Hierarchical Graph Augmented Deep Collaborative Dictionary Learning for Classification

Jianping Gou, *Senior Member, IEEE*, Xia Yuan, Lan Du, Shuyin Xia, *Member, IEEE*, and Zhang Yi, *Fellow, IEEE*

***Abstract*— Recently, deep dictionary learning (DDL) has aroused attention due to its abilities of learning multiple different dictionaries and extracting multi-level abstract feature representations for samples. It has been applied to many intelligent recognition tasks, such as vehicle detection, traffic sign recognition and driver monitoring. Nevertheless, the off-the-shelf DDL-based methods ignore the essential structural information of data in multi-layer dictionary learning. The learned hierarchical data representations are less discriminative. To address this issue, we develop a new DDL framework, called the hierarchical graph augmented deep collaborative dictionary learning (HGDCDL). Firstly, we propose a new deep collaborative dictionary learning (DCDL) that applies collaborative representation to the deepest-level representation learning. Most importantly, equipped with a simple yet effective hierarchal graph construction mechanism, our HGDCDL uses the structure of data to regularize dictionary learning, and generates more informative dictionaries and discriminative representations at different levels. Extensive experiments show that our HGDCDL performs significantly better than the state-of-the-art shallow and deep representation learning methods for classification.**

***Index Terms*— Deep dictionary learning, representation learning, graph construction, pattern classification.**

I. Introduction

DICTIONARY learning (DL) has been one of the most popular representation learning methods for decades in the field of artificial intelligence. Existing works in the literature have demonstrated that learning a good dictionary can achieve better outcomes in tasks, such as face recognition [1]–[3], denoising [4], [5], image clustering [6], [7], image super-resolution [8]–[10], object detection [11], [12] and person re-identification [13]. Dictionary learning has also been successfully applied to intelligent transportation [14], particularly including traffic sign recognition [15], license plate recognition [16], vehicle detection [17] and driver monitoring [18]. Among most off-the-shelf dictionary learning methods, the key idea is to learn informative dictionary atoms for data representation.

Traditional sparse representation (SR) [19] and collaborative representation (CR) [20] learning methods often learn coefficients by treating the training data matrix as a fixed dictionary. Although those approaches have no time complexity of computing the dictionary, the dependency on the data becomes a bottleneck. In real-world visual recognition tasks, like image classification, we prefer to learn unfixed dictionaries from the data automatically. For instance, K-SVD [21], J-RFDL [22] and RNMLF [23] are supervised dictionary learning that is better fit to the classification tasks. Specifically, D-KSVD [24], LC-KSVD [25] and FDDL [26] aim to learn a more discriminative dictionary via leveraging the label information.

The original high-dimensional data often contains rich discriminative information, like the geometry of the data, which has been found to play an important role in classifying images. Many manifold-based algorithms are designed to preserve the geometric properties of the original data when projecting the high-dimensional data into the low-dimensional space. For instance, locality preserving projection (LPP) [27] uses a heart kernel function to define an adjacent weight matrix of data points that can reflect the local manifold structure. Inspired by LPP and sparsity preserving projection [28], discriminative sparsity preserving graph embedding (DSPGE) [29] devises a new weight construction method which fully considers the class information and the geometric distribution of each data point. Discriminative globality and locality preserving graph embedding [30] further extends this idea to construct simultaneously a global adjacency graph and a local adjacency graph to preserve intra-class and inter-class manifold structures in the embedded subspace.

To consider the manifold structure of the original data, many existing studies use a single-layer dictionary learning approach, where a Laplacian graph regularization term is added to the learning objective. To be specific, Zheng *et al.* [31] proposed a graph-regularized sparse coding method, the objective function of which is regularized by a Laplacian graph term. Graph-regularized discriminative analysis-synthesis dictionary pair learning [32] takes into account the geometry structure and the label information

Manuscript received 18 November 2021; revised 18 March 2022 and 16 April 2022; accepted 21 May 2022. Date of publication 30 May 2022; date of current version 5 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61976107 and Grant 61502208 and in part by the Qing Lan Project of Colleges and Universities of Jiangsu Province in 2020. The Associate Editor for this article was H. Lu. *(Corresponding authors: Jianping Gou; Lan Du.)*

Jianping Gou and Xia Yuan are with the School of Computer Science and Communication Engineering and the Jiangsu Key Laboratory of Security Technology for Industrial Cyberspace, Jiangsu University, Zhenjiang, Jiangsu 212013, China (e-mail: goujianping@ujs.edu.cn; 2212008046@stmail.ujs.edu.cn).

Lan Du is with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia (e-mail: lan.du@monash.edu).

Shuyin Xia is with the Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: xiasy@cqupt.edu.cn).

Zhang Yi is with the School of Computer Science, Sichuan University, Chengdu 610017, China (e-mail: zhangyi@scu.edu.cn).

Digital Object Identifier 10.1109/TITS.2022.3177647

of data by introducing a graph-regularization term and a discriminative term. Adaptive graph regularized non-negative matrix factorization [33] integrates manifold learning and non-negative matrix factorization for image clustering, aiming to capture both the global and the local manifold structures. Rather than selecting features in original data space, Ding *et al.* [34] focused on learning dictionaries via unsupervised feature selection with a similarity graph. Rong *et al.* [35] proposed double graph regularized double dictionary learning that learns multiple class-specific sub-dictionaries and a class-shared dictionary to capture both the class-specific and the class-shared information. Charles *et al.* [36] proposed graph filtered temporal dictionary learning for calcium image analysis, where the pixel temporal correlations are reorganized into a data-driven Laplacian graph. Besides, dictionary learning has been adapted to graph data sets, where the graph is modeled as a combination of graph atoms (i.e., dictionary atoms) [37], [38].

Instead of learning a single-level dictionary, Tariyal *et al.* [39] proposed to learn multi-level dictionaries, known as deep dictionary learning (DDL), the idea of which is originated from the multi-level architecture used commonly in deep learning [40], [41]. The dictionaries can be learned in a greedy layer-by-layer manner in a deep structure. Thereafter, DDL has been extended in various ways. [42] achieved supervised deep dictionary learning with an extra regression term. Coupled deep dictionary learning [43] learns multi-level dictionaries from both the source and the target domains. The method of deep dictionary learning and the coding network (DDLCN) [44] replaces the convolutional layers with dictionary learning and coding layers. Multilayered K-singular value decomposition [45] first performs sparse coding, and then conducts multi-level dictionary learning that uses the class information to promote discrimination. The dictionary learning of [46] done in the encoder layer of a sparse deep autoencoder is able to describe the differences in the dynamic functional connectivity. Rodriguez-Dominguez *et al.* [47] embedded a hierarchical discriminant dictionary learning layer into a neural network. Yang *et al.* [48] proposed a collaborative representation method for the change detection of remote-sensing images, which uses l_2-norm instead of l_1-norm for the coefficient constraints.

All of the aforementioned methods extend the single-layer dictionary learning to the multi-level dictionary learning in order to obtain good feature representations of data. However, those methods nearly utilize the l_1-norm to learn multiple dictionaries that can lead subsequently to sparse representations. We argue that without considering the underlying rich discriminative and intrinsic structure exhibited by the original high-dimensional data, the learned dictionaries can be sub-optimal, possibly containing redundant and noisy information. Based on the theory of manifold learning, it is natural for us to assume that the learned sample feature representations should reflect the discriminative and geometric structure of the original high dimensional data. In other words, if two samples are close to each other in the original space, their representations are expected to be close to each other as well. To the best of our knowledge, almost all existing DDL-based methods neglect the geometric structure in the original data space in learning the multiple dictionaries, which can result in a less discriminative embedding space, as the geometric information and the class information of data are key elements for discrimination.

Motivated by those desiderata, we propose a novel hierarchical graph augmented deep collaborative dictionary learning (HGDCDL) model for effectively solving different image classification tasks. Our main contributions can be highlighted as follows:

- **Deep collaborative dictionary learning (DCDL)**: We propose a deep collaborative dictionary learning model that adapts collaborative representation to the deepest-level dictionary learning. Compared with DDL model, DCDL can obtain a more discriminative representation at the last layer, which is more conducive to image classification tasks.
- **Hierarchical graph augmented deep collaborative dictionary learning (HGDCDL)**: We further extend DCDL by constraining the dictionary learning at each layer in the multi-level setting. The constraints consist of a hierarchical graph constraint and a collaborative representation constraint. Although we use a supervised and more discriminative graph construction method to demonstrate the effectiveness of HGDCDL, it can also adopt other graph construction methods, either supervised or unsupervised. HGDCDL can obtain more informative dictionaries and more discriminating representations.
- **Extensive experiments containing both comparative analysis and ablation studies**: We compare HGDCDL with the standard representation-based methods, deep learning methods and deep dictionary learning methods on three facial recognition data sets and three non-facial recognition data sets. Our HGDCDL outperforms all the competitors with a notable margin. Meanwhile, we carried out in-depth analysis of the performance of HGDCDL with a set of ablation studies.

The remainder of this paper is organized as follows. Section II reviews the related works. Section III describes the proposed HCDCDL model in detail. Section IV presents the extensive experiments. Section V shows ablation studies. Finally, Section VI gives conclusions and future works.

II. The Related Works

We review collaborative representation learning, dictionary learning and deep dictionary learning as the backgrounds for our hierarchical graph augmented deep collaborative dictionary learning.

A. Collaborative Representation Learning

The collaborative representation (CR) learning method [20] represents each query sample as a weighted linear combination of all the training samples. Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M] \in \mathbb{R}^{d \times M}$ be a set of M training samples from C classes, each of which is represented as a d-dimensional feature vector. $\mathbf{Y}$ can be rewritten as $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_C]$, where $\mathbf{Y}_c$ indicates all samples from class c. CR substitutes the l_1-norm used in SR

with the l_2-norm. Thus, a new testing sample $\mathbf{y}_{test} \in \mathbb{R}^d$ can be reconstructed from $\mathbf{Y}$ via ridge regression:

$$\min_{\mathbf{z}} \|\mathbf{y}_{test} - \mathbf{Y}\mathbf{z}\|_2^2 + \alpha\|\mathbf{z}\|_2^2 \ , \tag{1}$$

where $\mathbf{z} \in \mathbb{R}^M$ is a representation coefficient vector and α is a regularization parameter of representation coefficients. Given the ridge form, CR favours a closed-form solution,

$$\mathbf{z} = (\mathbf{Y}^T\mathbf{Y} + \alpha\mathbf{I})^{-1}\mathbf{Y}^T\mathbf{y}_{test} \ , \tag{2}$$

where $\mathbf{z} = [z_1; z_2; \ldots; z_C]$ and $\mathbf{I}$ is an identity matrix. z_c is a vector of coefficients associated with the samples from class c, and can be seen as the class-specific representation coefficients for y_{test}. Let $\mathbf{G} = (\mathbf{Y}^T\mathbf{Y} + \alpha\mathbf{I})^{-1}\mathbf{Y}^T$. In practice, $\mathbf{G}$ can be pre-computed with $\mathbf{Y}$, which significantly reduces the time complexity of computing Eq. (2).

B. Dictionary Learning

Dictionary learning (DL) is usually used for matrix factorization [49], [50] and sparse coding [51] problems. DL-based techniques have also been explored for classification tasks.

The basic formulation for DL is given as:

$$\mathbf{Y} = \mathbf{D}\mathbf{Z} \ , \tag{3}$$

where $\mathbf{D} \in \mathbb{R}^{d\times K}$ is the learned dictionary (or known as a component matrix in factor analysis) and $\mathbf{Z} \in \mathbb{R}^{K\times M}$ contains the corresponding coefficients (or known as a factor loading matrix). As one of the most popular DL algorithms, K-SVD [21] possesses a sparse constraint on the coefficient matrix $\mathbf{Z}$, leading to the following formulation:

$$\min_{\mathbf{D},\mathbf{Z}} \|\mathbf{Y} - \mathbf{D}\mathbf{Z}\|_F^2 \ s.t. \|\mathbf{Z}\|_0 \le T \ , \tag{4}$$

where T represents the degree of sparsity. K-SVD iteratively performs sparse coding and dictionary updating.

C. Deep Dictionary Learning

Deep dictionary learning (DDL) [39] extends the idea of dictionary learning to a multi-layer structure similar to deep neural networks. Dictionary learning happens in a layer-by-layer fashion. The representation (i.e., the coefficient matrix $\mathbf{Z}$) learned at one level is often used as the input to learn the dictionary at the next level. The single-level DL can be seen as a basic construct of DDL. Fig. 1 shows as an example a three-level dictionary learning model.

