.D). By incorporating factorization strategies into the regression coefficient matrix and utilizing the Schatten-p norm with three distinct values of p, we enhance computational efficiency. Our formulation enables efficient subproblem solving through the introduction of auxiliary variables and the use of singular value threshold operators. We achieve closed-form solutions using the proposed multi-variable alternating direction method of multipliers (ADMM). Theoretical analysis establishes the local convergence properties and computational complexity of our optimization algorithm. Furthermore, we conduct numerical experiments on various image datasets, including face, object, and digital, to demonstrate the superior performance and computational efficiency of our methods compared to several related regression approaches. The source codes for our method are available at https://github.com/ZhangHengMin/TIFS_SLRMFR.**

***Index Terms*— Low-rank matrix regression, alternating direction method of multipliers (ADMM), Schatten-p norm, matrix factorization, theoretical analysis, image classification.**

I. Introduction

IT IS well-known that sparse coding and low-rank matrix recovery methods have made significant contributions to various domains, including pattern recognition, computer vision, and machine learning. These methods have found successful applications in diverse areas, such as image recognition [1], [2], [3], multimodal recognition [4], [5], finger-vein recognition [6], [7], and image restoration and inpainting [8], [9]. They have also been widely employed in tasks such as classification and subspace clustering. Among the linear regression techniques, Sparse Representation-based Classification (SRC) [10] and Low-Rank Representation (LRR) [11] have received considerable attention in this context. SRC aims to find the sparsest solution for each data point independently, while LRR seeks the lowest-rank solution for all the data jointly. These methods have been extended and improved in various ways, including Collaborative Representation Classification (CRC) [12], Nuclear norm Matrix Regression (NMR) and its faster version (FNMR) [13], Unifying Liner Regression (ULR) [14], Adaptive Low-Rank Representation (ALPR) [15], Discriminative Low-Rank Sparse Representation (DLRSR) [16], Nonconvex Linear Regression related to ℓ_{21}-norm (NLRℓ_{21}) [17], and Robust Supervised Low-Rank Discriminant Analysis (RSLDA) [18]. These regression methods have been extensively studied and applied, contributing to be an active area of research, especially for image classification.

The methods discussed above exhibit versatility, applicable to both one-dimensional vectors and two-dimensional matrices. These methods leverage various norms to achieve specific properties, including sparsity, collaboration, and low-rankness [12], [13], [14]. At the vector level, norms like the ℓ_1-norm, ℓ_2-norm, and $\ell_{2,1}$-norm are frequently employed. Similarly, at the matrix level, norms such as the Frobenius norm, $L_{2,1}$-norm, and nuclear norm are utilized as seen in works like [19], [20], and [21]. These norms play pivotal roles in shaping optimization objectives and guiding the representation process for coefficients and noise measurements. Alongside

Manuscript received 13 June 2023; revised 21 August 2023; accepted 14 November 2023. Date of publication 28 November 2023; date of current version 11 December 2023. This work was supported in part by the Ministry of Education, Republic of Singapore, through its Start-Up Grant and Academic Research Fund Tier 1 under Grant RG61/22; in part by the National Natural Science Fund (NSF) of China under Grant 61906067, Grant 61972212, and Grant 62176124; in part by the Peng Cheng Laboratory Key Project under Grant PCL2023A08; in part by the China Post-Doctoral Science Foundation under Grant 2019M651415 and Grant 2020T130191; and in part by the Fundamental Research Funds for the Central Universities under Grant 30918014108. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Roberto Caldelli. *(Corresponding author: Bihan Wen.)*

Hengmin Zhang, Zhiyuan Zha, and Bihan Wen are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: hengmin.zhang@ntu.edu.sg; zhiyuan.zha@ntu.edu.sg; bihan.wen@ntu.edu.sg).

Jian Yang and Jianjun Qian are with the PCA Laboratory and the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: csjyang@njust.edu.cn; csjqian@njust.edu.cn).

Guangwei Gao is with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China, and also with the Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo 101-0003, Japan (e-mail: csggao@gmail.com).

Xiangyuan Lan is with the Peng Cheng Laboratory, Shenzhen 518100, China (e-mail: lanxy@pcl.ac.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIFS.2023.3337717, provided by the authors.

Digital Object Identifier 10.1109/TIFS.2023.3337717

convex formulations, nonconvex approaches based on norms such as the ℓ_p-norm, $\ell_{2,p}$-norm, and Schatten p-norm with $p \in (0, 1)$ have been explored as alternatives for unbiased estimators. These nonconvex formulations act as substitutes for promoting sparsity, group sparsity, and low-rankness, and they have demonstrated advantages across various studies. These approaches have been applied in [22], [23], [24], [25], and [26]. A comprehensive analysis of the existing references has revealed three common aspects, as detailed below:

- Regression models related to one-dimensional vectors primarily focus on defining the loss function and regularization term under the assumption of independent and identically distributed (I.I.D) data [10], [13], [14]. These models handle testing samples individually by employing linear representations based on the training samples.
- Regression models related to two-dimensional matrices consider the residual function and regularization term while incorporating low-rank structural information [11], [13], [27]. The incorporation of additional information has been shown to enhance their performance, as demonstrated in studies such as [12], [15], [16], [18], [28], [29], and [30].
- Some linear regression methods utilize nonconvex relaxations of the ℓ_0-norm, $\ell_{2,0}$-norm, and rank function, which provide nearly unbiased estimators [19], [20], [22], [23], [26]. However, these optimization algorithms often suffer from higher complexity, particularly when dealing with large-scale matrices, due to the computations involved in singular value decomposition (SVD).

In this work, we address the aforementioned challenges by introducing three main contributions. Firstly, we propose a novel approach that adopts a joint representation method inspired by the matrix variate distribution for testing samples, enabling us to capture sample relationships and incorporate additional information into the representation process. Instead of treating testing samples independently, we consider them collectively, allowing for a more comprehensive representation. Secondly, we incorporate low-rank structures into the representation of the residual and coefficient matrices. By enforcing a low-rank property on the coefficient matrix, we effectively capture the underlying structure of the data and improve representation quality. This incorporation of low-rankness enhances the discriminative power of the algorithm. Lastly, we enhance computational efficiency by factorizing the coefficient matrix into two factor matrices. This factorization significantly reduces computational complexity, thereby enhancing the algorithm's efficiency. Furthermore, the resulting theoretical analysis provides support for practical applications. To illustrate our design process, we provide a representation relation in Fig. 1, which demonstrates that the testing image can be represented as a linear combination of the training samples and an error image. Notably, we take into account each residual image matrix while leveraging the global structure of the coefficient matrix. This approach allows us to capture pixel dependencies and enforce the low-rank property on the coefficient matrix, as demonstrated in previous works [11], [13], [20], [23], [27], [31]. By preserving the relationships within these matrices, we introduce the concept

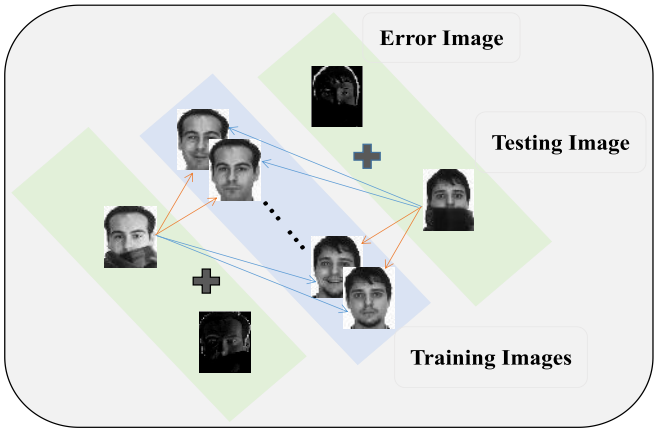


Fig. 1. Revisiting the matrix level in equation (1) through the interconnectedness of training, testing, and error images.

of matrix regression, as defined in ***Definition*** 1. This allows us to collectively represent the testing samples and integrate low-rank structures into the representation process, providing a solution.

Based on the flowchart depicted in Fig. 1, our approach involves jointly representing testing samples at the matrix level, successfully considering both the residual and coefficient matrices. This enables us to capture the underlying structure and dependencies present in the data, resulting in more informative and accurate representations. Meanwhile, we propose the Schatten-p Norm Factorized Low-rank Matrix Regression (S_pNFLMR) methods. These methods are optimized using the multi-variable ADMM algorithm [32], [33], [34], which involves computing the SVD and determining smaller factor matrices, which help in preserving the low-rank property and improving the computational efficiency. To better evaluate the performance, we apply them to image reconstruction and classification tasks and compare their performance against existing methods. This motivates us to assess the classification accuracy and computational efficiency of the methods and provide visual results from different viewpoints to facilitate a comprehensive comparison. Through this comprehensive evaluation, our main objective is to validate the effectiveness and efficiency of the proposed S_pNFLMR methods.

The subsequent sections are organized as follows: Section II introduces the problem formulations from a probabilistic distribution perspective and presents the factorized formulas that form the basis of our approach. Sections III and IV provide a detailed explanation of the iterative algorithm's development using the ADMM framework, along with provable analysis of its convergence properties. In Section V, we present experimental results to demonstrate the exceptional classification accuracy and computational efficiency achieved by our approach. Finally, in Section VI, we conclude the work by summarizing the key findings and discussing potential avenues for future research.

II. Problem Formulation

This section begins by introducing matrix regression and highlighting the relationship between the error matrix $\mathbf{E}$ and

the coefficient matrix $\mathbf{X}$, as depicted in Fig. 1. To capture the interactions between these matrices, we propose a novel representation framework inspired by the matrix variate power exponential distribution (M.P.E.D) [35], [36]. This framework enables us to effectively model and analyze the dependencies and relationships that exist within the data.

Definition 1: Let $\mathbf{A}_i$ be a column vector of dimension $lq \times 1$, where l is the number of rows and q is the number of columns. The training database is represented by $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m] \in \mathbb{R}^{lq \times m}$. For a testing sample $\mathbf{Y}_i \in \mathbb{R}^{lq \times 1}$, the linear representation is given by

$$\mathcal{M}(\mathbf{Y}_i) = \mathcal{A}(\mathbf{X}_i) + \mathcal{M}(\mathbf{E}_i), \tag{1}$$

where $\mathcal{M}(\cdot) : \mathbb{R}^{lq} \rightarrow \mathbb{R}^{l \times q}$ transforms a vector into a matrix. Here, $\mathcal{M}(\mathbf{E}_i)$ represents the residual matrix, $\mathbf{X}_i = [\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{im}]^\top \in \mathbb{R}^{m \times 1}$ is the coefficient vector, and $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n] \in \mathbb{R}^{m \times n}$ is the coefficient matrix. The reconstructed image matrix is given by $\mathcal{A}(\mathbf{X}_i) = \mathbf{X}_{i1}\mathcal{M}(\mathbf{A}_1) + \mathbf{X}_{i2}\mathcal{M}(\mathbf{A}_2) + \ldots + \mathbf{X}_{im}\mathcal{M}(\mathbf{A}_m)$.

To capture the characteristics of the residual and coefficient matrices, we depart from the assumption of independent and identically distributed (I.I.D.) elements and introduce a specific definition for the random matrices used in this study.

Definition 2: Consider a random matrix $\mathbf{Z} \in \mathbb{R}^{l \times n}$ that follows a $l \times n$ variate power exponential distribution with parameters $\mathbf{M} \in \mathbb{R}^{l \times n}$, $\Sigma \in \mathbb{R}^{l \times l}$, $\Phi \in \mathbb{R}^{n \times n}$, and $p, \beta > 0$. Then, the density function of $\mathbf{Z}$ can be defined as

$$\begin{aligned} &f(\mathbf{Z}, \mathbf{M}, \Sigma, \Phi, p, \beta) \\ &= \mathbf{C}|\Sigma|^{-\frac{n}{2}}|\Phi|^{-\frac{l}{2}}\mathbf{e}^{-\frac{1}{2}\left(\mathrm{tr}[(\mathbf{Z}-\mathbf{M})^\top \Sigma^{-1}(\mathbf{Z}-\mathbf{M})\Phi^{-1}]^{\frac{p}{2}}\right)^{\frac{\beta}{2}}}, \end{aligned} \tag{2}$$

where $\mathbf{C} = \frac{ln\Gamma\left(\frac{ln}{2}\right)}{\pi^{\frac{ln}{2}}\Gamma\left(1+\frac{ln}{2p}\right)2^{1+\frac{ln}{p}}}$ and $\Gamma(\cdot)$ denotes the Gamma function. Here, we use the notation $\mathbf{Z} \sim \Xi(\mathbf{M}, \Sigma, \Phi, p, \beta)$ to represent the distribution for simplicity.

