Low-Rank Preserving t-Linear Projection for Robust Image Feature Extraction

Xiaolin Xiao, Yongyong Chen, Yue-Jiao Gong, *Senior Member, IEEE*, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*— As the cornerstone for joint dimension reduction and feature extraction, extensive linear projection algorithms were proposed to fit various requirements. When being applied to image data, however, existing methods suffer from representation deficiency since the multi-way structure of the data is (partially) neglected. To solve this problem, we propose a novel Low-Rank Preserving t-Linear Projection (LRP-tP) model that preserves the intrinsic structure of the image data using t-product-based operations. The proposed model advances in four aspects: 1) LRP-tP learns the t-linear projection directly from the tensorial dataset so as to exploit the correlation among the multi-way data structure simultaneously; 2) to cope with the widely spread data errors, e.g., noise and corruptions, the robustness of LRP-tP is enhanced via self-representation learning; 3) LRP-tP is endowed with good discriminative ability by integrating the empirical classification error into the learning procedure; 4) an adaptive graph considering the similarity and locality of the data is jointly learned to precisely portray the data affinity. We devise an efficient algorithm to solve the proposed LRP-tP model using the alternating direction method of multipliers. Extensive experiments on image feature extraction have demonstrated the superiority of LRP-tP compared to the state-of-the-arts.**

***Index Terms*— Adaptive graph, low-rank tensor representation, robust feature extraction, t-linear projection learning, tensor-product (t-product).**

I. Introduction

IN FIELDS of image processing and computer vision, it is observed that the real-world data drawn from the high-dimensional ambient spaces are likely to approximately reside in low-dimensional intrinsic subspaces [1]. To well discover the underlying affinity of data, the projection learning approaches learn explicit projection bases to map the high-dimensional data into the low-dimensional subspaces, for joint dimension reduction and feature extraction. Along this research direction, the linear projection models are particularly attractive owning to the effectiveness with low computational costs [2]. Typically, Principal Component Analysis (PCA) [3] and linear discriminant analysis [4] were designed with different assumptions on the criteria of the subspaces.

In practice, the similarity graph plays a key role in revealing the intrinsic relationship of data [5]. In this respect, the graph-based projection learning witnesses enhanced performance by exploring the data similarity [6], [7]. Many previous methods pre-compute graphs from the data drawn from the ambient spaces, which may suffer from the side effect of the corruptions, e.g., noise and occlusions. To solve this problem, the recent works adopt the adaptive graph learning scheme in both unsupervised [5], [8] and supervised scenarios [9]. As such, the similarity and locality of the data are learned in a flexible manner to portray the relationship of data precisely.

Focusing on general data processing, classical projection learning methods primarily use vectors to represent samples. That is, to process the image set that naturally has a three-way structure,[1] these algorithms flatten the image data into vectors, breaking down the spatial coherency of images. 2D Projection (2DP) learning was proposed to alleviate this side effect to some extent. The essence is to learn the linear projection on either the row or the column space of the images [10]. In this respect, 2DPCA [11] and 2D Locality Preserving Projection (2DLPP) [12] generalize the PCA [3] and LPP [6] models, respectively. Later, the models in [13]–[16] substitute robust norms for the classical Euclidean metric to process corrupted data. For 2DP, however, only partial of the image structure is retained in the corresponding row/column direction. Considering this limitation, the Multi-Linear Projection (MLP) learning models [17]–[22] use the tensor representation to preserve the multi-way data structure. By learning projections from the unfoldings of tensors along all modes, not only the row and column spaces of images are considered, but also the cross-sample structure is exploited. In the sense, 2DP can be considered as a special case of MLP where only one unfolding along the row/column direction is considered. However, learning the projections separately from different unfoldings of tensors, existing MLP methods suffer from representation sub-optimality since they fail to exploit the correlation among the multi-way structure simultaneously [23].

Manuscript received April 27, 2020; revised September 22, 2020; accepted October 8, 2020. Date of publication October 22, 2020; date of current version November 19, 2020. This work was supported in part by the China Postdoctoral Science Foundation under Grant 2019M662913 and in part by the National Natural Science Foundation of China under Grant 62006080 and Grant 61873095. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Emanuele Salerno. *(Corresponding author: Yue-Jiao Gong.)*

Xiaolin Xiao and Yue-Jiao Gong are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: shellyxiaolin@gmail.com; gongyuejiao@gmail.com).

Yongyong Chen is with the Bio-Computing Research Center, Harbin Institute of Technology, Shenzhen, Shenzhen 518055, China.

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Taipa, Macau.

Digital Object Identifier 10.1109/TIP.2020.3031813

[1] We use "way" to denote the structure of data, and the three-way structure of an image set indicates the two-way structure within images as well as the cross-sample structure.

On the other hand, most of the aforementioned methods are likely to encounter problems in real scenarios due to the ubiquitous corruptions, e.g., noise and occlusions. Recently, the representation-based feature extraction has drawn considerable attention since it can handle corruptions by exploiting the fact that the optimal representations of the data are often sparse with respect to an overcomplete dictionary [24]. Using a dictionary consisting all training samples, one sample can be expressed by a linear combination of the samples in current dataset, which is called the *self-representation property*. The sparse representation [24], [25] and low-rank representation [26] were then proposed by imposing different constraints on the representation matrix. They use the representation coefficients as the new features, and thus, the dimensions of the extracted features equal to the number of training samples. This will bring computation burdens when many training samples are included. Besides, the methods in [24]–[26] are *transductive* ones and cannot handle new samples that are not involved in the training phase. To solve these problems, the works in [27]–[29] bond the representation learning and linear projection, known as the *representation-based projection learning*. They provide a setting where corruptions are alleviated via the self-representation data, from which the linear projection is learned. Algorithms within this category work in an *inductive* way since they can project the unseen samples for feature extraction. Continuing along this vein, the algorithms in [30]–[33] work in the unsupervised setting, whereas the model in [34] can be applied to both unsupervised and semi-supervised learning. To improve the discriminative ability of the extracted features, the works in [35]–[39] make use of the prior information from class labels. Specifically, in [35], [36], the discrimination is explored by maximizing the scatter of the projected between-class samples while minimizing that of within-class samples, inspired from discriminant analysis [40]. In the meantime, different regression-type modules are designed to incorporate the empirical classification error for discrimination enhancement [37]–[39]. However, these algorithms necessity the vectorization of samples. Thus, they suffer limitations when being applied to images since the multi-way structure residing in the dataset is sacrificed, and this side effect is irreversible for subsequent tasks.

To simultaneously exploit the correlation among the multi-way data structure, the tensor-product (t-product) based operations [41], [42] were proposed to overcome the limitation of the unfolding operation. Based on t-product, the "t-linear" combination of the tensor data has shown advanced performance in multi-way data clustering [43]–[46] and sparse coding [47] when compared to the matrix-based and unfolding-based methods. In light of the aforementioned concerns, we propose a Low-Rank Preserving t-Linear Projection (LRP-tP) model within the category of representation-based projection learning for robust image feature extraction. The key contributions of this paper are summarized as follows.

1) We propose a novel LRP-tP model for robust image feature extraction. Using t-product-based operations, the t-linear projection is learned by simultaneously exploiting the correlation among the multi-way data structure. Moreover, LRP-tP provides a physical interpretation on the learned projection basis, which resembles the linear projection basis in the vector space.

2) LRP-tP preserves the low-rankness of the self-representation tensor to alleviate data corruptions. The self-representation and projection learning mutually promote each other so as to achieve the overall optimum.

3) To improve the discriminative ability of the extracted features, LRP-tP works in a supervised manner by introducing a regression-type module. Moreover, an adaptive graph is learned simultaneously to receive benefits from the similarity and locality information of samples.

4) We design an iterative algorithm to solve the LRP-tP model using the alternating direction method of multipliers. Extensive experiments have shown that LRP-tP largely improves the effectiveness of the representation-based projection learning and competes well with the state-of-the-arts.

The remainder of this paper is organized as follows. Section II reviews the related work and preliminaries. Then, the LRP-tP model is elaborated in Section III. Extensive experimental verifications and model analysis are provided in Section IV to lead a clear understanding. Finally, Section V concludes the paper.

II. Related Work and Preliminaries

A. Representation-Based Projection Learning

Generally, the representation-based projection learning exploits the self-representation data to alleviate corruptions, and devises different error terms and regularizers to learn the optimal projection from the self-representation data. Since the proposed LRP-tP also belongs to this category, we briefly review the closely-related works, i.e., the Low-Rank Embedding model (LRE) [29], the Latent Low-Rank and Sparse Embedding model (LLRSE) [38], and the Constrained Discriminative Projection Learning model (CDPL) [39].

Let $X = [x_1, \ldots, x_h] \in \mathbb{R}^{d\times h}$ be the data matrix where each column corresponds to a sample. LRE learns the projection considering the self-representation data as

$$\begin{aligned} &\min_{P,E,Z} \ \|E\|_{2,1} + \lambda\|Z\|_* \\ &s.t. \quad P'X - P'XZ = E, \ P'P = I, \end{aligned} \tag{1}$$

where $P \in \mathbb{R}^{d\times b}$ represents the projection basis, and $b \leq \min(d, h)$ is the number of basis vectors; $Z \in \mathbb{R}^{h\times h}$ denotes the self-representation matrix, constrained by the matrix nuclear norm $\|\cdot\|_*$; $E \in \mathbb{R}^{b\times h}$ represents the error matrix.

The success of LRE in dealing with data corruptions verifies the effectiveness of representation-based projection learning. Later, two supervised methods were proposed to improve the discriminative ability of LRE together with other concerns. LLRSE promotes the row-sparsity of the learned projection basis vectors such that only partial variables are retained, for joint feature selection and projection learning:

$$\begin{aligned} &\min_{P,E,Z,R} \|E\|_{2,1} + \lambda_1\|Z\|_* + \lambda_2\|H - P'X\|_F^2 + \lambda_3\|P\|_{2,1} \\ &s.t. \quad X - RP'XZ = E, \ R'R = I, \end{aligned} \tag{2}$$

TABLE I
SUMMARY OF BASIC TENSOR NOTATIONS

Notation	Description
$\mathcal{X}$	third-order tensor
$\mathcal{X}(i,j,:)$	(i,j)-th tube of $\mathcal{X}$
$\mathcal{X}(i,:,:)$	i-th horizontal slice of $\mathcal{X}$
$\mathcal{X}(:,j,:)=\mathcal{X}_{(j)}$	j-th lateral slice of $\mathcal{X}$
$\mathcal{X}(:,:,k)=\mathcal{X}^{(k)}$	k-th frontal slice of $\mathcal{X}$
$\|\mathcal{X}\|_F$	Frobenius norm (F-norm)
$\|\mathcal{X}\|_{FL1}$	sum of F-norms along the lateral slices
$\mathcal{X}_f$	$\mathcal{X}$ in the Fourier space (3-rd direction)
$\mathcal{X}'$	(conjugate) transpose

where $H \in \mathbb{R}^{c\times h}$ comes from data labels, and c is class number; R is the dual variable of the projection matrix $P \in \mathbb{R}^{d\times c}$, and $R'R = I$ removes the orthogonal constraint on P to facilitate optimization; $\|P\|_{2,1}$ is used for the row-sparsity of P.

