Dual Discriminative Low-Rank Projection Learning for Robust Image Classification

Tingting Su, Dazheng Feng, *Member, IEEE*, Meng Wang, and Mohan Chen

***Abstract*— Numerous methods have exploited projection learning to extract low-dimensional features for image classification. Some projection learning methods integrate low-rank matrix recovery into classification models to equip the projection subspace with discrimination and robustness against corruption. However, these methods cannot directly recover "clean" components from the new datum in a low-dimensional subspace. Additionally, they are sensitive to the selection of projection dimensions. To overcome these shortcomings, we propose a dual discriminative low-rank projection learning framework for robust image classification. Specifically, the proposed method learns a low-rank projection and a semi-orthogonal projection to recover "clean" components from the original data and simultaneously obtain a low-dimensional subspace. Thereafter, to preserve discriminative information in the low-dimensional subspace, an $L_{2,1}$-norm term is constructed by concentrating the projected intra-class samples around their adaptive class centroids. Regression-based terms are appended using the low-dimensional features extracted from the recovered clean data and the class centroids for more accurate classification. Experiments on five public databases with various corruptions demonstrate that the proposed method can robustly classify image data despite a small training sample sizes and gross corruption. The superiority of the proposed method is further verified on the large-scale PubFig83 database, on which it achieves an 87.58% classification accuracy.**

***Index Terms*— Projection learning, low-rank matrix recovery, robust image classification, discriminative projection.**

I. Introduction

IMMENSE amounts of high-dimensional data are extracted from applications such as computer vision [1], document analysis [2], brain imaging [3], and handwritten word [4]. However, a problem called the "curse of dimensionality" occurs when high-dimensional data are processed directly. To alleviate this problem, many feature-extraction methods that project high-dimensional data into low-dimensional subspaces have been proposed based on the assumption that high-dimensional data lie within a union of low-dimensional subspaces [5]. These methods play an important role in various image processing tasks [6], [7], [8], [9].

Manuscript received 3 January 2023; revised 22 April 2023; accepted 9 May 2023. Date of publication 22 May 2023; date of current version 7 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61971470. This article was recommended by Associate Editor M. Shojafar. *(Corresponding author: Dazheng Feng.)*

The authors are with the National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China (e-mail: sutingting@stu.xidian.edu.cn; dzfeng@xidian.edu.cn; mwang_3@stu.xidian.edu.cn; mhchen@stu.xidian.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCSVT.2023.3278571.

Digital Object Identifier 10.1109/TCSVT.2023.3278571

The most widely used of these methods is principal component analysis (PCA) [10], which seeks low-dimensional orthogonal projections by minimizing reconstruction errors. PCA can handle images with low Gaussian noise; however, it is susceptible to gross corruption. To overcome this problem, Wright et al. proposed robust PCA (RPCA) [11], which aims to recover a low-rank "clean" data matrix from a corrupted data matrix that is corrupted by sparse noise. Additionally, various representation-based methods have been proposed to enhance robustness. Wright et al. [12] recovered the "clean" component from a test image with corruptions by using sparse representation of the corrupted image with respect to an predefined dictionary. To better reconstruct the image, the dictionary can also be adaptively learned [13]. However, sparse representation may not be robust against noise and outliers when no additional "clean" training data are available. To overcome this drawback, a low-rank representation (LRR) [14] method was proposed to find the lowest-rank representation for the recovery of corrupted data. Unfortunately, because these robust methods are transductive, they cannot directly handle new data that are not involved in the training procedure. Some improved methods have been proposed by introducing projection learning into the aforementioned methods. For example, Bao et al. proposed inductive RPCA (IRPCA) [15] to leverage the advantages of PCA and RPCA by learning low-rank linear projections. Low-rank embedding (LRE) [16] integrates LRR and projection learning to obtain a linear orthogonal projection for robust low-dimensional feature extraction. On the other hand, double robust PCA (DRPCA) [17] uses a low-rank linear projection matrix to reconstruct the "clean" principal components of data that are recovered by using RPCA.

A limitation of the aforementioned methods is that they are unsupervised and thus, cannot use prior label information to extract effective discriminative features for classification. One workaround is to use linear discriminant analysis (LDA) [18], which was developed based on the Fisher criterion. LDA can extract low-dimensional discriminative features by minimizing intra-class scatter while maximizing inter-class scatter. However, the performance of LDA deteriorates sharply when processing data involving gross corruption. To overcome this shortcoming, various robust supervised methods, including discriminative low-rank preserving projection [19] and truncated nuclear norm on low-rank discriminant embedding (TNNL) [20], based on jointly optimizing LDA and LRR terms have been proposed. Robust sparse linear discriminant analysis (RSLDA) [21] introduces terms for LDA

and sparse noise to extract the discriminative features and enhance their robustness. However, these methods directly introduce LDA as a regularization term for the overall model used for classification; consequently, they do not function reliably when the distribution of samples in each class is not Gaussian [22]. Least-squares regression (LSR) is another widely used method for extracting discriminative features for supervised classification tasks. The main concept of LSR is to learn a projection matrix that transforms original data into a binary label matrix. In contrast to LDA, LSR is not limited by data distributions. Recently, some novel methods [23], [24], [25] have been proposed to learn slack regression targets instead of strict zero–one labels to enhance classification performance. Typically, these methods are sensitive to gross corruption. Therefore, some methods [23], [26] integrate LRR and LSR into a unified framework to extract discriminative features that are robust against gross corruption.

However, the projection dimensions of the above LSR-based methods are limited by the number of data classes; consequently, the extracted features cannot preserve sufficient discriminative information for classification. To obtain additional projections for feature extraction, some recent LSR-based methods such as constrained discriminative projection learning (CDPL) [27], locality preserving robust regression (LPRR) [28] and robust latent subspace learning [29] utilize two separate matrices for feature extraction and regression. On the other hand, sparse non-negative transition subspace learning [30] introduces a transitional transformation subspace to learn two discriminative projection matrices. These methods provide enhanced classification performance for noisy data. Additionally, many methods [20], [31] use features extracted by a single projection with orthogonal constraints for regression. They may have poor performance when the dimensions of the semi-orthogonal projection are significantly greater than the number of classes. To avoid this problem, some methods [21], [28], [32], [33] use a row-sparsity matrix rather than a semi-orthogonal matrix to select the most important features of the original data.

Because LDA-based and LSR-based terms use the Frobenius norm as a basic metric, they are sensitive to gross corruption. However, most of the supervised robust methods described above utilize low-dimensional features from corrupted data to construct LDA-based or LSR-based terms. This may degrade the overall robustness of models to corruption. To solve this problem, discriminative low-rank projection (DLRP) [34] and constrained LRR (CLRR) [26] project the "clean" data obtained by LRR onto a label-indicator matrix, whereas the supervised regularization-based robust subspace (SRRS) [35] method adopts the Fisher criterion to learn discriminative subspaces from "clean" data recovered by LRR. However, CLRR is not an inductive method for classification tasks because it needs to recalculate the LRR coefficients of test samples relative to all the training data, and although DLRP and SRRS are inductive, they transform corrupted testing data into low-dimensional subspaces using projections learned from recovered "clean" training data, which is logically inconsistent and fails to extract "clean" features for tesing. Alternatively, recently developed robust feature extraction methods such as low-rank linear embedding [36], robust discriminant regression (RDR) [31], and linear discriminant analysis for robust dimensionality reduction (RLDA) [22] introduce the $L_{2,1}$ norm, instead of the Frobenius norm used by LDA and LSR, as the basic metric for regression. However, feature extraction in these methods is not directly related to subsequent classification tasks, which may reduce the overall optimality of data classification [27].

To effectively distinguish the previous methods, we have listed their pros and cons in Table I. Motivated by these observations, we then propose a novel supervised robust image classification method called dual discriminative low-rank projection learning (DDLRPL). Specifically, DDLRPL integrates projection learning into the inductive low-rank recovery method to extract low-dimensional features that are robust against sample-specific corruption. To extract features with discrimination, DDLRPL constructs an $L_{2,1}$-norm regularization term by exploiting the compactness of intra-class samples. Additionally, LSR is used to transform low-dimensional clean components into a label matrix for classification. Furthermore, our model jointly optimizes feature extraction and classification to achieve enhanced classification performance in low-dimensional subspaces. The main contributions of this study are as follows.

1) DDLRPL utilizes dual projection matrices (i.e., a semi-orthogonal matrix and a low-rank matrix) for dimensionality reduction and "clean" data recovery, respectively. The rank of the low-rank matrix is learned using nuclear norm. This has two main advantages. First, DDLRPL is robust to the selection of projection dimension, because the semi-orthogonal projection with large reduced dimension cannot influence the overall dimensions of the dual projections used for subsequent regression. Second, for noisy testing data, DDLRPL can directly obtain "clean" components in the low-dimensional subspace for subsequent tasks.
2) To maintain robustness against gross corruption, DDLRPL uses low-dimensional features from the recovered "clean" data, in contrast to the use of corrupted features by most earlier methods, to construct an LSR-based term for classification. Furthermore, in this term, a transformation matrix is introduced to ensure that the selection of the reduced dimension is not limited by the number of classes.
3) To directly learn additional discrimination in the projection subspaces, a regularization term is constructed to concentrate the projected samples around their adaptive class centroids. Simultaneously, LSR is used to separate the different class centroids. Notably, the regularization term uses the $L_{2,1}$ norm as a metric for the removal of outliers and preservation of intrinsic structures of the data.

The remainder of this paper is organized as follows. Preliminary knowledge on the proposed method is presented in Section II. The proposed method and its solutions are detailed in Section III. An analysis of the proposed method is presented in Section IV. Experiments to evaluate the classification performance of the proposed method are presented in Section V.

TABLE I
CHARACTERISTICS OF DIFFERENT METHODS

Method	Measurement	inductive	Supervised	flexible feature dimension	"clean" features for discrimination	"clean" features for testing	robustness
RPCA	nuclear norm, $L_{2,1}$ norm	✗	✗	no DR	–	✓	weak
IRPCA	nuclear norm, $L_{2,1}$ norm	✓	✗	no DR	–	✓	weak
RSLDA	$L_{2,1}$ norm, LDA-based regression	✓	✓	✓	✓	✓	strong
TNNL	truncated nuclear norm, LDA-based regression	✓	✓	✗	✗	✗	weak
CDPL	nuclear norm, $L_{2,1}$ norm, LSR-based regression	✓	✓	✗	✗	✗	weak
CLRR	nuclear norm, $L_{2,1}$ norm, LSR-based regression	✗	✓	✗	✓	✓	strong
DLRP	nuclear norm, $L_{2,1}$ norm, LSR-based regression	✓	✓	✗	✓	✗	weak
DDLRPL	nuclear norm, $L_{2,1}$ norm, LSR-based regression adaptive class center	✓	✓	✓	✓	✓	strong

TABLE II
NOTATIONS USED IN THIS PAPER

Notations	Descriptions
N	The number of all samples
C	The number of classes
d	The original dimension of each sample
d'	The reduced dimension.
n_i	The number of the samples from the i-th class
$\mathbf{X}$	$\mathbf{X} \in \mathbb{R}^{d\times N}$ is the data matrix
$\mathbf{Y}$	$\mathbf{Y} \in \mathbb{R}^{C\times N}$ is the label-indicator binary matrix
$\mathbf{Q}$	$\mathbf{Q} \in \mathbb{R}^{d\times d}$ is the projection matrix
$\mathbf{P}$	$\mathbf{P} \in \mathbb{R}^{d\times d'}$ is the matrix that projects the original subspace into the low-dimensional subspace
$\mathbf{E}$	additive error matrix
$\mathbf{T}$	transformation matrix that fits the data into a label-indicator matrix
$\mathbf{B}$	class centriod matrix
$\mathbf{I}_{d'}$, $\mathbf{I}_C$	$d'\times d'$ identity matrix, $C\times C$ identity matrix
$\mathbf{1}_N$, $\mathbf{1}_C$	N-dimensional, C-dimensional column vectors of all ones

Finally, the conclusions of this study are summarized in Section VI.

II. PRELIMINARY

In this section, the main notations used in this paper are defined, whereafter brief introductions to IRPCA and LRR are presented. Finally, LSR is introduced.