Let $\mathbf{D}_n \in \mathbb{R}^{K_{n-1}\times K_n}, \forall n = 1, 2, \ldots, N$ be the dictionaries for N different levels (or layers) and $\mathbf{Z}^n \in \mathbb{R}^{K_n\times M}$, $\forall n = 1, 2, \ldots, N$ be the layer-wise representation matrices. The input of the model can be denoted as $\mathbf{Z}^0$ where $\mathbf{Z}^0 = \mathbf{Y}$.

The objective function of DDL can be written as:

$$\min_{\mathbf{D}_1,\ldots,\mathbf{D}_N,\mathbf{Z}^N} \|\mathbf{Z}^0 - \mathbf{D}_1\varphi(\mathbf{D}_2\varphi(\cdots\varphi(\mathbf{D}_N\mathbf{Z}^N)))\|_F^2 + \alpha\|\mathbf{Z}^N\|_1 \ , \tag{5}$$

where $\varphi(.)$ is an activation function and α is a regularization parameter. From 1^{st} to $(N-1)^{th}$ layer, we have $\mathbf{Z}^n = \varphi(\mathbf{D}_{n+1}\cdots\varphi(\mathbf{D}_N\mathbf{Z}^N))$ and $\mathbf{Z}^{n-1} = \varphi(\mathbf{D}_n\mathbf{Z}^n)$ (i.e., $\varphi^{-1}(\mathbf{Z}^{n-1}) = \mathbf{D}_n\mathbf{Z}^n$). Thus, given $\mathbf{Z}^{n-1}$ as input to the n-th layer, the corresponding learning objective can be formulated as:

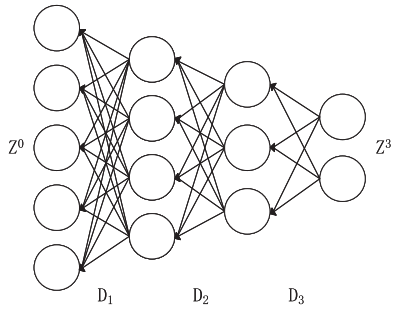


Fig. 1. The overview diagram of DDL [39].

$$\min_{\mathbf{D}_n,\mathbf{Z}^n} \|\varphi^{-1}(\mathbf{Z}^{n-1}) - \mathbf{D}_n\mathbf{Z}^n\|_F^2, n = 1, 2, \ldots, N-1 \ . \tag{6}$$

We could solve it using the alternating least square algorithm.

For the final layer, we have $\varphi^{-1}(\mathbf{Z}^{N-1}) = \mathbf{D}_N\mathbf{Z}^N$. Similarly, we can learn the dictionary and the corresponding representation at the last layer as:

$$\min_{\mathbf{D}_N,\mathbf{Z}^N} \|\varphi^{-1}(\mathbf{Z}^{N-1}) - \mathbf{D}_N\mathbf{Z}^N\|_F^2 + \alpha\|\mathbf{Z}^N\|_1 \ . \tag{7}$$

Indeed, the learning problem at the N-th layer has the exactly same form as Lasso, which can be solved efficiently.

The activation function $\varphi(\cdot)$ can be as simple as a copy function [39], i.e., $\varphi(x) = x$. Then in the testing stage, the dictionary $\mathbf{D}_{test}$ is equal to the product of the layer-wise dictionaries:

$$\mathbf{D}_{test} = \mathbf{D}_1\mathbf{D}_2\cdots\mathbf{D}_N \ . \tag{8}$$

The computation of the representation $\mathbf{Z}_{test}$ for a given set of testing samples $\mathbf{Y}_{test}$ is simplified as

$$\underset{\mathbf{Z}_{test}}{\operatorname{argmin}} \|\mathbf{Y}_{test} - \mathbf{D}_{test}\mathbf{Z}_{test}\|_2^2 + \alpha\|\mathbf{Z}_{test}\|_1 \ . \tag{9}$$

Finally, for classifying the testing samples, $\mathbf{Z}^N$ and $\mathbf{Z}_{test}$ are the input to the classifier that is the KNN method used in general.

III. The Proposed HGDCDL

Most existing DDL methods impose a sparse constraint only on the last level of the representation, which often leads to the problem of matrix singularity and further results in poor convergence. Moreover, those works often ignore the category information and the geometric distribution of the original data. The ignorance can cause the learned deep representations less discriminative, further degrading the classification performance.

To address the issues above, we introduce a new hierarchical graph augmented deep collaborative dictionary learning (HGDCDL). It constrains the representation learned at each level with the discriminative and geometric information of all the data samples uncovered by hierarchical Laplacian graphs, which are constructed by the layer-wise representations. We expect that the injection of the geometric and

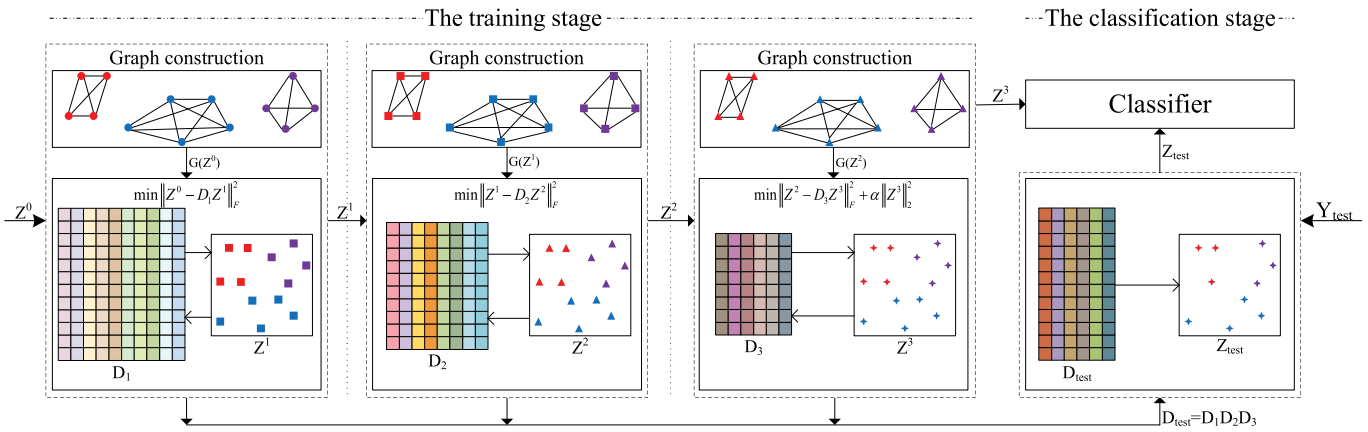


Fig. 2. The overview diagram of a three-level HGDCDL model as an example. $\mathbf{Z}^0$ and $\mathbf{Y}_{test}$ are the training data and the testing data, respectively. $\mathbf{D}_n$ $(n = 1, 2, 3)$ means the dictionary at the n-th level. $\mathbf{Z}^n$ $(n = 1, 2, 3)$ is the corresponding representation at the n-th level. $\mathbf{D}_{test}$ and $\mathbf{Z}_{test}$ are the dictionary and the representation in the classification stage. G($\mathbf{Z}^0$) is the graph constraint applied to $\mathbf{Z}^0$ where $G(\mathbf{Z}^0) = tr(\mathbf{Z}^1\mathbf{L}^1\mathbf{Z}^{1^T})$, and $\mathbf{L}^1$ is constructed by $\mathbf{Z}^0$. Meanwhile, $G(\mathbf{Z}^1) = tr(\mathbf{Z}^2\mathbf{L}^2\mathbf{Z}^{2^T})$, $G(\mathbf{Z}^3) = tr(\mathbf{Z}^3\mathbf{L}^3\mathbf{Z}^{3^T})$.

discriminative information can learn more informative dictionaries and discriminative representations. Fig. 2 shows a three-level HGDCDL architecture.

A. Deep Collaborative Dictionary Learning

DDL [39] assumes that the deepest-level representation $\mathbf{Z}^N$ is sparse (See Eq. (5)), the learning of which can be solved with an Iterative Soft Thresholding Algorithm (ISTA). However, sparse solutions can often cause the issue of matrix singularity and poor convergence. We instead propose to replace the l_1-norm constraint imposed on $\mathbf{Z}^N$ with an l_2-norm constraint, which gives us an extension of DDL, named deep collaborative dictionary learning (DCDL).

With l_1-norm in Eqs. (5), (7) and (9) replaced with l_2-norm and the assumption of the activation function $\varphi(x) = x$, the optimization problem of DCDL is formulated as:

$$\min_{\mathbf{D}_1,\ldots,\mathbf{D}_N,\mathbf{Z}^N} \|\mathbf{Z}^0 - \mathbf{D}_1\mathbf{D}_2\cdots\mathbf{D}_N\mathbf{Z}^N\|_F^2 + \alpha\|\mathbf{Z}^N\|_2^2 \ , \quad (10)$$

where α is a regularization parameter.

Similar to DDL, the dictionary learning at the n-th level $(n = 1, 2, \ldots, N-1)$, i.e.,

$$\min_{\mathbf{D}_n,\mathbf{Z}^n} \|\mathbf{Z}^{n-1} - \mathbf{D}_n\mathbf{Z}^n\|_F^2 \quad (11)$$

can be solved directly as a least square problem. As we argued above, we impose collaborative representation constraint on the feature representation learned at the last level in order to overcome the issues faced by the original DDL as follows:

$$\min_{\mathbf{D}_N,\mathbf{Z}^N} \|\mathbf{Z}^{N-1} - \mathbf{D}_N\mathbf{Z}^N\|_F^2 + \alpha\|\mathbf{Z}^N\|_2^2 \ . \quad (12)$$

There are two benefits brought by the l_2-norm constraint according to our empirical results (See Sections IV-E and IV-F): 1) it avoids singularity and further enhances convergence; 2) the learned dictionary can potentially contain richer information than that learned with l_1-norm.

It is noteworthy that the deep collaborative dictionary learning proposed above is unsupervised. A greedy learning algorithm can be used to learn dictionaries and corresponding representations in a layer-wise fashion. However, the geometries of data or data representations are still missing in the learning of both the dictionary matrices and representation matrices, which we believe will contribute significantly to the quality of the two types of matrices.

B. Hierarchical Graph Construction

There are many ways of quantifying the geometries of data, either supervised or unsupervised. The adjacent graphs constructed in [29] are discriminative amongst existing approaches, which can contribute positively to the classification performance. Thus, we combine the advantages of this graph construction method and DCDL in our hierarchical graph augmented deep collaborative dictionary learning method. Note that these hierarchical graphs are constructed by using either the original data or the learned layer-wise representations.