CDPL uses the data labels to construct a binary graph to exploit the locality of samples as:

$$\begin{aligned}\min_{P,E,Z,T,D}\ & \|E\|_{2,1}+\lambda_1\|Z\|_*+\lambda_2\|Z\|_1\\ &+\lambda_3\|H-TD\|_F^2+\lambda_4 tr(ZLZ')\\ s.t.\ & P'X-P'XZ=E,\ P'P=I,\ Z\geq 0,\ D=[P'X;\mathbf{1}'],\end{aligned} \tag{3}$$

where L is the Laplacian matrix of the binary graph, and $tr(ZLZ')$ measures the consistency between the label graph and the self-representation coefficients; a transformation matrix T is introduced to enable the flexibility of learning more basis vectors than the class number, and thus, the optimal projection P can alleviate the potential side effect of [29], [38] when samples are limited.

B. The Tensor Representation

In this section, we introduce the tensor representation and the t-product-based operations. Please refer to [41], [45], [46], [48] for details. Throughout this paper, the calligraphy letters denote the third-order tensors. It is convenient to split a tensor into different submodules, and to devise an index on each submodule. Specifically, given $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$, $\mathcal{X}(i,j,:) \in \mathbb{R}^{1\times 1\times n_3}$ represents a tube; $\mathcal{X}(i,:,:)$, $\mathcal{X}(:,j,:)$, and $\mathcal{X}(:,:,k)$ denote the horizontal, lateral, and frontal slices respectively; $\mathcal{X}(:,j,:)$ and $\mathcal{X}(:,:,k)$ are interchangeable with $\mathcal{X}_{(j)}$ and $\mathcal{X}^{(k)}$ respectively, for notation simplicity. The Frobenius norm (F-norm) of $\mathcal{X}$ is defined as $\|\mathcal{X}\|_F := (\sum_{i,j,k}|\mathcal{X}(i,j,k)|^2)^{\frac{1}{2}}$, and $\|\mathcal{X}\|_{FL1} := \sum_j(\sum_{i,k}|\mathcal{X}(i,j,k)|^2)^{\frac{1}{2}}$ denotes FL1-norm by summing the F-norms of the lateral slices. $\mathcal{X}_f := fft(\mathcal{X},[\,],3)$ applies fast Fourier transform (FFT) along the third direction of $\mathcal{X}$. The tensor (conjugate) transpose $\mathcal{X}' \in \mathbb{R}^{n_2\times n_1\times n_3}$ is obtained by (conjugate) transposing all frontal slices of $\mathcal{X}$ and then reversing the orders of the (conjugate) transposed slices from 2 to n_3. The notations are summarized in TABLE I.

For $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$, commonly-used tensor manipulations include the block circulant operator

$$bcirc(\mathcal{X}) := \begin{bmatrix}\mathcal{X}^{(1)} & \mathcal{X}^{(n_3)} & \dots & \mathcal{X}^{(2)}\\ \mathcal{X}^{(2)} & \mathcal{X}^{(1)} & \dots & \mathcal{X}^{(3)}\\ \vdots & \ddots & \ddots & \vdots\\ \mathcal{X}^{(n_3)} & \mathcal{X}^{(n_3-1)} & \dots & \mathcal{X}^{(1)}\end{bmatrix}, \tag{4}$$

the block vectorizing operator and its inverse

$$bvec(\mathcal{X}) := \begin{bmatrix}\mathcal{X}^{(1)}\\ \mathcal{X}^{(2)}\\ \vdots\\ \mathcal{X}^{(n_3)}\end{bmatrix}, \quad bvfold(bvec(\mathcal{X})) := \mathcal{X}, \tag{5}$$

and the block diagonalizing operator and its inverse

$$\begin{aligned}bdiag(\mathcal{X}) &:= \begin{bmatrix}\mathcal{X}^{(1)} & & & \\ & \mathcal{X}^{(2)} & & \\ & & \ddots & \\ & & & \mathcal{X}^{(n3)}\end{bmatrix},\\ bdfold(bdiag(\mathcal{X})) &:= \mathcal{X}.\end{aligned} \tag{6}$$

With these manipulations, the t-product is defined as follows.

Definition 1. [41] Let $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$ and $\mathcal{Y} \in \mathbb{R}^{n_2\times n_4\times n_3}$, the tensor-product (**t-product**) $\mathcal{X}*\mathcal{Y}$ is an $n_1\times n_4\times n_3$ tensor:

$$\begin{aligned}\mathcal{X}*\mathcal{Y} &:= bvfold(bcirc(\mathcal{X})\ bvec(\mathcal{Y}))\\ &:= bdfold(bcirc(\mathcal{X})\ bdiag(\mathcal{Y})).\end{aligned} \tag{7}$$

Some t-product-based operations are introduced as follows.

Definition 2 [41]: For $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$, the Singular Value Decomposition (**t-SVD**) of $\mathcal{X}$ is defined as

$$\mathcal{X} := \mathcal{W}*\mathcal{S}*\mathcal{V}',$$

where $\mathcal{W} \in \mathbb{R}^{n_1\times n_1\times n_3}$ and $\mathcal{V} \in \mathbb{R}^{n_2\times n_2\times n_3}$ are orthogonal, and $\mathcal{S} \in \mathbb{R}^{n_1\times n_2\times n_3}$ is f-diagonal (i.e., all frontal slices are diagonal matrices).

Definition 3 [46]: The **multi-rank** of $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$ is a vector in $\mathbb{R}^{n_3\times 1}$ where the k-th entry is the rank of $\mathcal{X}_f^{(k)}$.

Definition 4 ([45], [46]): The t-SVD-based Tensor Nuclear Norm (**t-TNN**) of $\mathcal{X}$ is defined as the sum of the singular values of all frontal slices of $\mathcal{X}_f$ as

$$\|\mathcal{X}\|_{\circledast} := \sum_{k=1}^{n_3}\|\mathcal{X}_f^{(k)}\|_* := \sum_{i=1}^{min\{n_1,n_2\}}\sum_{k=1}^{n_3}|\mathcal{S}_f^{(k)}(i,i)|, \tag{8}$$

where $\mathcal{S}_f^{(k)}$ is obtained from the complex-valued matrix SVD as $\mathcal{X}_f^{(k)} = \mathcal{W}_f^{(k)}\mathcal{S}_f^{(k)}\mathcal{V}_f^{(k)'}$.

The t-TNN is proved to be the tightest convex relaxation to the l_1-norm of the tensor multi-rank, and $\|\mathcal{X}\|_{\circledast}$ is computed from the rank of $bcirc(\mathcal{X})$ [46]. It has been validated that the t-TNN can exploit the structural information of a tensor better than the unfolding-based tensor nuclear norm [46].

The following fact (Fact 1, [41]) will be used to prove the Theorem 1 and to derive some optimization tricks.

Fact 1. Suppose $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$, F_{n_3} is the $n_3\times n_3$ normalized discrete Fourier transform (DFT) matrix (which

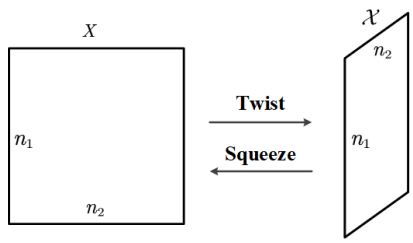

Fig. 1. The *twist* and *squeeze* operators.

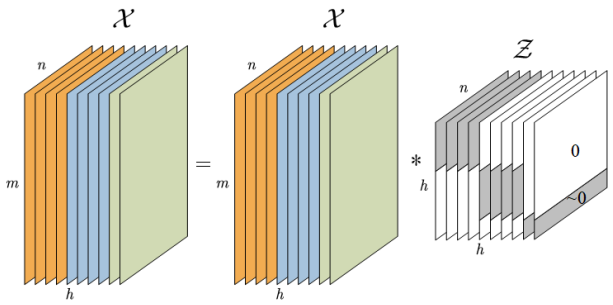

Fig. 2. The self-representation of the tensorial dataset via t-linear combination (samples within the same class are in the same color).

is unitary), F'_{n_3} is the conjugate transpose of F_{n_3}. The block-circulant matrix $bcirc(\mathcal{X})$ can be block-diagonalized as

$$(F_{n_3} \otimes I_{n_1}) \cdot bcirc(\mathcal{X}) \cdot (F'_{n_3} \otimes I_{n_2}) = bdiag(\mathcal{X}_f), \quad (9)$$

where $\otimes$ denotes the Kronecker product.

The *twist* and *squeeze* operators are used to arrange data as third-order tensors [45], [48]. As shown in Fig. 1, given a sample $X \in \mathbb{R}^{n_1 \times n_2}$, the *twist* operator $twist(X) = \mathcal{X} \in \mathbb{R}^{n_1 \times 1 \times n_2}$ transforms a sample into a third-order tensor where the sample is laterally oriented. The *squeeze* operator is the inverse of *twist* as $squeeze(\mathcal{X}) = X$. To obtain the tensorial representation of the dataset, the laterally oriented samples are stacked along the column direction.

C. Self-Representation Tensor

Using the twist operator to construct the data tensor, the self-representation model [43], [45] is expressed as follows.

$$\max_{\mathcal{Z}} \|\mathcal{X} - \mathcal{X} * \mathcal{Z}\|_F^2 + \lambda \mathfrak{R}(\mathcal{Z}), \quad (10)$$

where $\mathcal{X} \in \mathbb{R}^{m \times h \times n}$ is the twisted data tensor, m and n are the row and column numbers of images, and h is the number of samples; $\mathcal{Z} \in \mathbb{R}^{h \times h \times n}$ is the self-representation tensor; $\mathfrak{R}(\cdot)$ is used to regularize $\mathcal{Z}$. As shown in Fig. 2, the j-th laterally oriented sample can be represented by the t-linear combination of other samples as $\mathcal{X}_{(j)} = \mathcal{X} * \mathcal{Z}_{(j)}$, where the i-th tube of $\mathcal{Z}_{(j)}$, i.e., $\mathcal{Z}(i, j, :)$, is the encoding tube that uses the i-th sample to represent the j-th sample. The ideal $\mathcal{Z}$ holds a block diagonal property such that the coefficient tubes in $\mathcal{Z}$ are nonzero only for representing samples from the same class and remain zero otherwise. In practice, different regularizers $\mathfrak{R}(\cdot)$ (e.g., sparse [43], low-rank [45]) can be adopted to approximate the block diagonality of $\mathcal{Z}$.