By setting $\mathbf{M} = \mathbf{0}$, $\Sigma = \mathbf{I}_{l \times l}$, $\Phi = \mathbf{I}_{n \times n}$, and $\beta = 1$ in (2), we obtain $\mathbf{Z} \sim \Xi(\mathbf{0}, \mathbf{I}_{l \times l}, \mathbf{I}_{n \times n}, p, 1)$. Then, we have

$$f(\mathbf{Z}, \mathbf{0}, \mathbf{I}_{l \times l}, \mathbf{I}_{n \times n}, p, 1) = \mathbf{C}\mathbf{e}^{-\frac{1}{2}\mathrm{tr}(\mathbf{Z}^\top \mathbf{Z})^{\frac{p}{2}}}, \tag{3}$$

where $\mathbf{C}$ is a constant determined by (2). Taking the logarithm of both sides of (3), we obtain

$$\ln f(\mathbf{Z}, \mathbf{0}, \mathbf{I}_{l \times l}, \mathbf{I}_{n \times n}, p, 1) = \ln \mathbf{C} - \frac{1}{2}\mathrm{tr}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{\frac{p}{2}}. \tag{4}$$

Proof Based on the assumptions and definitions stated in ***Definitions*** 1 and 2, we can deduce the following

$$P(\mathcal{M}(\mathbf{E}_i)|\mathbf{X}_i) = \mathbf{C}_1\mathbf{e}^{-\frac{1}{2}\mathrm{tr}\left((\mathcal{M}(\mathbf{E}_i))^\top(\mathcal{M}(\mathbf{E}_i))\right)^{\frac{p}{2}}}, \tag{5}$$

$$P(\mathbf{X}) = \mathbf{C}_2\mathbf{e}^{-\frac{1}{2}\mathrm{tr}(\mathbf{X}^\top \mathbf{X})^{\frac{p}{2}}}, \tag{6}$$

where $\mathbf{C}_1$ and $\mathbf{C}_2$ are positive proportionality constants. Consequently, the estimation of the coefficient matrix $\mathbf{X}$ can be obtained by solving the Maximum a Posteriori (MAP) probability problem, which is formulated as follows

$$\begin{aligned} \mathbf{X}^* &= \mathrm{argmax}_{\mathbf{X}} \ln P(\mathbf{X}|\mathbf{E}) \\ &= \mathrm{argmax}_{\mathbf{X}} \ln P(\mathbf{E}|\mathbf{X}) + \ln P(\mathbf{X}) \\ &= \mathrm{argmax}_{\mathbf{X}} \ln \prod_{i=1}^{n}\prod_{j=1}^{n} P(\mathcal{M}(\mathbf{E}_i)|\mathbf{X}_j) + \ln P(\mathbf{X}) \\ &= \mathrm{argmax}_{\mathbf{X}} \sum_{i=1}^{n}\sum_{j=1}^{m} \ln P(\mathcal{M}(\mathbf{E}_i)|\mathbf{X}_j) + \ln P(\mathbf{X}) \\ &= \mathrm{argmax}_{\mathbf{X}} \sum_{i=1}^{n} \ln P(\mathcal{M}(\mathbf{E}_i)|\mathbf{X}_i) + \ln P(\mathbf{X}) \\ &= \mathrm{argmax}_{\mathbf{X}} -\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\left((\mathcal{M}(\mathbf{E}_i))^\top(\mathcal{M}(\mathbf{E}_i))\right)^{\frac{p}{2}} \\ &\quad -\frac{1}{2}\mathrm{tr}\left(\mathbf{X}^\top\mathbf{X}\right)^{\frac{p}{2}} + \ln(\mathbf{C}_1\mathbf{C}_2) \end{aligned} \tag{7}$$

where the fifth equality "=" in the statement holds because $P_{i \neq j}(\mathcal{M}(\mathbf{E}_i)|\mathbf{X}_j) = 0$. By using (2), (5), (6), and (7), and introducing a regularization parameter $\lambda > 0$, we arrive at the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{E}} \sum_{i=1}^{n}\mathrm{tr}\left((\mathcal{M}(\mathbf{E}_i))^\top(\mathcal{M}(\mathbf{E}_i))\right)^{\frac{p}{2}} + \lambda\mathrm{tr}\left(\mathbf{X}^\top\mathbf{X}\right)^{\frac{p}{2}} \\ &\text{s.t.} \quad \mathbf{E} = \mathbf{Y} - \mathbf{A}\mathbf{X}, \end{aligned} \tag{8}$$

which is equivalent to (9) based on the definition of the Schatten-p norm, i.e., $\|\mathbf{X}\|_{S_p}^p = \mathrm{tr}\left(\mathbf{X}^\top\mathbf{X}\right)^{\frac{p}{2}}$. □

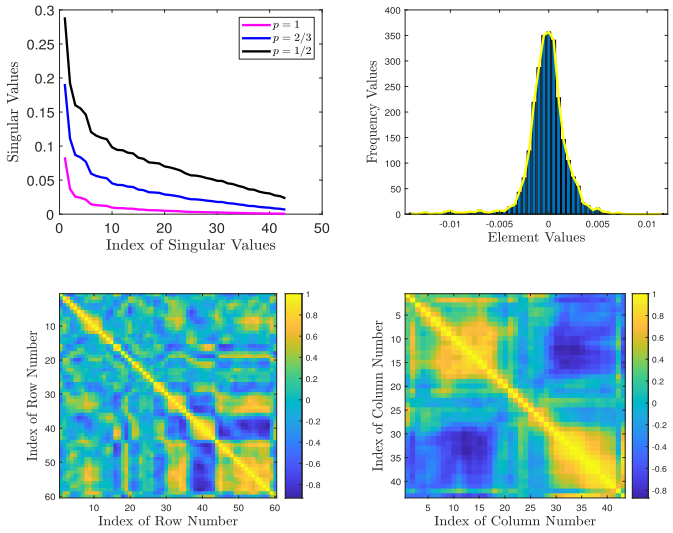
Theorem 1: For $1 \leq i \leq n$, let $\mathcal{M}(\mathbf{E}_i) = \mathcal{M}(\mathbf{Y}_i) - \mathcal{A}(\mathbf{X}_i)$ be a random matrix following an E. M. P. E distribution, specifically $\mathcal{M}(\mathbf{E}_i) \sim \Xi(\mathbf{0}, \mathbf{I}_{l \times l}, \mathbf{I}_{q \times q}, p, 1)$. In addition, let $\mathbf{X} \in \mathbb{R}^{m \times n}$ follow this probability distribution. Thus, the joint low-rank matrix regression model can be formulated as:

$$\min_{\mathbf{X}} \sum_{i=1}^{n} \|\mathcal{M}(\mathbf{Y}_i) - \mathcal{A}(\mathbf{X}_i)\|_{S_p}^p + \lambda\|\mathbf{X}\|_{S_p}^p. \tag{9}$$

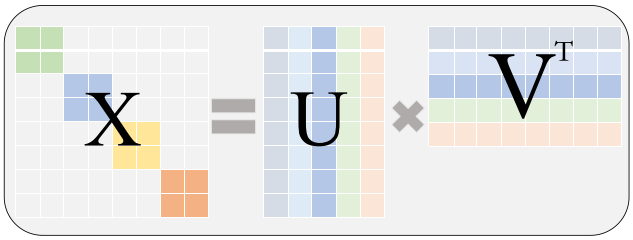
Remark 1: Problem (9) provides the flexibility to integrate low-rank structures simultaneously into the error measurements and representation coefficients, as shown in Fig. 2(a) and (b). By utilizing the Schatten-p norm, it gives improved correlation and adaptability to structural noises [13], [23] compared to convex norms that I.I.D. noise. For example, the L_1-norm assumes the Laplace distribution, while the Frobenius norm assumes the Gaussian distribution.

Remark 2: As the number of training and testing samples increases, the computational complexity of updating the matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ becomes higher, especially when $m > n$. This is primarily due to the computations involved in the SVD, which is commonly used for solving low-rank matrix minimization problems and has a complexity of $o(mn^2)$. It is important to discuss the distinctiveness of our factorization strategy in comparison to other acceleration techniques, such as:

- Utilization of nesterov's strategy and extragradient technique [22], [37]: These strategies are effective in algorithm designs aimed at curtailing the total iteration count.
- Analogous to the process in power methods and randomized SVD [33], [38]: These techniques follow a similar processing approach with the goal of diminishing computational complexity within each iteration.



(a) Residual Matrix



(b) Coefficient Matrix

Fig. 2. The illustrations consist of plotted curves depicting singular values and the distribution of element values for an error matrix sized 60×43 (top row). Additionally, correlation matrices are showcased as heatmaps in both row-wise and column-wise representations (bottom row) in (a), all related to a face image of the AR database. In (b), the prevalent factorization of the coefficient matrix is displayed, emphasizing its block-diagonal property.

Notably, this addition aims to enhance the understanding of the unique features of our approach while also acknowledging the effectiveness and efficiency of alternative acceleration strategies. Then, the utilization of a factorization technique for Schatten-p norm, as illustrated in Fig. 2 (b), is a natural step to address this challenge and is logically sound.

Definition 3: Let $\mathbf{X} = \mathbf{U}\mathbf{V}^\top \in \mathbb{R}^{m\times n}$, where $\mathbf{U} \in \mathbb{R}^{m\times d}$ and $\mathbf{V} \in \mathbb{R}^{n\times d}$, with $\text{rank}(\mathbf{X}) = r \le d \le \min(m, n)$. For any p, q_1, and $q_2 > 0$ satisfying $\frac{1}{p} = \frac{1}{q_1} + \frac{1}{q_2}$, we have

$$\frac{1}{p}\|\mathbf{X}\|_{S_p}^p = \min_{\mathbf{U},\mathbf{V}} \frac{1}{q_1}\|\mathbf{U}\|_{S_{q_1}}^{q_1} + \frac{1}{q_2}\|\mathbf{V}\|_{S_{q_2}}^{q_2}, \tag{10}$$

which leads to the following equivalences: (***i***) $\|\mathbf{X}\|_* = \min_{\mathbf{U},\mathbf{V}} \frac{1}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)$ for $p = 1$, $q_1 = 2$, and $q_2 = 2$, (***ii***) $\|\mathbf{X}\|_{S_{2/3}}^{2/3} = \min_{\mathbf{U},\mathbf{V}} \frac{1}{3}(\|\mathbf{U}\|_F^2 + 2\|\mathbf{V}\|_*)$ for $p = 2/3$, $q_1 = 2$, and $q_2 = 1$, and (***iii***) $\|\mathbf{X}\|_{S_{1/2}}^{1/2} = \min_{\mathbf{U},\mathbf{V}} \frac{1}{2}(\|\mathbf{U}\|_* + \|\mathbf{V}\|_*)$ for $p = 1/2$, $q_1 = 1$, and $q_2 = 1$.