Given the M training samples $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M] \in \mathbb{R}^{d\times M}$ from C classes where the class label of $\mathbf{y}_i$ is $c_i \in \{1, 2, \ldots, C\}$, HGDCDL learns N dictionaries and N representations for $\mathbf{Y}$ using a deep architecture with N layers. N dictionaries and representations are $\mathbf{D}_n$ and $\mathbf{Z}^n = [\mathbf{z}_1^n, \mathbf{z}_2^n, \ldots, \mathbf{z}_M^n] \in \mathbb{R}^{K_n\times M}(n = 1, 2, \ldots, N)$, respectively. At the n-th layer, the representation of $\mathbf{y}_i$ with label c_i is $\mathbf{z}_i^n$, the input is $\mathbf{Z}^{n-1}$ and the output is $\mathbf{Z}^n$. And the input at the first layer is $\mathbf{Z}^0 = \mathbf{Y}$. Given the input $Z^{n-1} \in \mathbb{R}^{K_{n-1}\times M}$, we treat each column of Z^{n-1} (i.e., z_i^{n-1}) as a vertex in a graph. The geometric distribution and class information of each sample are then used to determine the weight of each edge. Let $\mathbf{G}^{n+}$ be the intra-class adjacent graph at the n-th level and the corresponding edge weights be $\mathbf{W}^{n+}$. The edge weight w_{ij}^{n+} between two intra-class representations $\mathbf{z}_i^{n-1}$ of sample $\mathbf{y}_i$ and $\mathbf{z}_j^{n-1}$ of sample $\mathbf{y}_j$ can be defined as the average of the weight $\overline{w}_{ij}^{n+}$ from $\mathbf{z}_i^{n-1}$ to $\mathbf{z}_j^{n-1}$ and the weight $\overline{w}_{ji}^{n+}$ from $\mathbf{z}_j^{n-1}$ to $\mathbf{z}_i^{n-1}$. The weight $\overline{w}_{ij}^{n+}$ is defined as

$$\overline{w}_{ij}^{n+} = \exp\left(-\frac{\|\mathbf{z}_i^{n-1} - \mathbf{z}_j^{n-1}\|^2}{\rho_i^{n+}}\right) , \quad (13)$$

where the parameter ρ_i^{n+} represents the intra-class geometric distribution around $\mathbf{z}_i^{n-1}$. It is defined as

$$\rho_i^{n+} = \frac{1}{(N^+(c_i))^q} \sum_{t=1}^{N^+(c_i)} \|\mathbf{z}_i^{n-1} - \mathbf{z}_t^{n-1}\|^2 \ , \tag{14}$$

where q is an exponential regulated parameter which controls ρ_i^{n+}, and $N^+(c_i)$ is the number of the intra-class representations. Similarly, we can define the weight $\overline{w}_{ji}^{n+}$ as

$$\overline{w}_{ji}^{n+} = \exp(-\frac{\|\mathbf{z}_j^{n-1} - \mathbf{z}_i^{n-1}\|^2}{\rho_j^{n+}}) \ , \tag{15}$$

where the parameter ρ_j^{n+} represents the intra-class geometric distribution around $\mathbf{z}_j^{n-1}$ and it is defined as

$$\rho_j^{n+} = \frac{1}{(N^+(c_j))^q} \sum_{t=1}^{N^+(c_j)} \|\mathbf{z}_j^{n-1} - \mathbf{z}_t^{n-1}\|^2 \ . \tag{16}$$

Then, the weight w_{ij}^{n+} between two representations $\mathbf{z}_i^{n-1}$ and $\mathbf{z}_j^{n-1}$ is calculated as

$$w_{ij}^{n+} = \begin{cases} \frac{1}{2}\left(\overline{w}_{ij}^{n+} + \overline{w}_{ji}^{n+}\right), & c_i = c_j \\ 0, & c_i \neq c_j. \end{cases} \tag{17}$$

Let $\mathbf{G}^{n-}$ denote the inter-class adjacent graph at the n-th level, and $\mathbf{W}^{n-}$ indicate its corresponding adjacent weight matrix. Each edge weight w_{ij}^{n-} between two inter-class representations $\mathbf{z}_i^{n-1}$ of sample $\mathbf{y}_i$ and $\mathbf{z}_j^{n-1}$ of sample $\mathbf{y}_j$ can be defined as the average of the weight $\overline{w}_{ij}^{n-}$ from $\mathbf{z}_i^{n-1}$ to $\mathbf{z}_j^{n-1}$ and the weight $\overline{w}_{ji}^{n-}$ from $\mathbf{z}_j^{n-1}$ to $\mathbf{z}_i^{n-1}$. $\overline{w}_{ij}^{n-}$ is computed as

$$\overline{w}_{ij}^{n-} = \exp(-\frac{\|\mathbf{z}_i^{n-1} - \mathbf{z}_j^{n-1}\|^2}{\rho_i^{n-}}) \ , \tag{18}$$

where the parameter ρ_i^{n-} represents the inter-class geometric distribution around $\mathbf{z}_i^{n-1}$ and is defined as

$$\rho_i^{n-} = \frac{1}{(N^-(c_i))^q} \sum_{t=1}^{N^-(c_i)} \|\mathbf{z}_i^{n-1} - \mathbf{z}_t^{n-1}\|^2 \ . \tag{19}$$

$N^-(c_i)$ is the number of the inter-class representations associated with $\mathbf{z}_i^{n-1}$. In a similar manner, the weight $\overline{w}_{ji}^{n-}$ is defined as

$$\overline{w}_{ji}^{n-} = \exp(-\frac{\|\mathbf{z}_j^{n-1} - \mathbf{z}_i^{n-1}\|^2}{\rho_j^{n-}}), \tag{20}$$

where the parameter ρ_j^{n-} representing the inter-class geometric distribution around $\mathbf{z}_j^{n-1}$ is computed as

$$\rho_j^{n-} = \frac{1}{(N^-(c_j))^q} \sum_{t=1}^{N^-(c_j)} \|\mathbf{z}_j^{n-1} - \mathbf{z}_t^{n-1}\|^2. \tag{21}$$

Then, the weight w_{ij}^{n-} between two inter-class representations $\mathbf{z}_i^{n-1}$ and $\mathbf{z}_j^{n-1}$ is calculated as

$$w_{ij}^{n-} = \begin{cases} \frac{1}{2}\left(\overline{w}_{ij}^{n-} + \overline{w}_{ji}^{n-}\right), & c_i \neq c_j \\ 0, & c_i = c_j. \end{cases} \tag{22}$$

If two representations z_i^{n-1} and z_j^{n-1} learned at the $(n-1)$-th level from the same class are close to each other, their corresponding representations z_i^n and z_j^n learned at the n-th level are expected to be close to each other as well. To maintain the intra-class inherent properties, we prefer to minimize the intra-class objective function:

$$\begin{aligned} f^+ &= \frac{1}{2}\sum_{i,j} \|\mathbf{z}_i^n - \mathbf{z}_j^n\|^2 w_{ij}^{n+} \\ &= \frac{1}{2}\sum_{i,j} (\mathbf{z}_i^n - \mathbf{z}_j^n) w_{ij}^{n+} (\mathbf{z}_i^n - \mathbf{z}_j^n)^T \\ &= \frac{1}{2}\sum_{ij} (2w_{ij}^{n+} \mathbf{z}_i^n \mathbf{z}_i^{nT} - 2w_{ij}^{n+} \mathbf{z}_j^n \mathbf{z}_j^{nT}) \\ &= tr(\mathbf{z}_i^n \mathbf{H}^{n+} \mathbf{z}_i^{nT}) - tr(\mathbf{z}_j^n \mathbf{W}^{n+} \mathbf{z}_j^{nT}) \\ &= tr(\mathbf{Z}^n \mathbf{L}^{n+} \mathbf{Z}^{nT}), \end{aligned} \tag{23}$$

where $\mathbf{L}^{n+} = \mathbf{H}^{n+} - \mathbf{W}^{n+}$ is a Laplacian matrix, $\mathbf{H}^{n+}$ is a diagonal matrix with $\mathbf{H}_{ii}^{n+} = \sum_j w_{ij}^{n+}$ and $(\mathbf{W})_{ij}^{n+} = w_{ij}^{n+}$.

If two representations z_i^{n-1} and z_j^{n-1} at the $(n-1)$-th level from different classes are distant, the corresponding representations z_i^n and z_j^n at the n-th level should be farther away from each other. To maintain the inter-class inherent properties, we maximize the inter-class objective function:

$$\begin{aligned} f^- &= \frac{1}{2}\sum_{i,j} \|\mathbf{z}_i^n - \mathbf{z}_j^n\|^2 w_{ij}^{n-} \\ &= \frac{1}{2}\sum_{i,j} (\mathbf{z}_i^n - \mathbf{z}_j^n) w_{ij}^{n-} (\mathbf{z}_i^n - \mathbf{z}_j^n)^T \\ &= \frac{1}{2}\sum_{ij} (2w_{ij}^{n-} \mathbf{z}_i^n \mathbf{z}_i^{nT} - 2w_{ij}^{n-} \mathbf{z}_j^n \mathbf{z}_j^{nT}) \\ &= tr(\mathbf{z}_i^n \mathbf{H}^{n-} \mathbf{z}_i^{nT}) - tr(\mathbf{z}_j^n \mathbf{W}^{n-} \mathbf{z}_j^{nT}) \\ &= tr(\mathbf{Z}^n \mathbf{L}^{n-} \mathbf{Z}^{nT}), \end{aligned} \tag{24}$$

where $\mathbf{L}^{n-} = \mathbf{H}^{n-} - \mathbf{W}^{n-}$ is a Laplacian matrix, $\mathbf{H}^{n-}$ is a diagonal matrix with $\mathbf{H}_{ii}^{n-} = \sum_j w_{ij}^{n-}$ and $(\mathbf{W})_{ij}^{n-} = w_{ij}^{n-}$.

Considering simultaneously the geometric properties and discriminant properties of the intra-class and inter-class data representations, we minimize the intra-class objective function f^+ as well as maximize the inter-class objective function f^-. That is to say, at the n-th layer, we can minimize the graph regularized term, which is formulated as

$$\begin{aligned} f &= f^+ - f^- \\ &= tr(\mathbf{Z}^n \mathbf{L}^{n+} \mathbf{Z}^{nT}) - tr(\mathbf{Z}^n \mathbf{L}^{n-} \mathbf{Z}^{nT}) \\ &= tr(\mathbf{Z}^n \mathbf{L}^n \mathbf{Z}^{nT}), \end{aligned} \tag{25}$$

where $\mathbf{L}^n$ is also the Laplacian matrix of the n-th layer calculated as

$$\mathbf{L}^n = \mathbf{L}^{n+} - \mathbf{L}^{n-} \ . \tag{26}$$

C. The HGDCDL Framework

By incorporating deep collaborative dictionary learning and hierarchical graph constraints, our hierarchical graph augmented deep collaborative dictionary learning (HGDCDL)

model is defined as:

$$\min_{\mathbf{D}_1,\ldots,\mathbf{D}_N,\mathbf{Z}^N} \Sigma_{n=1}^{N}\Big(\|\varphi^{-1}(\mathbf{Z}^{n-1})-\mathbf{D}_n\mathbf{Z}^n\|_F^2+\beta tr\big(\mathbf{Z}^n\mathbf{L}^n\mathbf{Z}^{n^T}\big)\Big)+\alpha\|\mathbf{Z}^N\|_2^2\ , \tag{27}$$

where β is a graph regularization parameter and α is a regularization parameter of the deepest-level representation.

With the same assumption $\varphi(x)=x$, Eq. (27) can be further simplified to:

$$\min_{\mathbf{D}_1,\ldots,\mathbf{D}_N,\mathbf{Z}^N} \Sigma_{n=1}^{N}\Big(\|\mathbf{Z}^{n-1}-\mathbf{D}_n\mathbf{Z}^n\|_F^2+\beta tr\big(\mathbf{Z}^n\mathbf{L}^n\mathbf{Z}^{n^T}\big)\Big)+\alpha\|\mathbf{Z}^N\|_2^2\ . \tag{28}$$

Fig. 2 shows the three-level HGDCDL as an example. Eq. (28) can be divided into the following sub-problems:

1) The dictionary learning at the first $(N-1)$ levels: given the representation learned at the $(n-1)$-th level, i.e., $\mathbf{Z}^{n-1}$, the optimization problem at the n-th level is simplified to solve the following graph-regularized ridge regression problem:

$$\underset{\mathbf{D}_n,\mathbf{Z}^n}{\operatorname{argmin}}\ \|\mathbf{Z}^{n-1}-\mathbf{D}_n\mathbf{Z}^n\|_F^2+\beta tr(\mathbf{Z}^n\mathbf{L}^n\mathbf{Z}^{n^T}). \tag{29}$$

2) The l_2-norm constraint dictionary learning at the N-th level:

$$\underset{\mathbf{D}_N,\mathbf{Z}^N}{\operatorname{argmin}}\ \|\mathbf{Z}^{N-1}-\mathbf{D}_N\mathbf{Z}^N\|_F^2+\beta tr(\mathbf{Z}^N\mathbf{L}^N\mathbf{Z}^{N^T})+\alpha\|\mathbf{Z}^N\|_F^2. \tag{30}$$

D. The HGDCDL Training

Given the nested structure of deep dictionary learning, we develop a layer-wise alternating least square algorithm for learning both the dictionary and the sample representations at each level. The ridge forms shown in Eqs. (29) and (30) permit closed-form solutions to $\mathbf{D}_n$ and $\mathbf{Z}^n$, which can be efficiently learned from the training data.

1) Optimization at the first layer (i.e., the input layer): since $\mathbf{L}^1$ is constructed by training samples $\mathbf{Z}^0=\mathbf{Y}$ which can be computed by Eq. (26), $\mathbf{L}^1$ reflects discriminative and geometric distribution of original input samples. Then, the closed-form solution for the dictionary $\mathbf{D}_1$ and the representation $\mathbf{Z}^1$ are respectively:

$$\mathbf{D}_1=\mathbf{Z}^0\mathbf{Z}^{1^T}(\mathbf{Z}^1\mathbf{Z}^{1^T})^{-1} \tag{31}$$

and

$$\mathbf{Z}^1=(\mathbf{D}_1^T\mathbf{D}_1+\beta\mathbf{I})^{-1}\mathbf{D}_1^T\mathbf{Z}^0(\mathbf{I}+\mathbf{L}^1)^{-1}\ , \tag{32}$$

where $\mathbf{I}$ is an identity matrix.