III. Low-Rank Preserving t-Linear Projection

In this section, after identifying the objectives, we introduce the Low-Rank Preserving t-Linear Projection (LRP-tP) model. An iterative optimization algorithm is devised to solve LRP-tP under the framework of the Alternating Direction Method of Multipliers (ADMM). Then, we analyze the computational complexity of LRP-tP and compare it with existing works to lead a clear understanding.

A. Objectives

With the requirements of real-world image processing, a well-performing projection learning model should take into account the multi-way data structure while being robust to corruptions. In addition, the discriminative ability and the underlying relationship of data should be preserved. More specifically, our main concerns are introduced as follows, formulated using t-product-based operations.

1) Improving Robustness: Let $\{X_j \in \mathbb{R}^{m \times n}\}_{j=1}^h$ be the image set. A third-order data tensor can be constructed by first twisting the samples as shown in Fig. 1 and then stacking the twisted images along the column direction, i.e., $\mathcal{X} \in \mathbb{R}^{m \times h \times n}$. The fundament of LRP-tP is expressed by

$$\begin{aligned} &\min_{\mathcal{P},\mathcal{E},\mathcal{Z}} \ \|\mathcal{E}\|_{FL1} + \lambda_1 \|\mathcal{Z}\|_{\circledast} \\ &s.t. \quad \mathcal{P}' * \mathcal{X} = \mathcal{P}' * \mathcal{X} * \mathcal{Z} + \mathcal{E}, \quad \mathcal{P}' * \mathcal{P} = \mathcal{I}, \end{aligned} \quad (11)$$

where $\mathcal{P} \in \mathbb{R}^{m \times b \times n}$ is the tensorial projection basis (projection tensor) and b is the number of basis slices[2]; the error tensor $\mathcal{E}$ measures the disparity between the projected sample pairs. Usually, only a small fraction of data would be contaminated. As such, we model the sample-specific errors using the tensor FL1-norm $\|\cdot\|_{FL1}$; the low-rank constraint is imposed on $\mathcal{Z}$, and we use t-TNN $\|\cdot\|_{\circledast}$ as a surrogate of the rank function for computational tractability. Note that, 1) using the self-representation data $\mathcal{X} * \mathcal{Z}$, the robustness of the projection tensor can be expected compared to directly learning $\mathcal{P}$ from the raw data $\mathcal{X}$; 2) the physical interpretation of the "t-linear projection" onto $\mathcal{P}$ resembles the linear projection in the vector space by substituting t-product for traditional product operator.

2) Improving Discriminative Ability: While Eq. (11) is able to preserve the intrinsic structures of the image data, it works in the unsupervised setting and thus is inadequate to offer good discrimination across the extracted features. When being applied to subsequent tasks, e.g., classification and clustering, the discrimination of the features is highly preferred. To this end, we propose to minimize the empirical classification error on the training set for enhancing the discrimination of the learned projection tensor. Incorporating the empirical error for projection learning, we design a regression-type module as

$$\min_{\mathcal{P}} \sum_{j=1}^{h} l(\mathcal{H}_{(j)}, \phi(\mathcal{X}_{(j)}, \mathcal{P})), \quad (12)$$

where $\mathcal{X}_{(j)}$ is the j-th laterally oriented sample, $\mathcal{H}_{(j)}$ is constructed from the j-th data label, $\phi(\cdot)$ is the feature extractor, and $l(\cdot)$ is the loss function of the classifier imposed on the

[2]in analogy to basis vectors in linear projection learning.

training set. Specifically, 1) focusing on t-linear projection, the feature extractor is defined as $\phi(\mathcal{X}_{(j)}, \mathcal{P}) = \mathcal{P}' * \mathcal{X}_{(j)}$; 2) let the total number of classes be c and the label of the j-th samples be c_j. The only nonzero tube of $\mathcal{H}_{(j)} \in \mathbb{R}^{c\times 1\times n}$ is set to $\mathcal{H}(c_j, j, :) = [1, \ldots, 1] \in \mathbb{R}^{1\times 1\times n}$; 3) the squared loss $l(\mathcal{H}_{(j)}, \phi(\mathcal{X}_{(j)}, \mathcal{P})) = \|\mathcal{H}_{(j)} - \mathcal{P}' * \mathcal{X}_{(j)}\|_F^2$ is adopted for error measure. Minimizing the empirical classification error across all samples, Eq. (12) equals to $\min_{\mathcal{P}} \|\mathcal{H} - \mathcal{P}' * \mathcal{X}\|_F^2$, which improves the discriminative power of $\mathcal{P}' * \mathcal{X}$ in terms of the classification task.

3) Capturing Flexible Affinity: In real scenarios, the similarity and locality of data are important in identifying the affinity [32], [49]. However, this information is overlooked in Eqs. (11) and (12). To solve this limitation, a straightforward way is to introduce a graph regularizer $\sum_{i,j=1}^{h} S_{i,j}\|\mathcal{X}_{(i)} - \mathcal{X}_{(j)}\|_F^2$, where $S \in \mathbb{R}^{h\times h}$ measures the data similarity and $S_{i,j}$ is the (i, j)-th element of S. However, the raw data $\mathcal{X}$ and the corresponding graph S are susceptible to corruptions. Inspired by [5], [8], we impose the graph regularization on the self-representation tensor as

$$\min_{A} \sum_{i,j=1}^{h} A_{i,j}\|\mathcal{Z}_{(i)} - \mathcal{Z}_{(j)}\|_F^2 = \sum_{k=1}^{n} 2 * tr(\mathcal{Z}^{(k)} L_A \mathcal{Z}^{(k)'})$$
$$s.t.\ A' * \mathbf{1} = \mathbf{1},\ A \geq 0,\ A_{i,j} = 0\ for\ (i, j) \in \Omega, \quad (13)$$

where A indicates the adaptively learned affinity graph, which is more robust and flexible than directly adopting a fixed graph S; L_A is the Laplacian matrix by $L_A = D - A$ and $D = diag(sum(A, 1))$; the nonnegative constraint and the column-wise sum-to-one constraint on A guarantee that the affinity is a probability; Ω is the set of sample pairs (i, j) where sample i and sample j come from different classes, and thus, the constraint $A_{i,j} = 0$ for $(i, j) \in \Omega$ encourages the locality of samples using the prior knowledge from training labels. The motivation of introducing Eq. (13) lies in two folds: 1) the label-oriented locality can effectively purify the feature-oriented similarity learning. As the affinity matrix and the representation tensor are jointly optimized, a clean affinity matrix further promotes learning a precise representation tensor; 2) it has been pointed out that the discriminative power of the self-representation coefficients decreases when the class separability is small [5], [50]. This side effect can be reduced since the similarity learning is concentrated within samples from the same class. To summarize, Eq. (13) preserves the label-oriented locality so as to constraint the learning of the affinity matrix, and this, in return, encourages learning a precise self-representation tensor.

B. Model Formulation

Based on the above concerns, the LRP-tP model is formulated as

$$\min_{\mathcal{P},\mathcal{E},\mathcal{Z},A} \underbrace{\|\mathcal{E}\|_{FL1} + \lambda_1\|\mathcal{Z}\|_{\circledast}}_{\textbf{Self-Representation}} + \underbrace{\lambda_2\|\mathcal{H} - \mathcal{P}' * \mathcal{X}\|_F^2}_{\textbf{Classification Error}}$$
$$\underbrace{+ \lambda_3 \sum_{k=1}^{n} tr(\mathcal{Z}^{(k)} L_A \mathcal{Z}^{(k)'}) + \eta\|A\|_F^2}_{\textbf{Adaptive Graph}}$$
$$s.t.\quad \mathcal{P}' * \mathcal{X} = \mathcal{P}' * \mathcal{X} * \mathcal{Z} + \mathcal{E},\quad \mathcal{P}' * \mathcal{P} = \mathcal{I},$$
$$A' * \mathbf{1} = \mathbf{1},\ A \geq 0,\ A_{i,j} = 0\ for\ (i, j) \in \Omega, \quad (14)$$

where λ_1, λ_2 and λ_3 are the tradeoff parameters; $\|A\|_F^2$ is used to prevent the trivial solution of A. As LRP-tP works in a supervised manner, the parameter η can be determined according to the underlying number of neighbors.

The proposed model naturally overlays the strengths from self-representation, supervised projection and adaptive graph learning for robust feature extraction. More specifically,

- The t-linear projection is learned by considering the self-representation data to alleviate data corruptions; the learning procedure works in a supervised manner by minimizing the empirical classification error with a regression-type module; an adaptive graph learning scheme is incorporated to take advantage of the feature-oriented similarity and label-oriented locality. The three modules in Eq. (14) benefit from each other to achieve the overall optimum.
- Eq. (14) is formulated in the third-order tensor space. It learns the t-linear projection directly from the tensorial dataset. Thus, the two-way structure within images and the across-sample structure are simultaneously exploited. Eq. (14) therefore enables the flexibility of modeling the multi-way data.
- Introducing the regression-based classification term, the dimensions of the projected samples are fixed to the class number c. The optimal projection $\mathcal{P}$ can be considered as a feature extractor that maps a new sample $\mathcal{Y} \in \mathbb{R}^{m\times 1\times n}$ to $\mathcal{P}' * \mathcal{Y} \in \mathbb{R}^{c\times 1\times n}$ as the feature.