A. Notations

Matrices are denoted by bold uppercase letters, column vectors by bold lowercase letters, and scalars by italic not-bold characters. Throughout this paper, N and C represent the numbers of samples and classes, respectively. A data matrix is defined as $\mathbf{X} \in \mathbb{R}^{d\times N}$, and its corresponding label-indicator binary matrix is denoted as $\mathbf{Y} \in \mathbb{R}^{C\times N}$. The (i, j)-th element (denoted as y_{ij}) of matrix $\mathbf{Y}$ is set to 1 if the j-th sample belongs to the i-th class. The symbol $\mathrm{Tr}(\cdot)$ denotes the trace operator. $\boldsymbol{v}^T$ and $\mathbf{M}^T$ are the transposes of the vector $\boldsymbol{v}$ and matrix $\mathbf{M}$, respectively. $\mathbf{M}^{-1}$ represents the inverse of matrix $\mathbf{M}$. $\|\mathbf{M}\|_*$ is the nuclear norm of matrix $\mathbf{M}$. $\|\boldsymbol{v}\|_2$ denotes the l_2-norm of the vector $\boldsymbol{v}$, that is, $\|\boldsymbol{v}\|_2 = \sqrt{\boldsymbol{v}^T\boldsymbol{v}}$. $\|\mathbf{M}\|_1 = \sum_{i,j}|m_{ij}|$, $\|\mathbf{M}\|_{2,1} = \sum_j \|\boldsymbol{m}_j\|_2$, and $\|\mathbf{M}\|^2 = \sum_j \|\boldsymbol{m}_j\|_2^2 = \mathrm{Tr}(\mathbf{M}^T\mathbf{M})$ are the L_1 norm, $L_{2,1}$ norm, and Frobenius norm of matrix $\mathbf{M}$, respectively. Here, m_{ij} and $\boldsymbol{m}_j$ represent the (i, j)-th element and the j-th column vector of matrix $\mathbf{M}$, respectively. The main notations used in this paper are summarized in Table II.

B. Inductive Robust Principal Component Analysis (IRPCA)

IRPCA [15] argues that there is a linear projection $\mathbf{Q} \in \mathbb{R}^{d\times d}$ that projects corrupted data onto the underlying subspace to remove the corruption. According to the low-rankness of the projection matrix $\mathbf{Q}$ and sparsity of the noise matrix $\mathbf{E} \in \mathbb{R}^{d\times N}$, the objective function of IRPCA can be formulated as follows:

$$\min_{\mathbf{Q},\mathbf{E}} \mathrm{rank}\,(\mathbf{Q}) + \lambda\|\mathbf{E}\|_0, \quad s.t.\ \mathbf{X} = \mathbf{Q}\mathbf{X} + \mathbf{E}, \tag{1}$$

where λ is a positive regularization parameter, and $\|\mathbf{E}\|_0$ denotes the L_0 norm of matrix $\mathbf{E}$, which counts the number of nonzero entries in matrix $\mathbf{E}$. The above optimization problem is difficult to solve based on the discrete properties of the rank function and L_0 norm. Therefore, the nuclear norm and L_1 norm are used to replace the rank function and L_0 norm, respectively. Finally, (1) can be reformulated as the following nuclear-norm regularized minimization problem:

$$\min_{\mathbf{Q},\mathbf{E}} \|\mathbf{Q}\|_* + \lambda\|\mathbf{E}\|_1, \quad s.t.\ \mathbf{X} = \mathbf{Q}\mathbf{X} + \mathbf{E}. \tag{2}$$

The optimal projection $\mathbf{Q}^*$ in the convex problem above is determined using the inexact augmented Lagrangian multiplier method [37]. For a new datum $\boldsymbol{x}'$ not included in the training database, its noise-free principal components can be obtained using $\mathbf{Q}^*\boldsymbol{x}'$.

C. Low-Rank Representation (LRR) and Its Variants

Considering that data lie in multiple low-rank subspaces, LRR aims to find the lowest-rank representation of all data jointly. For databases with sample-specific corruptions, a matrix $\mathbf{E}$ is added to fit the noise using an $L_{2,0}$-norm regularized term. The $L_{2,0}$ norm counts the number of nonzero columns. To avoid the NP-hard problem in the optimization, the tightest convex relaxation of $L_{2,0}$ norm can be obtained;

this is referred to as the $L_{2,1}$ norm. Therefore, the LRR method can be formulated as

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \lambda\|\mathbf{E}\|_{2,1}, \quad s.t.\ \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \tag{3}$$

where $\mathbf{Z} \in \mathbb{R}^{N\times N}$ represents an LRR coefficient matrix. After obtaining the optimal result $(\mathbf{Z}^*, \mathbf{E}^*)$, clean data can be recovered using $\mathbf{X}-\mathbf{E}^*$ (or $\mathbf{XZ}^*$). LRR is effective and robust for data segmentation. However, this method cannot directly handle any new data that are not involved in the training procedure.

Various methods have been proposed to integrate projection learning and LRR into different models. Considering the LRE method [16] as an example, the objective function of LRE can be formulated as

$$\min_{\mathbf{P},\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \lambda\|\mathbf{E}\|_{2,1}, \quad s.t.\ \mathbf{P}^T\mathbf{X} = \mathbf{P}^T\mathbf{XZ} + \mathbf{E},\ \mathbf{P}^T\mathbf{P} = \mathbf{I}_{d'}, \tag{4}$$

where $\mathbf{P} \in \mathbb{R}^{d\times d'}(d > d')$, and $\mathbf{I}_{d'} \in \mathbb{R}^{d'\times d'}$ denotes the identity matrix. After obtaining the optimal result $(\mathbf{P}^*, \mathbf{Z}^*, \mathbf{E}^*)$, the result of $\mathbf{P}^{*T}\mathbf{X}$ is used for subsequent classification tasks. However, it would be unreasonable to use $\mathbf{P}^{*T}\mathbf{X}$, which can be considered as low-dimensional features $\mathbf{P}^{*T}\mathbf{XZ}$ corrupted by sparse noise $\mathbf{E}$, for classification.

D. Least-Squares Regression (LSR)

LSR for classification aims to learn a projection matrix $\mathbf{T} \in \mathbb{R}^{C\times d}$ to fit data into a label-indicator matrix. The objective function is formulated as follows:

$$\min_{\mathbf{T}} \|\mathbf{Y} - \mathbf{TX}\|^2. \tag{5}$$

The closed-form solution to this problem is $\mathbf{T} = \mathbf{YX}^T(\mathbf{XX}^T)^{-1}$. However, the matrix $\mathbf{XX}^T$ is typically singular for problems involving small samples sizes (i.e., the dimension of the samples is greater than their number). To avoid the instability of solutions caused by the singularity of matrix $\mathbf{XX}^T$, a small positive regularization parameter τ is added such that the solution can be rewritten as $\mathbf{T} = \mathbf{YX}^T(\mathbf{XX}^T + \tau\mathbf{I}_d)^{-1}$. Here, $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ denotes the identity matrix.

III. Proposed Method

In this section, we first present a detailed description of the proposed model. Thereafter, we present its optimization method. Finally, we define the corresponding algorithmic procedures.

A. Formulation

In IRPCA, "clean" testing data can be recovered using the optimal projection matrix $\mathbf{Q}$ learned from the training data. However, this method lacks dimensionality-reduction functionalities (i.e., this method cannot reduce a datum to a predefined dimension). To overcome this shortcoming, we introduce a semi-orthogonal matrix $\mathbf{P} \in \mathbb{R}^{d\times d'}$ $(d > d')$ into formulation (2) to project the recovered data into low-dimensional subspaces. Furthermore, to address sample-specific corruptions, the noise matrix $\mathbf{E} \in \mathbb{R}^{d'\times N}$ is supposed to be sparse in its columns. Finally, we solved the following optimization problem:

$$\min_{\mathbf{Q},\mathbf{E},\mathbf{P}} \|\mathbf{Q}\|_* + \gamma\|\mathbf{E}\|_{2,1} \quad s.t.\ \mathbf{P}^T\mathbf{X} = \mathbf{P}^T\mathbf{QX} + \mathbf{E},\ \mathbf{P}^T\mathbf{P} = \mathbf{I}_{d'}, \tag{6}$$

where γ is a positive parameter, and $\mathbf{I}_{d'} \in \mathbb{R}^{d'\times d'}$ denotes the identity matrix. The rank of matrix $\mathbf{P}^T\mathbf{Q}$ is lower than those of matrices $\mathbf{P}$ and $\mathbf{Q}$, that is, $\text{rank}(\mathbf{P}^T\mathbf{Q}) \le \min\{d', \text{rank}(\mathbf{Q})\}$, and problem (6) becomes (2) if $d = d'$. The objective function (6) exhibits both "clean" components recovery and dimensionality-reduction properties. We know that $\text{rank}(\mathbf{P}^T\mathbf{Q}) = \text{rank}(\mathbf{Q})$ if $d' > \text{rank}(\mathbf{Q})$ and $\text{rank}(\mathbf{Q})$ is learned adaptively. This demonstrates that selecting a large projection dimension (i.e., $d' > \text{rank}(\mathbf{Q})$) does not change the dimension of the features extracted by the dual projections for subsequent regression, such that the sensitivity to selection of projection dimensions can be reduced. Furthermore, for new corrupted data $\boldsymbol{x}'$, the low-dimensional features of the recovered "clean" components can be directly obtained using the optimal solutions $(\mathbf{P}^*, \mathbf{Q}^*)$ of (6), that is, $\mathbf{P}^{*T}\mathbf{Q}^*\boldsymbol{x}'$. This illustrates that problem (6) is inductive. However, $(\mathbf{P}^*, \mathbf{Q}^*)$ in (6) does not retain discriminative information; therefore, it performs poorly on classification tasks.

To utilize label information for enhanced classification performance, we construct an LSR-based regularization term in (6) to project the recovered clean data onto the label-indicator matrix. Because the projection dimension of classic LSR is limited by the number of classes, a transformation matrix $\mathbf{T} \in \mathbb{R}^{C\times d'}$ is introduced to connect the low-dimensional features of the recovered data and the indicator matrix. Finally, formulation (6) is transformed into the following supervised optimization problem:

$$\min_{\mathbf{Q},\mathbf{E},\mathbf{T},\mathbf{P}} \|\mathbf{Q}\|_* + \gamma\|\mathbf{E}\|_{2,1} + \alpha/2\|\mathbf{Y} - \mathbf{T}\left[\mathbf{P}^T\mathbf{QX}; \mathbf{1}_N^T\right]\|^2 \quad s.t.\ \mathbf{P}^T\mathbf{X} = \mathbf{P}^T\mathbf{QX} + \mathbf{E}, \quad \mathbf{P}^T\mathbf{P} = \mathbf{I}_{d'}, \tag{7}$$

where α is a positive regularization parameter, and $\mathbf{1}_N \in \mathbb{R}^{N\times 1}$ denotes a column vector of all ones. The LSR classifier and feature extraction were jointly optimized in (7). This combined model typically performs better than learning two independent tasks. However, although the reduced dimension selection in problem (7) is free based on the addition of the transformation matrix $\mathbf{T}$, problem (7) does not directly and adequately preserve the discrimination properties of the low-dimensional subspace $\mathbf{P}^T\mathbf{Q}$ used for subsequent tasks.

LDA is an effective method for preserving the discrimination of low-dimensional subspaces. It concentrates samples of the same class around their class centroids as closely as possible and separates different class centroids to the greatest extent possible in low-dimensional subspaces. Inspired by the concept of LDA, we attempt to learn discriminative projections by concentrating "clean" samples around their class centroids. Additionally, to separate different class centroids, we use the transformation matrix $\mathbf{T}$ to fit the class centroids to their label matrix, which is the identity matrix $\mathbf{I}_C \in \mathbb{R}^{C\times C}$. Let $\mathbf{B} \in \mathbb{R}^{d'\times N}$ denote the class centroid matrix, and its i-th

column represent the class centroid of the i-th class. Finally, the complete objective function for DDLRPL is formulated as the following optimization problem:

$$\begin{aligned}\min_{\mathbf{Q},\mathbf{E},\mathbf{T},\mathbf{B},\mathbf{P}} \ & \|\mathbf{Q}\|_* + \gamma\|\mathbf{E}\|_{2,1} + \alpha/2\|\mathbf{Y} - \mathbf{T}\left[\mathbf{P}^T\mathbf{Q}\mathbf{X}; \mathbf{1}_N^T\right]\|^2 \\ & + \beta\|\mathbf{P}^T\mathbf{Q}\mathbf{X} - \mathbf{B}\mathbf{Y}\|_{2,1} + \lambda/2\|\left(\mathbf{I}_C - \mathbf{T}\left[\mathbf{B}; \mathbf{1}_C^T\right]\right)\mathbf{W}\|^2 \\ s.t. \ & \mathbf{P}^T\mathbf{X} = \mathbf{P}^T\mathbf{Q}\mathbf{X} + \mathbf{E}, \ \mathbf{P}^T\mathbf{P} = \mathbf{I}_{d'}, \end{aligned} \tag{8}$$

where β and λ are the positive regularization parameters, and vector $\mathbf{1}_C \in \mathbb{R}^{C\times 1}$ denotes a column vector of all ones. $\mathbf{W} \in \mathbb{R}^{C\times C}$ is a diagonal matrix, whose element w_{ii} denotes the square root of the number of samples from the i-th class, that is, $w_{ii} = \sqrt{n_i}$. The fourth term in (8) exploits the $L_{2,1}$ norm as a basic metric instead of using the Frobenius norm. The $L_{2,1}$ norm is the sum of unsquared residuals. Using $L_{2,1}$ as the basic metric can reduce the contribution from large residuals that may caused by outliers. In conlusion, the construction of the fourth term with $L_{2,1}$ norm in (8) has two advantages: the intrinsic structures of data are automatically preserved; and (8) is more robust against outliers compared to problem (7). Additionally, LDA uses the mean of each class as its centroid, whereas in our method, the class centroid matrix $\mathbf{B}$ is adaptively learned. Therefore, our method can handle data that are not normally distributed.