The cases outlined in ***Definition*** 3 correspond to specific values of p. These values have also been investigated in the case of $p = 1$ in studies like [21], [39], [40], and for $p = 1/2$ and $2/3$ in [41], [42]. Then, we can introduce a substitution of the Schatten-p norm with the minimization of the sum of two norms, denoted as $g_{q_1,q_2}(\mathbf{U}, \mathbf{V})$, for cases (***i***)-(***iii***). This substitution is not only meaningful but also aligned with the motivation presented in ***Remark*** 2, as it effectively reduces computational complexity. The sizes of the factor matrices $\mathbf{U}$ and $\mathbf{V}$ depend on the number of training and testing samples. Without loss of generality, by combining (9) and (8) with (10), we formulate the optimization problem:

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V},\mathbf{X},\mathbf{E}} \sum_{i=1}^{n} \|\mathcal{M}(\mathbf{E}_i)\|_{S_p}^p + \lambda g_{q_1,q_2}(\mathbf{U}, \mathbf{V}), \\ &\text{s.t.} \quad \mathbf{E} = \mathbf{Y} - \mathbf{A}\mathbf{X}, \quad \mathbf{X} = \mathbf{U}\mathbf{V}^\top, \end{aligned} \tag{11}$$

where to analyze the solution to problem (11) for different values of p in the Schatten-p norm, we consider specific values of p and their corresponding choices of q_1 and q_2. The factorization formulations can be expressed as follows:

- When $p = 1$, we choose $g_{2,2}(\mathbf{U}, \mathbf{V}) = \frac{1}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)$ in (11), and then (11) becomes the minimization problem based on nuclear norm factorization, i.e.,

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V},\mathbf{X},\mathbf{E}} \sum_{i=1}^{n} \|\mathcal{M}(\mathbf{E}_i)\|_{S_1}^1 + \frac{\lambda}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) \\ &\text{s.t.} \quad \mathbf{E} = \mathbf{Y} - \mathbf{A}\mathbf{X}, \ \mathbf{X} = \mathbf{U}\mathbf{V}^\top. \end{aligned} \tag{12}$$

- When $p = 2/3$, we choose $g_{2,1}(\mathbf{U}, \widehat{\mathbf{V}}) = \frac{1}{3}(\|\mathbf{U}\|_F^2 + 2\|\widehat{\mathbf{V}}\|_*)$ in (11), and also introduce $\widehat{\mathbf{V}} = \mathbf{V}$ in the constraints, then (11) becomes the minimization problem based on Schatten-2/3 norm factorization, i.e.,

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V},\widehat{\mathbf{V}},\mathbf{X},\mathbf{E}} \sum_{i=1}^{n} \|\mathcal{M}(\mathbf{E}_i)\|_{S_{2/3}}^{2/3} + \frac{\lambda}{3}(\|\mathbf{U}\|_F^2 + 2\|\widehat{\mathbf{V}}\|_*) \\ &\text{s.t.} \quad \mathbf{E} = \mathbf{Y} - \mathbf{A}\mathbf{X}, \quad \mathbf{X} = \mathbf{U}\mathbf{V}^\top, \quad \widehat{\mathbf{V}} = \mathbf{V}. \end{aligned} \tag{13}$$

- When $p = 1/2$, we choose $g_{1,1}(\widehat{\mathbf{U}}, \widehat{\mathbf{V}}) = \frac{1}{2}(\|\widehat{\mathbf{U}}\|_* + \|\widehat{\mathbf{V}}\|_*)$ in (11), and also introduce $\widehat{\mathbf{U}} = \mathbf{U}$ and $\widehat{\mathbf{V}} = \mathbf{V}$ in the constraints, then (11) becomes the minimization problem based on Schatten-1/2 norm factorization, i.e.,

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V},\widehat{\mathbf{V}},\widehat{\mathbf{U}},\mathbf{X},\mathbf{E}} \sum_{i=1}^{n} \|\mathcal{M}(\mathbf{E}_i)\|_{S_{1/2}}^{1/2} + \frac{\lambda}{2}(\|\widehat{\mathbf{U}}\|_* + \|\widehat{\mathbf{V}}\|_*) \\ &\text{s.t.} \ \mathbf{E} = \mathbf{Y} - \mathbf{A}\mathbf{X}, \mathbf{X} = \mathbf{U}\mathbf{V}^\top, \widehat{\mathbf{V}} = \mathbf{V}, \widehat{\mathbf{U}} = \mathbf{U}. \end{aligned} \tag{14}$$

To obtain a closed-form solution for each subproblem in solving problems (13) and (14), the introduction of auxiliary variables such as $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{U}}$ is advantageous. This approach allows us to utilize $g_{q_1,q_2}(\mathbf{U}, \mathbf{V})$ and apply the singular value thresholding operator [43], which is associated with the nuclear norm. Alternatively, when auxiliary variables are not introduced as splitting variables, linearized strategies, as employed in [17], [37], and [44], are often used for the square term. In general, the optimization of most constrained minimization problems involves iterative optimization methods, such as modified ADMM [42], [45], [46], [47]. These algorithms offer a substantial reduction in computational complexity through the factorization of the low-rank matrix. However, it is essential to engage in a comprehensive discussion regarding the potential challenges and intricacies associated with integrating complexity reduction into our proposed algorithms. This can be achieved by examining two key perspectives:

- On the one hand, the adoption of complexity reduction strategies entails the incorporation of matrix factorization, specifically $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$. In our approach, we have opted for the widely used Schatten p-norm [41], [42], as other nonconvex rank relaxation functions may lack readily available factorization formulas. The presence of closed-form or analytic solutions for the associated subproblems greatly enhances computational efficiency. Thus, both factorization formulas and closed-form/analytic solutions are fundamental prerequisites for the successful implementation of our proposed approach.
- On the other hand, implementing the factorization strategy for low-rank matrices inherently introduces the complexity of managing multiple variables, a challenge crucial for ensuring convergence guarantees, as highlighted in [34]. Importantly, the incorporation of multiple variables naturally intensifies the computational workload, particularly during the optimization process involving SVD for large-scale matrix computations. Nevertheless, it is essential to acknowledge that the dimensions of the factor matrices are significantly smaller when compared to the learned low-rank matrix. This distinct attribute has substantially influenced our decision to capitalize on it, aiming to curtail computational complexity.

III. The Optimization Scheme

In this section, we will employ the ADMM approach to solve problems (12), (13), and (14). To simplify and introduce commonly used functions, we define the following:

$$\begin{cases} f_{1,\mu^k}(\Gamma_1, \mathbf{E}, \mathbf{X}) = \langle \Gamma_1, \mathbf{E} + \mathbf{A}\mathbf{X} - \mathbf{Y} \rangle \\ \quad + \dfrac{\mu^k}{2} \|\mathbf{E} + \mathbf{A}\mathbf{X} - \mathbf{Y}\|_F^2, & (15) \\ f_{2,\mu^k}(\Gamma_2, \mathbf{X}, \mathbf{U}, \mathbf{V}) = \langle \Gamma_2, \mathbf{X} - \mathbf{U}\mathbf{V}^\top \rangle \\ \quad + \dfrac{\mu^k}{2} \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|_F^2, & (16) \\ f_{3,\mu^k}(\Gamma_3, \mathbf{V}, \widehat{\mathbf{V}}) = \langle \Gamma_3, \widehat{\mathbf{V}} - \mathbf{V} \rangle + \dfrac{\mu^k}{2} \|\widehat{\mathbf{V}} - \mathbf{V}\|_F^2, & (17) \\ f_{4,\mu^k}(\Gamma_4, \mathbf{U}, \widehat{\mathbf{U}}) = \langle \Gamma_4, \widehat{\mathbf{U}} - \mathbf{U} \rangle + \dfrac{\mu^k}{2} \|\widehat{\mathbf{U}} - \mathbf{U}\|_F^2, & (18) \end{cases}$$

where $\langle \cdot, \cdot \rangle$ represents the inner product operator. The dual variables Γ_1, Γ_2, Γ_3, and Γ_4 are introduced, and the penalty parameter $\mu^{k+1} = \rho\mu^k > 0$ is commonly used to accelerate the convergence speed, where $\rho > 1$. Empirical analysis suggests that larger values of ρ can lead to fewer iterations but lower accuracy, while smaller values of ρ can result in more iterations but higher accuracy. The augmented Lagrangian functions (ALFs) for solving problems (12), (13), and (14) are then defined as follows:

- Using (15) and (16) for $p = 1$, we have

$$\begin{aligned} &\mathcal{L}_{1,\mu^k}(\mathbf{U}, \mathbf{V}, \mathbf{X}, \mathbf{E}, \Gamma_1, \Gamma_2) \\ &\quad = \sum_{i=1}^{n} \|\mathcal{M}(\mathbf{E}_i)\|_{S_1}^1 + f_{1,\mu^k}(\Gamma_1, \mathbf{E}, \mathbf{X}) \\ &\qquad + \frac{\lambda}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) + f_{2,\mu^k}(\Gamma_2, \mathbf{X}, \mathbf{U}, \mathbf{V}), \end{aligned} \tag{19}$$

- Using (15)-(17) for $p = 2/3$, we obtain

$$\begin{aligned} &\mathcal{L}_{2/3,\mu^k}(\mathbf{U}, \mathbf{V}, \widehat{\mathbf{V}}, \mathbf{X}, \mathbf{E}, \Gamma_1, \Gamma_2, \Gamma_3) \\ &\quad = \sum_{i=1}^{n} \|\mathcal{M}(\mathbf{E}_i)\|_{S_{2/3}}^{2/3} + f_{1,\mu^k}(\Gamma_1, \mathbf{E}, \mathbf{X})) \\ &\qquad + \frac{\lambda}{3}(\|\mathbf{U}\|_F^2 + 2\|\widehat{\mathbf{V}}\|_*) + f_{2,\mu^k}(\Gamma_2, \mathbf{X}, \mathbf{U}, \mathbf{V}) \\ &\qquad + f_{3,\mu^k}(\Gamma_3, \mathbf{V}, \widehat{\mathbf{V}}), \end{aligned} \tag{20}$$

- Using (15)-(18) for $p = 1/2$, we achieve

$$\begin{aligned} &\mathcal{L}_{1/2,\mu^k}(\mathbf{U}, \mathbf{V}, \widehat{\mathbf{V}}, \widehat{\mathbf{U}}, \mathbf{X}, \mathbf{E}, \Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4) \\ &\quad = \sum_{i=1}^{n} \|\mathcal{M}(\mathbf{E}_i)\|_{S_{1/2}}^{1/2} + f_{1,\mu^k}(\Gamma_1, \mathbf{E}, \mathbf{X})) \\ &\qquad + \frac{\lambda}{2}(\|\widehat{\mathbf{U}}\|_* + \|\widehat{\mathbf{V}}\|_*) + f_{2,\mu^k}(\Gamma_2, \mathbf{X}, \mathbf{U}, \mathbf{V}) \\ &\qquad + f_{3,\mu^k}(\Gamma_3, \mathbf{V}, \widehat{\mathbf{V}}) + f_{4,\mu^k}(\Gamma_4, \mathbf{U}, \widehat{\mathbf{U}}), \end{aligned} \tag{21}$$

where in the subsequent iterations of the nonconvex multi-variable ADMM, we address a sequence of equations (19) to (21) to update primal variables, dual variables, and penalty parameters sequentially. We place particular emphasis on elucidating our rationale behind the utilization of the ADMM algorithm, aiming to foster a transparent understanding of our choice. In this context, we provide a concise overview: ADMM was selected due to its proven effectiveness in dealing with constrained optimization problems. The algorithm's ability to derive closed-form solutions for each subproblem significantly contributes to its computational efficiency. This selection aligns with our optimization objectives, striking a balance between robustness and computational simplicity. Therefore, it is essential to comprehensively explain our selection to employ the ADMM algorithm, highlighting its compatibility with our specific problem's characteristics. Moreover, to solve the individual subproblems arising from equations (19) to (21), we seek closed-form solutions by minimizing optimization problems grounded in the Schatten-p norm, where $p = 1$, $2/3$, and $1/2$. This approach is crucial for managing residual descriptions and factorization strategies for coefficient matrix. To achieve this, we introduce singular value thresholding operators, a technique extensively utilized in previous studies [23], [41], [42], [48] and presented below for reference.

Proposition 1: Let $\Lambda\Sigma\Theta^\top$ be the SVD of a matrix $\mathbf{D} \in \mathbb{R}^{l\times q}$ satisfying $\Lambda\Lambda^\top = \mathbf{I}$ and $\Theta\Theta^\top = \mathbf{I}$ for the identity matrix $\mathbf{I}$, and assume $\Sigma = diag(\{\sigma_j\}_{1\le j\le r})$ with $r = \min(l, q)$. For each of positive singular value σ_j and $\kappa > 0$, we provide the singular value function shrinkage operators of the problem

$$\min_{\mathbf{E}} \kappa \|\mathbf{E}\|_{S_p}^p + \frac{1}{2}\|\mathbf{E} - \mathbf{D}\|_F^2, \tag{22}$$

where the optimal solver, denoted as $\mathbf{E}^*$, for $p = 1$, $2/3$, and $1/2$, will be respectively represented as below.