C. Optimization

In this section, we devise an iterative optimization algorithm to solve the proposed LRP-tP model under the framework of ADMM [51]. An auxiliary variable $\mathcal{U}$ with a constraint $\mathcal{Z} = \mathcal{U}$ is introduced to make Eq. (14) separable:

$$\min_{\mathcal{P},\mathcal{E},\mathcal{U},\mathcal{Z},A} \|\mathcal{E}\|_{FL1} + \lambda_1\|\mathcal{U}\|_{\circledast} + \lambda_2\|\mathcal{H} - \mathcal{P}' * \mathcal{X}\|_F^2$$
$$+ \lambda_3 \sum_{k=1}^{n} tr(\mathcal{Z}^{(k)} L_A \mathcal{Z}^{(k)'}) + \eta\|A\|_F^2$$
$$s.t.\quad \mathcal{P}' * \mathcal{X} = \mathcal{P}' * \mathcal{X} * \mathcal{Z} + \mathcal{E},\quad \mathcal{P}' * \mathcal{P} = \mathcal{I},$$
$$A' * \mathbf{1} = \mathbf{1},\ A \geq 0,\ A_{i,j} = 0\ for\ (i, j) \in \Omega,$$
$$\mathcal{Z} - \mathcal{U} = 0. \quad (15)$$

Accordingly, the augmented Lagrange function of Eq. (15) is formulated as

$$\mathcal{L}(\mathcal{P}, \mathcal{E}, \mathcal{U}, \mathcal{Z}, A) = \|\mathcal{E}\|_{FL1} + \lambda_1\|\mathcal{U}\|_{\circledast} + \lambda_2\|\mathcal{H} - \mathcal{P}' * \mathcal{X}\|_F^2$$
$$+ \lambda_3 \sum_{k=1}^{n} tr(\mathcal{Z}^{(k)} L_A \mathcal{Z}^{(k)'}) + \eta\|A\|_F^2 + \frac{\rho}{2}(\|\mathcal{Z} - \mathcal{U} + \frac{\mathcal{C}_1}{\rho}\|_F^2$$
$$+ \|\mathcal{P}' * \mathcal{X} - \mathcal{P}' * \mathcal{X} * \mathcal{Z} - \mathcal{E} + \frac{\mathcal{C}_2}{\rho}\|_F^2)$$
$$s.t.\quad \mathcal{P}' * \mathcal{P} = \mathcal{I}, \quad (16)$$

where $\mathcal{C}_1$, $\mathcal{C}_2$ are the Lagrange multipliers and $\rho > 0$ is the penalty parameter. Eq. (16) can be alternatively optimized with respect to five subproblems as follows.

1) $\mathcal{P}$-Subproblem: Fixing other variables except $\mathcal{P}$, the optimization of Eq. (16) reduces to

$$\min_{\mathcal{P}} \ \lambda_2\|\mathcal{H}-\mathcal{P}'*\mathcal{X}\|_F^2+\frac{\rho}{2}\|\mathcal{P}'*\mathcal{X}-\mathcal{P}'*\mathcal{X}*\mathcal{Z}-\mathcal{E}+\frac{\mathcal{C}_2}{\rho}\|_F^2 \quad s.t. \quad \mathcal{P}'*\mathcal{P}=\mathcal{I}. \tag{17}$$

We introduce the following Theorem 1 to solve Eq. (17), and the proof of Theorem 1 is given in APPENDIX.

Theorem 1: Given third-order tensors $\mathcal{N}_1$ and $\mathcal{N}_2$ with matched dimensions, the solution of

$$\min_{\mathcal{R}} \ \|\mathcal{R}'*\mathcal{N}_1-\mathcal{N}_2\|_F^2 \quad s.t. \quad \mathcal{R}'*\mathcal{R}=\mathcal{I} \tag{18}$$

is obtained as follows. Let the economy-size t-SVD [45] of $\mathcal{N}_1*\mathcal{N}_2'$ be $\mathcal{W}*\mathcal{S}*\mathcal{V}'$. Then, $\mathcal{R}^*=\mathcal{W}*\mathcal{V}'$.

With Theorem 1, let $\mathcal{M}_1=\lambda_2\mathcal{X}*\mathcal{H}'+\frac{\rho}{2}(\mathcal{X}-\mathcal{X}*\mathcal{Z})*(\mathcal{E}-\frac{\mathcal{C}_2}{\rho})'$, we can obtain the optimal $\mathcal{P}$ from the t-SVD of $\mathcal{M}_1$.

2) $\mathcal{E}$-Subproblem: The optimization of Eq. (16) with respect to $\mathcal{E}$ becomes

$$\min_{\mathcal{E}} \|\mathcal{E}\|_{FL1}+\frac{\rho}{2}\|\mathcal{P}'*\mathcal{X}-\mathcal{P}'*\mathcal{X}*\mathcal{Z}-\mathcal{E}+\frac{\mathcal{C}_2}{\rho}\|_F^2, \tag{19}$$

which can be solved via the soft-thresholding operator [44] on the tensor FL1-norm. Let $\mathcal{M}_2=\mathcal{P}'*\mathcal{X}-\mathcal{P}'*\mathcal{X}*\mathcal{Z}+\frac{\mathcal{C}_2}{\rho}$, the solution to Eq. (19) is obtained at

$$\mathcal{E}^*(:,j,:)=\begin{cases}\dfrac{\|\mathcal{M}_2(:,j,:)\|_F-1/\rho}{\|\mathcal{M}_2(:,j,:)\|_F}\mathcal{M}_2(:,j,:), & \text{if } \|\mathcal{M}_2(:,j,:)\|_F>\dfrac{1}{\rho}\\ 0, & \text{otherwise.}\end{cases} \tag{20}$$

3) $\mathcal{U}$-Subproblem: Fixing $\mathcal{P}$, $\mathcal{E}$, $\mathcal{Z}$, and A, the optimal $\mathcal{U}$ is computed from

$$\min_{\mathcal{U}} \lambda_1\|\mathcal{U}\|_{\circledast}+\frac{\rho}{2}\|\mathcal{Z}-\mathcal{U}+\frac{\mathcal{C}_1}{\rho}\|_F^2, \tag{21}$$

which is a t-TNN minimization problem. Let $\mathcal{M}_3=\mathcal{Z}+\frac{\mathcal{C}_1}{\rho}$, Eq. (21) can be solved by applying the tensor tubal-shrinkage operator on $\mathcal{M}_3$, according to [46] (Theorem 2).

4) $\mathcal{Z}$-Subproblem: The subproblem associated with $\mathcal{Z}$ is formulated as

$$\min_{\mathcal{Z}} \lambda_3\sum_{k=1}^{n} tr(\mathcal{Z}^{(k)}L_A\mathcal{Z}^{(k)'})+\frac{\rho}{2}(\|\mathcal{Z}-\mathcal{U}+\frac{\mathcal{C}_1}{\rho}\|_F^2 + \|\mathcal{P}'*\mathcal{X}-\mathcal{P}'*\mathcal{X}*\mathcal{Z}-\mathcal{E}+\frac{\mathcal{C}_2}{\rho}\|_F^2). \tag{22}$$

Recall that $\sum_{k=1}^{n3} tr(\mathcal{Z}^{(k)}L_A\mathcal{Z}^{(k)'})=\frac{1}{2}\sum_{i,j=1}^{h}A_{i,j}\|\mathcal{Z}_{(i)}-\mathcal{Z}_{(j)}\|_F^2$. Based on Fact 1, Eq. (22) can be transformed into the Fourier domain as independent problems with respect to the frontal slices:

$$\min_{\mathcal{Z}_f^{(k)}} \lambda_3 tr(\mathcal{Z}_f^{(k)}L_A\mathcal{Z}_f^{(k)'})+\frac{\rho}{2}(\|\mathcal{Z}_f^{(k)}-\mathcal{U}_f^{(k)}+\frac{\mathcal{C}_{1f}^{(k)}}{\rho}\|_F^2 + \|\mathcal{P}_f^{(k)'}\mathcal{X}_f^{(k)}-\mathcal{P}_f^{(k)'}\mathcal{X}_f^{(k)}\mathcal{Z}_f^{(k)}-\mathcal{E}_f^{(k)}+\frac{\mathcal{C}_{2f}^{(k)}}{\rho}\|_F^2), \tag{23}$$

where $\mathcal{Z}_f$, $\mathcal{U}_f$, $\mathcal{P}_f$, $\mathcal{X}_f$, $\mathcal{E}_f$, $\mathcal{C}_{1f}$ and $\mathcal{C}_{2f}$ are in the Fourier domain, and $k=1,\ldots,n$. Setting the derivation of Eq. (23) to zero,[3] the optimal k-th frontal slice of $\mathcal{Z}_f$ is obtained at

$$\mathcal{Z}_f^{(k)*}=((2\lambda_3+\rho)I+\rho M_5'M_5)^{-1}(\rho M_4+\rho M_5'M_6)(L_A+2I)^{-1}, \tag{24}$$

where $M_4=\mathcal{U}_f^{(k)}-\frac{\mathcal{C}_{1f}^{(k)}}{\rho}$, $M_5=\mathcal{P}_f^{(k)'}\mathcal{X}_f^{(k)}$, $M_6=\mathcal{P}_f^{(k)'}\mathcal{X}_f^{(k)}-\mathcal{E}_f^{(k)}+\frac{\mathcal{C}_{2f}^{(k)}}{\rho}$. Then, we can recover $\mathcal{Z}$ via the inverse FFT as $\mathcal{Z}=ifft(\mathcal{Z}_f,[\,],3)$.

5) A-Subproblem: To find the optimal A, we solve the following constrained problem

$$\min_{A} \frac{1}{2}\sum_{i,j=1}^{h}A_{i,j}\|\mathcal{Z}_{(i)}-\mathcal{Z}_{(j)}\|_F^2+\eta\|A\|_F^2 \quad s.t. \quad A'*\mathbf{1}=\mathbf{1},\ A\geq 0,\ A_{i,j}=0\ for\ (i,j)\in\Omega, \tag{25}$$

where $\eta\|A\|_F^2$ is used to prevent the trivial solution. When η is set to zero, only the affinity of the nearest neighbor is preserved on A, whereas all training samples are equally treated as neighbors when η is set to infinite. In this view, by tuning η, we can adaptively preserve the neighbors for each sample. However, η is difficult to tune since its value could be continuous from zero to infinite. To relieve the computation burden of parameter tuning, the work in [8] proposed to adjust η by tuning the number of sample neighbors since the latter is an integer and has explicit meaning.

As LRP-tP works in a supervised manner, the underlying number of neighbors is known in advance. Thus, we do not need to specify or tune the value of η in practical implementation. Instead, it is always feasible to know the number of sample neighbors using the prior knowledge from training labels, and accordingly, the number of neighbors is set to the number of samples in current class minus one.

Note that without the constraint $A_{i,j}=0$ for $(i,j)\in\Omega$, Eq. (25) can be solved column-by-column using the off-the-shelf quadratic programming method [8]. In our setting, we apply the optimization scheme on the nonzero subset of each column of A such that only samples with the same class labels are considered. Specifically, denote $F\in\mathbb{R}^{h\times h}$ where $f_{i,j}=\|\mathcal{Z}_{(i)}-\mathcal{Z}_{(j)}\|_F^2$ is the (i,j)-th entry; $\tilde{a}_j=A_{pind,j}$ where $pind$ collects a subset from $\{1,\ldots,h\}$ for sample pairs $i\in pind$ and j belonging to the same class. The optimization of the j-th column of A can be solved by optimizing $\tilde{a}_j$ using

$$\min_{\tilde{a}_j} \|\tilde{a}_j+\frac{1}{4\eta}\tilde{f}_j\|_F^2 \quad s.t. \quad \tilde{a}_j'*\mathbf{1}=1,\ \tilde{a}_j\geq 0, \tag{26}$$

where $\tilde{f}_j=F_{pind,j}$. Eq. (26) can be solved using the off-the-shelf solver. Once $\tilde{a}_j$ is obtained, $A_{pind,j}$ can be recovered, and other entries $A_{i,j}$ for $i\notin pind$ remain zero.