B. Optimization

In this subsection, we exploit the alternating direction method of multipliers (ADMM) [38] to optimize problem (8). First, an auxiliary variable $\mathbf{R}$ is introduced to represent $\mathbf{Q}^T\mathbf{P}$ to simplify optimization. Therefore, problem (8) can be rewritten as

$$\begin{aligned}\min_{\mathbf{Q},\mathbf{E},\mathbf{T},\mathbf{B},\mathbf{P},\mathbf{R}} \ & \|\mathbf{Q}\|_* + \gamma\|\mathbf{E}\|_{2,1} + \beta\|\mathbf{R}^T\mathbf{X} - \mathbf{B}\mathbf{Y}\|_{2,1} \\ & + \alpha/2\|\mathbf{Y} - \mathbf{T}\left[\mathbf{R}^T\mathbf{X}; \mathbf{1}_N^T\right]\|^2 \\ & + \lambda/2\|\left(\mathbf{I}_C - \mathbf{T}\left[\mathbf{B}; \mathbf{1}_C^T\right]\right)\mathbf{W}\|^2 \\ \text{s.t. } & \mathbf{P}^T\mathbf{X} = \mathbf{R}^T\mathbf{X} + \mathbf{E}, \ \mathbf{R} = \mathbf{Q}^T\mathbf{P}, \ \mathbf{P}^T\mathbf{P} = \mathbf{I}_{d'}. \end{aligned} \tag{9}$$

Its augmented Lagrangian function is expressed as follows:

$$\begin{aligned} L_\rho = & \|\mathbf{Q}\|_* + \gamma\|\mathbf{E}\|_{2,1} + \beta\|\mathbf{R}^T\mathbf{X} - \mathbf{B}\mathbf{Y}\|_{2,1} \\ & + \alpha/2\|\mathbf{Y} - \mathbf{T}\left[\mathbf{R}^T\mathbf{X}; \mathbf{1}_N^T\right]\|^2 \\ & + \lambda/2\|\left(\mathbf{I}_C - \mathbf{T}\left[\mathbf{B}; \mathbf{1}_C^T\right]\right)\mathbf{W}\|^2 \\ & + \mu/2\|\mathbf{P}^T\mathbf{X} - \mathbf{R}^T\mathbf{X} - \mathbf{E} + \mathbf{L}_1/\mu\|^2 \\ & + \mu/2\|\mathbf{R} - \mathbf{Q}^T\mathbf{P} + \mathbf{L}_2/\mu\|^2, \end{aligned} \tag{10}$$

where $\mathbf{L}_1 \in \mathbb{R}^{d'\times N}$ and $\mathbf{L}_2 \in \mathbb{R}^{d\times d'}$ are the Lagrangian multiplier matrices. μ is the penalty parameter, and (10) satisfies the constraint $\mathbf{P}^T\mathbf{P} = \mathbf{I}_{d'}$. The optimal solution $(\mathbf{Q}, \mathbf{E}, \mathbf{T}, \mathbf{R}, \mathbf{B}, \mathbf{P})$ for the objective function (10) can be derived individually by fixing the other variables.

(1) Update $\mathbf{Q}$: By fixing $\mathbf{E}, \mathbf{T}, \mathbf{R}, \mathbf{B}, \mathbf{P}$, problem (10) is converted into

$$\min_{\mathbf{Q}} \|\mathbf{Q}\|_* + \mu/2\|\mathbf{R} - \mathbf{Q}^T\mathbf{P} + \mathbf{L}_2/\mu\|^2. \tag{11}$$

However, this function does not have a closed-form solution. According to linearized ADMM [38], function (11) can be solved by linearizing its second term. Therefore, the solution of (11) in the $(k+1)$-th iteration can be obtained as follows:

$$\mathbf{Q}_{k+1} = \Phi_{1/(\mu\eta_Q)}\left(\mathbf{Q}_k - \frac{\nabla_{\mathbf{Q}}(\mathbf{Q}_k)}{\mu\eta_Q}\right), \tag{12}$$

where $\nabla_{\mathbf{Q}}(\mathbf{Q}_k) = \mu\mathbf{P}\left(\mathbf{Q}_k^T\mathbf{P} - \mathbf{R} - \mathbf{L}_2/\mu\right)^T$, and the proximal parameter $\eta_Q = 1.01\|\mathbf{P}\|^2$. Further, $\eta_Q = 1.01\ d'$ because $\|\mathbf{P}\|^2 = \mathrm{Tr}(\mathbf{P}^T\mathbf{P}) = \mathrm{Tr}(\mathbf{I}_{d'}) = d'$. Φ is a singular value shrinkage operator [39], with a threshold $(\mu\eta_Q)^{-1}$.

(2) Update $\mathbf{E}$: By keeping $\mathbf{Q}, \mathbf{T}, \mathbf{R}, \mathbf{B}, \mathbf{P}$ fixed, (10) is reduced to the following problem:

$$\min_{\mathbf{E}} \gamma\|\mathbf{E}\|_{2,1} + \mu/2\|\mathbf{P}^T\mathbf{X} - \mathbf{R}^T\mathbf{X} - \mathbf{E} + \mathbf{L}_1/\mu\|^2. \tag{13}$$

Supposing that $\mathbf{A} = \mathbf{P}^T\mathbf{X} - \mathbf{R}^T\mathbf{X} + \mathbf{L}_1/\mu$ and $\boldsymbol{a}_i$ is the i-th column vector of $\mathbf{A}$, according to Lemma 4.1 in [14], the i-th column vector of $\mathbf{E}$ in the $(k+1)$-th iteration can be calculated as follows:

$$\left[\mathbf{E}_{k+1}\right]_{:,i} = \begin{cases} \dfrac{\|\boldsymbol{a}_i^k\|_2 - \gamma/\mu_k}{\|\boldsymbol{a}_i^k\|_2}\boldsymbol{a}_i^k, & \text{if } \|\boldsymbol{a}_i^k\|_2 > \gamma/\mu \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

(3) Update $\mathbf{T}$: By fixing $\mathbf{Q}, \mathbf{E}, \mathbf{R}, \mathbf{B}, \mathbf{P}$, the transformation matrix $\mathbf{T}$ can be obtained as follows:

$$\min_{\mathbf{T}} \alpha\|\mathbf{Y} - \mathbf{T}\left[\mathbf{R}^T\mathbf{X}; \mathbf{1}_N^T\right]\|^2 + \lambda\|\left(\mathbf{I}_C - \mathbf{T}\left[\mathbf{B}; \mathbf{1}_C^T\right]\right)\mathbf{W}\|^2. \tag{15}$$

Clearly, (15) is an LSR problem, and thus, it has a closed-form solution. Suppose $\mathbf{D}_1 = \left[\mathbf{R}^T\mathbf{X}; \mathbf{1}_N^T\right]$ and $\mathbf{D}_2 = \left[\mathbf{B}; \mathbf{1}_C^T\right]\mathbf{W}$. Then, the optimal $\mathbf{T}$ in the current iteration can be obtained as follows:

$$\mathbf{T} = \left(\alpha\mathbf{Y}\mathbf{D}_1^T + \lambda\mathbf{W}\mathbf{D}_2^T\right)\left(\alpha\mathbf{D}_1\mathbf{D}_1^T + \lambda\mathbf{D}_2\mathbf{D}_2^T + \tau\mathbf{I}_{(d'+1)}\right)^{-1}, \tag{16}$$

where τ is a small positive constant, in case the singularity for $(\alpha\mathbf{D}_1\mathbf{D}_1^T + \lambda\mathbf{D}_2\mathbf{D}_2^T)$ leads to an ill-posed solution.

(4) Update $\mathbf{R}$: By fixing $\mathbf{Q}, \mathbf{E}, \mathbf{T}, \mathbf{B}, \mathbf{P}$, problem (10) can be converted into the following formulation:

$$\begin{aligned}\min_{\mathbf{R}} \ & \beta\|\mathbf{R}^T\mathbf{X} - \mathbf{B}\mathbf{Y}\|_{2,1} + \alpha/2\|\mathbf{Y} - \mathbf{T}\left[\mathbf{R}^T\mathbf{X}; \mathbf{1}_N^T\right]\|^2 \\ & + \mu/2\|\mathbf{P}^T\mathbf{X} - \mathbf{R}^T\mathbf{X} - \mathbf{E} + \mathbf{L}_1/\mu\|^2 \\ & + \mu/2\|\mathbf{R} - \mathbf{Q}^T\mathbf{P} + \mathbf{L}_2/\mu\|^2. \end{aligned} \tag{17}$$

The first term in (17) is convex. However, it is non-smooth because its derivation with respect to $\mathbf{R}$ does not exist when $\|[\mathbf{R}^T\mathbf{X} - \mathbf{B}\mathbf{Y}]_{:,i}\|_2 = 0$ for $i = 1, 2, \cdots, N$. Thus, supposing that $\mathbf{T} = [\mathbf{T}_{d'}, \boldsymbol{t}]$, $\mathbf{H}_1 = \mathbf{Y} - \boldsymbol{t}\mathbf{1}_N^T$, $\mathbf{H}_2 = \mathbf{P}^T\mathbf{X} - \mathbf{E} + \mathbf{L}_1/\mu$ and $\mathbf{H}_3 = \mathbf{Q}^T\mathbf{P} - \mathbf{L}_2/\mu$, (17) can be reformulated as the following smooth optimization problem:

$$\begin{aligned}\min_{\mathbf{R}} \ & \beta\,\mathrm{Tr}\left(\left(\mathbf{R}^T\mathbf{X} - \mathbf{B}\mathbf{Y}\right)\mathbf{M}\left(\mathbf{R}^T\mathbf{X} - \mathbf{B}\mathbf{Y}\right)^T\right) \\ & + \alpha/2\|\mathbf{H}_1 - \mathbf{T}_{d'}\mathbf{R}^T\mathbf{X}\|^2 \\ & + \mu/2\|\mathbf{H}_2 - \mathbf{R}^T\mathbf{X}\|^2 + \mu/2\|\mathbf{R} - \mathbf{H}_3\|^2, \end{aligned} \tag{18}$$

where $\mathbf{M}$ is a diagonal matrix, and its i-th element is $m_{ii} = 1/(2\|[\mathbf{R}^T\mathbf{X}-\mathbf{B}\mathbf{Y}]_{:,i}\|_2+\xi)$. ξ is an extremely small value that prevents numerical instability caused by $\|[\mathbf{R}^T\mathbf{X}-\mathbf{B}\mathbf{Y}]_{:,i}\|_2 \to 0$. $\mathbf{M}$ is assumed to be a constant and is calculated using $\mathbf{R}$ in the last iteration. By taking the partial derivative of (18) with respect to $\mathbf{R}$ and setting it to zero, the optimal solution $\bar{\mathbf{R}}$ of problem (18) can be obtained by solving the following equation:

$$\begin{aligned} &2\beta\mathbf{X}\mathbf{M}\mathbf{X}^T\bar{\mathbf{R}} + \alpha\mathbf{X}\mathbf{X}^T\bar{\mathbf{R}}\mathbf{T}_{d'}^T\mathbf{T}_{d'} + \mu\mathbf{X}\mathbf{X}^T\bar{\mathbf{R}} + \mu\bar{\mathbf{R}} \\ &\quad = 2\beta\mathbf{X}\mathbf{M}\mathbf{Y}^T\mathbf{B}^T + \alpha\mathbf{X}\mathbf{H}_1^T\mathbf{T}_{d'} + \mu\mathbf{X}\mathbf{H}_2^T + \mu\mathbf{H}_3. \end{aligned} \tag{19}$$

Equation (19) can be reformulated as a Stein equation:

$$\begin{aligned} &\mathbf{F}\mathbf{X}\mathbf{X}^T\bar{\mathbf{R}}\left(\alpha\mathbf{T}_{d'}^T\mathbf{T}_{d'} + \mu\mathbf{I}_{d'}\right) + \bar{\mathbf{R}} \\ &= \mathbf{F}\left(2\beta\mathbf{X}\mathbf{M}\mathbf{Y}^T\mathbf{B}^T + \alpha\mathbf{X}\mathbf{H}_1^T\mathbf{T}_{d'} + \mu\mathbf{X}\mathbf{H}_2^T + \mu\mathbf{H}_3\right), \end{aligned} \tag{20}$$

where $\mathbf{F} = \left(2\beta\mathbf{X}\mathbf{M}\mathbf{X}^T + \mu\mathbf{I}_d\right)^{-1}$. We use the classic Hessenberg–Schur method [40] to solve problem (20).