2) Optimization at the intermediate layers, i.e., $n\in\{2,3,\ldots,N-1\}$: the output of the previous layer $\mathbf{Z}^{n-1}$ is the input of the next layer. As such, $\mathbf{D}_n$ and $\mathbf{Z}^n$ are optimized according to:

$$\mathbf{D}_n=\mathbf{Z}^{n-1}\mathbf{Z}^{n^T}(\mathbf{Z}^n\mathbf{Z}^{n^T})^{-1} \tag{33}$$

$$\mathbf{Z}^n=(\mathbf{D}_n^T\mathbf{D}_n+\beta\mathbf{I})^{-1}\mathbf{D}_n^T\mathbf{Z}^{n-1}(\mathbf{I}+\mathbf{L}^n)^{-1}. \tag{34}$$

Here, $\mathbf{L}^n$ is constructed by the output from the previous layer $\mathbf{Z}^{n-1}$ and can be computed by Eq. (26). It preserves the class information and geometric structure of the previous layer.

3) Optimization of the N-th layer: as we discussed above, the representation $\mathbf{Z}^N$ does not have to be sparse, we apply the l_2-norm constraint on $\mathbf{Z}^N$, which makes the least square solution stable and reduces computational cost. The dictionary and representation are optimized by:

$$\mathbf{D}_N=\mathbf{Z}^{N-1}\mathbf{Z}^{N^T}(\mathbf{Z}^N\mathbf{Z}^{N^T})^{-1} \tag{35}$$

$$\mathbf{Z}^N=(\mathbf{D}_N^T\mathbf{D}_N+(\alpha+\beta)\mathbf{I})^{-1}\mathbf{D}_N^T\mathbf{Z}^{N-1}(\mathbf{I}+\mathbf{L}^N)^{-1}. \tag{36}$$

Meanwhile, $\mathbf{L}^N$ is constructed by $\mathbf{Z}^{N-1}$ using Eq. (26) when $n=N$. Thus, the hierarchical graph constraint on each layer using the output representation of the previous layer make the relationships between layers closer. Finally, we obtain a trained HGDCDL model.

The HGDCDL training for iteratively learning the layer-wise dictionaries and representations is given in Algorithm 1.

Algorithm 1 The Training Algorithm in HGDCDL

Input: The training sample matrix $\mathbf{Z}^0=\mathbf{Y}\in\mathbb{R}^{d\times M}$, the given dictionary layers N and the parameters α and β;

Initialization: $\mathbf{D}_1,\mathbf{D}_2,\ldots,\mathbf{D}_N$

For the first level

1. Construct $\mathbf{L}^1$ with Eq. (26) when $n=1$ whose inputs are training sample matrix $\mathbf{Z}^0$.

2. Repeat until convergence

2.1 Update $\mathbf{D}_1$ using Eq. (31)

2.2 Update $\mathbf{Z}^1$ using Eq. (32)

From 2^{nd} to penultimate levels

3. Construct $\mathbf{L}^n$ with Eq. (26) whose inputs are previous output $\mathbf{Z}^{n-1}$.

4. Repeat until convergence

4.1 Update $\mathbf{D}_n$ using Eq. (33)

4.2 Update $\mathbf{Z}^n$ using Eq. (34)

For the final level

5. Construct $\mathbf{L}^N$ with Eq. (26) when $n=N$ whose inputs are previous output $\mathbf{Z}^{N-1}$.

6. Repeat until convergence

6.1 Update $\mathbf{D}_N$ using Eq. (35)

6.2 Update $\mathbf{Z}^N$ using Eq. (36)

Output: $\mathbf{D}_1,\mathbf{D}_2,\ldots,\mathbf{D}_N,\ \mathbf{Z}^N$

E. The HGDCDL Classification

Through the HGDCDL training, we can obtain the multi-level dictionaries $\mathbf{D}_1,\mathbf{D}_2,\ldots,\mathbf{D}_N$ and the final representation $\mathbf{Z}^N$ at the N-th layer for all the training samples $\mathbf{Y}$. In the testing phase, the final dictionary used for classification is a product of the multiple dictionaries we learn at different levels during the training phase:

$$\mathbf{D}_{test}=\mathbf{D}_1\mathbf{D}_2\cdots\mathbf{D}_N\ . \tag{37}$$

Thus, for the given testing sample matrix $\mathbf{Y}_{test}$, we can compute the feature representation $\mathbf{Z}_{test}$ using:

$$\min_{\mathbf{Z}_{test}} \|\mathbf{Y}_{test} - \mathbf{D}_{test}\mathbf{Z}_{test}\|_2^2 + \alpha\|\mathbf{Z}_{test}\|_2^2 . \quad (38)$$

Algorithm 2 The Classification Algorithm in HGDCDL

Input: The given testing sample matrix $\mathbf{Y}_{test}$ and the learned $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_N$, $\mathbf{Z}^N$ in Algorithm 1.
1. Collapse multiple levels of dictionaries into a single one using Eq. (37)
2. Compute $\mathbf{Z}_{test}$ using Eq. (38)
3. Compute classification accuracy by inputting $\mathbf{Z}^N$ and $\mathbf{Z}_{test}$ into the classifier.
Output: Classification results

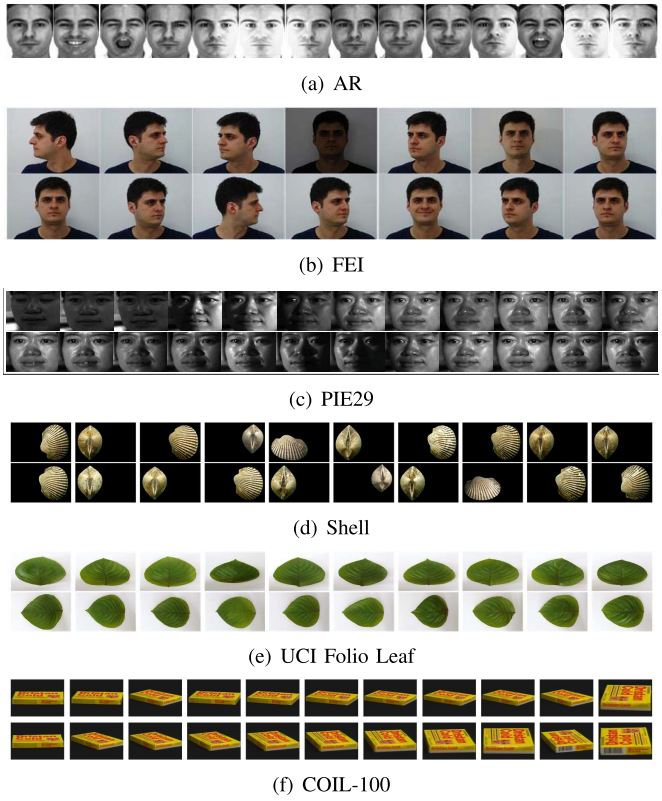
Once obtained $\mathbf{Z}_{test}$, inputting $\mathbf{Z}^N$ of all the training samples and $\mathbf{Z}_{test}$ of the testing samples into the classifier, we can calculate the classification accuracy. For simplification, the non-parametric k-nearest neighbor classifier (KNN), which is always used for deep dictionary learning-based classification, is employed to verify the pattern discrimination in the feature representation space learned by our method. It should be noted that our DCDL method has the same classification procedure above. Algorithm 2 summarizes the classification procedure of our proposed HGDCDL.

IV. Experiments

We conducted a series of experiments on six image data sets to verify the effectiveness and robustness of our HGDCDL. Those data sets are three face data sets including AR [54], FEI [55] and PIE [56] and three non-face data sets including Shell [57], UCI Folio Leaf [58] and COIL-100 [59]. The following experiments are based on three-level HGDCDL. By default, we set the numbers of the dictionary atoms in three dictionaries of the first, second and third layers to 400, 196, 100, respectively. The number of iterations for each layer is set to 20. Note that our DCDL had the same settings. In the classification stage, we chose KNN as the final classifier, and the neighborhood size of the KNN classification was searched from the range of 1 to 30 with step size 1.

A. Data Sets

The three public face data sets are AR, FEI, and PIE29. The AR data set contains approximately 4000 face images from 126 subjects, including 70 men and 56 women. We used a subset of AR with 1400 face images from 50 men and 50 women. Each subject has 14 face images taken from different facial expressions and illumination conditions. The FEI face data set contains 2800 face images from 200 subjects, each of which has 14 samples. All the face images were rotated by 180 degrees in a vertical front position under a white uniform background. The CMU PIE data set contains 41368 images of faces from 68 people. We used a subset of the CMU PIE data set (C29) (abbreviated as PIE29), which contains 24 images of 68 people, totaling 1632 images.



(a) AR (b) FEI (c) PIE29 (d) Shell (e) UCI Folio Leaf (f) COIL-100

Fig. 3. Some or all images samples from one subject on each used data set.

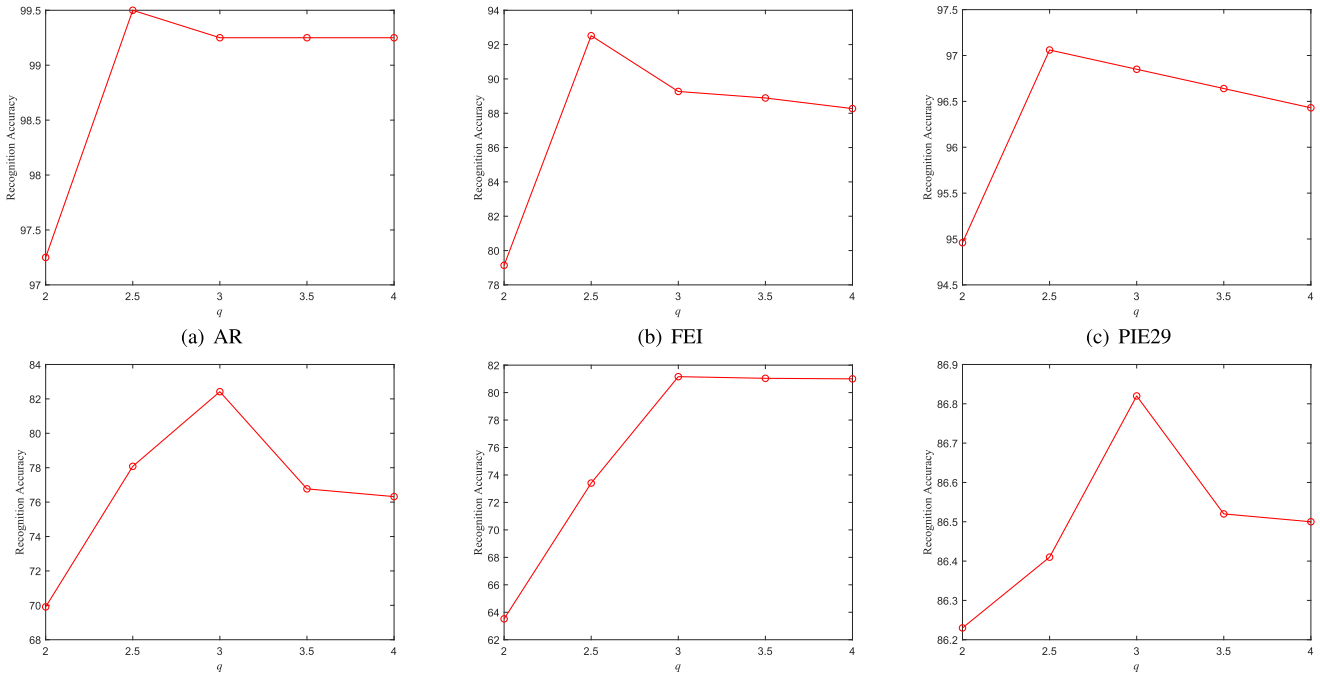
The three non-facial data sets are Shell, UCI Folio Leaf and COIL-100. The Shell data set contains 29622 images of shells from 7894 categories. All images were collected at different angles considering each shell's color, shape and texture. In our experiments, we used a subset of 2680 images from 134 classes, each of which contains 20 images. UCI Folio Leaf (abbreviated as Leaf) is an image data set of different types of leaves, including 32 different leaf types, each of which has 20 images. The COIL-100 data set is composed of different objects collected at different angles in 360 degree rotation. It contains 100 objects, and each object has 72 different postures.