[3]Eq. (23) is optimized in the complex domain. Please refer to [52] for calculating the complex derivations.

Algorithm 1: Solving the LRP-tP Model Eq. 14

Input : Training data tensor $\mathcal{X}$, label tensor $\mathcal{H}$ and parameters λ_1, λ_2, λ_3.
Output: Projection tensor $\mathcal{P}$.
Initialize $\mathcal{P}$, $\mathcal{E}$, $\mathcal{U}$, $\mathcal{Z}$, $\mathcal{C}_1$, $\mathcal{C}_2$ to zero; initialize A using data labels; initialize ρ to 0.01; $\rho^{max} = 10^5$, $\mu = 1.9$, $\epsilon = 10^{-5}$.
repeat
  Update $\mathcal{P}$ according to Theorem 1.
  Update $\mathcal{E}$ by Eq. (20).
  Update $\mathcal{U}$ using the tensor tubal-shrinkage operator.
  Update $\mathcal{Z}$ by Eq. (24) and inverse FFT.
  Update A by Eq. (26).
  Update multipliers and penalty parameter by Eq. (27).
until *Residuals* $r_1 < \epsilon$ *and* $r_2 < \epsilon$;
Optimal projection tensor $\mathcal{P}$.

6) Multipliers and penalty parameter: The multipliers and penalty parameter are adjusted according to

$$\begin{aligned} \mathcal{C}_1^* &= \mathcal{C}_1 + \rho(\mathcal{Z} - \mathcal{U}), \\ \mathcal{C}_2^* &= \mathcal{C}_2 + \rho(\mathcal{P}' * \mathcal{X} - \mathcal{P}' * \mathcal{X} * \mathcal{Z} - \mathcal{E}), \\ \rho^* &= \min\{\mu * \rho,\ \rho_{max}\}, \end{aligned} \tag{27}$$

where the penalty parameter ρ is iteratively updated to improve the convergence behavior and to make the performance less dependent on the initial value of ρ [51]; the constant μ is empirically set to 1.9 throughout this paper.

The stopping criterion of the iterative algorithm is met when the residuals defined below are small enough.

$$\begin{aligned} r_1 &= \|\mathcal{Z} - \mathcal{U}\|_F / \|\mathcal{X}\|_F, \\ r_2 &= \|\mathcal{P}' * \mathcal{X} - \mathcal{P}' * \mathcal{X} * \mathcal{Z} - \mathcal{E}\|_F / \|\mathcal{X}\|_F. \end{aligned} \tag{28}$$

The optimization procedure of LRP-tP is summarized in Algorithm 1. Afterward, the twisted image samples can be projected onto the optimized projection tensor for feature extraction.

D. Discussion

1) Complexity Analysis: The complexities of basic operations are computed as follows. The (inverse) FFT of $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_4}$ along the third direction takes the order of $\mathcal{O}(n_1 n_2 n_4 \log(n_4))$; the t-product of $\mathcal{X}$ and $\mathcal{Y} \in \mathbb{R}^{n_2 \times n_3 \times n_4}$ takes the order of $\mathcal{O}(\max(n_1, n_3) n_2 n_4 \log(n_4) + n_1 n_2 n_3 n_4)$; the economic-size t-SVD of $\mathcal{X}$ requires $\mathcal{O}(n_1 n_2 n_4 \log(n_4) + \min(n_1, n_2) n_1 n_2 n_4)$ operations.

Assume that $h > m$, $c > m$, m and n are comparable. As the optimization of LRP-tP involves an iterative scheme, we first investigate the computational cost of a single iteration: 1) the main cost of the $\mathcal{P}$-subproblem comes from t-product and t-SVD that require a complexity of order $\mathcal{O}(h^2 n \log(n) + m h^2 n)$; 2) for the $\mathcal{E}$-subproblem, the computation of the temporary tensor $\mathcal{M}_2$ takes the complexity of order $\mathcal{O}(h^2 n \log(n) + c h^2 n)$, and the slice-by-slice thresholding is negligible; 3) when solving the $\mathcal{U}$-subproblem, the t-TNN minimization consumes $\mathcal{O}(h^2 n \log(n) + h^3 n)$ operations; 4) the $\mathcal{Z}$-subproblem consists FFT, matrix product, and matrix inverse, costing $\mathcal{O}(h^2 n \log(n) + h^3 n)$ operations; 5) the optimization of the A-subproblem applies quadratic programming on subsets of the columns of A, and its cost is negligible compared with that of other subproblems; 6) the cost of updating multipliers and penalty parameter is negligible. Since $h > c$ and we can always assume that $h > \log(n)$, one iteration of the iterative scheme costs $\mathcal{O}(h^3 n)$ operations. Let the number of iterations be T, the total computation complexity of LRP-tP will be $\mathcal{O}(T h^3 n)$.

2) Comparison With Related Methods: Two state-of-the-art models, i.e., LLRSE [38] and CDPL [39], provide insights into our LRP-tP model on that they all conduct the representation-based projection learning in a supervised manner. However, the most fundamental difference lies in the fact that LLRSE and CDPL use vectors to represent samples, while LRP-tP does not resort to the vectorization operation such that enhanced performance can be expected when dealing with data that naturally have multi-way structures. Stacking the vectorized samples into a matrix, LLRSE and CDPL cope with the dataset using matrix manipulations, while LRP-tP employs the t-product-based operations. That is, these models are formulated in different spaces and their optimization procedures are totally different accordingly.

In addition, LLRSE does not take the data graph into consideration, and this may lead to performance degradation in correctly identifying data affinity; CDPL solves this limitation by using the labels to construct a binary graph. Although the locality information of data is preserved, the pre-defined binary graph is fixed and overlooks the similarity of data. Besides, the computational complexities of LLRSE and CDPL are $\mathcal{O}(T h^3)$, where T and h are the number of iterations and that of samples. In contrast, the optimization of LRP-tP takes the complexity of order $\mathcal{O}(T h^3 n)$, where n is the column number of image matrices. This is because, LRP-tP employs the tensor-based operators to capture the multi-way data structure such that the complexity is affected by the intrinsic data dimensions.

IV. Experiments

In this section, we evaluate the performance of LRP-tP in extracting features from images via the classification task. After introducing the experimental configurations, LRP-tP is thoroughly compared with the state-of-the-arts. Then, we conduct the ablation study, parameter sensitive analysis, and convergence analysis to promote the understanding of LRP-tP.

A. Experimental Settings

1) Datasets: Five commonly-used image databases are chosen for model evaluation. *AR database*[4] provides a popular cropped version of the raw AR face database. The images are captured in two sessions under varying expressions, illumination, and occlusions from scarfs and sunglasses; *Extended YaleB database*[5] contains frontal face images with expression, pose, and illumination changes; *FERET database*[6] consists a cropped and scaled version of the raw FERET face dataset with

[4]http://www2.ece.ohio-state.edu/ aleix/ARdatabase.html/
[5]http://vision.ucsd.edu/ iskwak/ExtYaleDatabase/ExtYaleB.html/
[6]https://www.nist.gov/itl/products-and-services/color-feret-database/

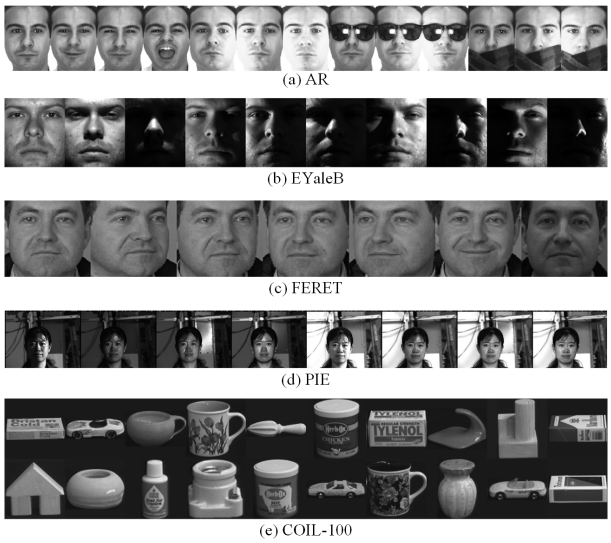

Fig. 3. Samples from different databases.

TABLE II
STATISTICS OF THE DATABASES

Data	Content	Class	Samples
AR	face	100	2600
EYaleB	face	38	2414
FERET	face	200	1400
PIE	face	68	4284
COIL-100	object	100	7200

variations in expression, pose, and illumination; *CMU PIE database*[7] is composed of face images with different poses, illumination, expressions, and talking sequences. We collect a subset of CMU PIE with illumination changes, resulting in 43 images per subject; *COIL-100 database*[8] contains objects with a wide variety of complex geometric and reflectance characteristics (toys, cups, etc). Images of each objects are taken at pose intervals of five degrees, corresponding to 72 images per class. The statistics of the databases are summarized in TABLE II, and typical examples of samples are shown in Fig. 3. We downsample all images to 32*32 pixels for efficiency.

2) Competitors: To examine the performance of LRP-tP, we compare it with nine projection learning models. Among them, PCA [3] and 2DPCA [11] are baselines; Multilinear Principal Component Analysis (MPCA) [17] and Multilinear Discriminant Analysis (MDA) [18] use the unfoldings of tensors, in unsupervised and supervised manners respectively; five state-of-the-art representation-based projection learning approaches are compared, in which LRE [29] and Low-Rank Sparse preserving Projections (LRSP) [30] are unsupervised models, and Extended Approximate Low-rank Projection Learning (EALPL) [37], LLRSE [38], CDPL [39] are supervised ones.

3) Parameters: LRP-tP has three tradeoff parameters λ_1, λ_2, and λ_3. In our implementation, they are tuned from $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. For the competitors, the model parameters are tuned according to the recommendations.

4) Classification Configurations: For quantitative comparison, we adopt the classification task to evaluate different models. Specifically, the database is divided into the disjoint training and testing sets. The projection bases are learned from the training set by solving different models. Afterward, images in both sets are projected onto the learned projection bases. The accuracies of correctly classifying the projected testing samples into the corresponding projected training samples are recorded. In all experiments, the nearest neighbor classifier is adopted for classification. All tests are repeated ten times for statistical stability.