(5) Update $\mathbf{B}$: Matrices $\mathbf{Q}$, $\mathbf{E}$, $\mathbf{T}$, $\mathbf{R}$, $\mathbf{P}$ are fixed, and problem (10) is reduced to the following optimization problem:

$$\min_{\mathbf{B}} \|\mathbf{R}^T\mathbf{X} - \mathbf{B}\mathbf{Y}\|_{2,1} + \lambda/2\|\left(\mathbf{I}_C - \mathbf{T}[\mathbf{B}; \mathbf{1}_C^T]\right)\mathbf{W}\|^2. \tag{21}$$

Similar to in the updating process for $\mathbf{R}$, (21) can be rewritten as follows:

$$\begin{aligned} \min_{\mathbf{B}}\ &\beta\,\mathrm{Tr}\left(\left(\mathbf{R}^T\mathbf{X} - \mathbf{B}\mathbf{Y}\right)\mathbf{M}\left(\mathbf{R}^T\mathbf{X} - \mathbf{B}\mathbf{Y}\right)^T\right) \\ &+ \lambda/2\|\left(\mathbf{I}_C - \mathbf{T}_{d'}\mathbf{B} - t\mathbf{1}_C^T\right)\mathbf{W}\|^2, \end{aligned} \tag{22}$$

where the definition of $\mathbf{M}$ is identical to that in (18), and $\mathbf{T} = [\mathbf{T}_{d'}, t]$. By taking the partial derivative of (22) with respect to $\mathbf{B}$ and setting it to zero, the optimal $\bar{\mathbf{B}}$ can be obtained by solving the following Sylvester equation:

$$2\beta\bar{\mathbf{B}}\mathbf{Y}\mathbf{M}\mathbf{Y}^T\left(\mathbf{W}\mathbf{W}^T\right)^{-1} + \lambda\mathbf{T}_{d'}^T\mathbf{T}_{d'}\bar{\mathbf{B}} = \mathbf{S}. \tag{23}$$

Here, $\mathbf{S} = 2\beta\mathbf{R}^T\mathbf{X}\mathbf{M}\mathbf{Y}^T\left(\mathbf{W}\mathbf{W}^T\right)^{-1} + \lambda\mathbf{T}_{d'}^T\left(\mathbf{I}_C - t\mathbf{1}_C^T\right)$.

Lemma 1: [41] A real symmetric matrix can be orthogonally diagonalizable.

According to Lemma 1, we can obtain $\mathbf{T}_{d'}^T\mathbf{T}_{d'} = \mathbf{U}\tilde{\Lambda}\mathbf{U}^T$ because $\mathbf{T}_{d'}^T\mathbf{T}_{d'}$ is a symmetric matrix. Here, $\mathbf{U}$ is an orthogonal matrix. $\tilde{\Lambda}$ is a diagonal matrix, and its elements are the eigenvalues $\{\tilde{\lambda}_i | i = 1, \ldots, d'\}$ of $\mathbf{T}_{d'}^T\mathbf{T}_{d'}$. Moreover, it can be easily obtained that $\mathbf{Y}\mathbf{M}\mathbf{Y}^T\left(\mathbf{W}\mathbf{W}^T\right)^{-1}$ is a diagonal matrix, and its elements are denoted as $\{\bar{\lambda}_j | j = 1, \ldots, C\}$. Assuming that $\bar{\mathbf{S}} = \mathbf{U}^T\mathbf{S}$ and matrix $\mathbf{K}$, the (i, j)-th element of $\mathbf{K}$ can be calculated as $[\mathbf{K}]_{i,j} = [\bar{\mathbf{S}}]_{i,j}/(\lambda\tilde{\lambda}_i + 2\beta\bar{\lambda}_j)$. Ultimately, the optimal $\bar{\mathbf{B}}$ can be obtained as $\bar{\mathbf{B}} = \mathbf{U}\mathbf{K}$.

(6) Update $\mathbf{P}$: By fixing $\mathbf{Q}$, $\mathbf{E}$, $\mathbf{T}$, $\mathbf{R}$, $\mathbf{B}$ to solve $\mathbf{P}$, (10) can be transformed into the following minimization problem:

$$\begin{aligned} &\min_{\mathbf{P}} \|\mathbf{P}^T\mathbf{X} - \mathbf{R}^T\mathbf{X} - \mathbf{E} + \mathbf{L}_1/\mu\|^2 + \|\mathbf{R} - \mathbf{Q}^T\mathbf{P} + \mathbf{L}_2/\mu\|^2 \\ &s.t.\ \mathbf{P}^T\mathbf{P} = \mathbf{I}_{d'}. \end{aligned} \tag{24}$$

This problem is a combination of two Frobenius functions. According to the definition of the Frobenius functions, (24) can be converted into

$$\begin{aligned} \min_{\mathbf{P}} \mathrm{Tr}\ &\left\{\mathbf{P}^T\left(\mathbf{X}\mathbf{X}^T + \mathbf{Q}\mathbf{Q}^T\right)\mathbf{P} - 2\mathbf{P}^T\left[\mathbf{Q}\left(\mathbf{R} + \mathbf{L}_2/\mu\right)\right.\right. \\ &\left.\left. + \mathbf{X}\left(\mathbf{R}^T\mathbf{X} + \mathbf{E} - \mathbf{L}_1/\mu\right)^T\right]\right\}, \quad s.t.\ \mathbf{P}^T\mathbf{P} = \mathbf{I}_{d'}. \end{aligned} \tag{25}$$

A generalized power iteration (GPI) [42] method is adopted to determine the objective function (25), which is a classic quadratic problem on the Stiefel manifold.

The detailed iterative process of the proposed method is summarized in Algorithm 1.

Algorithm 1 The DDLRPL Method

Require: The training sample matrix $\mathbf{X} \in \mathbb{R}^{d\times N}$, its corresponding label indicator matrix $\mathbf{Y} \in \mathbb{R}^{C\times N}$, reduced dimension d', and regularization parameters α, β, λ, γ.

Ensure: Projection matrices $\mathbf{P}$ and $\mathbf{Q}$.

Initialization: Orthogonal matrix $\mathbf{P}$, $\mathbf{E} = \mathbf{L}_1 = \mathbf{0}^{d'\times N}$, $\mathbf{Q} = \mathbf{0}^{d\times d}$, $\mathbf{R} = \mathbf{L}_2 = \mathbf{0}^{d\times d'}$, $\mathbf{B} = \mathbf{0}^{d'\times C}$, $\tau = 10^{-4}$, $\rho = 1.1$, $\mu = 0.01$, $\mu_{\max} = 10^6$, $\varepsilon = \xi = 10^{-5}$.
while not converged **do**
  Update $\mathbf{Q}$ by using (12);
  Update $\mathbf{E}$ by using (14);
  Update $\mathbf{T}$ by using (16);
  Update $\mathbf{R}$ by solving (20);
  Update $\mathbf{B}$ by solving (23);
  Update $\mathbf{P}$ by solving (25);
  Update Lagrangian multipliers $\mathbf{L}_1$, $\mathbf{L}_2$:
    $\mathbf{L}_1 = \mathbf{L}_1 + \mu\left(\mathbf{P}^T\mathbf{X} - \mathbf{R}^T\mathbf{X} - \mathbf{E}\right)$,
    $\mathbf{L}_2 = \mathbf{L}_2 + \mu\left(\mathbf{R} - \mathbf{Q}^T\mathbf{P}\right)$;
  Update penalty parameter μ:
    $\mu = \min\left(\mu_{\max}, \rho\mu\right)$;
  Check the convergence conditions: if
  $\max\left(\|\mathbf{P}^T\mathbf{X} - \mathbf{R}^T\mathbf{X} - \mathbf{E}\|/\|\mathbf{X}\|, \|\mathbf{R} - \mathbf{Q}^T\mathbf{P}\|\right) < \varepsilon$
  then stop
end while

C. Classification

The optimal projection matrices $\mathbf{P}^*$ and $\mathbf{Q}^*$ are obtained by solving the overall objective function (8). Assuming that $\boldsymbol{x}_{\text{test}}$ is a testing sample, its noise can be removed by $\mathbf{Q}^*\boldsymbol{x}_{\text{test}}$. Thereafter, the low-dimensional representation of the "clean" testing sample can be obtained using the matrix $\mathbf{P}^*$ to project $\mathbf{Q}^*\boldsymbol{x}_{\text{test}}$ into the low-dimensional subspaces as $\mathbf{P}^{*T}\mathbf{Q}^*\boldsymbol{x}_{\text{test}}$. According to the constraint $\mathbf{R}^* = \mathbf{Q}^{*T}\mathbf{P}^*$ in (9), the low-dimensional features of the "clean" test sample can also be directly obtained using $\mathbf{R}^{*T}\boldsymbol{x}_{\text{test}}$. Finally, the obtained low-dimensional features are inputted into the subsequent classifier.

IV. Algorithm Analysis

In this section, we present a convergence analysis and discuss the computational complexity of the proposed method. Additionally, we detail the differences between the proposed method and related methods.

A. Convergence Analysis

Because the overall model in (8) is not convex, it is difficult to prove the strong convergence in the proposed method.

Therefore, a weak convergence property of the proposed method is presented by proving that the proposed method can reach a stationary point that satisfies the Karush-Kuhn-Tucker (KKT) condition. The procedure for deriving $\mathbf{B}$ and $\mathbf{P}$ is not involved in the Lagrangian multipliers; therefore, we have not proven their KKT conditions here.

Theorem 1: Let the sequence $\Theta_1^k = \left(\mathbf{Q}_k, \mathbf{E}_k, \mathbf{T}_k, \mathbf{R}_k, \mathbf{L}_1^k, \mathbf{L}_2^k\right)$ and $\{\Theta_1^k\}_{k=1}^{\infty}$ be generated by Algorithm 1. The sequence $\left\{\Theta_1^k\right\}_{k=1}^{\infty}$ is assumed to be bounded and satisfies $\lim_{k\to\infty}\{\Theta_1^{k+1} - \Theta_1^k\} \to 0$. If penalty parameter μ is non-decreasing and upper bounded, and $\eta_Q > \|\mathbf{P}\|^2$, then every limit point of the sequence $\{\Theta_1^k\}_{k=1}^{\infty}$ satisfies the following KKT conditions:

$$\begin{aligned}
&\mathbf{P}^T\mathbf{X} = \mathbf{R}^T\mathbf{X} + \mathbf{E}, \qquad \mathbf{R} = \mathbf{Q}^T\mathbf{P},\\
&\mathbf{P}\mathbf{L}_2^T \in \partial_{\mathbf{Q}}\|\mathbf{Q}\|_*, \quad \mathbf{L}_1/\gamma \in \partial_{\mathbf{E}}\|\mathbf{E}\|_{2,1},\\
&\mathbf{T}\left(\alpha\mathbf{D}_1\mathbf{D}_1^T + \lambda\mathbf{D}_2\mathbf{D}_2^T\right) = \alpha\mathbf{Y}\mathbf{D}_1^T + \lambda\mathbf{W}\mathbf{D}_2^T,\\
&\alpha\mathbf{X}\left(\mathbf{Y}-\mathbf{T}\mathbf{D}_1\right)^T\mathbf{T}_{d'} + \mathbf{X}\mathbf{L}_1^T - \mathbf{L}_2 \in \beta\partial_{\mathbf{R}}\|\mathbf{R}^T\mathbf{X}-\mathbf{B}\mathbf{Y}\|_{2,1}, \quad (26)
\end{aligned}$$

where $\mathbf{D}_1 = \left[\mathbf{R}^T\mathbf{X}; \mathbf{1}_N^T\right]$ and $\mathbf{D}_2 = \left[\mathbf{B}; \mathbf{1}_C^T\right]$.

Theorem 2: Let $\mathbf{B}_k$ be generated by step 7 in Algorithm 1. If Θ_1^k is fixed, then the updated $\{\mathbf{B}_k\}^{\infty}$ monotonically reduces the objective function value in (8) for each iteration.

The detailed proofs of Theorems 1 and 2 are presented in the Appendix of supplementary material, to maintain the flow of this paper.

The convergence property of $\mathbf{P}_k$ generated by step 8 in Algorithm 1 was proven in [42]. Consequently, consider that $\Theta_k = \left(\Theta_1^k, \mathbf{B}_k, \mathbf{P}_k\right)$, and assume that Θ_k is bounded and satisfies the condition of $\lim_{k\to\infty}\{\Theta_{k+1} - \Theta_k\} \to 0$. Based on Theorems 1 and 2, the following conclusion can be drawn: $\Theta_k = \left(\mathbf{Q}_k, \mathbf{E}_k, \mathbf{T}_k, \mathbf{R}_k, \mathbf{L}_1^k, \mathbf{L}_2^k, \mathbf{B}_k, \mathbf{P}_k\right)$ converges to a local optimum that satisfies the KKT condition.

B. Computational Complexity

The most computationally expensive steps for DDLRPL presented in Algorithm 1 are steps 3 and 6. According to [38], the main computational complexity of step 3 is approximately $O(d^3 + 2d^2d')$. The main computational load of step 6 stems from calculating matrix $\mathbf{F}$, calculating matrix multiplications in (20), and solving the Stein equation, whose computational complexities are approximately $O(\frac{2}{3}d^3 + d^2N + dN)$, $O(d^3 + d^2N + d'^2C + d^2d' + dNd')$, and $O(\frac{5}{3}d^3 + 10d'^3 + 3d^2d' + 2dd'^2)$ according to [40], respectively. Therefore, the total computational complexity of step 6 is $O(\frac{10}{3}d^3 + 2d^2N + 4d^2d' + (2d + C)d'^2 + 10d'^3 + dN(d' + 1))$. Additionally, the computational complexities of steps 4, 5 and 7 are approximately $O(dNd' + Nd')$, $O(\frac{2}{3}d'^3 + (d' + 1)^2N + 2C(d' + 1)^2)$ and $O(d'^3 + 2d'^2C)$, respectively. According to the GPI method [42], the computational complexity of solving $\mathbf{P}$ is $O(d^2(d+3d') + d^2d'k')$. Here, k' denotes the iteration number in the GPI method, which is typically small. Furthermore, we assume that $d, N \gg d', C$. Therefore, the computational complexities of solving $\mathbf{E}$, $\mathbf{T}$ and $\mathbf{B}$ (i.e., steps 4, 5 and 7) can be negligible. Finally, the overall computational complexity of Algorithm 1 in each iteration is approximately $O(d^3 + d^2N)$.