- For $p = 1$, we have

$$\mathbf{E}_* = \mathcal{S}_\kappa(\mathbf{D}) = \Lambda diag(\{\max(\sigma_i - \kappa, 0)\})\Theta^\top. \tag{23}$$

- For $p = 2/3$, we have

$$\mathbf{E}_* = \mathcal{T}_\kappa(\mathbf{D}) = \Lambda diag(\{\theta(\sigma_i)\cdot\xi\})\Theta^\top, \tag{24}$$

where

$$\begin{cases} \theta(\sigma_i) = ((\varpi + \sqrt{2\sigma_i/\varpi - \varpi^2})/2)^3, \\ \varpi = \dfrac{2}{\sqrt{3}}(2\beta)^{1/4}\cosh(\phi/3)^{1/2}, \\ \phi = \operatorname{arccosh}\left(\dfrac{27\sigma_i^2}{16}(2\beta)^{-1.5}\right), \end{cases}$$
$$\begin{cases} \xi = 1, & \sigma_i > \dfrac{2}{3}(3(2\beta)^3)^{1/4}, \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

- For $p = 1/2$, we have

$$\mathbf{E}_* = \mathcal{W}_\kappa(\mathbf{D}) = \Lambda diag(\{\theta(\sigma_i)\cdot\xi\})\Theta^\top, \tag{26}$$

where

$$\begin{cases} \theta(\sigma_i) = \dfrac{2}{3}\sigma_i\left(1+\cos\left(\dfrac{2\pi}{3} - \dfrac{2}{3}\varphi\right)\right), \\ \varphi = \arccos\left(\dfrac{\beta}{4}(\dfrac{\sigma_i}{3})^{-1.5}\right), \end{cases}$$
$$\begin{cases} \xi = 1, & \sigma_i > \dfrac{\sqrt[3]{54}}{4}(2\eta)^{2/3}, \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

where the closed-form solutions for the involved Schatten-p norm subproblems can be obtained by using equations (23), (24), and (26). These equations provide the necessary operations to compute the updated variables, namely $\mathbf{E}$, $\widehat{\mathbf{V}}$, and $\widehat{\mathbf{U}}$. By leveraging these closed-form solutions, an effective optimization process is achieved, which facilitates faster convergence and improves computational efficiency.

Given the variables at the k-th iteration, i.e., $\{\mathbf{U}^k, \mathbf{V}^k, \widehat{\mathbf{V}}^k, \widehat{\mathbf{U}}^k, \mathbf{X}^k, \mathbf{E}^k\}$, we update the variables at the $(k+1)$-th iteration by minimizing (19)-(21) with respect to $\mathbf{U}$, $\mathbf{V}$, $\widehat{\mathbf{V}}$, $\widehat{\mathbf{U}}$, $\mathbf{E}$, and $\mathbf{X}$. Subsequently, we update the dual variables according to the following rules:

$$\begin{cases} \Gamma_1^{k+1} = \Gamma_1^k + \mu^k(\mathbf{E}^{k+1} + \mathbf{A}\mathbf{X}^{k+1} - \mathbf{Y}), & (28) \\ \Gamma_2^{k+1} = \Gamma_2^k + \mu^k\left(\mathbf{X}^{k+1} - \mathbf{U}^{k+1}(\mathbf{V}^{k+1})^\top\right), & (29) \\ \Gamma_3^{k+1} = \Gamma_3^k + \mu^k\left(\widehat{\mathbf{V}}^{k+1} - \mathbf{V}^{k+1}\right), & (30) \\ \Gamma_4^{k+1} = \Gamma_4^k + \mu^k\left(\widehat{\mathbf{U}}^{k+1} - \mathbf{U}^{k+1}\right), & (31) \end{cases}$$

where the closed-form solutions for the $(k+1)$-th iteration can be easily obtained using ***Proposition*** 1. These solutions involve matrix computations that include derivatives and multiplications based on equations (28)-(31).

A. Updating $\mathbf{U}^{k+1}$ and $\mathbf{V}^{k+1}$

By optimizing (19), we compute the derivatives with respect to $\mathbf{U}$ and $\mathbf{V}$ and set them to zero. This allows us to find the updated variables $\mathbf{U}^{k+1}$ and $\mathbf{V}^{k+1}$ by solving

$$\begin{aligned} &\arg\min_{\mathbf{U}} \frac{\lambda}{2}\|\mathbf{U}\|_F^2 + f_{2,\mu^k}(\Gamma_2^k, \mathbf{X}^k, \mathbf{U}, \mathbf{V}^k) \\ &= \left(\mathbf{X}^k + \frac{\Gamma_2^k}{\mu^k}\right)\mathbf{V}^k\left(\frac{\lambda\mathbf{I}}{\mu^k} + (\mathbf{V}^k)^\top\mathbf{V}^k\right)^{-1}, \end{aligned} \tag{32}$$

$$\begin{aligned} &\arg\min_{\mathbf{V}} \frac{\lambda}{2}\|\mathbf{V}\|_F^2 + f_{2,\mu^k}(\Gamma_2^k, \mathbf{X}^k, \mathbf{U}^{k+1}, \mathbf{V}) \\ &= \left(\mathbf{X}^k + \frac{\Gamma_2^k}{\mu^k}\right)^\top\mathbf{U}^{k+1}\left(\frac{\lambda\mathbf{I}}{\mu^k} + (\mathbf{U}^{k+1})^\top\mathbf{U}^{k+1}\right)^{-1}. \end{aligned} \tag{33}$$

Similar to (32) and (33), we optimize (20) and obtain the updates for $\mathbf{U}^{k+1}$ and $\mathbf{V}^{k+1}$ through

$$\begin{aligned} &\arg\min_{\mathbf{U}} \frac{\lambda}{3}\|\mathbf{U}\|_F^2 + f_{2,\mu^k}(\Gamma_2^k, \mathbf{X}^k, \mathbf{U}, \mathbf{V}^k) \\ &= \left(\mathbf{X}^k + \frac{\Gamma_2^k}{\mu^k}\right)\mathbf{V}^k\left(\frac{2\lambda\mathbf{I}}{3\mu^k} + (\mathbf{V}^k)^\top\mathbf{V}^k\right)^{-1}, \end{aligned} \tag{34}$$

$$\begin{aligned} &\arg\min_{\mathbf{V}} f_{2,\mu^k}(\Gamma_2^k, \mathbf{X}^k, \mathbf{U}^{k+1}, \mathbf{V}) + f_{3,\mu^k}(\Gamma_3^k, \mathbf{V}, \widehat{\mathbf{V}}^k) \\ &= \left[\left(\widehat{\mathbf{V}}^k + \frac{\Gamma_3^k}{\mu^k}\right) + \left(\mathbf{X}^k + \frac{\Gamma_2^k}{\mu^k}\right)^\top\mathbf{U}^{k+1}\right] \\ &\quad\times\left(\mathbf{I} + (\mathbf{U}^{k+1})^\top\mathbf{U}^{k+1}\right)^{-1}. \end{aligned} \tag{35}$$

By optimizing (21), we update $\mathbf{U}^{k+1}$ based on the computational method described in (35), resulting in

$$\begin{aligned} &\arg\min_{\mathbf{U}} f_{2,\mu^k}(\Gamma_2^k, \mathbf{X}^k, \mathbf{U}, \mathbf{V}^k) + f_{4,\mu^k}(\Gamma_4^k, \mathbf{U}, \widehat{\mathbf{U}}^k) \\ &= \left[\left(\widehat{\mathbf{U}}^k + \frac{\Gamma_4^k}{\mu^k}\right) + \left(\mathbf{X}^k + \frac{\Gamma_2^k}{\mu^k}\right)\mathbf{V}^{k+1}\right] \\ &\quad\times\left(\mathbf{I} + (\mathbf{V}^k)^\top\mathbf{V}^k\right)^{-1}. \end{aligned} \tag{36}$$

In particular, when updating both $\mathbf{U}^{k+1}$ and $\mathbf{V}^{k+1}$, it is advantageous to incorporate an adaptive rank tuning technique [41], [49]. This technique helps estimate the value of d, which corresponds to the rank of the low-rank matrix. By selecting a suitable rank value, we can strike a balance between computational complexity and clustering performance.

B. Updating $\widehat{\mathbf{U}}^{k+1}$ and $\widehat{\mathbf{V}}^{k+1}$

In the optimization problem (19), the variables $\widehat{\mathbf{U}}^{k+1}$ and $\widehat{\mathbf{V}}^{k+1}$ are not explicitly involved. However, in the optimization problem (20), only $\widehat{\mathbf{V}}^{k+1}$ is included. Furthermore, in the optimization problem (21), both $\widehat{\mathbf{U}}^{k+1}$ and $\widehat{\mathbf{V}}^{k+1}$ are present. These auxiliary variables are introduced to simplify the optimizations and enable closed-form solutions.

Fixing other unrelated variables and combining (23) with (20), we can observe that updating $\widehat{\mathbf{V}}^{k+1}$ involves

$$\begin{aligned} &\arg\min_{\widehat{\mathbf{V}}} \frac{2\lambda}{3}\|\widehat{\mathbf{V}}\|_* + f_{3,\mu^k}(\Gamma_3^k, \mathbf{V}^{k+1}, \widehat{\mathbf{V}}) \\ &= \mathcal{S}_{2\lambda/3\mu^k}\left(\mathbf{V}^{k+1} - \frac{\Gamma_3^k}{\mu^k}\right). \end{aligned} \tag{37}$$

By optimizing (21) to update $\widehat{\mathbf{V}}^{k+1}$ and $\widehat{\mathbf{U}}^{k+1}$, we can derive the following solution formulas

$$\arg\min_{\widehat{\mathbf{V}}} \frac{\lambda}{2}\|\widehat{\mathbf{V}}\|_* + f_{3,\mu^k}(\Gamma_3^k, \mathbf{V}^{k+1}, \widehat{\mathbf{V}})$$

$$= \mathcal{S}_{\lambda/2\mu^k}\left(\mathbf{V}^{k+1} - \frac{\Gamma_3^k}{\mu^k}\right), \tag{38}$$

$$\arg\min_{\widehat{\mathbf{U}}} \frac{\lambda}{2}\|\widehat{\mathbf{U}}\|_* + f_{4,\mu^k}(\Gamma_4^k, \mathbf{U}^{k+1}, \widehat{\mathbf{U}})$$
$$= \mathcal{S}_{\lambda/2\mu^k}\left(\mathbf{U}^{k+1} - \frac{\Gamma_4^k}{\mu^k}\right). \tag{39}$$

C. Updating $\mathbf{X}^{k+1}$ and $\mathbf{E}^{k+1}$

Fixing the unrelated variables and computing the derivatives with respect to $\mathbf{X}$, we can set the derivative value to zero to obtain the following closed-form solutions for (19)-(21) based on the $(k+1)$-th iterate for $\mathbf{X}^{k+1}$:

$$\arg\min_{\mathbf{X}} f_{1,\mu^k}(\Gamma_1, \mathbf{E}, \mathbf{X}) + f_{2,\mu^k}(\Gamma_2^k, \mathbf{X}, \mathbf{U}^{k+1}, \mathbf{V}^{k+1})$$
$$= \left(\mathbf{I} + \mathbf{A}^\top\mathbf{A}\right)^{-1} \times \left[\mathbf{A}^\top\left(Y - \mathbf{E}^k - \frac{\Gamma_1^k}{\mu^k}\right) + \left(\mathbf{U}^{k+1}(\mathbf{V}^{k+1})^\top - \frac{\Gamma_2^k}{\mu^k}\right)\right], \tag{40}$$

where $\mathbf{A} = [\mathbf{A}_1(:), \mathbf{A}_2(:), \ldots, \mathbf{A}_n(:)] \in \mathbb{R}^{lq\times m}$ is defined as the data matrix generated from all the training samples. Furthermore, we observe that $\mathbf{I} + \mathbf{A}^\top\mathbf{A}$ remains unchanged when the databases are fixed. Therefore, it can be computed outside the iteration loop to avoid repeated computations.