All the images in these data sets were cropped and resized to 28×28 pixels. The number of features is 784. The grey-values of each image have been normalized to [0,1]. It should be noted that the numbers of training samples l per class were set as follows: $l = 8$ for AR, $l = 10$ for FEI, $l = 16$ for PIE29, $l = 14$ for Shell, $l = 14$ for Leaf, and $l = 28$ for COIL-100. As an example, Fig. 3 shows some or all images from one subject on each used data set.

B. Baseline Methods

We considered the following baselines:

- **Representation-based learning methods** include collaborative representation-based classification (CRC) [20] and discriminative sparsity preserving graph embedding (DSPGE) [29]. We fine-tuned the following parameters: the regularization parameter of the representation coefficients for CRC and a regularization parameter of



(a) AR (b) FEI (c) PIE29

(d) Shell (e) Leaf (f) COIL-100

Fig. 4. The classification results (%) of HGDCDL with different values of q on each data set.

adjusting the degree of the geometric distributions of data points for DSPGE.

- **Shallow dictionary learning methods** include D-KSVD [24] and LC-KSVD [25]. D-KSVD has four parameters that are associated with its regularization penalty term, the discriminative term, sparsity and the number of dictionary atoms, respectively. LC-KSVD also has four parameters that are associated with its discrimination coefficient encoding term, the classification error term, sparsity and the number of dictionary atoms, respectively.
- **Deep learning methods** include DBN [52] and SAE [53]. In the two deep models, the number of layers was set to three, and the numbers of hidden units for three layers were set to 400, 196 and 100, respectively.
- **Deep dictionary learning methods** include DDL [39] and DDLCN [44]. DDL has these parameters: the number of layers set to three, the numbers of multi-layer dictionary atoms respectively set to 400, 196 and 100, a regularization parameter of representation coefficient empirically set to be optimal. DDLCN has two parameters: the number of the categorical training dictionary samples and the number of the categorical dictionary atoms of the first layer.

Except for the parameters specified above, the other parameters were set strictly in accordance with the original papers of the compared methods, and they were empirically tuned to reach the optimal performance for fair comparison in the experiments.

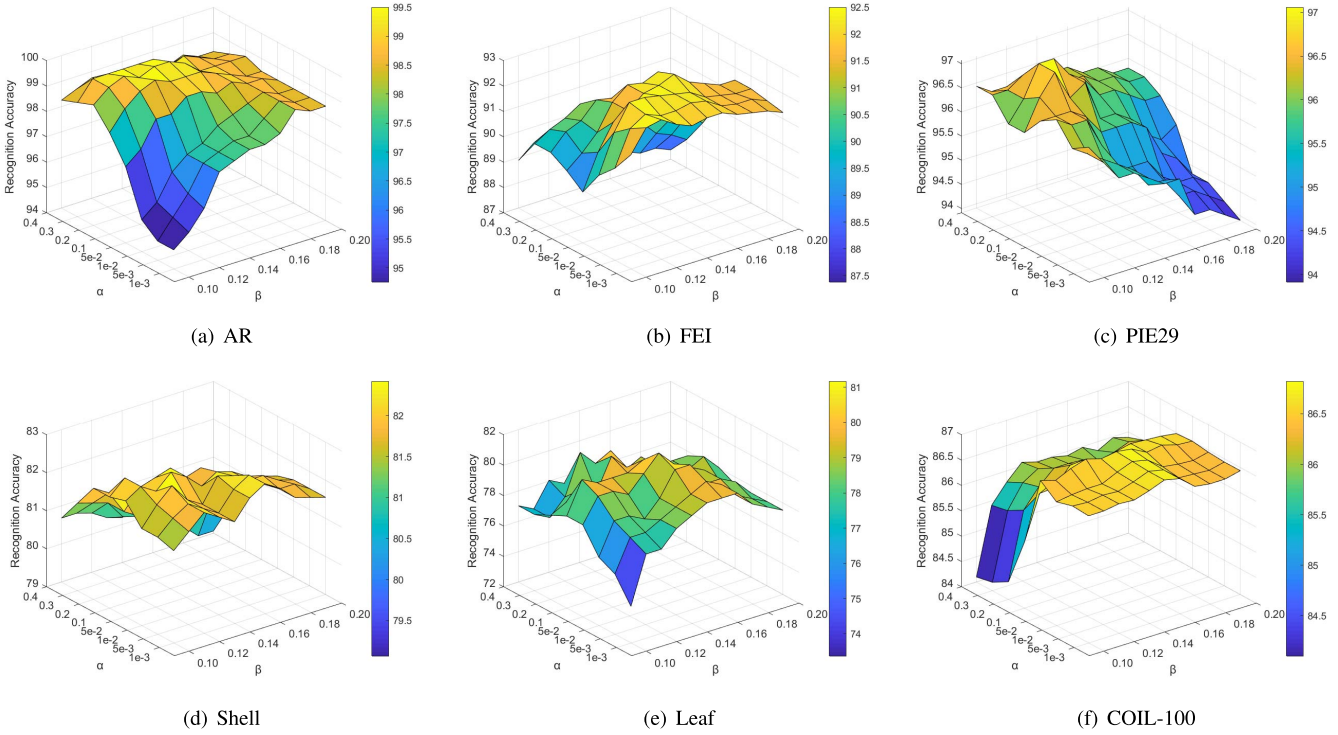
C. Empirical Parameter Settings

We investigated how the performance of the proposed HGDCDL method varies in terms of parameter selection on six data sets. There are three parameters to be determined in our model, which are α, β and q. We set those parameters as follows. The values of q range from 2 to 4 with step size 0.5, the values of α are in $\{1e-3, 5e-3, 1e-2, 5e-2, 0.1, 0.2, 0.3, 0.4\}$, and the values of β are in the range between 0.10 and 0.20 with a step of 0.01. We greedily searched for the optimal parameter settings to be used in the subsequent experiments based on the empirical results derived in this section.

Fig. 4 shows the classification results of HGDCDL by varying the q value and fixing α and β to their optimal values. It is interesting that the three-level HGDCDL always achieves the best classification accuracy across all the six datasets when q is equal to 2.5 or 3, which implies that the optimal q value tuned on one data set could also be applied to the other data sets without losing much performance. Moreover, the experimental results show how the impact of the geometric distributions of the layer-wise representations on the classification performance is controlled by q. Fig. 5 displays the classification results with different values of α and β when q is fixed to its optimal values. The experimental results show that one can easily identify the optimal values of both α and β via greedy search. Table I summarizes the tuned parameter settings of HGDCDL on the six data sets. It should be noted that the parameter α in our DCDL was also searched from the range $\{1e-3, 5e-3, 1e-2, 5e-2, 0.1, 0.2, 0.3, 0.4\}$.

D. Comparative Results for Image Classification

We compared HGDCDL with the representation-based learning methods (i.e., CRC [20] and DSPGE [29]), the traditional shallow dictionary learning methods (i.e., D-KSVD [24] and LC-KSVD [25]), the deep learning methods (i.e., DBN [52] and SAE [53]), and DDL-based methods



(a) AR (b) FEI (c) PIE29

(d) Shell (e) Leaf (f) COIL-100

Fig. 5. The classification results (%) of HGDCDL with different values of α and β on each data set.

TABLE I

PARAMETER SETTINGS ON EACH DATA SET

Datasets	q	α	β
AR	2.5	0.1	0.13
FEI	2.5	0.01	0.12
PIE29	2.5	0.2	0.12
Shell	3	0.05	0.13
UCI Folio Leaf	3	0.05	0.15
COIL-100	3	0.05	0.13

TABLE II

THE CLASSIFICATION ACCURACY RESULTS (%) OF THE COMPETING METHODS ON EACH DATA SET

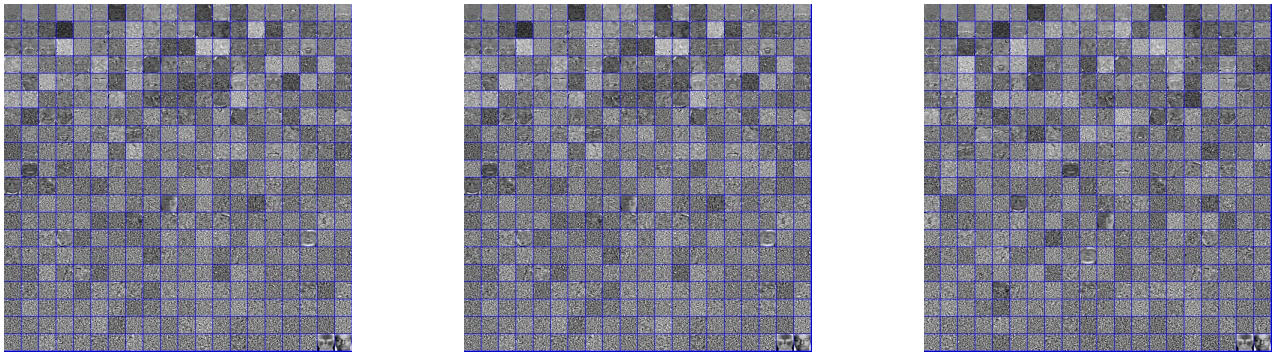
Methods	AR	FEI	PIE29	Shell	Leaf	COIL-100
CRC	98.42	88.38	94.18	61.48	62.15	74.53
DSPGE	98.21	88.75	94.88	73.38	67.40	79.32
D-KSVD	94.43	70.93	94.81	56.59	57.81	76.42
LC-KSVD	96.35	72.05	94.08	62.54	64.11	78.63
DBN	97.72	84.42	94.54	77.16	77.25	82.27
SAE	98.17	88.50	94.96	77.99	71.35	70.66
DDLCN	99.25	80.96	95.42	67.51	73.59	84.67
DDL	97.13	86.45	95.81	72.86	74.94	72.18
DCDL	98.46	87.25	95.85	73.24	77.71	73.42
HGDCDL	**99.50**	**92.50**	**97.06**	**82.42**	**81.17**	**86.82**

(i.e., DDL [39], DDLCN [44] and our DCDL). For each data set, we randomly generated 10 training-testing splits, and reported the average classification accuracy. The average classification accuracy results of all the methods on each data set are shown in Table II.

Compared with representation-based methods such as CRC and DSPGE, and traditional dictionary learning algorithms such as D-KSVD and LC-KSVD, deep dictionary learning-based methods in general perform much better. Our HGDCDL performs best across all the six data sets, and significantly outperforms CRC, DSPGE, D-KSVD and LC-KSVD with a large margin. Meanwhile, our DCDL also obtains the competitive or better classification performance, compared to the other existing competitors. This set of results show that the dictionaries learned with a nested factorization structure consists of abstract but more discriminative feature information, which links directly to the classification performance.

Both SAE and DBN use a three-level deep learning structure, similar to the three-level HGDCDL. The performance gain of our HGDCDL over these two models is substantial, which can be attributed to the HGDCDL's capability of dealing with the few-shot setting, where the number of training samples per class is relatively small. In contrast, SAE and DBN need a large number of training samples in order to achieve decent classification accuracy.

Table II also shows that our HGDCDL outperforms those DDL-based methods. DDLCN, DDL and DCDL ignore the manifold structure of the original data during the multi-level dictionary learning. While HGDCDL integrates the discriminative and geometric structure of data preserved on the basis of DDL, which considers class information and geometric structure of the upper layer, HGDCDL further constructs adjacent graphs based on current input, then updates the dictionary and the representation, and finally takes the learned representation as the input to the next layer, thus improving the relationships between layers. The significant performance differences between the proposed HGDCDL and the other three deep dictionary learning methods (i.e., DDLCN, DDL



(a) The first-level dictionary of DDL (b) The first-level dictionary of DCDL (c) The first-level dictionary of HGDCDL

Fig. 6. The first-level dictionary of the AR data set obtained by DDL, DCDL and HGDCDL.