C. Differences Between Proposed Method and Some Related Methods

In this subsection, we highlight the main differences between our method and other related methods to illustrate the advantages of the proposed method.

1) Comparison to CDPL [27]: CDPL uses low-rank/sparsity representations and linear regression to construct an overall model for subspace learning and classification. The objective function is defined as follows:

$$\begin{aligned}
\min_{\mathbf{Z},\mathbf{E},\mathbf{P},\mathbf{T},\mathbf{D}} \; & \|\mathbf{Z}\|_* + \lambda\|\mathbf{Z}\|_1 + \gamma\|\mathbf{E}\|_{2,1} + \beta\,\mathrm{Tr}(\mathbf{Z}\mathbf{L}\mathbf{Z}^T)\\
& + \alpha/2\|\mathbf{Y} - \mathbf{T}\mathbf{D}\|^2\\
s.t.\ & \mathbf{P}^T\mathbf{X} = \mathbf{P}^T\mathbf{X}\mathbf{Z} + \mathbf{E}, \mathbf{Z} \geq 0,\\
& \mathbf{P}^T\mathbf{P} = \mathbf{I}_{d'}, \mathbf{D} = [\mathbf{P}^T\mathbf{X}; \mathbf{1}_N^T], \qquad (27)
\end{aligned}$$

where $\mathbf{L}$ is a Laplacian matrix, and $\lambda, \gamma, \beta, \alpha$ are positive regularization parameters. $\mathbf{P}^T\mathbf{X}\mathbf{Z}$ represents the "clean" recovered data in a union of low-dimensional subspaces, and $\mathbf{P}^T\mathbf{X}$ can be considered to be the low-dimensional features of the data corrupted by sparse noise. Both CDPL and DDLRPL jointly optimize low-rank recovery, LSR, and projection learning to equip models with noise suppression, classification, and dimensionality reduction functionalities. However, in the fifth term of (27), the noisy data $\mathbf{P}^T\mathbf{X}$ are used directly in the LSR term, which is sensitive to corruptions. This may degrade the overall robustness of the model. By contrast, DDLRPL uses recovered "clean" data to construct the LSR term to avoid this problem. Additionally, whereas the fourth term of CDPL learns the discriminative representation by concentrating the LRR of intra-class neighbors, thus avoiding the influence of outliers, CDPL requires the manual selection of neighbors, the number of which is difficult to determine. In DDLRPL, the $L_{2,1}$ norm is used as a metric to concentrate samples around their class centroids to learn the discriminative projections $(\mathbf{P}, \mathbf{Q})$ and the class centroids $\mathbf{B}$. Because using $L_{2,1}$ norm can reduce the contribution from outliers to the objective function, our method can achieve robustness against outliers without the manual selection of neighbors.

2) Connections to CLRR [26]: CLRR aims to identify the lowest-rank representation $\mathbf{Z} \in \mathbb{R}^{N\times N}$ and discriminative robust projection subspace $\mathbf{P} \in \mathbb{R}^{d\times N}$. The objective function can be formulated as

$$\begin{aligned}
\min_{\mathbf{Z},\mathbf{E},\mathbf{P}} \; & \|\mathbf{Z}\|_* + \lambda\|\mathbf{E}\|_{2,1} + \alpha\,\mathrm{Tr}\left(\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}\right) + \beta\|\mathbf{V} - \mathbf{Y}\|^2\\
s.t.\ & \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}, \mathbf{V} = \mathbf{P}^T\mathbf{X}\mathbf{Z}, \mathbf{1}_N^T\mathbf{Z} = \mathbf{1}_N, \qquad (28)
\end{aligned}$$

where λ, α, β are positive regularization parameters, and $\mathbf{L}$ is a Laplacian matrix containing the intrinsic properties of the data. CLRR enables the recovered "clean" low-dimensional representation to fit into the label matrix, similar to in the proposed method. However, its projection dimension is fixed to the number of classes because it uses a single matrix $\mathbf{P}$ for projection and regression. By contrast, the proposed method introduces $\mathbf{T}$ to transform the recovered data $\mathbf{Q}\mathbf{X}$ in low-dimensional subspaces $\mathbf{P}$ into a label-indicator matrix. Therefore, the projection $\mathbf{P}$ in the proposed method can have more dimensions for feature extraction, further improving

classification performance. Furthermore, CLRR learns the low-rank matrix for testing data to recover its "clean" components. By contrast, our method learns a low-rank matrix $\mathbf{Q}$ from the training dataset and uses it to project new testing samples onto the corresponding "clean" components.

3) Relationship With RLDA [22]: RLDA uses the $L_{2,1}$ norm to reduce the influence of outliers on the objective function value. The objective function is defined as follows:

$$\min_{\boldsymbol{b}_k,\, \mathbf{P}^T\mathbf{S}_t\mathbf{P}=\mathbf{I}_{d'}} \sum_{k=1}^{C} \sum_{\boldsymbol{x}_i \in \boldsymbol{\pi}_k} \|\mathbf{P}^T (\boldsymbol{x}_i - \boldsymbol{b}_k)\|_2, \tag{29}$$

where $\pi_k \in \mathbb{R}^{d\times n_k}$ and n_k denote the dataset of class k and the number of samples from class k, respectively. $\mathbf{P} \in \mathbb{R}^{d\times d'}$ is the projection matrix. The total-class scatter matrix is defined as $\mathbf{S}_t = \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$, where $\bar{\boldsymbol{x}} = 1/N \sum_{i=1}^{N} \boldsymbol{x}_i$. Similar to in the proposed method, the class centroids $\boldsymbol{b}_k$ in RLDA are learned adaptively to avoid the influence of outliers and noise. However, the objective function in RLDA does not have constraint terms separating different class centroids to the greatest extent; therefore, its classification performance may be influenced negatively. By using the fifth term in (8), our method can separate the different class centroids by exploiting LSR.

4) Comparison to Orthogonal Low-Rank Projection Learning (OLRPL) [33]: OLRPL uses the weighted truncated Schatten p-norm $\|\cdot\|_{w,r}^p$ to replace the nuclear norm used to approximate the rank function, and meanwhile, the correntropy $\varphi(\cdot)$ is applied to replace the L_2 norm and $L_{2,1}$ norm used for the classification and removal of sparse noise. The objective function is defined as follows:

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{E},\hat{\mathbf{E}},\mathbf{Q},\mathbf{P}} \; & \frac{1}{2}\sum_{i,j} \varphi\left(\hat{e}_{ij}\right) + \lambda \sum_{i,j} \varphi\left(e_{ij}\right) + \gamma \|\mathbf{Q}\|_{2,1} + \beta \|\mathbf{X}\|_{w,r}^p \\ s.t. \; & \mathbf{X} = \mathbf{P}\mathbf{Q}^T\mathbf{X} + \mathbf{E}, \mathbf{Q}^T\mathbf{X} = \mathbf{Q}^T\mathbf{X}\mathbf{Z}, \\ & \mathbf{Y} = \mathbf{Q}^T\mathbf{X} + \hat{\mathbf{E}}, \mathbf{P}^T\mathbf{P} = \mathbf{I}_C \end{aligned} \tag{30}$$

where $\hat{e}_{ij}$ and e_{ij} are the (i, j)-th elements of $\hat{\mathbf{E}}$ and $\mathbf{E}$. OLRPL first selects the most discriminative features of the data by the row-sparsity projection $\mathbf{Q}$, and then removes the noise by reconstructing the data. For comparison, our method first uses a low-rank projection to extract "clean" components from the noisy data, and then creates new low-dimensional features by using a semi-orthogonal projection to combine all the features of the "clean" data. In OLRPL, projection $\mathbf{Q}$ and the label-indicator matrix $\mathbf{Y}$ have the same ambient dimension; thus, the dimensions of projection $\mathbf{Q}$ and "clean" data $\mathbf{P}\mathbf{Q}^T\mathbf{X}$ are limited by the number of classes. By contrast in our method, the addition of transformation matrix $\mathbf{T}$ can enable the dimension of projection $\mathbf{P}$ to be flexible, and the dimension of "clean" components $\mathbf{Q}\mathbf{X}^T$ is adaptively learned.

V. Experiments

In this section, experiments were conducted to evaluate the performance of the proposed method on four face databases: CMU-PIE, ORL, AR, and PubFig83 databases; and two object databases: COIL20 and Caltech101 databases.

A. Descriptions of Databases

The AR face database [43] contains over 4000 face images from 126 subjects. A subset of AR containing 1400 face images from 50 men and 50 women was used in our experiments. Each subject had 14 face images in this subset, with varying facial expressions and illumination conditions. All images were cropped and resized to 43×60 pixels. Sample images of one person from this database are presented in Fig. 1(a).

The CMU PIE face database [44] contains 41368 face images from 68 subjects with different poses, illumination conditions, and expressions. A subset (C29, rightward pose) of the face database was selected for our experiments. This subset contained 1632 images from 68 subjects, with each subject having 24 images. The images were resized to 64×64 pixels. To verify the robustness of the proposed method, experiments were conducted on three CMU PIE databases with the inclusion of salt-and-pepper noise at densities of 0.03, 0.06, and 0.09. Sample images of one person, with different noise densities, are presented in Fig. 1 (b).

The YALE database [18] contains 165 images from 15 individuals, with each individual providing 11 images for various facial experiments and lighting conditions. For our experiments, each image was cropped and resized to 32×32 pixels. To verify the robustness of the proposed method, two YALE databases with continuous occlusion were generated by adding a baboon face with random intensity to half of the images and all images, respectively. Fig. 1 (c) shows sample images of one person from this database.

The COIL20 database [45] contains 1440 images of 20 artificial objects. The objects were placed on a motorized turntable against a black background. The turntable was rotated by 360°, and images of each object were captured every 5° using a fixed gray camera. This corresponds to 72 gray images per object. For our experiments, each image was resized to 32×32 pixels. All images in the database were randomly added to the gray images with different levels. Thereafter, block occlusions of different sizes were added. The sizes of the blocks were 6×6 and 12×12 pixels. Corrupted sample images of one object from this database are shown in Fig. 1 (e).

The Caltech101 database [46] contains 9146 images from 101 object classes and an additional background class. Each class has 31–800 images, and the size of each image is roughly 300×200 pixels. Sample images are shown in Fig. 1 (f). In the experiments, the ImageNet pre-trained VGG16 [47] was used to extract deep features (4096-dimensional) of images from the Caltech101 database.

The PubFig83 database [48] is a large-scale face database collected from the Internet. The database contains 13002 faces of 83 individuals, divided into 8720 images for training and 4282 images for testing. Each individual has 46–231 images, and the size of each image is 250×250. Sample images are shown in Fig. 1 (d). Following the experimental settings presented in [48], the histogram of oriented gradients (HOG), local binary patterns (LBP), and Gabor wavelet features were extracted from the aligned images, and reduced to 2048 dimensions with PCA. The first 1536 dimensions of the descriptors were used in the experiments.