By fixing other unrelated terms with $\mathbf{E}_i^k$, we reformulate equation (15) as follows

$$\sum_{i=1}^{n} f_{1,\mu^k}(\Gamma_{1,i}^k, \mathbf{E}_i^k, \mathbf{X}_i) \hat{=} \frac{\mu^k}{2}\|\mathbf{E}^k + \mathbf{A}\mathbf{X} - \mathbf{Y} + \frac{\Gamma_1^k}{\mu^k}\|_F^2. \tag{41}$$

where we can rewrite $f_{1,\mu^k}(\Gamma_{1,i}^k, \mathbf{E}_i, \mathbf{X}_i^{k+1}) = \frac{\mu^k}{2}\|\mathcal{M}(\mathbf{E}_i) - h_{\mu^k}(\mathbf{X}_i^{k+1}, \Gamma_{1,i}^k)\|_F^2$ with $h_{\mu^k}(\mathbf{X}_i^{k+1}, \Gamma_{1,i}^k) = \mathcal{M}(\mathbf{Y}_i) - \mathcal{A}(\mathbf{X}_i^{k+1}) - \frac{\mathcal{M}(\Gamma_{1,i}^k)}{\mu^k}$. By combining equations (19)-(21) and (41), we can transform the sub-problems for iteratively updating $\mathbf{E}^{k+1}$ into the following optimization problem

$$\arg\min_{\mathcal{M}(\mathbf{E}_i)} \|\mathcal{M}(\mathbf{E}_i)\|_{S_p}^p + f_{1,\mu^k}(\Gamma_{1,i}^k, \mathbf{E}_i, \mathbf{X}_i^{k+1}), \tag{42}$$

where, based on ***Proposition*** 1, we can obtain the analytic solvers for (42) with three different p-values as follows

$$\mathcal{M}(\mathbf{E}_i^{k+1}) = \begin{cases} \mathcal{S}_{1/\mu^k}\left(h_{\mu^k}(\mathbf{X}_i^{k+1}, \Gamma_{1,i}^k)\right), & p=1, \\ \mathcal{T}_{1/\mu^k}\left(h_{\mu^k}(\mathbf{X}_i^{k+1}, \Gamma_{1,i}^k)\right), & p=2/3, \\ \mathcal{W}_{1/\mu^k}\left(h_{\mu^k}(\mathbf{X}_i^{k+1}, \Gamma_{1,i}^k)\right), & p=1/2. \end{cases} \tag{43}$$

After obtaining the vectors $\mathbf{E}_i^{k+1}$ for $1 \le i \le n$ from (43) in parallel to improve computational efficiency, we can construct the residual matrix $\mathbf{E}^{k+1}$ by concatenating these individual vectors. Specifically, we form $\mathbf{E}^{k+1}$ by horizontally combining the vectors: $\mathbf{E}^{k+1} = [\mathbf{E}_1^{k+1}, \mathbf{E}_2^{k+1}, \ldots, \mathbf{E}_n^{k+1}]$. This updated matrix $\mathbf{E}^{k+1}$ will be utilized in the subsequent iterations.

In summary, the iteration procedure for solving problems (12)-(14) can be outlined in **Algorithm** 1. The algorithm initializes the Lagrange multipliers, primal variables, and parameters according to the technique suggested in [41], [42], and [44]. The algorithm continues iterating until the stopping criteria are met, which is determined by the following condition:

$$\frac{\|\mathbf{E}^{k+1} + \mathbf{A}\mathbf{X}^{k+1} - \mathbf{Y}\|_F}{\|\mathbf{Y}\|_F} < \epsilon, \tag{44}$$

where $0 < \epsilon \ll 1$ is a pre-defined threshold value. To evaluate the classification accuracy, we design a classification criterion using the Schatten-p norm with three different p-values, as proposed in [12] and [13]. Based on the coefficient matrix $\mathbf{X}^*$, we calculate the class-wise error matrices for all the testing samples and assign the final label to the i-th testing sample $\mathbf{Y}_i$. The final label can be determined by

$$Label(\mathbf{Y}_i) = \text{argmin}_j \frac{\|\text{Mat}(\mathbf{Y}_i) - \mathcal{A}_{c_j}(\mathbf{X}_{i,c_j}^*)\|_{S_p}^p}{\|\mathbf{X}_{i,c_j}^*\|_2}, \tag{45}$$

where $\mathbf{X}_{i,c_j}^*$ represents the coefficient vector associated with the i-th testing sample for the j-th class, denoted as c_j, training samples, and $\mathcal{A}_{c_j}(\mathbf{X}_{i,c_j}^*)$ represents the reconstructed image matrix for class c_j. The computable $Label(\mathbf{Y}_i)$ is determined by selecting the value on the right side of the "=" sign that is the smallest among all $1 \le j \le n$, indicating the class with the lowest reconstruction error for $\mathbf{Y}_i$.

Algorithm 1 Optimization for Problem (11)

Input: $\mathbf{Y}$, $\mathbf{X}^1$, $\mathbf{U}^1$, $\mathbf{V}^1$, $\widehat{\mathbf{V}}^1$, $\widehat{\mathbf{U}}^1$, and $\{\Gamma_i^0\}_{1\le i\le 4}$
Parameter: λ, p, $d \ge r$, $\rho = 1.1$, μ^0 and $k = 0$
Output: $\mathbf{X}^* \leftarrow \mathbf{X}^{k+1}$

while not converged **do**
  if $p = 1$
    update $\mathbf{U}^{k+1}$ and $\mathbf{V}^{k+1}$ by (32) and (33),
    update $\mathbf{X}^{k+1}$ by (22),
    update $\mathbf{E}^{k+1}$ by (43) with $p = 1$,
    update Γ_1^{k+1} and Γ_2^{k+1} by (28) and (29),
  elseif $p = 2/3$
    update $\mathbf{U}^{k+1}$ and $\mathbf{V}^{k+1}$ by (34) and (35),
    update $\widehat{\mathbf{V}}^{k+1}$ by (37),
    update $\mathbf{X}^{k+1}$ by (22),
    update $\mathbf{E}^{k+1}$ by (43) with $p = 2/3$,
    update Γ_1^{k+1}, Γ_2^{k+1} and Γ_3^{k+1} by (28)-(30),
  else
    update $\mathbf{U}^{k+1}$ and $\mathbf{V}^{k+1}$ by (36) and (35),
    update $\widehat{\mathbf{V}}^{k+1}$ and $\widehat{\mathbf{U}}^{k+1}$ by (38) and (39),
    update $\mathbf{X}^{k+1}$ by (22),
    update $\mathbf{E}^{k+1}$ by (43) with $p = 1/2$,
    update Γ_1^{k+1}, Γ_2^{k+1}, Γ_3^{k+1} and Γ_4^{k+1} by (28)-(31),
  end
end while

IV. Theoretical Analysis

In this section, we present the computational complexity of Algorithm 1, and then present the algorithmic convergence property. The detailed analysis was given as below.

- For the computational complexity in (32)-(43)., it mainly depends on the computations of the SVD of the error matrix Mat($\mathbf{E}_i$) and the factorized matrix $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ along with the matrix multiplications for updating $\mathbf{X}$, $\mathbf{U}$, and $\mathbf{V}$, respectively. For the studied cases, the complexity of updating $\mathbf{E}$ and $\mathbf{X}$ is $o(nlq^2)$ and $o(mn\varsigma + m^2n + mdn)$ for $l \geq q$ and $\varsigma = lq$, For $p = 2/3$, the complexity of computing $\widehat{\mathbf{V}}$ is $o(nd^2)$, while for $p = 1/2$, the total complexity of computing $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{U}}$ is $o((m+n)d^2)$. Then, we give the complexity of updating $\mathbf{U}$ and $\mathbf{V}$ as follows
  - the computational complexity of updating $\mathbf{U}$ is the same, i.e., $o(d^3 + 2nd^2 + mdn)$, for the three p-values.
  - the computational complexity of updating $\mathbf{V}$ is $o(d^3 + 2md^2 + mdn)$ for $p = 1$ as well as $o(d^3 + (n+m)d^2 + mdn)$ for both $p = 2/3$ and $p = 1/2$, respectively.
- For the convergence property, we theoretically prove the boundedness of generated variable sequences and the satisfied Karush-Kuhn-Tucker (KKT) conditions.
  - We prove the boundedness of the generated variable sequences, both the dual variables $\{\Gamma_i^{k+1}\}_{i=1}^4$ and the primal variables involved in (12)-(14). This shows that these variable sequences remain within certain bounds, ensuring the stability of the optimization process.
  - The global convergence of the generated variable subsequences are proved theoretically, which can be exploited to characterize the cluster point.

We next present a brief overview of two important concepts: the dual norm [33] and the sub-differential of the singular value function [9], [50]. Understanding these concepts is essential for analyzing the convergence of our proposed methods.

Lemma 1: Let $\mathcal{H}$ be a real Hilbert space endowed with an inner product $\langle \cdot, \cdot \rangle$ and a corresponding norm $\| \cdot \|$, and $\mathbf{z} \in \partial \|\mathbf{y}\|$, where $\partial f(\mathbf{y})$ denotes the sub-gradient of $f(\mathbf{y})$. Then $\|\mathbf{z}\|^* = 1$ if $\mathbf{X} \neq 0$, and $\|\mathbf{z}\|^* \leq 1$ if $\mathbf{X} = 0$, where $\|\mathbf{z}\|^*$ is the dual norm of $\|\mathbf{z}\|$. For example, the dual norm of the nuclear norm is the spectral norm $\| \cdot \|_2$, i.e., the largest singular value of given matrix.

Lemma 2: Let $F(\mathbf{X}) : \mathbb{R}^{l \times m} \to \mathbb{R}$ be the corresponding singular value function $f \circ \sigma(\mathbf{X})$ at a matrix $\mathbf{X}$, and $f(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is an absolutely symmetric function, then assume that $\mathbf{X} = \Lambda \Sigma \Theta^\top$ is the SVD of $\mathbf{X}$, then the sub-differential of $F(\cdot)$ at $\mathbf{X}$ is given by the following formula

$$\frac{\partial F(\mathbf{X})}{\partial \mathbf{X}} = \partial(f \circ \sigma)(\mathbf{X}) = \Lambda \widehat{\Sigma} \Theta^\top, \tag{46}$$

where $\widehat{\Sigma}_{ii} = \frac{\partial f(\mathbf{y})}{\partial \mathbf{y}}|_{\mathbf{y}=\Sigma_{ii}}$ holds for $1 \leq ii \leq m$.

Proposition 2: Let $(\mathbf{U}^k, \mathbf{V}^k, \widehat{\mathbf{V}}^k, \widehat{\mathbf{U}}^k, \mathbf{X}^k, \mathbf{E}^k)$ be the multi-variable sequence as well as the dual variable sequence $\{\Gamma_i^{k+1}\}_{i=1}^4$ generated from Algorithm 1, and we assume that $\mu^k(\mathbf{E}^k - \mathbf{E}^{k+1})$ is bounded. Then, we have the assertions:

(***i***) the dual variables $\{\Gamma_i^{k+1}\}_{i=1}^4$ are bounded, and

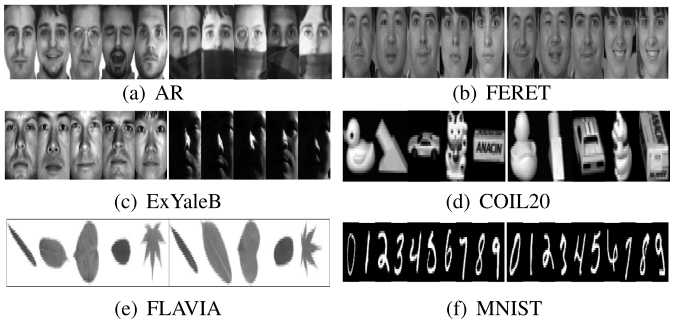
(***ii***) the primal ones $(\mathbf{U}^k, \mathbf{V}^k, \widehat{\mathbf{V}}^k, \widehat{\mathbf{U}}^k, \mathbf{X}^k, \mathbf{E}^k)$ involved in the problems (12)-(14) are also bounded for three cases.