In addition to the model parameters, the classification accuracies also relate to the number of projection vectors/slices b. For PCA, LRE, and LRSP, b is chosen from five to $\min(h, 300)$ with a step of five; considering 2DPCA, MPCA, and MDA, b is chosen from $\{2, 4, \ldots, 32\}$; the number of projection vectors/slices of EALPL, LLRSE and the proposed LRP-tP is fixed to the class number c; CDPL can extract more projection vectors than the class number, and b is therefore selected from $\{5, 10, \ldots, \min(4c, h)\}$. Afterward, the best-performing classification results over all candidates of b are used for comparison.

B. Results and Analysis

In this section, we provide comprehensive experimental comparisons between LRP-tP and its competitors. Generally, LRP-tP shows superiority and good generalization ability in dealing with clean images, commonly-observed data corruptions (i.e., Gaussian noise, impulse noise, block occlusions), and real-world occlusions. The detailed experiments and related analysis are presented as follows.

1) Image Classification Accuracy: We use different algorithms to learn the projection bases from the training set, and then project the testing samples to extract features for classification. All five databases are adopted for comparison. The experimental results are reported in TABLE III.

- For AR, the training and testing sets are constructed by randomly splitting the clean images in half; for EYaleB and PIE, we randomly select ten images per subject for training and the remainders for testing; similarly, FERET is separated into disjoint training and testing sets randomly, with five samples per subject for training and the left two for testing. We find that, on the face databases, LRP-tP and LLRSE consistently obtain the best and second-best performance with the margins of around 0.5%-6.6%, validating the effectiveness of the representation-based projection learning with supervised information. It is interesting to notice that, while closely-related, LLRSE achieves relatively better performance than CDPL. This may come from the function of feature selection in LLRSE by encouraging the row-sparsity of projection basis. This observation also provides a potential research direction to improve the performance of LRP-tP with constraints on the projection slices.
- For COIL-100, the training set consists ten over 72 images per class by random selection, and the

[7]http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html/
[8]https://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php/

TABLE III

PERFORMANCE COMPARISON (I.E., CLASSIFICATION ACCURACIES, STANDARD DEVIATIONS, AND THE NUMBERS OF BASIS VECTORS/SLICES) OF COMPETING ALGORITHMS ON DIFFERENT DATABASES

		PCA	2DPCA	MPCA	MDA	LRE	LRSP	EALPL	LLRSE	CDPL	LRP-tP
AR	*rate*	66.03±2.12	69.33±2.14	85.05±1.51	94.88±0.48	90.57±1.25	90.76±1.28	95.31±0.49	***96.44***±0.49	96.18±0.55	**97.69**±0.55
	num	160	18	24+16	18+18	200	190	100	100	260	100
EYaleB	*rate*	40.59±0.93	44.66±1.47	70.89±2.79	78.18±0.85	83.50±4.04	84.15±1.64	84.69±1.12	***88.64***±0.86	85.45±0.85	**93.64**±0.74
	num	200	14	20+16	20+12	200	200	38	38	140	38
FERET	*rate*	49.30±1.73	68.25±2.26	68.25±2.37	65.77±1.66	74.97±1.79	78.55±1.98	79.88±1.23	***82.35***±1.63	77.15±2.23	**87.77**±1.55
	num	180	24	22+12	14+16	180	185	200	200	220	200
PIE	*rate*	64.93±0.93	69.96±0.92	87.32±0.94	91.52±0.90	85.78±1.47	88.33±1.88	92.56±0.98	***94.84***±0.43	93.71±0.45	**95.24**±0.42
	num	350	20	18+16	22+10	300	200	68	68	210	68
COIL	*rate*	80.64±0.58	81.37±0.47	85.18±0.55	***86.44***±0.60	82.10±1.21	81.39±1.75	75.55±0.92	76.98±0.45	75.78±0.98	**87.72**±0.54
	num	50	20	20+12	14+18	140	165	100	100	120	100

Bold numbers represent the best performance, and bold-italic numbers denote the second best-performing reuslts.

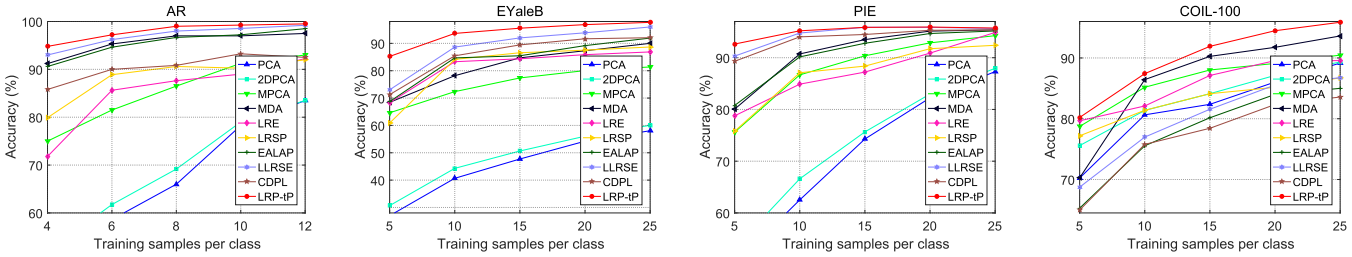

Fig. 4. Robustness of competing algorithms to varying training sizes.

remainders are used for testing. LRP-tP obtains the highest classification accuracy, followed by MDA. Since the COIL-100 database consists objects with complex geometric structures, this observation validates the benefits from the multi-way data structure when dealing with the data containing rich structures.

2) Robustness to Training Size: Many competing projection learning models denote samples as vectors, except for 2DPCA, MPCA, MDA. The vectorization operation not only sacrifices the spatial structure of images, but also suffers problems when the number of training images is small. To show the influence of the training size, we compare the classification results of different algorithms over varying numbers of training samples per class. For the clean AR dataset, the number of training samples per class is selected from {4, 6, 8, 10, 12}; as to EYaleB, PIE, and COIL-100, we compare different models with {5, 10, 15, 20, 25} samples per class in the training sets. According to the results in Fig. 4, we have the following observations:

- Compared to LRE, LRSP, EALPL, LLRSE and CDPL, the performance of LRP-tP is less affected when the numbers of training samples are extremely small. Similarly, the classification accuracies of MPCA are more stable than PCA and 2DPCA. This shows that the multi-way data structure can bring significant gains when samples are limited.
- LRP-tP obtains satisfactory results with five training samples per class on AR, EYaleB, and PIE. For COIL-100, the performance of LRP-tP is greatly improved by expanding the training size from five to ten per class. This is because images in COIL-100 have relatively large variations such that more samples are needed to correctly model the structures of the subspaces.
- For face databases, LRP-tP consistently obtains the best performance. LLRSE and CDPL are comparable to LRP-tP on AR and PIE. The accuracies of LLRSE and CDPL decrease on EYaleB, but they are much better than other competitors.
- On COIL-100, LRP-tP ranks in the first place and MDA obtains good performance in most cases. The performance of LLRSE and CDPL witnesses the steep descent. This is because there are large variations between different subjects in the COIL-100 database, and thus, the multi-way structure of images will play an important role in projection learning.

3) Robustness to Synthetic Corruptions: As images are likely to be contaminated, the robustness of models is important. We impose different kinds of data errors on the EYaleB database to simulate commonly observed data errors, i.e., Gaussian noise, impulse noise, and block occlusions, where only partial of the training sets are contaminated. In each trail, we randomly select ten images per class for training, in which half images are imposed with varying strengths of errors. The variance of the Gaussian noise varies from 0.01 to 0.15, and the density of the impulse noise varies from 0.05 to 0.4. The size of block occlusions varies

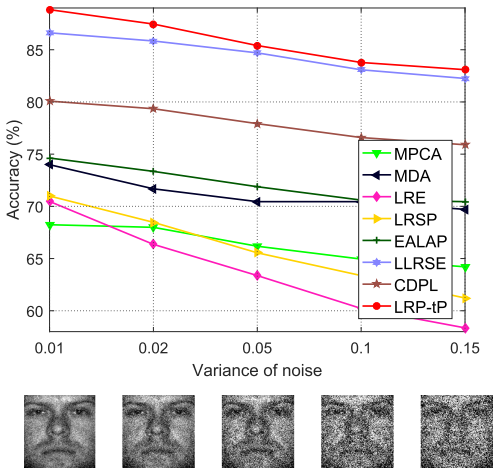

Fig. 5. Robustness to varying variances of Gaussian noise.

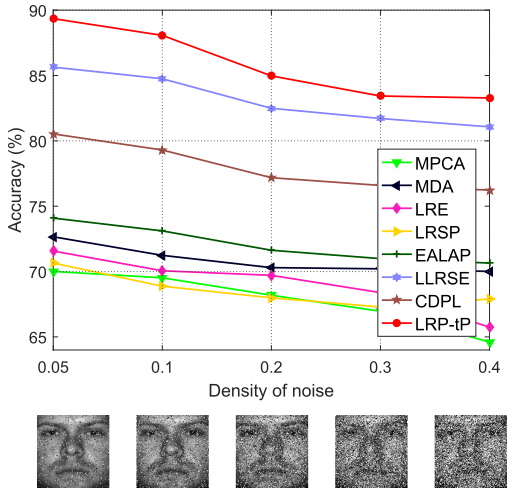

Fig. 6. Robustness to varying densities of impulse noise.

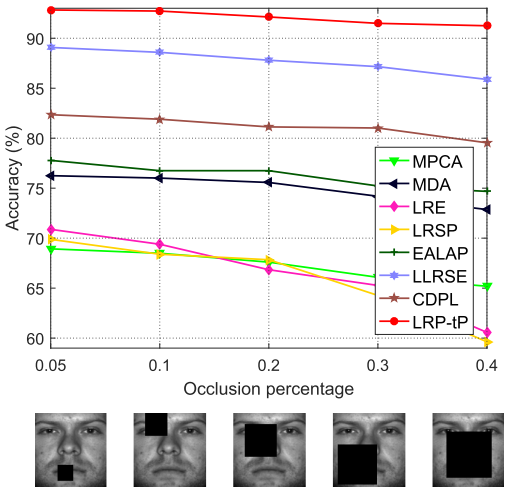

Fig. 7. Robustness to varying percentages of block occlusion.