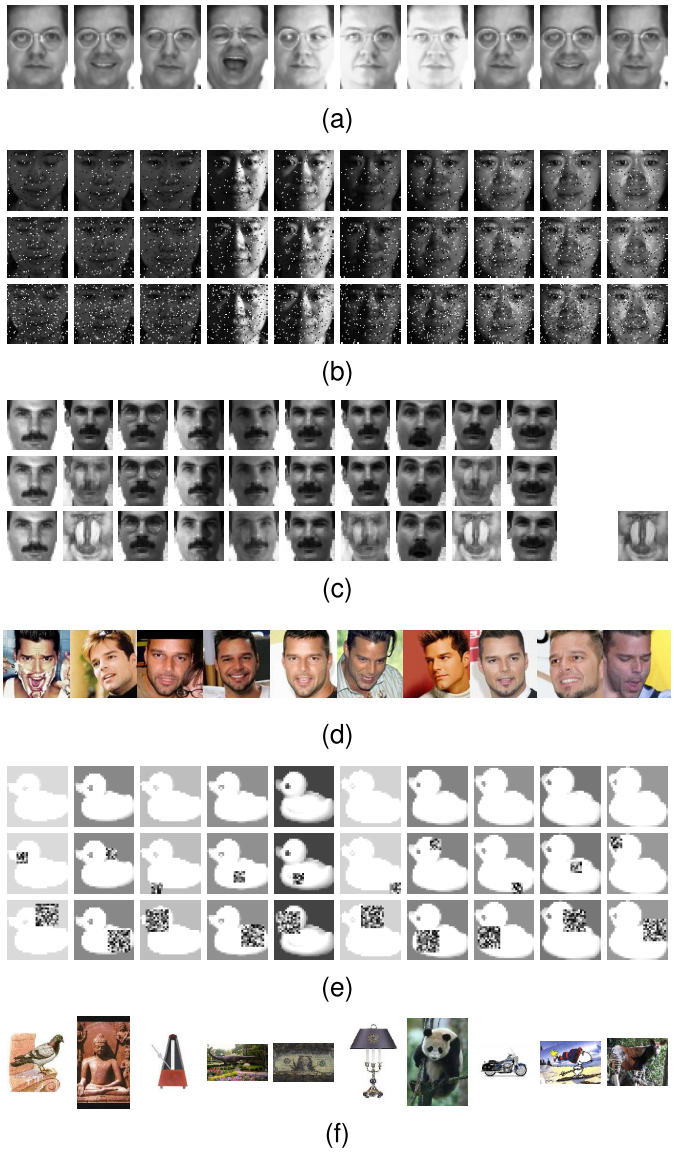

Fig. 1. Image samples from different databases. (a) AR; (b) CMU PIE with salt-and-pepper noise, first row: noise density = 0.03, second row: noise density = 0.06, third row: noise density = 0.09; (c) YALE, first row: original images, second row: half of images corrupted by baboon image (right), third row: all images corrupted by baboon image (right); (d) PubFig83; and (e) COIL20 corrupted by gray images, first row: no block, second row: 6×6 block, third row: 12×12 block; (f) Caltech101.

B. Experimental Setup

In our experiments, the proposed method was compared to various projection learning-based robust feature extraction methods. The compared methods included unsupervised methods such as RPCA [11] and low-rank adaptive graph embedding (LRAGE) [49], and supervised methods such as LRDE [50], RDR [31], RSLDA [21], CDPL [27], TNNL [20], and RDPDG [51]. Additionally, a pre-trained VGG16 network was also compared with these methods on the Caltech101 and the PubFig83 databases. Following the experimental settings in [21], the other two deep learning methods, i.e., DeepLDA [52] and Alexnet [53], are also evaluated on the PubFig83 database. To improve computational efficiency, PCA was utilized for all methods, on all databases except PubFig83, to extract low-dimensional features with 99% high-dimensional data energy in the preprocessing stage. For classification, all methods used a nearest-neighbor (NN) classifier, except for CDPL, which used its own classification method [27].

TABLE III
EXPERIMENTAL CONFIGURATION

database	input dimension	reduced dimension	L
AR	2580	5:5:200	3, 4, 5
CMU PIE	4096	5:5:160	3,5
YALE	1024	2:2:70	6
COIL20	1024	2:2:100	10
Caltech101	4096	102	5, 10, 15, 20
PubFig83	1536	100	-

In our experiments, L images of each class were randomly selected as training samples, and the remaining images were used for testing. L was set to 6 for the YALE face database, and 10 for the COIL20 object database. For the AR and CMU PIE face databases, L was set to 3, 4, and 5; and 3 and 5, respectively. For the Caltech101 database, L was set to 5, 10, 15 and 20. The subspace dimensions were varied from 5 to 200 and 5 to 160 in steps of 5 for the AR and CMU PIE face databases, respectively. For the YALE face database and COIL 20 object database, the subspace dimensions were set to the ranges 2–70 and 2–100, respectively, in steps of 2. For the Caltech101 object database, the subspace dimension was set to 102, which is the number of classes. For the PubFig83 database, the subspace dimension was set to 100. The experimental setup is summarized in Table III. Additionally, the neighborhood parameters for the intrinsic graphs involved in LRAGE, LRDE, RDR, CDPL and RDPDG were set to $L - 1$ for all databases, except PubFig83, for which a value of 40 was used. The neighborhood parameter for the penalty graph involved in LRDE was selected from $\{1C, 2C, 3C\}$. All the compared methods selected the optimal regularization parameters using a grid-search strategy. For LRAGE, LRDE, and RSLDA, the regularization parameters were selected from $[10^{-5}, 10^{-4}, \ldots, 10^{5}]$. For RDR, CDPL, TNNL, RDPDG and the proposed method, the regularization parameters (i.e., $\gamma, \alpha, \beta, \lambda$ in DDLRPL) were selected from $[10^{-3}, 10^{-2}, \ldots, 10^{3}]$.

C. Experimental Results and Analysis

To evaluate the effectiveness of the proposed method, we conducted experiments on the AR, CMU PIE, YALE, COIL20, Caltech101, and PubFig83 databases with different types and levels of corruption. All compared methods databases were executed 10 times. Tables IV, V, VI, VII, VIII report the experimental results, which include the highest average classification accuracies with corresponding standard deviations and dimensions. Table IX shows the classification accuracy on the large-scale PubFig83 face database. Fig. 2 shows the variation in the highest average classification accuracy with the training sample size on the AR database and levels of corruption on the CMU PIE, YALE, and COIL20 databases. We also evaluated the classification accuracies of the different methods by varying the projection dimensions,

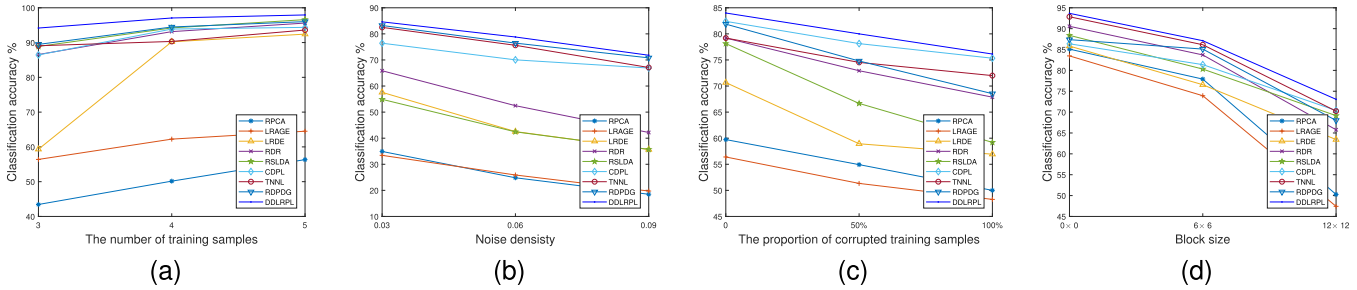

Fig. 2. Highest average classification accuracies versus variation in (a) training sample size on AR database, (b) density of salt-and-pepper noise on CMU PIE database, (c) proportion of corrupted training samples in YALE database, (d) size of block corruption in COIL20 database.

as shown in Fig. 3. The observations, results, and conclusions are listed hereunder.

1) In most cases, the supervised methods exhibited higher classification accuracies than those of the unsupervised methods (i.e., RPCA and LRAGE). Although the unsupervised methods had similar performance to those of some of the supervised methods on the COIL20 database without block occlusion and with 6×6 block occlusion, the performance of the unsupervised methods degraded significantly more sharply than those of the supervised methods on the COIL20 database with 12×12 block occlusion. This demonstrates that supervised information enhances robustness against gross corruptions.
2) In all experiments, our method achieved higher average classification accuracies than those of the other methods. Additionally, the standard deviation of our method was competitive compared to those of the other methods. Particularly, according to the classification results reported in Table IV and VIII, when 3, 4, and 5 training samples of each class were selected from the AR database, the classification accuracies of the proposed method were 4.71%, 2.60% and 1.34% higher than those of the second-best methods, respectively. When 5, 10, 15, and 20 training samples of each class were selected from the Caltech101 database, the classification accuracies of the proposed method database were 2.93%, 1.79%, 0.82%, and 0.76% higher than those of the second-best method, respectively. These results illustrates that the performance of our method was superior when the training samples are small. The main reason for this result may be that our method can reduce the influence of outliers and preserve the intrinsic structure by using the $L_{2,1}$ norm to learn adaptive class centroids.
3) Various corruptions were added to the different databases in our experiments to evaluate the robustness of our method. It can be observed in Tables V, VI, VII and Fig. 2 that our method demonstrated highest classification accuracies at any level of corruption on all the databases. As the level of corruption increasesd, the classification accuracies of DDLRPL degraded more slowly than most compared supervised methods. These results demonstrated that our method is robust against various corruptions, which is mainly due to the improvements in three aspects. First, nuclear and $L_{2,1}$ norms are used to recover "clean" low-rank components and remove the sparse noise. Second, in DDLRPL, the features extracted from "clean" data are used for the least-square regression term to avoid incurred noise error. Meanwhile, "clean" low-dimensional features used for testing can be obtained by using dual projections. Third, using the $L_{2,1}$ norm to learn adaptive class centroids reduces the influence of outliers.
4) As shown in Fig. 3, RDR and TNNL exhibited worse performance when the projection dimension was significantly greater than the number of classes. However, the performance of our method did not degrade as the projection dimension increased. This may be because RDR and TNNL use a single semi-orthogonal projection for classification, and a greater number of dimensions may cause over-fitting or fail to remove abundant information and noise that interfere with discrimination. By contrast, in our method, the dual projections $\mathbf{P}^T\mathbf{Q}$ used for feature extraction are composed of a low-rank projection $\mathbf{Q}$ and a semi-orthogonal projection $\mathbf{P}$. The dimensions of low-rank projection are adaptively learned by using the nuclear norm. The dimension of extracted features $\mathbf{P}^T\mathbf{QX}$ depends on the smaller one between the dimensions of the dual projections; thus, when selecting large dimensions for projection $\mathbf{P}$, the proposed method can still obtain adaptive lower-dimensional features that can remove abundant information and noise. Additionally, the LSR-based method, namely CDPL, performs worse if the projection dimension is smaller than the number of classes, whereas our method can obtain high classification

TABLE IV

CLASSIFICATION PERFORMANCE (HIGHEST AVERAGE CLASSIFICATION ACCURACIES (%), STANDARD DEVIATIONS (%), AND PROJECTION DIMENSIONS) OF DIFFERENT METHODS AND THE CORRESPONDING STANDARD DEVIATIONS (%) ON THE AR FACE DATABASE WITH DIFFERENT TRAINING SAMPLE SIZES

L	3	4	5
RPCA	43.44±1.66(180)	50.15±1.49(195)	56.34±0.81(185)
LRAGE	56.40±2.68(175)	62.23±2.22(195)	64.51±4.84(200)
LRDE	59.33±6.34(70)	90.22±1.13(75)	92.42±1.23(55)
RDR	86.56±1.23(50)	93.14±0.85(50)	95.61±0.86(70)
RSLDA	88.88±1.69(145)	94.21±1.06(105)	96.60±0.69(110)
CDPL	86.44±1.04(155)	93.88±0.74(200)	94.37±1.45(190)
TNNL	89.08±3.57(55)	90.32±1.17(105)	93.60±0.89(135)
RDPDG	89.47±1.42(160)	94.47±0.95(200)	96.02±0.42(180)
DDLRPL	**94.18±1.24(60)**	**97.07±0.72(90)**	**97.94±0.50(150)**

TABLE V

CLASSIFICATION PERFORMANCE (HIGHEST AVERAGE CLASSIFICATION ACCURACIES (%), STANDARD DEVIATIONS (%), AND PROJECTION DIMENSIONS) OF DIFFERENT METHODS ON CMU PIE FACE DATABASE WITH SALT-AND-PEPPER NOISE AT DIFFERENT DENSITIES (DEN.)