TABLE III

THE OBJECTIVE FUNCTIONS AT DIFFERENT LEVELS OF LEANING STRUCTURES OF DDL, DCDL AND HGDCDL. NOTE THAT NUMBERS 1, 2, AND 3 INDICATE THE FIRST LAYER, THE SECOND LAYER AND THE THIRD LAYER IN THE TRAINING PHASE RESPECTIVELY, WHILE NUMBERS 1*, 2*, AND 3* RESPECTIVELY REPRESENT THE CORRESPONDING COUNTERPARTS IN THE TESTING PHASE

No.	DDL	DCDL	HGDCDL
1	$\min_{\mathbf{D}_1,\mathbf{Z}^1} \|\mathbf{Z}^0 - \mathbf{D}_1\mathbf{Z}^1\|_F^2$	$\min_{\mathbf{D}_1,\mathbf{Z}^1} \|\mathbf{Z}^0 - \mathbf{D}_1\mathbf{Z}^1\|_F^2$	$\min_{\mathbf{D}_1,\mathbf{Z}^1} \|\mathbf{Z}^0 - \mathbf{D}_1\mathbf{Z}^1\|_F^2 + \beta tr(\mathbf{Z}^1\mathbf{L}^1\mathbf{Z}^{1^T})$
2	$\min_{\mathbf{D}_2,\mathbf{Z}^2} \|\mathbf{Z}^1 - \mathbf{D}_2\mathbf{Z}^2\|_F^2$	$\min_{\mathbf{D}_2,\mathbf{Z}^2} \|\mathbf{Z}^1 - \mathbf{D}_2\mathbf{Z}^2\|_F^2$	$\min_{\mathbf{D}_2,\mathbf{Z}^2} \|\mathbf{Z}^1 - \mathbf{D}_2\mathbf{Z}^2\|_F^2 + \beta tr(\mathbf{Z}^2\mathbf{L}^2\mathbf{Z}^{2^T})$
3	$\min_{\mathbf{D}_3,\mathbf{Z}^3} \|\mathbf{Z}^2 - \mathbf{D}_3\mathbf{Z}^3\|_F^2 + \alpha\|\mathbf{Z}^3\|_1$	$\min_{\mathbf{D}_3,\mathbf{Z}^3} \|\mathbf{Z}^2 - \mathbf{D}_3\mathbf{Z}^3\|_F^2 + \alpha\|\mathbf{Z}^3\|_2^2$	$\min_{\mathbf{D}_3,\mathbf{Z}^3} \|\mathbf{Z}^2 - \mathbf{D}_3\mathbf{Z}^3\|_F^2 + \alpha\|\mathbf{Z}^3\|_2^2 + \beta tr(\mathbf{Z}^3\mathbf{L}^3\mathbf{Z}^{3^T})$
1*	$\mathbf{Z}^1_{test} = \min_{\mathbf{Z}^1_{test}} \|\mathbf{Y}_{test} - \mathbf{D}_1\mathbf{Z}^1_{test}\|_2^2 + \alpha\|\mathbf{Z}^1_{test}\|_1$	$\mathbf{Z}^1_{test} = \min_{\mathbf{Z}^1_{test}} \|\mathbf{Y}_{test} - \mathbf{D}_1\mathbf{Z}^1_{test}\|_2^2 + \alpha\|\mathbf{Z}^1_{test}\|_2^2$	$\mathbf{Z}^1_{test} = \min_{\mathbf{Z}^1_{test}} \|\mathbf{Y}_{test} - \mathbf{D}_1\mathbf{Z}^1_{test}\|_2^2 + \alpha\|\mathbf{Z}^1_{test}\|_2^2$
2*	$\mathbf{Z}^2_{test} = \min_{\mathbf{Z}^2_{test}} \|\mathbf{Y}_{test} - \mathbf{D}_1\mathbf{D}_2\mathbf{Z}^2_{test}\|_2^2 + \alpha\|\mathbf{Z}^2_{test}\|_1$	$\mathbf{Z}^2_{test} = \min_{\mathbf{Z}^2_{test}} \|\mathbf{Y}_{test} - \mathbf{D}_1\mathbf{D}_2\mathbf{Z}^2_{test}\|_2^2 + \alpha\|\mathbf{Z}^2_{test}\|_2^2$	$\mathbf{Z}^2_{test} = \min_{\mathbf{Z}^2_{test}} \|\mathbf{Y}_{test} - \mathbf{D}_1\mathbf{D}_2\mathbf{Z}^2_{test}\|_2^2 + \alpha\|\mathbf{Z}^2_{test}\|_2^2$
3*	$\mathbf{Z}^3_{test} = \min_{\mathbf{Z}^3_{test}} \|\mathbf{Y}_{test} - \mathbf{D}_1\mathbf{D}_2\mathbf{D}_3\mathbf{Z}^3_{test}\|_2^2 + \alpha\|\mathbf{Z}^3_{test}\|_1$	$\mathbf{Z}^3_{test} = \min_{\mathbf{Z}^3_{test}} \|\mathbf{Y}_{test} - \mathbf{D}_1\mathbf{D}_2\mathbf{D}_3\mathbf{Z}^3_{test}\|_2^2 + \alpha\|\mathbf{Z}^3_{test}\|_2^2$	$\mathbf{Z}^3_{test} = \min_{\mathbf{Z}^3_{test}} \|\mathbf{Y}_{test} - \mathbf{D}_1\mathbf{D}_2\mathbf{D}_3\mathbf{Z}^3_{test}\|_2^2 + \alpha\|\mathbf{Z}^3_{test}\|_2^2$

and DCDL) demonstrate the advantage of using hierarchical adjacent graphs in regularizing the learned representations during the multi-level dictionary learning. Therefore, it can be concluded that our proposed HGDCDL can learn informative dictionaries and discriminative representations for favorable classification.

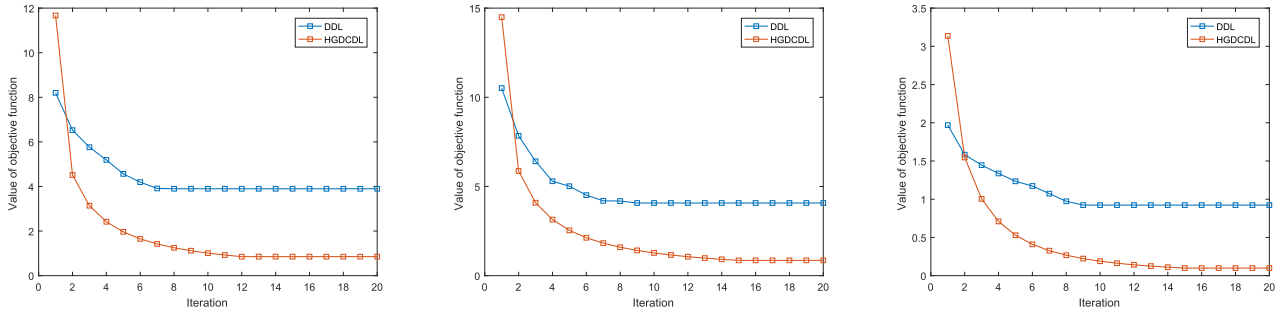
E. Multi-Level Dictionary Analysis

We further analyzed the multi-level dictionary structures of DDL, DCDL and HGDCDL. Here we took the three-level structures of DDL, DCDL and HGDCDL on the AR data set as an example. Firstly, we visualized the first-level dictionary of DDL, DCDL and HGDCDL, respectively. Then, we made classification decision based on the first-level representation $\mathbf{Z}^1_{test}$, the second-level representation $\mathbf{Z}^2_{test}$, and the third-level representation $\mathbf{Z}^3_{test}$ generated by DDL, DCDL and HGDCDL. Experimental results show that the deeper representations the better classification performance. Table III summarizes the optimization training and classification problems at different levels of learning structures of DDL, DCDL and HGDCDL.

For the first-level dictionary, we used a deterministic initialization based on the QR decomposition of the training data matrix $\mathbf{Z}^0$ to initialize $\mathbf{D}_1$. The second-level dictionary $\mathbf{D}_2$ and the third-level dictionary $\mathbf{D}_3$ were randomly initialized instead. Then, we transferred the divided training set and initialized three-level dictionaries $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}_3$ into three-level DDL, DCDL and HGDCDL models respectively, and trained each layer iteratively until convergence. Finally, we obtained the trained three-level dictionaries of DDL, DCDL and HGDCDL and the corresponding representations of each layer.

Fig. 6 shows the first-level dictionary of DDL, DCDL and HGDCDL, respectively. Most dictionary atoms have learned the key features of the faces in the AR data set. For example, some dictionary atoms have learned the contour of the faces, some have learned the mouth, and some have learned the eyes and other features. Also, some dictionary atoms have learned more abstract features. At the same time, we can find that the first-level dictionary of DDL and DCDL are identical. The reason is that the first-level dictionary and the corresponding first-level representation of both DDL and DCDL are the same, as shown in Table III. Compared with the first-level dictionary of DDL and DCDL, some dictionary atoms in the first-level dictionary of HGDCDL have obviously changed. Through our experiments in Section IV-D, using the dictionary atoms of HGDCDL can learn more discriminative representations for better classification. Thus, the dictionary atoms learned by our HGDCDL are more informative.

Table IV shows the classification results of the first-level representation $\mathbf{Z}^1_{test}$, the second-level representation $\mathbf{Z}^2_{test}$ and the third-level representation $\mathbf{Z}^3_{test}$ for the AR testing set learned by DDL, DCDL and HGDCDL, respectively. The



(a) FEI (b) PIE29 (c) Leaf

Fig. 7. Convergence of DDL and HGDCDL on FEI, PIE29 and Leaf.

TABLE IV

THE CLASSIFICATION RESULTS (%) OF $\mathbf{Z}^1_{test}$, $\mathbf{Z}^2_{test}$ AND $\mathbf{Z}^3_{test}$ ON THE AR DATA SET OBTAINED BY DDL, DCDL AND HGDCDL, RESPECTIVELY

Methods	$\mathbf{Z}^1_{test}$	$\mathbf{Z}^2_{test}$	$\mathbf{Z}^3_{test}$
DDL	96.54	97.01	97.13
DCDL	97.75	98.23	98.46
HGDCDL	**98.71**	**99.00**	**99.50**

results show that the deeper the level the more discriminative the learned representation, and thus $\mathbf{Z}^3_{test}$ has higher classification accuracy compared with $\mathbf{Z}^1_{test}$ and $\mathbf{Z}^2_{test}$. Compared with DDL, DCDL imposes collaborative representation constraint on the deepest representation rather than sparse representation constraint, accordingly DCDL also revises the corresponding classification model in the testing stage, so that the deepest dictionary $\mathbf{D}_3$ contains rich information and the corresponding representation $\mathbf{Z}^3_{test}$ is more discriminative. HGDCDL introduces hierarchical graph constraints on the basis of DCDL, which makes the representations more discriminative, so as to achieve better classification results.

F. Convergence Analysis

In this section, we further analyzed and compared the convergence properties of DDL and HGDCDL on the FEI, PIE29 and Leaf data sets respectively as an example. The objective functions of DDL and HGDCDL are shown in Eq. (5) and Eq. (28), respectively.

We fixed the three-level initialization dictionary $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}_3$ and preset the number of iterations for each layer to be 20. We plotted the values of the corresponding losses as a function of iterations in Fig. 7. We have the following observations: On the FEI data set, the objective function value of DDL algorithm remains unchanged after eight iterations, and the value of HGDCDL algorithm becomes stable after twelve iterations. On the PIE29 data set, DDL algorithm achieves the stability of the objective function after nine iterations, and HGDCDL algorithm achieves the stability of the objective function value after fifteen iterations. On the Leaf data set, the value of DDL's objective function remains unchanged after nine iterations, but the result of HGDCDL reaches the best after fifteen iterations. Overall, our algorithm converges better than DDL algorithm.

V. ABLATION STUDIES

A. Hierarchical Graph Ablations

This experiment aims to explore the effectiveness of hierarchical graph constraints. We introduce seven variants of HGDCDL. Table V shows the objective functions of different variants of HGDCDL, which include

- **HGDCDL**$_1$ only explores the geometric and discriminative structure of the original data at the first-level.
- **HGDCDL**$_2$ only explores the geometric and discriminative structure of the first-level representation at the second-level.
- **HGDCDL**$_3$ only explores the geometric and discriminative structure of the second-level representation at the third-level.
- **HGDCDL**$_{12}$ explores the geometric and discriminative structures of the original data at the first-level and of the first-level representation at the second-level.
- **HGDCDL**$_{13}$ explores the geometric and discriminative structures of the original data at the first-level and of the second-level representation at the third-level.
- **HGDCDL**$_{23}$ explores the geometric and discriminative structures of the first-level representation at the second-level and of the second-level representation at the third-level
- **HGDCDL**$_4$ only explores the geometric and discriminative structure of the original data at the first-level, second-level and third-level.