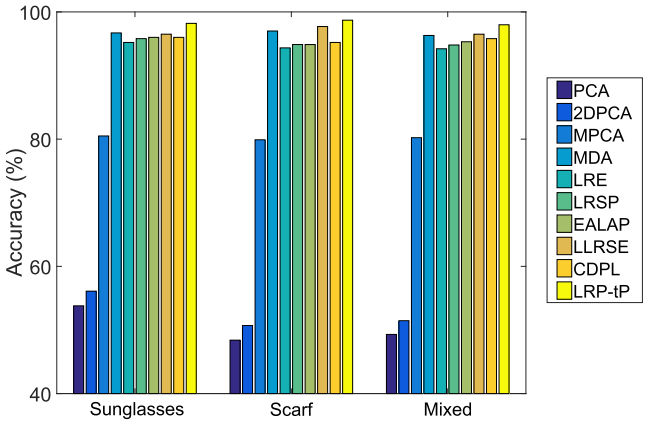

Fig. 8. Robustness to different natural occlusions.

from $\{0.05, 0.1, 0.2, 0.3, 0.4\}$ of the original images, and the occlusions are imposed at random positions. Examples of the contaminated images and the classification results are shown in Figs. 5-7, where the results of PCA and 2DPCA are ignored for clear visualization.

We notice that: 1) LRP-tP shows advanced robustness to different kinds of data corruptions; 2) although LLRSE and CDPL have relatively good performance compared with other models, their performance is not adequate when compared to LRP-tP. This validates the strength of LRP-tP by using t-product-based operations in multi-way data modeling; 3) the performance of MDA decreases rapidly with data corruptions since MDA lacks of the robust measure.

4) Robustness to Natural Occlusions: To compare the performance of different algorithms thoroughly, we use the naturally-occluded AR dataset to investigate the robustness over real-world occlusions. Since the AR database contains images occluded by scarfs and sunglasses, we construct three testing cases by including one kind of occlusion in the first two trials respectively and using the full database in the third trial. The corresponding results are plotted in Fig. 8. We find that, LRP-tP consistently achieves the best performance, followed by LLRSE and MDA. In addition to the supervised information, LLRSE takes advantage of self-representation learning and sparse feature selection. Meanwhile, MDA makes use of the multi-linear structure of images as well as the between-class separability for improving the discriminative ability. Generally, the representation-based projection learning models achieve satisfactory results, validating their robustness in dealing with natural occlusions.

C. Model Analysis

In this section, we experimentally analysis the LRP-tP model for enhanced understanding.

1) Ablation Study: The ablation study is conducted to examine the effectiveness of each module of LRP-tP. Specifically, the $\mathcal{U}$ subproblem uses t-TNN to encourage the block-diagonality of the self-representation tensor; the $\mathcal{H}$ subproblem introduces a regression-type module to improve the discriminative ability; the similarity and locality of samples are exploited via the A subproblem. We conduct experiments by randomly selecting ten samples per subject for both databases, and the results are reported in TABLE IV. This test shows that the three modules of LRP-tP work collaboratively to obtain

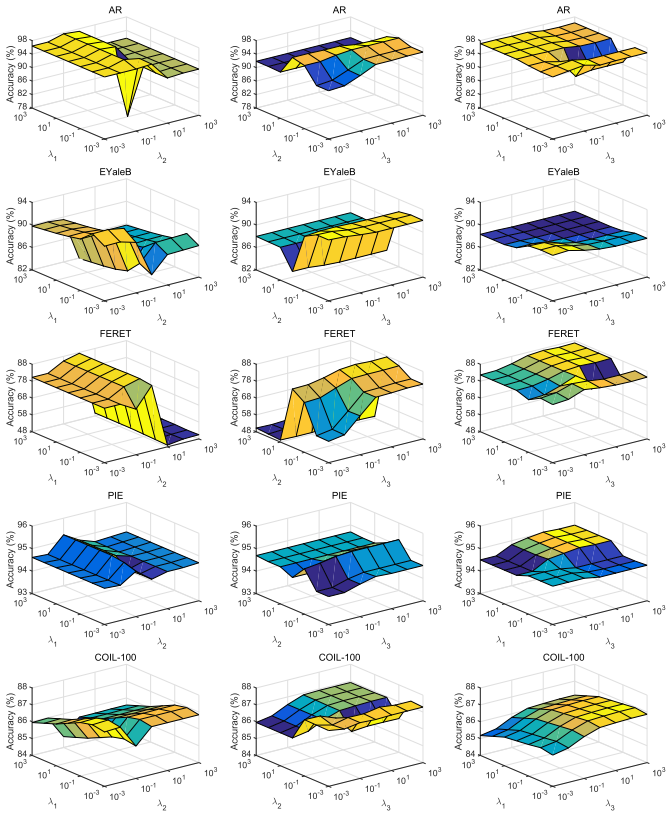

Fig. 9. Performance of LRP-tP over different settings of parameters.

the overall optimality, and the classification term receives high importance compared to the other two modules.

2) Parameter Sensitivity: LRP-tP has three tradeoff parameters λ_1, λ_2, and λ_3. To show the performance of LRP-tP over different parameters, we fix one parameter to one and examine the effects of the other two parameters. We notice that:

- As plotted in the first and second columns of Fig. 9, the performance of LRP-tP relies primarily on the choice of λ_2. This is because, incorporating the empirical classification error into projection learning, the regression-type module is beneficial for the classification task. Compared to AR, EYaleB, and FERET, the performance of LRP-tP is less affected by the values of λ_2 on PIE and COIL-100. This indicates that the discriminative power obtained from the regression-type module is closely related to the intrinsic properties of the datasets, and thus, when being applied to new databases, λ_2 should be tuned carefully.
- Comparing the last column of Fig. 9 to the first two columns, we find that the performance margins are relatively small by adjusting λ_1 and λ_3. In particular, LRP-tP obtains stable performance with different settings of λ_1 in most cases, showing the consistent reliability of the low-rank tensor based self-representation module. When λ_2 is fixed, the performance of LRP-tP is generally stable over varying values of λ_3. This is because, both modules explore the supervision from the training labels, and thus, the effects are related to some extent.

TABLE IV
ABLATION STUDY ON DIFFERENT MODULES OF LRP-TP

	LRP-tP–~~$\mathcal{U}$~~	LRP-tP–~~$\mathcal{H}$~~	LRP-tP–~~$\mathcal{A}$~~	LRP-tP
EYaleB	91.77	88.20	91.87	93.64
COIL-100	84.47	80.37	82.65	87.72

The ablated modules are denoted by dashed lines.

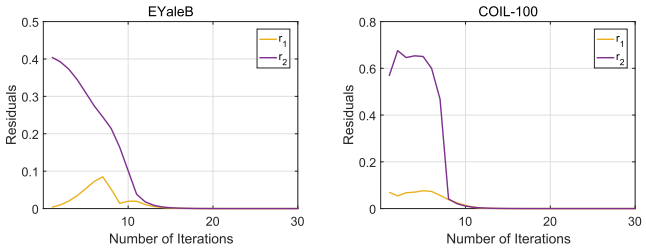

Fig. 10. Empirical convergence of LRP-tP.

3) Convergence Analysis: Since the theoretical convergence of the ADMM framework is not guaranteed with more than two block variables [53], we investigate the empirical convergence of LRP-tP instead. The residuals of variables are plotted in Fig. 10. We can see that the residuals drop quickly within few iterations. Generally, LRP-tP can reach the smallest residuals within 20 iterations, showing that LRT-2DP holds a fast convergence property in real scenarios.

V. CONCLUSION

In this paper, we proposed a Low-Rank Preserving t-Linear Projection (LRP-tP) model for robust image feature extraction. LRP-tP overlays the advantages of the robustness of self-representation learning, the discrimination information from data labels, and the flexibility of adaptive graph learning. More importantly, LRP-tP can simultaneously exploit the correlation among the multi-way data structure by taking advantage of the t-product-based operations. The development of LRP-tP opens up new avenues for developing powerful projection learning methods that are specialized for the multi-way data.

Motivated by the good performance of LLRSE and other sparse feature selection models [54], [56], it is highly expected to introduce some constraints on the basis slices of LRP-tP for joint feature selection and projection learning. In light of the fact that the t-product can be carried out efficiently in the Fourier domain, it might be possible to learn the constrained t-linear basis slices in the Fourier domain. Yet, how to associate the constraints in the Fourier with those in the original domain is an open problem. Our future work will investigate this topic.

APPENDIX

Proof: Owning to the conjugate symmetry of the Fourier transform, we have the observation $\mathit{fft}(\mathcal{R}', [\ ], 3) = (\mathit{fft}(\mathcal{R}, [\ ], 3))'$. Note that this equation is also used for the proof of t-SVD (Theorem 4.1, [41]).

The proof of Theorem 1 is by construction. First, let the third dimension of the variable tensors be n_3, we can

diagonalize $\mathcal{R}' * \mathcal{N}_1 - \mathcal{N}_2$ based on Fact 1 as

$$\begin{bmatrix} R_f^{(1)'} & & & \\ & R_f^{(2)'} & & \\ & & \ddots & \\ & & & R_f^{(n_3)'} \end{bmatrix} \cdot \begin{bmatrix} N_{1f}^{(1)} & & & \\ & N_{1f}^{(2)} & & \\ & & \ddots & \\ & & & N_{1f}^{(n_3)} \end{bmatrix} - \begin{bmatrix} N_{2f}^{(1)} & & & \\ & N_{2f}^{(2)} & & \\ & & \ddots & \\ & & & N_{2f}^{(n_3)} \end{bmatrix}, \quad (29)$$

where $\mathcal{R}_f$, $\mathcal{N}_{1f}$, and $\mathcal{N}_{2f}$ are computed by applying FFT on $\mathcal{R}$, $\mathcal{N}_1$, and $\mathcal{N}_2$ along the third dimension.