Theorem 2: Let $(\mathbf{U}^k, \mathbf{V}^k, \widehat{\mathbf{V}}^k, \widehat{\mathbf{U}}^k, \mathbf{X}^k, \mathbf{E}^k)$ be a variable sequence along with the dual variable sequence $\{\Gamma_i^{k+1}\}_{i=1}^4$ generated by **Algorithm** 1, and assume the same conditions as stated in ***Proposition*** 2. Under the additional condition that $\mu^k(\mathcal{T}^k - \mathcal{T}^{k+1}) \to \mathbf{0}$ as $k \to +\infty$, where $\mathcal{T}^k = (\mathbf{V}^k, \mathbf{X}^k, \mathbf{E}^k)$, any cluster point, denoted as $(\mathbf{U}^*, \mathbf{V}^*, \widehat{\mathbf{V}}^*, \widehat{\mathbf{U}}^*, \mathbf{X}^*, \mathbf{E}^*)$ and $\{\Gamma_i^*\}_{i=1}^4$, obtained from the variable sequence, is a stationary point with respect to (19)-(21). Moreover, the choices of these variables satisfy the KKT conditions for the cases of $p = 1, 2/3$, and $1/2$.

It should be mentioned that we provide the detailed proofs of the theoretical results, i.e., ***Proposition*** 2 and ***Theorem*** 2, in the supplementary materials.

TABLE I
STATISTIC DESCRIPTIONS OF INVOLVED SIX EXPERIMENTAL DATABASES

Dataset	Styles	Subjects	Size	Training	Testing
(a) AR	face	100	60×43	2580×700	2580×600
(b) FERET	face	200	40×40	1600×1000	1600×400
(c) ExYaleB	face	38	96×84	8064×266	8064×722
(d) COIL20	object	20	32×32	1024×300	1024×300
(e) FLAVIA	object	32	30×40	1200×800	1200×800
(f) MNIST	digital	10	28×28	784×1000	784×1000



(a) AR (b) FERET (c) ExYaleB (d) COIL20 (e) FLAVIA (f) MNIST

Fig. 3. Partial images from six databases (faces, objects, and digital images) with various variations such as occlusions and illuminations. The images on the left are for training, and the ones on the right are for testing.

V. NUMERICAL EXPERIMENTS

This section starts by providing statistics on six widely used image databases, which are summarized in TABLE I. The selected images from these databases are also showcased in Fig. 3 (a)-(f). Subsequently, we provide comprehensive descriptions of each experimental image database, highlighting their unique characteristics and applications.

- The AR database contains images of 100 individuals, including samples with scarf occlusions, making it a challenging dataset for face recognition tasks. The FERET database consists of images with frontal, left, or right profile views, showcasing variations in pose, expression, and lighting conditions. We specifically selected a subset of 1400 images, representing 200 individuals. The ExYaleB database comprises images of 38 subjects captured under various illuminations, providing valuable insights into illumination robustness in face recognition.
- The COIL20 object database consists of 600 images depicting 20 different objects. Each image features the main body of the object centered against a black

background, and multiple horizontal angles are captured, providing diverse views for each object. The FLAVIA database focuses on leaf classification and contains images of leaves from various plant species. Each leaf species exhibits unique variations in shape, length, and width, making it a suitable dataset for studying leaf classification tasks. We specifically selected 1600 images representing 32 different leaf species. The MNIST database is a well-known dataset widely used for handwriting digit recognition. It includes multiple handwritten digits ranging from 0 to 9, making it a valuable resource for training and evaluating digit recognition algorithms.

The selected databases used in our experiments are publicly available and can be easily loaded. We conducted the experiments without any pre-processing, using the training and testing samples directly. The experiments were performed on a 64-bit PC with an Intel(R) Core(TM) i7-7700 CPU@3.6GHz and 8.0GB RAM, utilizing MATLAB R2021b. To compare our proposed SpNFLMR approach, we implemented and compared several mostly related methods, including CRC [12], NMR with its faster version (FNMR) [13], ULR$_*$ [14], ULR$_{**}$ [14], ALPR [15], DLRSR [16], NLRℓ_{21} [17], RSLDA [18], and generalized iterated shrinkage algorithm (GISA) with three p-values [24]. Note that ULR$_*$ and ULR$_{**}$ are specific instances of the ULR model [14], and GISA$_1$ and SRC share the same model formulations while utilizing different optimization algorithms. For conducting the comparisons, we meticulously fine-tuned the parameters of these methods to achieve the best possible performance and computational efficiency. In contrast, our proposed SpNFLMR approach, for $p = 1$, $2/3$, and $1/2$, consistently demonstrated effectiveness and robustness across all three scenarios.

A. Experiments on Face Databases

The results presented in TABLE II provide a comprehensive comparison between our proposed methods and ten other approaches across three face databases. Upon analyzing the results, it becomes apparent that CRC, ULR, and GISA exhibit lower timing costs due to their differentiable objective functions and overall approach to processing testing data. In contrast, methods such as NMR and FNMR incur higher time costs as they compute the coefficient representation for each individual testing data independently. Our proposed methods, along with NLRℓ_{21}, NMR, and FNMR, effectively leverage the low-rank structure information, resulting in improved classification accuracy while consistently demonstrating lower timing costs. This highlights the higher efficiency of our factorization strategies and validates the superior classification accuracy they offer. On the other hand, ALPR, DLSR, and RSLDA achieve lower classification accuracy compared to some of the comparative methods due to their reliance on inexact measurements of the residual function in these datasets. This suggests that our methods are better suited for capturing the inherent structure and leading to enhanced accuracy.

To further validate the effectiveness of our proposed methods, Fig. 4 (a) shows the curves of the stopping function on the AR database, which exhibit a desirable non-increasing convergence property. To gain deeper insights into the nonconvex nature of our methods, Fig. 4 (b) presents the distribution of classification accuracy across 500 runs on the FERET database, considering random initial variables. Analyzing the spread of accuracy values allows us to assess the robustness and stability in different initialization scenarios. Furthermore, in Fig. 4 (c), we visualize the timing costs associated with various parameter settings, focusing on the ExYaleB database. This analysis helps us evaluate the computational efficiency

TABLE II

CLASSIFICATION ACCURACY (%) AND COMPUTATION TIMES OF THE METHODS USED ON THREE FACE DATABASES

Methods	AR		FERET		ExYaleB	
	Acc	Time	Acc	Time	Acc	Time
CRC	67.50	3.9s	70.50	2.2s	48.50	4.6s
NMR	70.67	323.8s	**78.25**	205.4s	48.07	550.5s
FNMR	70.17	210.5s	*77.75*	143.6s	46.81	294.8s
ALPR	52.00	156.7s	69.25	226.2s	24.59	698.9s
DLRSR	60.17	27.7s	72.00	11.3s	26.55	330.5s
ULR$_*$	66.83	10.6s	70.50	8.5s	49.31	11.9s
ULR$_{**}$	67.33	13.7s	71.50	13.1s	49.17	15.7s
NLRℓ_{21}	63.33	295.6s	71.00	378.9s	**95.71**	204.5s
GISA$_1$	65.50	5.8s	77.25	3.9s	27.29	6.1s
GISA$_{2/3}$	65.67	5.7s	73.50	4.1s	27.29	6.4s
GISA$_{1/2}$	66.00	5.8s	70.50	4.3s	27.29	6.1s
RSLDA	59.18	210.6s	73.25	139.7s	*57.75*	1263.6s
S$_1$NFLMR	73.50	15.7s	69.75	8.9s	42.52	61.3s
S$_{2/3}$NFLMR	**81.50**	45.4s	76.75	23.1s	*63.44*	126.1s
S$_{1/2}$NFLMR	*76.33*	69.5s	72.25	35.5s	51.94	168.9s

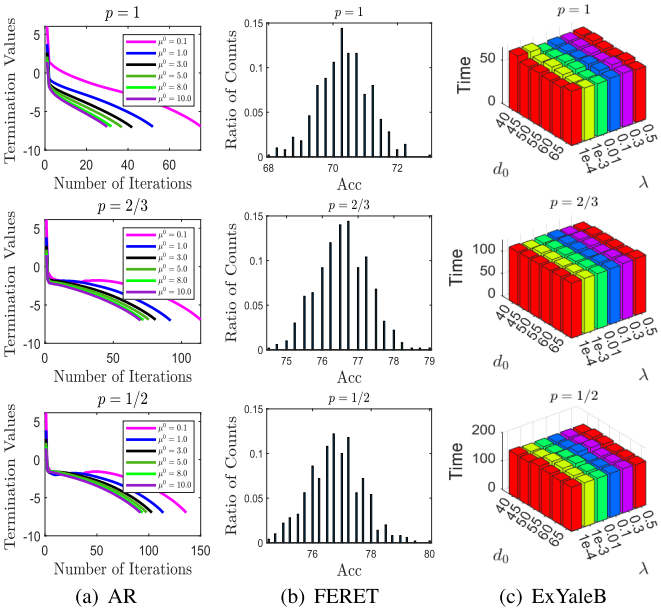


Fig. 4. Visual comparisons of the influences of our methods with three different p-values are shown for three face databases. The comparisons take into account plotted curves, random initial variables, and model parameters.

TABLE III

CLASSIFICATION ACCURACY (%) AND COMPUTATION TIMES OF THE METHODS USED ON THE OBJECT AND DIGITAL DATABASES

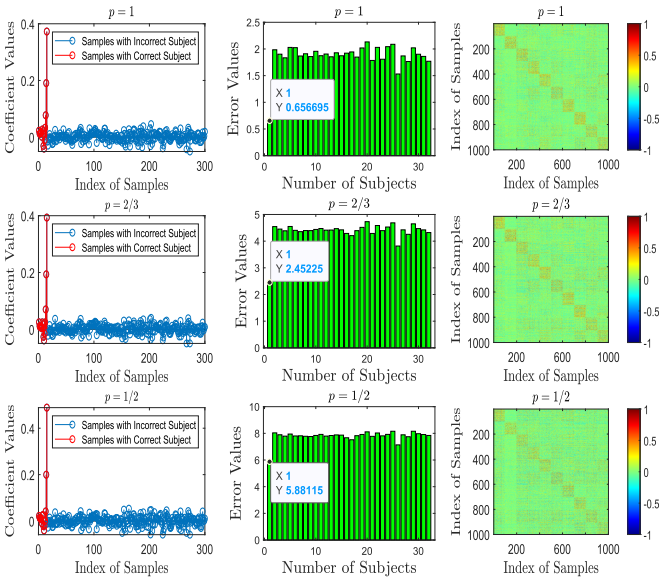
Methods	COIL20		FLAVIA		MNIST	
	Acc	Time	Acc	Time	Acc	Time
CRC	62.33	<1s	58.88	3.2s	87.70	2.1s
NMR	62.67	22.1s	54.50	169.6s	89.20	124.5s
FNMR	62.00	15.6s	54.50	121.1s	85.10	115.6s
ALPR	*65.00*	14.0s	61.50	58.2s	86.60	17.1s
DLRSR	57.00	<1s	55.88	1.8s	83.90	1.5s
ULR_{*}	62.00	3.7s	61.00	24.1s	90.50	37.1s
ULR_{**}	62.33	4.6s	61.63	29.5s	90.60	38.2s
NLRℓ_{21}	69.00	19.2s	51.13	431.1s	87.40	193.5
GISA_1	61.33	<1s	60.75	3.0s	90.20	4.1s
$\text{GISA}_{2/3}$	61.00	<1s	61.38	3.0s	90.10	3.4s
$\text{GISA}_{1/2}$	62.00	<1s	59.25	3.1	89.70	3.8s
RSLDA	58.00	20.1s	62.00	79.3s	**90.70**	46.2s
S_1NFLMR	61.33	4.3s	**76.50**	16.1s	89.80	11.1s
$\text{S}_{2/3}$NFLMR	64.00	9.7s	*64.13*	53.6s	87.40	26.4s
$\text{S}_{1/2}$NFLMR	*65.00*	11.2s	60.75	82.2s	81.80	35.3s

with examining the timing costs under different parameter configurations, we can make informed decisions to strike a balance between accuracy and computational efficiency.