Table VI reports the classification accuracy of HGDCDL and its variants on different data sets. In the three-level collaborative dictionary learning, when we only consider geometric and discriminative structure of the original data or of the first-level representation or of the second-level representation, such as HGDCDL$_1$, HGDCDL$_2$ and HGDCDL$_3$, the classification accuracy is not very high. But when considering geometric and discriminative structures of the original data and the first-level representation or of the original data and the second-level representation or of the first-level and the second-level representations at the same time, such as HGDCDL$_{12}$, HGDCDL$_{13}$ and HGDCDL$_{23}$, we can obtain a higher recognition accuracy. Our HGDCDL takes into account geometric and discriminative structures of the original data, the first-level and the second-level representations, a higher recognition accuracy than multiple variants of HGDCDL is acquired.

Furthermore, HGDCDL$_4$ always performs better than the variants with the graph constraints at one level or two levels,

TABLE V
THE OBJECTIVE FUNCTIONS OF DIFFERENT HGDCDL VARIANTS AT EACH LAYER

Variants	The first-level	The second-level	The third-level
HGDCDL_1	$\min_{\mathbf{D}_1,\mathbf{Z}^1} \|\mathbf{Z}^0-\mathbf{D}_1\mathbf{Z}^1\|_F^2+\beta tr(\mathbf{Z}^1\mathbf{L}^1\mathbf{Z}^{1^T})$	$\min_{\mathbf{D}_2,\mathbf{Z}^2} \|\mathbf{Z}^1-\mathbf{D}_2\mathbf{Z}^2\|_F^2$	$\min_{\mathbf{D}_3,\mathbf{Z}^3} \|\mathbf{Z}^2-\mathbf{D}_3\mathbf{Z}^3\|_F^2+\alpha\|\mathbf{Z}^3\|_2^2$
HGDCDL_2	$\min_{\mathbf{D}_1,\mathbf{Z}^1} \|\mathbf{Z}^0-\mathbf{D}_1\mathbf{Z}^1\|_F^2$	$\min_{\mathbf{D}_2,\mathbf{Z}^2} \|\mathbf{Z}^1-\mathbf{D}_2\mathbf{Z}^2\|_F^2+\beta tr(\mathbf{Z}^2\mathbf{L}^2\mathbf{Z}^{2^T})$	$\min_{\mathbf{D}_3,\mathbf{Z}^3} \|\mathbf{Z}^2-\mathbf{D}_3\mathbf{Z}^3\|_F^2+\alpha\|\mathbf{Z}^3\|_2^2$
HGDCDL_3	$\min_{\mathbf{D}_1,\mathbf{Z}^1} \|\mathbf{Z}^0-\mathbf{D}_1\mathbf{Z}^1\|_F^2$	$\min_{\mathbf{D}_2,\mathbf{Z}^2} \|\mathbf{Z}^1-\mathbf{D}_2\mathbf{Z}^2\|_F^2$	$\min_{\mathbf{D}_3,\mathbf{Z}^3} \|\mathbf{Z}^2-\mathbf{D}_3\mathbf{Z}^3\|_F^2+\alpha\|\mathbf{Z}^3\|_2^2+\beta tr(\mathbf{Z}^3\mathbf{L}^3\mathbf{Z}^{3^T})$
HGDCDL_{12}	$\min_{\mathbf{D}_1,\mathbf{Z}^1} \|\mathbf{Z}^0-\mathbf{D}_1\mathbf{Z}^1\|_F^2+\beta tr(\mathbf{Z}^1\mathbf{L}^1\mathbf{Z}^{1^T})$	$\min_{\mathbf{D}_2,\mathbf{Z}^2} \|\mathbf{Z}^1-\mathbf{D}_2\mathbf{Z}^2\|_F^2+\beta tr(\mathbf{Z}^2\mathbf{L}^2\mathbf{Z}^{2^T})$	$\min_{\mathbf{D}_3,\mathbf{Z}^3} \|\mathbf{Z}^2-\mathbf{D}_3\mathbf{Z}^3\|_F^2+\alpha\|\mathbf{Z}^3\|_2^2$
HGDCDL_{13}	$\min_{\mathbf{D}_1,\mathbf{Z}^1} \|\mathbf{Z}^0-\mathbf{D}_1\mathbf{Z}^1\|_F^2+\beta tr(\mathbf{Z}^1\mathbf{L}^1\mathbf{Z}^{1^T})$	$\min_{\mathbf{D}_2,\mathbf{Z}^2} \|\mathbf{Z}^1-\mathbf{D}_2\mathbf{Z}^2\|_F^2$	$\min_{\mathbf{D}_3,\mathbf{Z}^3} \|\mathbf{Z}^2-\mathbf{D}_3\mathbf{Z}^3\|_F^2+\alpha\|\mathbf{Z}^3\|_2^2+\beta tr(\mathbf{Z}^3\mathbf{L}^3\mathbf{Z}^{3^T})$
HGDCDL_{23}	$\min_{\mathbf{D}_1,\mathbf{Z}^1} \|\mathbf{Z}^0-\mathbf{D}_1\mathbf{Z}^1\|_F^2$	$\min_{\mathbf{D}_2,\mathbf{Z}^2} \|\mathbf{Z}^1-\mathbf{D}_2\mathbf{Z}^2\|_F^2+\beta tr(\mathbf{Z}^2\mathbf{L}^2\mathbf{Z}^{2^T})$	$\min_{\mathbf{D}_3,\mathbf{Z}^3} \|\mathbf{Z}^2-\mathbf{D}_3\mathbf{Z}^3\|_F^2+\alpha\|\mathbf{Z}^3\|_2^2+\beta tr(\mathbf{Z}^3\mathbf{L}^3\mathbf{Z}^{3^T})$
HGDCDL_4	$\min_{\mathbf{D}_1,\mathbf{Z}^1} \|\mathbf{Z}^0-\mathbf{D}_1\mathbf{Z}^1\|_F^2+\beta tr(\mathbf{Z}^1\mathbf{L}^1\mathbf{Z}^{1^T})$	$\min_{\mathbf{D}_2,\mathbf{Z}^2} \|\mathbf{Z}^1-\mathbf{D}_2\mathbf{Z}^2\|_F^2+\beta tr(\mathbf{Z}^2\mathbf{L}^1\mathbf{Z}^{2^T})$	$\min_{\mathbf{D}_3,\mathbf{Z}^3} \|\mathbf{Z}^2-\mathbf{D}_3\mathbf{Z}^3\|_F^2+\alpha\|\mathbf{Z}^3\|_2^2+\beta tr(\mathbf{Z}^3\mathbf{L}^1\mathbf{Z}^{3^T})$
HGDCDL	$\min_{\mathbf{D}_1,\mathbf{Z}^1} \|\mathbf{Z}^0-\mathbf{D}_1\mathbf{Z}^1\|_F^2+\beta tr(\mathbf{Z}^1\mathbf{L}^1\mathbf{Z}^{1^T})$	$\min_{\mathbf{D}_2,\mathbf{Z}^2} \|\mathbf{Z}^1-\mathbf{D}_2\mathbf{Z}^2\|_F^2+\beta tr(\mathbf{Z}^2\mathbf{L}^2\mathbf{Z}^{2^T})$	$\min_{\mathbf{D}_3,\mathbf{Z}^3} \|\mathbf{Z}^2-\mathbf{D}_3\mathbf{Z}^3\|_F^2+\alpha\|\mathbf{Z}^3\|_2^2+\beta tr(\mathbf{Z}^3\mathbf{L}^3\mathbf{Z}^{3^T})$

TABLE VI
THE CLASSIFICATION ACCURACY RESULTS (%) OF HGDCDL AND ITS VARIANTS ON EACH DATA SET

Variants	AR	FEI	PIE29	Shell	Leaf	COIL-100
HGDCDL_1	91.25	83.63	93.08	64.55	68.96	81.48
HGDCDL_2	91.04	84.16	94.18	69.65	64.15	83.52
HGDCDL_3	90.36	83.50	94.39	67.41	67.72	82.61
HGDCDL_{12}	97.00	90.75	94.54	75.89	76.58	84.55
HGDCDL_{13}	96.25	89.88	95.59	75.14	74.41	83.19
HGDCDL_{23}	96.50	87.50	96.84	73.66	76.83	83.64
HGDCDL_4	98.50	90.63	94.76	73.31	75.17	83.76
HGDCDL	**99.50**	**92.50**	**97.06**	**82.42**	**81.17**	**86.82**

but it performs less than the proposed HGDCDL. The graph constraints of HGDCDL_4 used in each layer are based on original samples, that is, the original samples are first used for graph construction, the class information and geometric structure of original samples are always kept unchanged during the following dictionary learning process. Although HGDCDL_4 always keeps class information and geometric structure of original samples at each layer, the relationships between layers are ignored in the process of multi-level dictionary learning. The original data usually contains noises and outliers in practical applications, results in inaccuracy adjacent graphs, and ultimately leads to performance degradation. HGDCDL makes full use of the geometric and discriminative structure of the original data and its representations between different layers in the process of hierarchical graph constraints, so as to obtain a higher recognition accuracy than HGDCDL_4.

B. Graph Construction Ablations

In the proposed HGDCDL framework, we use the graph construction method designed in DSPGE [29], and the adjacent graphs change dynamically with the current input at each layer. However, our framework can adopt different graph construction methods including unsupervised and supervised ones. In this experiment, we compared the effects of different Laplacian graph construction methods on the final recognition accuracy of HGDCDL on six data sets.

We introduce several variants of HGDCDL from the perspective of graph constructions. Among these variants, we only change the graph construction methods in HGDCDL, and the rest remains unchanged.

HGDCDL-1: It replaces the Laplacian matrix $\mathbf{L}^n$ with $\mathbf{L}_1^n$, the corresponding weight matrix in [27] is defined as:

$$w_{ij}^n=\begin{cases}1, & \mathbf{z}_j^{n-1}\in K(\mathbf{z}_i^{n-1})\ or\ \mathbf{z}_i^{n-1}\in K(\mathbf{z}_j^{n-1})\\ 0, & otherwise\ ,\end{cases} \tag{39}$$

where $K(\mathbf{z}_i^{n-1})$ represents the set of k nearest neighbours of $\mathbf{z}_i^{n-1}$. Then, the Laplacian matrix is calculated as $\mathbf{L}_1^n=\mathbf{H}_1^n-\mathbf{W}_1^n$ where $(\mathbf{H}_1^n)_{ii}=\sum_j w_{ij}^n$ and $(\mathbf{W}_1^n)_{ij}=w_{ij}^n$. Note that the Laplacian matrices $\mathbf{L}_g^n$ $(g=2,3,4)$ in what follows are calculated as $\mathbf{L}_1^n$.

HGDCDL-2: It replaces the Laplacian matrix $\mathbf{L}^n$ with $\mathbf{L}_2^n$, the corresponding weight matrix in [27] is defined as follows:

$$w_{ij}^n=\begin{cases}\exp\left(-\frac{\|\mathbf{z}_i^{n-1}-\mathbf{z}_j^{n-1}\|^2}{t}\right), & \mathbf{z}_j^{n-1}\in K(\mathbf{z}_i^{n-1})\ or\\ & \mathbf{z}_i^{n-1}\in K(\mathbf{z}_j^{n-1})\\ 0, & otherwise\end{cases} \tag{40}$$

where t is an adjusted parameter.