L	3 train			5 train		
noise	Den.=0.03	Den.=0.06	Den.=0.09	Den.=0.03	Den.=0.06	Den.=0.09
RPCA	34.92±0.83(120)	24.82±1.02(30)	18.51±1.30(20)	52.95±2.07(55)	52.89±1.14(35)	32.48±2.10(25)
LRAGE	33.42±1.34(110)	25.89±1.31(75)	19.74±1.35(50)	60.91±4.82(110)	54.30±1.41(50)	33.44±1.47(45)
LRDE	57.58±11.92(60)	42.54±6.64(160)	35.67±7.85(145)	70.26±9.09(130)	57.76±7.01(160)	44.43±9.99(50)
RDR	65.86±1.67(20)	52.46±1.58(15)	42.19±1.76(10)	80.96±2.02(25)	72.62±2.13(20)	63.05±1.92(15)
RSLDA	54.89±2.49(70)	42.44±1.85(80)	35.72±2.60(150)	82.14±2.31(5)	75.26±1.48(35)	53.61±1.63(130)
CDPL	76.41±2.65(160)	70.04±2.50(160)	66.90±1.64(155)	84.98±1.53(100)	82.63±0.95(140)	77.51±0.76(150)
TNNL	82.49±1.98(65)	75.62±2.60(75)	67.11±1.06(50)	91.03±0.70(70)	87.76±0.90(95)	83.54±0.78(75)
RDPDG	83.21±1.41(50)	76.45±2.69(55)	70.81±1.60(55)	90.48±0.67(55)	86.05±1.35(60)	81.87±1.04(45)
DDLRPL	**84.61±1.59(160)**	**78.80±2.42(155)**	**71.81±1.65(160)**	**91.44±0.97(160)**	**87.84±1.15(155)**	**84.81±0.40(160)**

TABLE VI

CLASSIFICATION PERFORMANCE (HIGHEST AVERAGE CLASSIFICATION ACCURACIES (%), STANDARD DEVIATIONS (%), AND PROJECTION DIMENSIONS) ON YALE FACE DATABASE. OCC. REPRESENTS THE PERCENTAGE OF IMAGES THAT ARE CORRUPTED

Occ.	0%	50%	100%
RPCA	59.73±3.60(62)	54.93±5.96(64)	50.00±4.80(42)
LRAGE	56.40±4.83(50)	51.33±3.99(66)	48.27±4.25(42)
LRDE	70.67±5.66(32)	58.93±7.27(62)	56.93±6.35(38)
RDR	79.20±3.09(10)	72.93±4.36(12)	67.87±7.58(12)
RSLDA	78.13±3.22(16)	66.67±5.90(56)	59.20±4.41(18)
CDPL	82.40±3.49(64)	78.13±4.54(64)	75.33±5.56(60)
TNNL	79.20±5.23(26)	74.53±4.81(14)	72.00±4.83(20)
RDPDG	81.87±3.62(64)	74.80±5.78(70)	68.53±5.27(58)
DDLRPL	**84.00±3.20(28)**	**80.00±4.12(48)**	**76.13±4.72(24)**

TABLE VII

CLASSIFICATION PERFORMANCE (HIGHEST AVERAGE CLASSIFICATION ACCURACIES (%), STANDARD DEVIATIONS (%), AND PROJECTION DIMENSIONS) ON COIL20 OBJECT DATABASE WITH DIFFERENT SIZES OF BLOCK CORRUPTION

Block size	0×0	6×6	12×12
RPCA	85.09±1.53(62)	77.91±1.45(54)	50.27±1.71(66)
LRAGE	83.47±1.28(42)	73.93±2.22(48)	47.42±2.62(56)
LRDE	85.82±1.61(20)	76.56±2.08(98)	63.42±0.73(82)
RDR	90.56±1.46(14)	83.70±1.42(16)	65.77±2.17(14)
RSLDA	88.40±1.07(26)	80.31±1.39(30)	69.15±1.40(48)
CDPL	86.37±1.31(68)	81.38±2.00(82)	70.40±1.34(96)
TNNL	92.85±1.23(12)	86.15±1.65(28)	70.23±1.79(24)
RDPDG	87.36±1.17(100)	85.10±0.84(84)	68.00±1.44(76)
DDLRPL	**93.67±1.34(16)**	**87.11±1.74(26)**	**73.06±1.30(12)**

TABLE VIII

CLASSIFICATION PERFORMANCE (HIGHEST AVERAGE CLASSIFICATION ACCURACIES (%), STANDARD DEVIATIONS (%)) ON DEEP LEARNING FEATURES OF CALTECH101 OBJECT DATABASE

L	5	10	15	20
VGG16	75.25±1.73	81.65±0.70	83.81±0.57	85.71±0.59
RPCA	75.10±1.01	78.12±0.51	80.00±0.46	81.32±0.33
LRAGE	75.50±1.08	78.89±0.53	80.93±0.39	82.28±0.46
LRDE	76.42±1.05	83.45±0.48	86.82±0.38	88.52±0.43
RDR	78.06±1.10	83.76±0.56	86.88±0.41	88.48±0.35
RSLDA	77.79±1.08	80.17±0.84	84.10±0.75	86.17±0.61
CDPL	77.97±0.77	80.64±0.86	81.45±1.11	81.99±1.01
TNNL	75.51±1.05	78.39±0.51	80.14±0.49	81.43±0.37
RDPDG	77.04±1.33	82.73±0.78	85.14±0.65	86.53±0.31
DDLRPL	**80.99±1.19**	**85.55±0.57**	**87.70±0.40**	**89.28±0.31**

TABLE IX

CLASSIFICATION ACCURACIES (%) ON PUBFIG83 FACE DATABASE

Method	Acc.	Method	Acc.
RPCA	49.79	LRAGE	51.07
LDA	77.95	LRDE	86.59
RDR	86.45	RSLDA	84.78
TNNL	84.35	RDPDG	74.10
DeepLDA	44.35	Alexnet	44.35
VGG	**96.25**	DDLRPL	87.58

accuracies with a low projection dimension. Evidently, the robustness of our method against the selection of the projection dimension is superior to those of the other methods.

5) For the Caltech101 database, the VGG16 convolutional network had lower classification accuracies than those of some of the conventional supervised learning methods, such as LRDE, RDR, RDPDG and the proposed method. For PubFig83 databse, two deep learning methods, i.e., DeepLDA and Alexnet, performed worse than conventional methods. Pre-trained VGG achieved the highest classification accuracy. However, the VGG without a pre-trained model cannot converge in the training procedure. The reason for these experimental results is that the deep networks contain large numbers of parameters, and thus the limited training samples may have caused the model to overfit or fail to convergenve.

D. Analytical Experiments

To evaluate the effectiveness of each component of the proposed DDLRPL, we considered five special cases of our method. (1) DDLRPL$_{\beta,\lambda}$ denotes DDLRPL without the discriminative terms by setting $\beta, \lambda = 0$, resulting in problem (7). (2) DDLRPL$_{\alpha,\lambda}$ denotes DDLRPL without the LSR terms by setting $\alpha, \lambda = 0$. (3) DDLRPL$_{\text{dual}}$ denotes DDLRPL without the low-rank recovery term, which implies the low-rank projection matrix $\mathbf{Q}$ and noise $\mathbf{E}$ do not exist. DDLRPL$_{\text{dual}}$ uses a single semi-orthogonal matrix $\mathbf{P}$ for feature extraction, and the objective function of DDLRPL$_{\text{dual}}$ is rewritten in (31). (4) DDLRPL$_{\lambda}$ denotes DDLRPL without the LSR term for

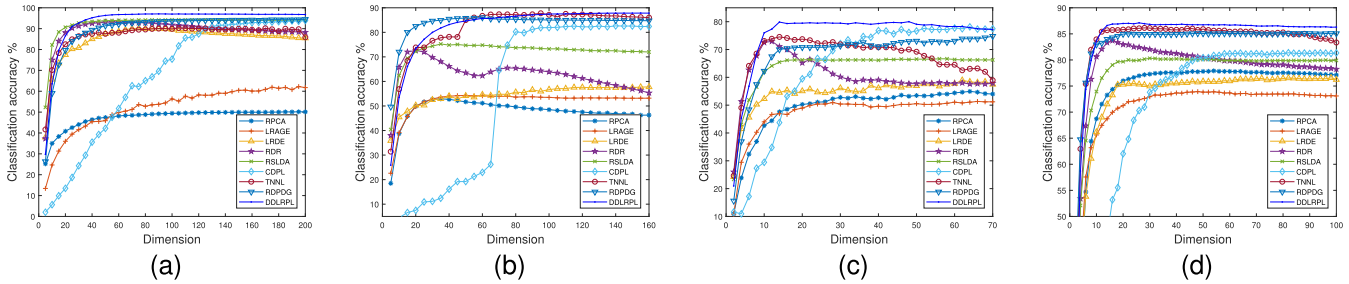

Fig. 3. Average classification accuracies versus the number of subspace dimensions for different methods on (a) AR face database, (b) CMU PIE face database, (c) YALE face database, and (d) COIL20 object database.

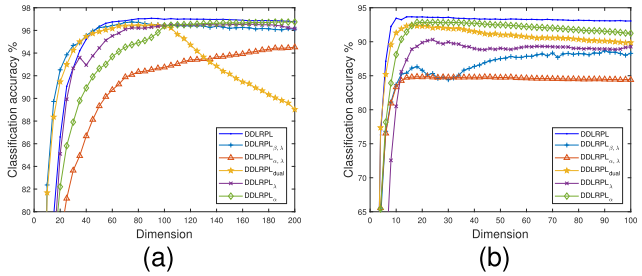

Fig. 4. Ablation study of the proposed method on the (a) AR face database and (b) COIL20 object database.

TABLE X
RUNNING TIME COMPARISON OF DIFFERENT METHODS ON CALTECH101 OBJECT DATABASE

Method	Training Time (s)	Test Time (s)
RPCA	3.8301	0.6554
LRAGE	4.2766	0.6498
RSLDA	11.6235	0.6116
LRDE	96.1081	0.6594
RDR	8.8955	0.6451
CDPL	206.3853	0.0219
TNNL	254.2283	0.6562
RDPDG	6.3517	0.6333
DDLRPL	67.2428	0.6550

classifying class centroids $\mathbf{B}$ by setting $\lambda = 0$. (5) DDLRPL$_\alpha$ denotes DDLRPL without the LSR term for classifying projected data by setting $\alpha = 0$. These baselines apply the strategy of previous parameter selection for DDLRPL.

$$\begin{aligned}\min_{\mathbf{T},\mathbf{B},\mathbf{P}} \; & \|\mathbf{P}^T\mathbf{X} - \mathbf{B}\mathbf{Y}\|_{2,1} + \alpha/2\|\mathbf{Y} - \mathbf{T}\left[\mathbf{P}^T\mathbf{X}; \mathbf{1}_N^T\right]\|^2 \\ & + \lambda/2\|\left(\mathbf{I}_C - \mathbf{T}\left[\mathbf{B}; \mathbf{1}_C^T\right]\right)\mathbf{W}\|^2 \\ s.t. \; & \mathbf{P}^T\mathbf{P} = \mathbf{I}_{d'} \end{aligned} \tag{31}$$

The experimental results for the AR face database and COIL20 object database are presented in Fig. 4. Clearly, DDLRPL$_{\alpha,\lambda}$ performs worse than DDLRPL$_\alpha$ and DDLRPL$_\lambda$, whereas DDLRPL performs better. Therefore, the use of LSR terms is highly effective for classification tasks and the separation of class centroids, whereas the latter is important for preserving the discriminability of extracted features. The classification accuracy of DDLRPL$_{\text{dual}}$ is lower than that of DDLRPL and degrades at larger projection dimensions for both databases. This result demonstrates that using the dual projections reduces the sensitivity of the method to the selection of projection dimensions. Furthermore, the poorer performance of DDLRPL$_{\beta,\lambda}$ demonstrates that directly preserving the discriminability and intrinsic structure of low-dimensional features can improve performance to some extent.

E. Parameter Analysis

In this subsection, we analyze the proposed method by studying variations in its parameters. Four critical regularization parameters exist in the proposed method: γ, β, α, and λ. To study the effectiveness of these parameters, experiments were conducted to examine variations in the classification accuracy of the proposed method on the AR database by setting different parameter values. Specifically, the values of two parameters were tuned in the range $[10^{-3}, 10^3]$, while the other two parameters were fixed to 1. The classification accuracy of the proposed method with respect to variations in the parameters was presented in Fig. 5. The results demonstrate that DDLRPL can select the suitable parameters in a wide range to achieve high classification performance.

F. Convergence Study

The convergence properties of the proposed method were proven in Section IV-A, indicating that the proposed method can converge to a local optimum under mild conditions. We conducted additional experiments on the AR face and Caltech101 object databases to validate the convergence of the proposed method. The convergence curves presented in Fig. 6 represent the value of the objective function (10) and classification accuracy versus the number of iterations. As shown in Fig. 6, the proposed method can ultimately converge to stable values for the objective function and classification accuracy on both databases.

G. Runtime Evaluation

The computational complexity of the proposed method has been analyzed in Section IV-B. In this subsection, we conducted experiments to compare the running times of different methods on the Caltech101 object database ($L = 20$). All experiments were conducted using MATLAB R2017B on Windows 10 with an Intel Core i7-7700 CPU and 32GB of RAM. Table X lists the experimental results. RPCA and LRAGE are unsupervised methods, and they had faster training times. The supervised methods using LRR, such as CDPL and

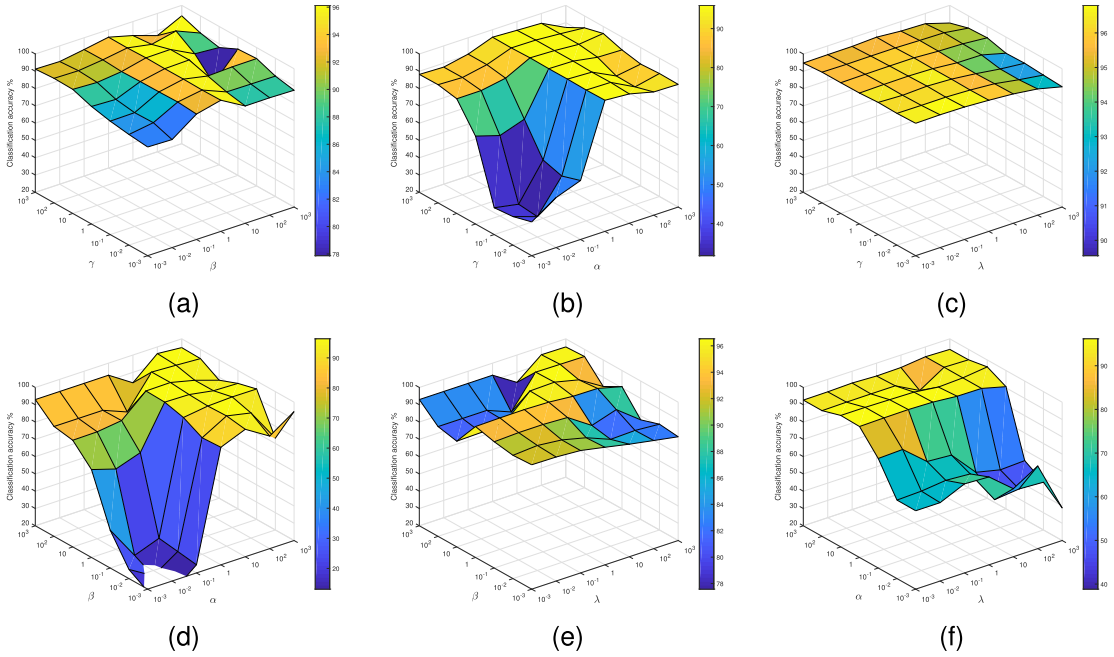

Fig. 5. Classification accuracy versus variations of parameters (a) β and γ, (b) α and γ, (c) λ and γ, (d) α and β, (e) λ and β, (f) λ and α on AR face database.