B. Experiments on Object and Digital Databases

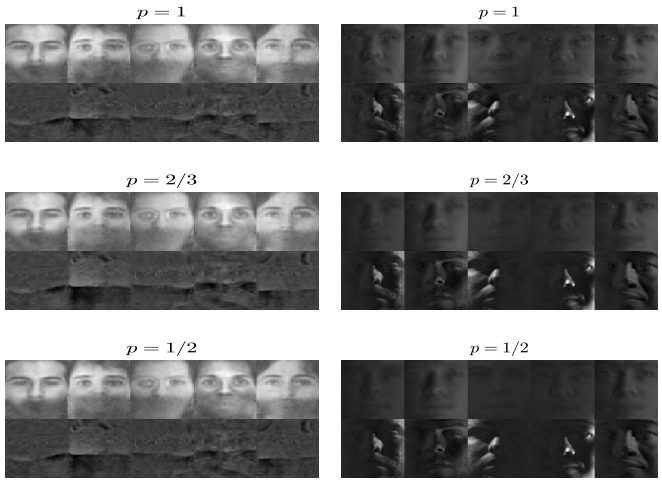
The results listed in TABLE III provide quantitative comparisons for object and digital databases. The compared methods, including CRC, DLSR, ULR, and GISA, have higher computational efficiency. Here, ALPR, NMR, NLRℓ_{21}, and FNMR exhibit lower computational efficiency. Overall, our methods consistently achieve lower timing costs while improving accuracy compared to NMR and FNMR. However, there are instances where our methods do not surpass all the comparison methods, and this can be attributed to the utilization of additional information, such as ALPR, ULR, NLRℓ_{21}, DLRSR, and RSLDA. This phenomenon could be attributed to the absence of occlusions or illuminations in the testing images, which are the factors in which our methods excel to a certain extent.

To gain further insights, we conducted additional investigations to explore different perspectives. In Fig. 5 (a), we illustrate the distributions of representation coefficients obtained by our proposed methods for three different values of p on the COIL20 dataset. This analysis allows us to understand the impact of varying p on the coefficient matrix. Furthermore, in Fig. 5 (b), we analyze the reconstructed errors on the FLAVIA database to validate the effectiveness of our methods. The reconstructed errors for the same subjects as the testing samples exhibit lower values, while errors for different subjects show higher values. This demonstrates the ability of our methods to capture subject-specific information. Moreover, Fig. 5 (c) showcases the block-diagonal structures



(a) COIL20 (b) FLAVIA (c) MNIST

Fig. 5. Visual comparisons derived from the coefficient matrices generated by our methods using three different p-values on object and digital databases.



(a) AR (b) ExYaleB

Fig. 6. Visual comparisons of the reconstruction and residual images for three p-values, focusing on (a) the AR database and (b) the ExtYaleB database.

with normalization on the MNIST database, providing insights into the influence of the coefficient matrix.

C. Further Analysis and Discussion

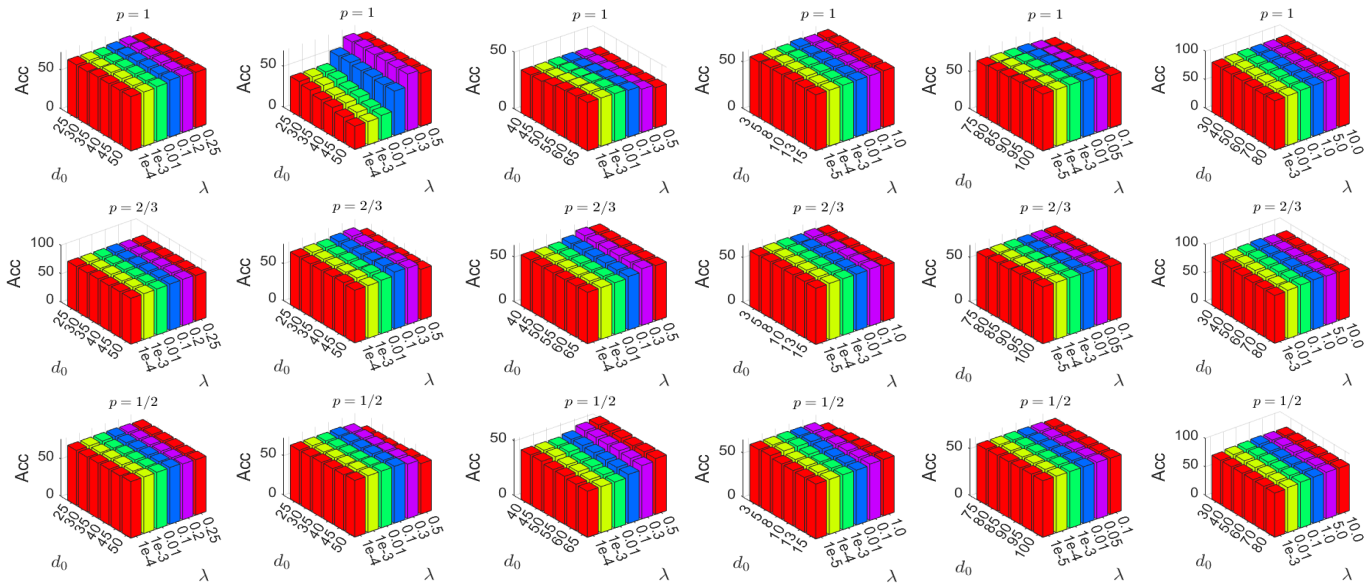
This subsection first performed an ablation analysis on six databases to assess the effectiveness and efficiency of the modules used in our objective function. Additionally, we investigated the capabilities of our methods in face reconstruction and noise removal tasks, while also exploring the impact of parameter sensitivity on classification accuracy. These findings were thoroughly analyzed and discussed below.

- *Reconstruction Validation*: To demonstrate the effectiveness of our S_pNFLMR methods, we provide visual examples in Fig. 6 (a) and (b). These examples showcase a series of reconstructed images and their corresponding error images from the AR and ExYaleB databases.

TABLE IV

COMPARISONS OF ACCURACY (%) AND COMPUTATION TIMES OF THE METHODS UNDER VARIOUS ABLATION SETTINGS ACROSS SIX DATABASES

//	AR		FERET		ExtYaleB		COIL20		FLAVIA		MNIST	
	Acc	Time	Acc	Time	Acc	Time	Acc	Time	Acc	Time	Acc	Time
(a)	73.50	15.7s	69.75	8.9s	42.52	61.3s	61.33	4.3s	**76.50**	16.1s	**89.80**	11.1s
(b1)	*78.83*	35.9s	71.25	20.5s	*52.91*	121.9s	*65.67*	8.1s	*74.13*	33.2s	*88.40*	26.9s
(b2)	76.67	43.0s	*72.25*	23.1s	**57.48**	149.1s	**67.00**	10.3s	71.10	40.6s	87.50	33.6s
(c1)	**79.17**	21.9s	**75.25**	10.2s	39.19	61.5s	61.67	4.5s	65.13	26.2s	87.90	11.9s
(c2)	73.83	26.7s	65.75	13.5s	35.87	66.5s	62.67	5.5s	61.50	35.1s	83.70	13.6s



(a) AR (b) FERET (c) ExtYaleB (d) COIL20 (e) FLAVIA (f) MNIST

Fig. 7. Visual comparisons of proposed methods across six databases to analyze the effects of different parameter pairs (p, d_0, λ) on classification accuracy.

By visually inspecting these samples, it becomes evident that our methods successfully remove occlusions and illuminations from the testing images, attesting to their capability in preserving image fidelity.

- *Ablation Analysis*: The results of the ablation studies are presented in TABLE IV. The baseline method, denoted as (a) and represented by equation (12), serves as the reference. In the first set of experiments, we fixed the coefficient regularization term and replaced the residual term with equations (13) and (14), denoted as (b1) and (b2), respectively. In the second set, we fixed the residual term and substituted the regularization term with equations (13) and (14), denoted as (c1) and (c2), respectively. The results indicate that each module has a minor influence on both the classification accuracy and timing cost. Notably, when compared to the baseline (a), both (c1) and (c2) exhibit more significant improvements in timing costs compared to (b1) and (b2).
- *Parameter Sensitivity*: We conducted a thorough parameter analysis, and the results are presented in Fig. 7 (a)-(f). This analysis aimed to investigate the effects of different values of (p, λ, d_0) across six databases. For the p parameter, we specifically selected three values and adjusted the regularization parameter λ accordingly for each set. The choice of the rank number of the coefficient matrix, denoted as $d = d_0 \times r_e$, can vary depending on the number of subjects and provides flexibility in the selection process. Automatic estimation of the rank number, denoted as r_e in previous studies [41], [49], becomes crucial in scenarios where the exact rank of the coefficient matrix is unknown. Furthermore, by analyzing the classification accuracy of our S$_p$NFLMR methods, we observed that accuracy varied across different experimental settings while remaining stable overall. This comprehensive analysis of the (p, λ, d_0) values provides valuable insights into the validation and exploration of the potential advantages. Carefully selecting appropriate values for p, λ, and d_0 further enhances the performance of our S$_p$NFLMR methods across various databases.

VI. CONCLUSION AND FUTURE WORK

This study focuses on investigating a structured nonconvex and nonsmooth low-rank matrix regression model, which utilizes the extended matrix variate power exponential distribution. The main objective is to address structured noise, such as occlusions and continuous illustration, by incorporating the residual function and capturing the block-structures of

the coefficient matrix. To achieve this, we introduce the Schatten p-norm and its factorization for three different p-values. This allows us to formulate structured regression problems and develop efficient iteration procedures using the augmented Lagrange function within the multi-variate ADMM framework. Theoretical analysis is performed to establish convergence properties under mild assumptions and the computational complexity is also provided. Additionally, extensive experiments are conducted to demonstrate the superior performance and lower timing cost of our methods compared to several related linear regression approaches.

In our upcoming research endeavors, we have identified two primary avenues for future exploration. Firstly, we plan to extend our efficient optimization algorithms to tackle tensor recovery problems, as examined in [51] and [52]. Secondly, our focus will be on incorporating additional information into our models, including the integration of graph structures [21], [40], [46], discriminative features [18], [53], [54], and latent attributes [55]. These enhancements will be built upon the groundwork laid by prior studies. We anticipate that these efforts will lead to improved performance and greater robustness across a variety of applications.

ACKNOWLEDGMENT

The authors would like to thank the editors and the anonymous reviewers for their critical and constructive comments and suggestions.

REFERENCES

[1] Y. Chen, Z. Lai, W. K. Wong, L. Shen, and Q. Hu, "Low-rank linear embedding for image recognition," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3212–3222, Dec. 2018.

[2] M. Pang, B. Wang, S. Huang, Y. M. Cheung, and B. Wen, "A unified framework for bidirectional prototype learning from contaminated faces across heterogeneous domains," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1544–1557, 2022.

[3] Y. Li, J. Liu, H. Lu, and S. Ma, "Learning robust face representation with classwise block-diagonal structure," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2051–2062, Dec. 2014.

[4] M. Abavisani and V. M. Patel, "Multimodal sparse and low-rank subspace clustering," *Inf. Fusion*, vol. 39, pp. 168–177, Jan. 2018.

[5] S. Li and B. Zhang, "Joint discriminative sparse coding for robust hand-based multimodal recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 3186–3198, 2021.

[6] S. Li, R. Ma, L. Fei, and B. Zhang, "Learning compact multirepresentation feature descriptor for finger-vein recognition," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1946–1958, 2022.

[7] L. Yang, G. Yang, K. Wang, F. Hao, and Y. Yin, "Finger vein recognition via sparse reconstruction error constrained low-rank representation," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4869–4881, 2021.

[8] Z. Zha, B. Wen, X. Yuan, J. Zhou, and C. Zhu, "Image restoration via reconciliation of group sparsity and low-rank models," *IEEE Trans. Image Process.*, vol. 30, pp. 5223–5238, 2021.

[9] H. Zhang, C. Gong, J. Qian, B. Zhang, C. Xu, and J. Yang, "Efficient recovery of low-rank matrix via double nonconvex nonsmooth rank minimization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 2916–2925, Oct. 2019.

[10] J. Wang, C. Lu, M. Wang, P. Li, S. Yan, and X. Hu, "Robust face recognition via adaptive sparse representation," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2368–2378, Dec. 2014.

[11] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

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

[13] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 156–171, Jan. 2017.