Similarly, $\mathcal{R}' * \mathcal{R} = \mathcal{I}$ is diagonalized as

$$\begin{bmatrix} R_f^{(1)'} & & & \\ & R_f^{(2)'} & & \\ & & \ddots & \\ & & & R_f^{(n_3)'} \end{bmatrix} \cdot \begin{bmatrix} R_f^{(1)} & & & \\ & R_f^{(2)} & & \\ & & \ddots & \\ & & & R_f^{(n_3)} \end{bmatrix} = I. \quad (30)$$

The original problem Eq. (18) is therefore transformed into independent problems with respect to the frontal slices

$$\min_{\mathcal{R}_f^{(k)}} \|\mathcal{R}_f^{(k)'} \mathcal{N}_{1f}^{(k)} - \mathcal{N}_{2f}^{(k)}\|_F^2 \quad s.t. \quad \mathcal{R}_f^{(k)'} \mathcal{R}_f^{(k)} = I, \quad (31)$$

for $k = 1, \ldots, n_3$. Eq. (31) can be solved by the orthogonal Procrustes problem in the complex space (APPENDIX B, [57]). Let the economy-size complex-valued SVD of $\mathcal{N}_{1f}^{(k)} \mathcal{N}_{2f}^{(k)'}$ be WSV'. Then, $\mathcal{R}_f^{(k)*} = WV'$, and the solution of Eq. (18) is recovered from $\mathcal{R} = ifft(\mathcal{R}_f, [\,], 3)$

On the other hand, since the t-SVD is implemented in the Fourier domain (Theorem 4.1, [41]), the optimization of Eq. (31) for $k = 1, \ldots, n_3$ is equivalent to calculating t-SVD as $\mathcal{N}_1 * \mathcal{N}_2' = \mathcal{W} * \mathcal{S} * \mathcal{V}'$, and then setting $\mathcal{R}^* = \mathcal{W} * \mathcal{V}'$. ■

REFERENCES

[1] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[2] J. P. Cunningham and Z. Ghahramani, "Linear dimensionality reduction: Survey, insights, and generalizations," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 2859–2900, 2015.

[3] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1991, pp. 586–591.

[4] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Müllers, "Fisher discriminant analysis with kernels," in *Proc. Neural Netw. Signal Process. Workshop*, 1999, pp. 41–48.

[5] M. Yin, S. Xie, Z. Wu, Y. Zhang, and J. Gao, "Subspace clustering via learning an adaptive low-rank graph," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3716–3728, Aug. 2018.

[6] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 153–160.

[7] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[8] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Min.*, 2014, pp. 977–986.

[9] K. Xiong, F. Nie, and J. Han, "Linear manifold regularization with adaptive graph for semi-supervised dimensionality reduction," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3147–3153.

[10] D. Zhang and Z.-H. Zhou, "(2D) 2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition," *Neurocomputing*, vol. 69, nos. 1–3, pp. 224–231, Dec. 2005.

[11] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.

[12] S. Chen, H. Zhao, M. Kong, and B. Luo, "2D-LPP: A two-dimensional extension of locality preserving projections," *Neurocomputing*, vol. 70, nos. 4–6, pp. 912–921, Jan. 2007.

[13] F. Zhang, J. Yang, J. Qian, and Y. Xu, "Nuclear norm-based 2-DPCA for extracting features from images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2247–2260, Oct. 2015.

[14] Z. Zhang, F. Li, M. Zhao, L. Zhang, and S. Yan, "Robust neighborhood preserving projection by Nuclear/l2,1-norm regularization for image feature extraction," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1607–1622, Apr. 2017.

[15] Y. Lu, C. Yuan, Z. Lai, X. Li, W. K. Wong, and D. Zhang, "Nuclear norm-based 2DLPP for image classification," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2391–2403, Nov. 2017.

[16] Y. Lu, C. Yuan, X. Li, Z. Lai, D. Zhang, and L. Shen, "Structurally incoherent low-rank 2DLPP for image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 6, pp. 1701–1714, Jun. 2019.

[17] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "MPCA: Multilinear principal component analysis of tensor objects," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 18–39, Jan. 2008.

[18] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang, "Multilinear discriminant analysis for face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 212–220, Jan. 2007.

[19] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

[20] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "A survey of multilinear subspace learning for tensor data," *Pattern Recognit.*, vol. 44, no. 7, pp. 1540–1551, Jul. 2011.

[21] S. Yuan, X. Mao, and L. Chen, "Multilinear spatial discriminant analysis for dimensionality reduction," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2669–2681, Jun. 2017.

[22] Y.-J. Deng, H.-C. Li, K. Fu, Q. Du, and W. J. Emery, "Tensor low-rank discriminant embedding for hyperspectral image dimensionality reduction," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7183–7194, Dec. 2018.

[23] O. Semerci, N. Hao, M. E. Kilmer, and E. L. Miller, "Tensor-based formulation and nuclear norm regularization for multienergy computed tomography," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1678–1693, Apr. 2014.

[24] J. Wright, A. Y. Yang, A. Ganesh, S. Shankar 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.

[25] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[26] 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.

[27] Y. Zhang, M. Xiang, and B. Yang, "Low-rank preserving embedding," *Pattern Recognit.*, vol. 70, pp. 112–125, Oct. 2017.

[28] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, and C. Yuan, "Low-rank preserving projections," *IEEE Trans. Cybern.*, vol. 46, no. 8, pp. 1900–1913, Aug. 2016.

[29] 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.

[30] L. Xie, M. Yin, X. Yin, Y. Liu, and G. Yin, "Low-rank sparse preserving projections for dimensionality reduction," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5261–5274, Nov. 2018.

[31] Y. Chen, Z. Lai, W. Keung 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.

[32] J. Wen, N. Han, X. Fang, L. Fei, K. Yan, and S. Zhan, "Low-rank preserving projection via graph regularized reconstruction," *IEEE Trans. Cybern.*, vol. 49, no. 4, pp. 1279–1291, Apr. 2019.
[33] X. Fang, Y. Xu, Z. Zhang, Z. Lai, and L. Shen, "Orthogonal self-guided similarity preserving projections," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 344–348.
[34] X. Fang, Y. Xu, X. Li, Z. Lai, S. Teng, and L. Fei, "Orthogonal self-guided similarity preserving projection for classification and clustering," *Neural Netw.*, vol. 88, pp. 1–8, Apr. 2017.
[35] S. Li and Y. Fu, "Robust subspace discovery through supervised low-rank constraints," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2014, pp. 163–171.
[36] 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.
[37] X. Fang *et al.*, "Approximate low-rank projection learning for feature extraction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5228–5241, Nov. 2018.
[38] Z. Ren, Q. Sun, B. Wu, X. Zhang, and W. Yan, "Learning latent low-rank and sparse embedding for robust image feature extraction," *IEEE Trans. Image Process.*, vol. 29, pp. 2094–2107, 2020.
[39] 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.
[40] 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.
[41] M. E. Kilmer and C. D. Martin, "Factorization strategies for third-order tensors," *Linear Algebra Appl.*, vol. 435, no. 3, pp. 641–658,
[42] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover, "Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging," *SIAM J. Matrix Anal. Appl.*, vol. 34, no. 1, pp. 148–172, Jan. 2013.
[43] E. Kernfeld, S. Aeron, and M. Kilmer, "Clustering multi-way data: A novel algebraic approach," 2014, *arXiv:1412.7056*. [Online]. Available: http://arxiv.org/abs/1412.7056
[44] M. Cheng, L. Jing, and M. K. Ng, "Tensor-based low-dimensional representation learning for multi-view clustering," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2399–2414, May 2019.
[45] P. Zhou, C. Lu, J. Feng, Z. Lin, and S. Yan, "Tensor low-rank representation for data recovery and clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 21, 2019, doi: 10.1109/TPAMI.2019.2954874.
[46] Y. Xie, D. Tao, W. Zhang, Y. Liu, L. Zhang, and Y. Qu, "On unifying multi-view self-representations for clustering by tensor multi-rank minimization," *Int. J. Comput. Vis.*, vol. 126, no. 11, pp. 1157–1179, Nov. 2018.
[47] F. Jiang, X.-Y. Liu, H. Lu, and R. Shen, "Efficient multi-dimensional tensor sparse coding using T-linear combination," in *Proc. Amer. Assoc. Artif. Intell.*, 2018, pp. 3326–3333.
[48] M. Yin, J. Gao, S. Xie, and Y. Guo, "Multiview subspace clustering via tensorial t-Product representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 851–864, Mar. 2019.
[49] Z. Zhang *et al.*, "Adaptive structure-constrained robust latent low-rank coding for image recovery," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 846–855.
[50] W. Deng, J. Hu, and J. Guo, "Face recognition via collaborative representation: Its discriminant nature and superposed representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2513–2521, Oct. 2018.
[51] C. Song, S. Yoon, and V. Pavlovic, "Fast ADMM algorithm for distributed optimization with adaptive penalty," in *Proc. Amer. Assoc. Artif. Intell.*, 2016, pp. 753–759.
[52] A. Hjorungnes and D. Gesbert, "Complex-valued matrix differentiation: Techniques and key results," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2740–2746, Jun. 2007.
[53] 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.
[54] A. Y. Ng, "Feature selection, L_1 vs. L_2 regularization, and rotational invariance," in *Proc. Int. Conf. Mach. Intell.*, 2004, p. 78.
[55] X. Xiao and Y. Zhou, "Two-dimensional quaternion PCA and sparse PCA," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 2028–2042, Jul. 2019.
[56] X. Xiao, Y. Chen, Y.-J. Gong, and Y. Zhou, "2D quaternion sparse discriminant analysis," *IEEE Trans. Image Process.*, vol. 29, pp. 2271–2286, 2020.
[57] L. Li, X. Wang, and G. Wang, "Alternating direction method of multipliers for separable convex optimization of real functions in complex variables," *Math. Problems Eng.*, vol. 2015, pp. 1–14, Jan. 2015.

Xiaolin Xiao received the B.E. degree from Wuhan University, China, in 2013, and the Ph.D. degree from the University of Macau, Macau, in 2019. She is currently a Postdoctoral Fellow with the School of Computer Science and Engineering, South China University of Technology, China. Her research interests include multi-view learning and color image processing and understanding.

Yongyong Chen received the B.S. and M.S. degrees from the Shandong University of Science and Technology, Qingdao, China, in 2014 and 2017, respectively, and the Ph.D. degree from the University of Macau, Macau, in 2020. He is currently an Assistant Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. His research interests include (non-convex) low-rank and sparse matrix/tensor decomposition models, with applications to image processing, data mining, and computer vision.

Yue-Jiao Gong (Senior Member, IEEE) received the B.S. and Ph.D. degrees in computer science from Sun Yat-sen University, China, in 2010 and 2014, respectively. She is currently a Full Professor with the School of Computer Science and Engineering, South China University of Technology, China. Her research interests include evolutionary computation, swarm intelligence, machine learning, and their applications to image processing and smart city. She has published over 80 articles, including more than 40 IEEE TRANSACTIONS articles, in her research areas.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree in electrical engineering from Hunan University, Changsha, China, and the M.S. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA.

He is currently an Associate Professor and the Director of the Vision and Image Processing Laboratory, Department of Computer and Information Science, University of Macau. His research interests include image processing, computer vision, machine learning, and multimedia security.

Dr. Zhou is a Senior Member of the International Society for Optical Engineering (SPIE). He was a recipient of the Third Price of Macao Natural Science Award in 2014 and 2020. He is the Co-Chair of Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society. He serves as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and four other journals.