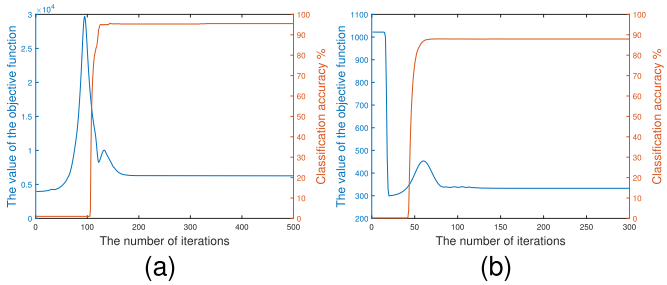

Fig. 6. Convergence properties of proposed method on (a) AR face and (b) Caltech101 object database.

TNNL, required much more training times than the other compared methods. On the other hand, although the test time of DDLRPL was similar to those of the other compared methods, the training time is slower than some supervised methods, such as RSLDA, RDR, and RDPDG. According to the analysis in Section IV-B, this is because dual projections lead to higher computational complexities. In future work, we will develop methods to reduce the training time of DDLRPL. One way to improve computational speed is to use acceleration methods for solving the non-smooth convex problem and the Stein equation included in the proposed method.

VI. Conclusion

In this paper, we propose a novel supervised feature extraction method called dual discriminative low-rank projections learning for robust image classification. The proposed method can use dual projections to extract "clean" low-dimensional features from testing data in real time. Moreover, dual projection enables the performance of the proposed method less affected by the predefined projection dimensions in the real system. The LSR-based regression and $L_{2,1}$-norm regularization constructions guarantee the discrimination and robustness of the proposed method. Comparative experiments were conducted on three databases (i.e., CMU PIE, YALE, and COIL20 databases) designed manually with various levels and types of noise and occlusion, and the classification accuracies of the proposed method increased by an average of 1.74%, 1.42%, and 1.54% compared to those of the closest competitors, respectively. These results demonstrate that the proposed method is robust to corruption. When a few training samples of each class were selected from the AR and Caltech101 databases, the proposed method improved the classification accuracies by 4.71% and 2.96%, respectively, compared to its closest competitors. Additionally, the proposed method exhibited 87.58% recognition accuracy on the larger-scale PubFig83 database. Thus, the proposed method is inferred to be able to produce improved performance for real-world image classification systems. In the future, the proposed method can be extended to the other real-time high-dimensional data processing tasks, such as hyperspectral image (HSI) processing [54] and handwritten word recognition [55].

References

[1] T. Georgiou, Y. Liu, W. Chen, and M. Lew, "A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision," *Int. J. Multimedia Inf. Retr.*, vol. 9, no. 3, pp. 135–170, Sep. 2020.

[2] K. Min, Z. Zhang, J. Wright, and Y. Ma, "Decomposing background topics from keywords by principal component pursuit," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2010, pp. 269–278.

[3] R. Liu and N.-M. Cheung, "Joint estimation of low-rank components and connectivity graph in high-dimensional graph signals: Application to brain imaging," *Signal Process.*, vol. 182, May 2021, Art. no. 107931.

[4] R. Tavoli and M. Keyvanpour, "A method for handwritten word spotting based on particle swarm optimisation and multi-layer perceptron," *IET Softw.*, vol. 12, no. 2, pp. 152–159, Apr. 2018.

[5] J. Liu, Y. Chen, J. Zhang, and Z. Xu, "Enhancing low-rank subspace clustering by manifold regularization," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4022–4030, Sep. 2014.

[6] X. Li, J. Pan, M. Gao, A. Souri, and J. Shang, "An improved blind/referenceless image spatial quality evaluator algorithm for image quality assessment," *Int. J. Comput. Sci. Eng.*, vol. 1, no. 1, p. 1, 2022.

[7] J. Tang, X. Shu, Z. Li, Y. Jiang, and Q. Tian, "Social anchor-unit graph regularized tensor completion for large-scale image retagging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 2027–2034, Aug. 2019.

[8] R. Tavoli, E. Kozegar, M. Shojafar, H. Soleimani, and Z. Pooranian, "Weighted PCA for improving document image retrieval system based on keyword spotting accuracy," in *Proc. 36th Int. Conf. Telecommun. Signal Process. (TSP)*, Jul. 2013, pp. 773–777.

[9] X. Shu, J. Tang, Z. Li, H. Lai, L. Zhang, and S. Yan, "Personalized age progression with bi-level aging dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 905–917, Apr. 2018.

[10] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, Jul. 2010.

[11] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 1, pp. 1–37, 2009.

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

[13] X. Shu, J. Tang, G.-J. Qi, Z. Li, Y.-G. Jiang, and S. Yan, "Image classification with tailored fine-grained dictionaries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 2, pp. 454–467, Feb. 2018.

[14] 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.

[15] B.-K. Bao, G. Liu, C. Xu, and S. Yan, "Inductive robust principal component analysis," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3794–3800, Aug. 2012.

[16] W. K. Wong, Z. Lai, J. Wen, X. Fang, and Y. Lu, "Low-rank embedding for robust image feature extraction," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2905–2917, Jun. 2017.

[17] Q. Wang, Q. Gao, G. Sun, and C. Ding, "Double robust principal component analysis," *Neurocomputing*, vol. 391, pp. 119–128, May 2020.

[18] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[19] Z. Liu, J. Wang, G. Liu, and L. Zhang, "Discriminative low-rank preserving projection for dimensionality reduction," *Appl. Soft Comput.*, vol. 85, Dec. 2019, Art. no. 105768.

[20] W. Wang et al., "TNNL: A novel image dimensionality reduction method for face image recognition," *Digit. Signal Process.*, vol. 115, Aug. 2021, Art. no. 103082.

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

[22] H. Zhao, Z. Wang, and F. Nie, "A new formulation of linear discriminant analysis for robust dimensionality reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 629–640, Apr. 2019.

[23] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, "Discriminative transfer subspace learning via low-rank and sparse representation," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 850–863, Feb. 2016.

[24] N. Han et al., "Double relaxed regression for image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 307–319, Feb. 2020.

[25] W. Wang, L. Fang, and W. Zhang, "Robust double relaxed regression for image classification," *Signal Process.*, vol. 203, Feb. 2023, Art. no. 108796.

[26] P. Li, J. Yu, M. Wang, L. Zhang, D. Cai, and X. Li, "Constrained low-rank learning using least squares-based regularization," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4250–4262, Dec. 2017.

[27] M. Meng, M. Lan, J. Yu, J. Wu, and D. Tao, "Constrained discriminative projection learning for image classification," *IEEE Trans. Image Process.*, vol. 29, pp. 186–198, 2020.

[28] N. Liu et al., "Locality preserving robust regression for jointly sparse subspace learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2274–2287, Jun. 2021.

[29] X. Fang, S. Teng, Z. Lai, Z. He, S. Xie, and W. K. Wong, "Robust latent subspace learning for image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2502–2515, Jun. 2018.

[30] Z. Chen, X.-J. Wu, Y.-H. Cai, and J. Kittler, "Sparse non-negative transition subspace learning for image classification," *Signal Process.*, vol. 183, Jun. 2021, Art. no. 107988.

[31] Z. Lai, D. Mo, W. K. Wong, Y. Xu, D. Miao, and D. Zhang, "Robust discriminant regression for feature extraction," *IEEE Trans. Cybern.*, vol. 48, no. 8, pp. 2472–2484, Aug. 2018.

[32] W. Yan, M. Yang, and Y. Li, "Robust low rank and sparse representation for multiple kernel dimensionality reduction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 1–15, Jan. 2023.

[33] X. Zhang, Z. Tan, H. Sun, Z. Wang, and M. Qin, "Orthogonal low-rank projection learning for robust image feature extraction," *IEEE Trans. Multimedia*, vol. 24, pp. 3882–3895, 2022.

[34] Z. Lai, J. Bao, H. Kong, M. Wan, and G. Yang, "Discriminative low-rank projection for robust subspace learning," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 10, pp. 2247–2260, Oct. 2020.

[35] S. Li and Y. Fu, "Learning robust and discriminative subspace with low-rank constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2160–2173, Nov. 2016.

[36] 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.

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

[38] 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.*, vol. 24, 2011, pp. 1–22.

[39] 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.

[40] G. H. Golub, S. G. Nash, and C. V. Loan, "A Hessenberg–Schur method for the problem $AX + XB = C$," *IEEE Trans. Autom. Control*, vol. AC-24, no. 6, pp. 913–990, Dec. 1979.

[41] H. C. Pinkham, *Linear Algebra, Version 10*. New York, NY, USA: Columbia University in the City of New York, Jul. 2015. [Online]. Available: https://www.math.columbia.edu/ pinkham/HCP_LinearAlgebra.pdf

[42] F. Nie, R. Zhang, and X. Li, "A generalized power iteration method for solving quadratic problem on the Stiefel manifold," *Sci. China Inf. Sci.*, vol. 60, no. 11, pp. 1–10, Nov. 2017.

[43] A. M. Martinez, "The AR face database," CVC, New Delhi, India, Tech. Rep., #24, 1998.

[44] 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, Dec. 2003.

[45] S. Nene et al. (1996). *Columbia Object Image Library (COIL-20)*. [Online]. Available: http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php

[46] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Comput. Vis. Image Understand.*, vol. 106, no. 1, pp. 59–70, Apr. 2007.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[48] B. C. Becker and E. G. Ortiz, "Evaluating open-universe face identification on the web," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 904–911.

[49] J. Lu, H. Wang, J. Zhou, Y. Chen, Z. Lai, and Q. Hu, "Low-rank adaptive graph embedding for unsupervised feature extraction," *Pattern Recognit.*, vol. 113, May 2021, Art. no. 107758.

[50] J. Li, Y. Wu, J. Zhao, and K. Lu, "Low-rank discriminant embedding for multiview learning," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3516–3529, Nov. 2017.

[51] H. Qu, L. Li, Z. Li, J. Zheng, and X. Tang, "Robust discriminative projection with dynamic graph regularization for feature extraction and classification," *Knowl.-Based Syst.*, vol. 253, Oct. 2022, Art. no. 109563.

[52] M. Dorfer, R. Kelz, and G. Widmer, "Deep linear discriminant analysis," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–12.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[54] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[55] A. M. Al-Shatnawi, F. Al-Saqqar, and A. Souri, "Arabic handwritten word recognition based on stationary wavelet transform technique using machine learning," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 3, pp. 1–21, May 2022.

Tingting Su received the B.S. degree from the School of Electronic Engineering, Xidian University, Xi'an, China, in 2016, where she is currently pursuing the Graduate degree. Her current research interests include pattern recognition and machine learning.

Dazheng Feng (Member, IEEE) received the Diploma degree from the Xi'an University of Technology, Xi'an, China, in 1982, the M.S. degree from Xi'an Jiaotong University, Xi'an, in 1986, and the Ph.D. degree in electronic engineering from Xidian University, Xi'an, in 1995. From May 1996 to May 1998, he was a Post-Doctoral Research Affiliate with Xi'an Jiaotong University. From May 1998 to June 2000, he was an Associate Professor with Xidian University. Since July 2000, he has been a Professor with Xidian University. His current research interests include signal processing, intelligence and brain information processing, and radar techniques.

Meng Wang received the B.S. degree from the School of Electronic Engineering, Xidian University, Xi'an, China, in 2016, where he is currently pursuing the Graduate degree. His current research interests include array signal processing, radar techniques, deep learning, and speaker recognition.

Mohan Chen received the bachelor's and master's degrees from the School of Information Science and Engineering, Lanzhou University, Lanzhou, China, in 2013 and 2016, respectively. She is currently pursuing the Ph.D. degree with the School of Electronic Engineering, Xidian University, Xi'an, China. Her current research interests include computational neuroscience and machine learning.