[14] H. Zhang, F. Qian, B. Zhang, W. Du, J. Qian, and J. Yang, "Incorporating linear regression problems into an adaptive framework with feasible optimizations," *IEEE Trans. Multimedia*, vol. 25, pp. 4041–4051, 2022, doi: 10.1109/TMM.2022.3171088.

[15] J. Wen, Z. Zhong, Z. Zhang, L. Fei, Z. Lai, and R. Chen, "Adaptive locality preserving regression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 75–88, Jan. 2020.

[16] J. Wen, Y. Xu, Z. Li, Z. Ma, and Y. Xu, "Inter-class sparsity based discriminative least square regression," *Neural Netw.*, vol. 102, pp. 36–47, Jun. 2018.

[17] H. Zhang, J. Gao, J. Qian, J. Yang, C. Xu, and B. Zhang, "Linear regression problem relaxations solved by nonconvex ADMM with convergence analysis," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jul. 4, 2023, doi: 10.1109/TCSVT.2023.3291821.

[18] J. Wen et al., "Robust sparse linear discriminant analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 390–403, Feb. 2019.

[19] Z. Hu, F. Nie, R. Wang, and X. Li, "Low rank regularization: A review," *Neural Netw.*, vol. 136, pp. 218–232, Apr. 2021.

[20] H. Zhang et al., "Generalized nonconvex nonsmooth low-rank matrix recovery framework with feasible algorithm designs and convergence analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 5342–5353, Sep. 2023, doi: 10.1109/TNNLS.2022.3183970.

[21] Y. Wang and L. Wu, "Beyond low-rank representations: Orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering," *Neural Netw.*, vol. 103, pp. 1–8, Jul. 2018.

[22] H. Zhang, F. Qian, F. Shang, W. Du, J. Qian, and J. Yang, "Global convergence guarantees of (A)GIST for a family of nonconvex sparse learning problems," *IEEE Trans. Cybern.*, vol. 52, no. 5, pp. 3276–3288, May 2022.

[23] L. Luo, J. Yang, J. Qian, Y. Tai, and G. F. Lu, "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.

[24] W. Zuo, D. Meng, L. Zhang, X. Feng, and D. Zhang, "A generalized iterated shrinkage algorithm for nonconvex sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern. Recogn (CVPR)*, Dec. 2013, pp. 217–224.

[25] 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.

[26] C. Lu, J. Tang, S. Yan, and Z. Lin, "Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 829–839, Feb. 2016.

[27] X. Lan, M. Ye, S. Zhang, H. Zhou, and P. C. Yuen, "Modality-correlation-aware sparse representation for RGB-infrared object tracking," *Pattern Recognit. Lett.*, vol. 130, pp. 12–20, Feb. 2020.

[28] S. Yin, Y. Sun, J. Gao, Y. Hu, B. Wang, and B. Yin, "Robust image representation via low rank locality preserving projection," *ACM Trans. Knowl. Discovery From Data*, vol. 15, no. 4, pp. 1–22, Aug. 2021.

[29] H. Zhang, J. Yang, J. Xie, J. Qian, and B. Zhang, "Weighted sparse coding regularized nonconvex matrix regression for robust face recognition," *Inf. Sci.*, vols. 394–395, pp. 1–17, Jul. 2017.

[30] H. Zhang et al., "Efficient and effective nonconvex low-rank subspace clustering via SVT-free operators," *IEEE Trans. Circuits Syst. Video Technol.*, early access, May 11, 2023, doi: 10.1109/TCSVT.2023.3275299.

[31] 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, Oct. 2015.

[32] S. Boyd, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.

[33] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," 2010, *arXiv:1009.5055*.

[34] 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, nos. 1–2, pp. 57–79, Jan. 2016.

[35] E. G. Manzano, M. A. G. Villegas, and J. M. Marín, "Multivariate exponential power distributions as mixtures of normal distributions with Bayesian applications," *Commun. Statist. Theory Methods*, vol. 37, no. 6, pp. 972–985, Feb. 2008.

[36] E. G. S. Manzano, M. A. G. Villegas, and J. M. M. Diazaraque, "A matrix variate generalization of the power exponential family of distributions," *Commun. Statist. Theory Methods*, vol. 31, no. 12, pp. 2167–2182, Dec. 2002.

[37] H. Zhang, J. Qian, J. Gao, J. Yang, and C. Xu, "Scalable proximal Jacobian iteration method with global convergence analysis for nonconvex unconstrained composite optimizations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2825–2839, Sep. 2019.

[38] Q. Yao, J. T. Kwok, T. Wang, and T. Y. Liu, "Large-scale low-rank matrix learning with nonconvex regularizers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2628–2643, Nov. 2019.

[39] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2488–2495.

[40] Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4833–4843, Oct. 2018.

[41] F. Shang, J. Cheng, Y. Liu, Z. Q. Luo, and Z. Lin, "Bilinear factor matrix norm minimization for robust PCA: Algorithms and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2066–2080, Sep. 2018.

[42] H. Zhang, J. Yang, F. Shang, C. Gong, and Z. Zhang, "LRR for subspace segmentation via tractable Schatten-p norm minimization and factorization," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1722–1734, May 2019.

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

[44] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Adv. Neural. Inf. Process. Syst. (NIPS)*, 2011, pp. 612–620.

[45] Q. Shen, Y. Liang, S. Yi, and J. Zhao, "Fast universal low rank representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1262–1272, Mar. 2022.

[46] Y. Wang, W. Zhang, L. Wu, X. Lin, M. Fang, and S. Pan, "Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering," in *Proc. Int. Joint. Conf. Artif. Intell. (IJCAI)*, 2016, pp. 2153–2159.

[47] Y. Liu et al., "Laplacian smoothing stochastic ADMMs with differential privacy guarantees," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1814–1826, 2022.

[48] Z. Xu, X. Chang, F. Xu, and H. Zhang, "$L_{1/2}$ regularization: A thresholding representation theory and a fast solver," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1013–1027, Jul. 2012.

[49] T. H. Oh, Y. W. Tai, J. C. Bazin, H. Kim, and I. S. Kweon, "Partial sum minimization of singular values in robust PCA: Algorithm and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 744–758, Apr. 2016.

[50] A. S. Lewis and H. S. Sendov, "Nonsmooth analysis of singular values. Part I: Theory," *Set-Valued Anal.*, vol. 13, no. 3, pp. 213–241, Sep. 2005.

[51] Y. Liu, Z. Long, H. Huang, and C. Zhu, "Low CP rank and tucker rank tensor completion for estimating missing components in image data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 944–954, Apr. 2020.

[52] Y. Chen, X. Xiao, C. Peng, G. Lu, and Y. Zhou, "Low-rank tensor graph learning for multi-view subspace clustering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 92–104, Jan. 2022.

[53] Y. P. Zhao, L. Chen, and C. L. P. Chen, "Laplacian regularized nonnegative representation for clustering and dimensionality reduction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 1–14, Jan. 2021.

[54] G. Gao, J. Yang, X.-Y. Jing, F. Shen, W. Yang, and D. Yue, "Learning robust and discriminative low-rank representations for face recognition with occlusion," *Pattern Recognit.*, vol. 66, pp. 129–143, Jun. 2017.

[55] B. Wang, H. Niu, J. Zeng, G. Bai, S. Lin, and Y. Wang, "Latent representation learning model for multi-band images fusion via low-rank and sparse embedding," *IEEE Trans. Multimedia*, vol. 23, pp. 3137–3152, 2021.

Hengmin Zhang (Member, IEEE) received the Ph.D. degree from the School of Computer Science and Engineering, Nanjing University of Science and Technology (NJUST), Nanjing, China, in 2019.

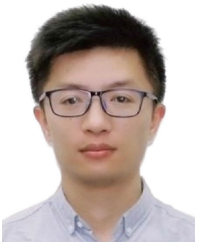
He was a Post-Doctoral Research Fellow with the School of Information Science and Engineering, East China University of Science and Technology (ECUST), Shanghai, China, and also a Post-Doctoral Research Fellow with the PAMI Research Group, Department of Computer and Information Science, University of Macau (UM), Macau, from 2019 to 2022. He is currently a Post-Doctoral Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include sparse coding and low-rank matrix recovery, nonconvex optimizations, and large-scale representation learning methods.

Jian Yang (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NJUST), Nanjing, China, in 2002.

In 2003, he was a Post-Doctoral Researcher with the University of Zaragoza, Zaragoza, Spain. From 2004 to 2006, he was a Post-Doctoral Fellow with the Biometrics Centre, The Hong Kong Polytechnic University, Hong Kong. From 2006 to 2007, he was a Post-Doctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA. He is currently a Chang-Jiang Professor with the School of Computer Science and Engineering, NJUST. He has authored more than 380 scientific papers in pattern recognition and computer vision. His papers have been cited more than 39 000 times in the Scholar Google. His research interests include pattern recognition, computer vision, biometrics, and machine learning. He is a fellow of International Association for Pattern Recognition (IAPR). He is/was an Associate Editor of *Pattern Recognition*, *Pattern Recognition Letters*, *Neurocomputing*, and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Jianjun Qian (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NJUST), Nanjing, China, in 2014.

In 2018, he was selected as a Hong Kong Scholar in China. He is currently an Associate Processor with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, NJUST. His research interests include pattern recognition, computer vision, and machine learning.



Guangwei Gao (Senior Member, IEEE) received the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology, Nanjing, in 2014.

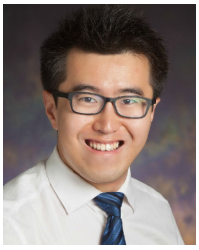
He was a Visiting Student with the Department of Computing, The Hong Kong Polytechnic University, in 2011 and 2013. From 2019 to 2021, he was a Project Researcher with the National Institute of Informatics, Japan. He is currently an Associate Professor with the Nanjing University of Posts and Telecommunications. He has published more than 70 scientific articles in IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, *ACM TOIT*, *TOMM*, PR, AAAI, and IJCAI. His research interests include pattern recognition and computer vision. For more information visit the link: https://guangweigao.github.io.

Xiangyuan Lan received the B.Eng. degree from the South China University of Technology, China, in 2012, and the Ph.D. degree from Hong Kong Baptist University, Hong Kong, in 2016.

He was a Visiting Scholar with the University of Maryland, College Park, MD, USA, in 2015, and also a Visiting Researcher with the University of California at Merced, Merced, CA, USA, in 2017. He was a Post-Doctoral Research Fellow from 2016 to 2018 and then a Research Assistant Professor from 2018 to 2021, with Hong Kong Baptist University. He joined the Peng Cheng Laboratory, Shenzhen, China, in 2022, where he is currently an Associate Professor. He is a Ph.D. Supervisor affiliated with the Southern University of Science and Technology, China. He is currently an Associate Editor of *Signal, Image and Video Processing* and *Frontiers in Signal Processing*. His current research interests include pretrained foundation models and their applications in multi-modal perception.

Zhiyuan Zha (Senior Member, IEEE) received the Ph.D. degree from the School of Electronic Science and Engineering, Nanjing University, Nanjing, China, in 2018.

He is currently a Senior Post-Doctoral Research Fellow with Nanyang Technological University, Singapore. His research interests include inverse problems in image/video processing, sparse signal representation, and machine learning. He was a recipient of the Platinum Best Paper Award and the Best Paper Runner-Up Award at the IEEE International Conference on Multimedia and Expo (ICME) in 2017 and 2020, respectively. He has been an Associate Editor of *The Visual Computer* since 2023.



Bihan Wen (Senior Member, IEEE) received the B.S. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2012, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2015 and 2018, respectively.

He is currently a Nanyang Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University. His research interests include machine learning, computational imaging, computer vision, image and video processing, and big data applications. He was a recipient of the 2016 Yee Fellowship and the 2012 Professional Engineers Board Gold Medal, Singapore. He was also a recipient of the Best Paper Runner-Up Award at the IEEE International Conference on Multimedia and Expo in 2020. He has been an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY since 2022 and an Associate Editor of *Micromachines* (MDPI) since 2021. He was a Guest Editor of *IEEE Signal Processing Magazine* from 2021 to 2023 and a Guest Editor of IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING from 2023 to 2025.