HGDCDL-3: It replaces the Laplacian matrix $\mathbf{L}^n$ with $\mathbf{L}_3^n$, the corresponding weight matrix in [30] can be in (41), as shown at the bottom of the next page, where

$$w_{ij}^{n+}=\begin{cases}\exp(-\frac{\|\mathbf{z}_i^{n-1}-\mathbf{z}_j^{n-1}\|^2}{\rho_i^k})\Big(1+\exp\big(-\frac{\|\mathbf{z}_i^{n-1}-\mathbf{z}_j^{n-1}\|^2}{\rho_i^k}\big)\Big), & \\ \qquad \mathbf{z}_j^{n-1}\in K(\mathbf{z}_i^{n-1})\ and\ c_i=c_j & \\ 0, \qquad otherwise\ , & \end{cases} \tag{42}$$

$$w_{ji}^{n+}=\begin{cases}\exp(-\frac{\|\mathbf{z}_j^{n-1}-\mathbf{z}_i^{n-1}\|^2}{\rho_j^k})\Big(1+\exp\big(-\frac{\|\mathbf{z}_j^{n-1}-\mathbf{z}_i^{n-1}\|^2}{\rho_j^k}\big)\Big), & \\ \qquad \mathbf{z}_i^{n-1}\in K(\mathbf{z}_j^{n-1})\ and\ c_j=c_i & \\ 0, \qquad otherwise\ , & \end{cases} \tag{43}$$

$$w_{ij}^{n-}=\begin{cases}\exp(-\frac{\|\mathbf{z}_i^{n-1}-\mathbf{z}_j^{n-1}\|^2}{\rho_i^k})\Big(1-\exp\big(-\frac{\|\mathbf{z}_i^{n-1}-\mathbf{z}_j^{n-1}\|^2}{\rho_i^k}\big)\Big), & \\ \qquad \mathbf{z}_j^{n-1}\in K(\mathbf{z}_i^{n-1})\ and\ c_i\neq c_j & \\ 0, \qquad otherwise\ , & \end{cases} \tag{44}$$

$$w_{ji}^{n-}=\begin{cases}\exp(-\frac{\|\mathbf{z}_j^{n-1}-\mathbf{z}_i^{n-1}\|^2}{\rho_j^k})\Big(1-\exp\big(-\frac{\left\|\mathbf{z}_j^{n-1}-\mathbf{z}_i^{n-1}\right\|^2}{\rho_j^k}\big)\Big), \\ \qquad\qquad \mathbf{z}_i^{n-1}\in K(\mathbf{z}_j^{n-1})\ and\ c_j\neq c_i \\ 0, \qquad\qquad otherwise\ . \end{cases} \tag{45}$$

ρ_i^k and ρ_j^k are the parameters that represent the local geometry distributions around $\mathbf{z}_i^{n-1}$ and $\mathbf{z}_j^{n-1}$, respectively.

HGDCDL-4: It replaces the Laplacian matrix $\mathbf{L}^n$ with $\mathbf{L}_4^n$, the corresponding weight matrix in [30] can be written as:

$$w_{ij}^n=\begin{cases}\frac{1}{2}\big(w_{ij}^{n+}+w_{ji}^{n+}\big), & c_i=c_j \\ -\frac{1}{2}(w_{ij}^{n-}+w_{ji}^{n-}), & c_i\neq c_j,\end{cases} \tag{46}$$

where

$$w_{ij}^{n+}=\begin{cases}\frac{1}{2}\exp(-\frac{\|\mathbf{z}_i^{n-1}-\mathbf{z}_j^{n-1}\|^2}{\rho_i^+})\Big(1+\exp\big(-\frac{\|\mathbf{z}_i^{n-1}-\mathbf{z}_j^{n-1}\|^2}{\rho_i^+}\big)\Big) \\ \qquad\qquad c_i=c_j \\ 0, \qquad\qquad otherwise\ , \end{cases} \tag{47}$$

$$w_{ji}^{n+}=\begin{cases}\frac{1}{2}\exp(-\frac{\|\mathbf{z}_j^{n-1}-\mathbf{z}_i^{n-1}\|^2}{\rho_j^+})\Big(1+\exp\big(-\frac{\|\mathbf{z}_j^{n-1}-\mathbf{z}_i^{n-1}\|^2}{\rho_j^+}\big)\Big) \\ \qquad\qquad c_i=c_j \\ 0, \qquad\qquad otherwise\ , \end{cases} \tag{48}$$

$$w_{ij}^{n-}=\begin{cases}\frac{1}{2}\exp(-\frac{\|\mathbf{z}_i^{n-1}-\mathbf{z}_j^{n-1}\|^2}{\rho_i^-})\Big(1-\exp\big(-\frac{\|\mathbf{z}_i^{n-1}-\mathbf{z}_j^{n-1}\|^2}{\rho_i^-}\big)\Big) \\ \qquad\qquad c_i\neq c_j \\ 0, \qquad\qquad otherwise\ , \end{cases} \tag{49}$$

and

$$w_{ji}^{n-}=\begin{cases}\frac{1}{2}\exp(-\frac{\|\mathbf{z}_j^{n-1}-\mathbf{z}_i^{n-1}\|^2}{\rho_j^-})\Big(1-\exp\big(-\frac{\|\mathbf{z}_j^{n-1}-\mathbf{z}_i^{n-1}\|^2}{\rho_j^-}\big)\Big) \\ \qquad\qquad c_i\neq c_j \\ 0, \qquad\qquad otherwise\ . \end{cases} \tag{50}$$

ρ_i^{n+} and ρ_j^{n+} are the parameters that represent the intra-class geometric distributions around $\mathbf{z}_i^{n-1}$ and $\mathbf{z}_j^{n-1}$, respectively. And ρ_i^{n-} and ρ_j^{n-} are the parameters that represent the inter-class geometric distributions around $\mathbf{z}_i^{n-1}$ and $\mathbf{z}_j^{n-1}$, respectively.

Table VII reports the classification accuracy of HGDCDL and its variants with different graph constructions on different data sets. Compared to the classification results of the competitors in Table II, the HGDCDL variants with different graph constructions have achieved the competitive or better performance. This verifies the advantage of our proposed HGDCDL framework using hierarchical adjacent graphs for deep dictionary learning. Among these graph constructions, our HGDCDL with the graph construction method in Section III-B performs best. Thus, our strategy of hierarchical graph-augmented deep dictionary learning is effective for representation and classification.

TABLE VII
THE CLASSIFICATION ACCURACY RESULTS (%) OF HGDCDL AND ITS VARIANTS WITH DIFFERENT GRAPH CONSTRUCTIONS ON EACH DATA SET

Variants	AR	FEI	PIE29	Shell	Leaf	COIL-100
HGDCDL-1	96.75	86.50	94.33	71.52	61.56	71.06
HGDCDL-2	95.14	88.38	93.91	67.47	71.07	70.43
HGDCDL-3	97.01	89.88	96.64	75.75	77.81	76.20
HGDCDL-4	98.43	90.38	96.04	78.39	73.17	80.36
HGDCDL	**99.50**	**92.50**	**97.06**	**82.42**	**81.17**	**86.82**

VI. CONCLUSION

In this paper, we have proposed a novel hierarchical graph augmented deep collaborative dictionary learning (HGDCDL) model for image classification. The proposed HGDCDL combines collaborative representation learning with the constraints originated from the hierarchical graph augmented structure into DDL. We use l_2-norm instead of l_1-norm to constrain the deepest-level representation to design the deep collaborative dictionary learning (DCDL). In the training stage, we first construct the hierarchical adjacent graphs, which are able to reflect the discriminative and geometric information derived from both the original data and the multi-level data representations. Then, the graph regularized deep dictionary learning is used to generate dictionaries and representations at different abstract levels. The representation of the current layer is used as the input of the next layer, so as to get more informative dictionaries and more discriminative representation of the deepest level. In the classification stage, when the testing samples are input, we multiply the multi-level dictionaries learned in the training phase, and obtain the deepest feature representation for image classification.

We can use both supervised and unsupervised graph construction methods within our framework, the adjacent graphs change dynamically based upon the current input, rather than being static. Not only can our method get more informative dictionaries and more discriminative representations, but also strengthen the connection between layers. To verify the effectiveness of the proposed HGDCDL, we conducted

$$w_{ij}^n=\begin{cases}+\frac{1}{2}(w_{ij}^{n+}+w_{ji}^{n+}),\ \mathbf{z}_j^{n-1}\in K(\mathbf{z}_i^{n-1})\ or\ \mathbf{z}_i^{n-1}\in K(\mathbf{z}_j^{n-1}) \\ \qquad\qquad and\ c_i=c_j \\ -\frac{1}{2}(w_{ij}^{n-}+w_{ji}^{n-}),\ \mathbf{z}_j^{n-1}\in K(\mathbf{z}_i^{n-1})\ or\ \mathbf{z}_i^{n-1}\in K(\mathbf{z}_j^{n-1}) \\ \qquad\qquad and\ c_i\neq c_j \\ 0,\ otherwise, \end{cases} \tag{41}$$

extensive experiments on six image data sets. The experimental results and ablations have verified that the proposed method outperforms the state-of-the-art representation learning methods. In future works, we plan to apply HGDCDL on more visual recognition tasks, such as hyper-spectral recognition and object detection.

REFERENCES

[1] M. Liao and X. Gu, "Face recognition based on dictionary learning and subspace learning," *Digit. Signal Process.*, vol. 90, pp. 110–124, Jul. 2019.

[2] Z. Zhao, G. Feng, L. Zhang, J. Zhu, and Q. Shen, "Novel orthogonal based collaborative dictionary learning for efficient face recognition," *Knowl.-Based Syst.*, vol. 163, pp. 533–545, Jan. 2019.

[3] G. Zhang, F. Porikli, H. Sun, Q. Sun, G. Xia, and Y. Zheng, "Cost-sensitive joint feature and dictionary learning for face recognition," *Neurocomputing*, vol. 391, pp. 177–188, May 2020.

[4] X. Wang and J. Ma, "Adaptive dictionary learning for blind seismic data denoising," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 7, pp. 1273–1277, Jul. 2020.

[5] X. Gong, W. Chen, and J. Chen, "A low-rank tensor dictionary learning method for hyperspectral image denoising," *IEEE Trans. Signal Process.*, vol. 68, pp. 1168–1180, 2020.

[6] C.-Z. You, Z.-Q. Shu, and H.-H. Fan, "Low-rank sparse subspace clustering with a clean dictionary," *J. Algorithms Comput. Technol.*, vol. 15, Jan. 2021, Art. no. 174830262098369.

[7] J. Bruton and H. Wang, "Dictionary learning for clustering on hyperspectral images," *Signal, Image Video Process.*, vol. 15, no. 2, pp. 255–261, Mar. 2021.

[8] J. Wang, W. Xu, J.-F. Cai, Q. Zhu, Y. Shi, and B. Yin, "Multi-direction dictionary learning based depth map super-resolution with autoregressive modeling," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1470–1484, Jun. 2020.

[9] P. Song, X. Deng, J. F. C. Mota, N. Deligiannis, P. L. Dragotti, and M. R. D. Rodrigues, "Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 57–72, 2020.

[10] G. Gao, D. Zhu, H. Lu, Y. Yu, H. Chang, and D. Yue, "Robust facial image super-resolution by kernel locality-constrained coupled-layer regression," *ACM Trans. Internet Technol.*, vol. 21, no. 3, pp. 1–15, Jun. 2021.

[11] T. Guo, F. Luo, L. Zhang, B. Zhang, X. Tan, and X. Zhou, "Learning structurally incoherent background and target dictionaries for hyperspectral target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3521–3533, 2020.

[12] S. LI, J. Q. Huang, X. S. Hua, and L. Zhang, "Category dictionary guided unsupervised domain adaptation for object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, 2021, pp. 1949–1957.

[13] X. Li, Q. Li, W. Wang, and L. Guo, "An unsupervised multi-shot person re-identification method via mutual normalized sparse representation and stepwise learning," *IEEE Trans. Intell. Transp. Syst.*, early access, Apr. 28, 2021, doi: 10.1109/TITS.2021.3073936.

[14] Y. Zhang, Y. Li, R. Wang, M. S. Hossain, and H. Lu, "Multi-aspect aware session-based recommendation for intelligent transportation services," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4696–4705, Jul. 2021.

[15] C. Liu, F. Chang, Z. Chen, and D. Liu, "Fast traffic sign recognition via high-contrast region extraction and extended sparse representation," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 79–92, Jan. 2016.

[16] Z. Yao and W. Yi, "Bionic vision system and its application in license plate recognition," *Natural Comput.*, vol. 19, no. 1, pp. 199–209, Mar. 2020.

[17] Z. Chen *et al.*, "Vehicle detection in high-resolution aerial images based on fast sparse representation classification and multiorder feature," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 8, pp. 2296–2309, Aug. 2016.

[18] C.-Y. Chiou, W.-C. Wang, S.-C. Lu, C.-R. Huang, P.-C. Chung, and Y.-Y. Lai, "Driver monitoring using sparse representation with part-based temporal face descriptors," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 346–361, Jan. 2020.

[19] 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.

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

[21] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[22] J. Ren *et al.*, "Learning hybrid representation by robust dictionary learning in factorized compressed space," *IEEE Trans. Image Process.*, vol. 29, pp. 3941–3956, 2020.

[23] J. Fan, C. Yang, and M. Udell, "Robust non-linear matrix factorization for dictionary learning, denoising, and clustering," *IEEE Trans. Signal Process.*, vol. 69, pp. 1755–1770, 2021.

[24] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2691–2698.

[25] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 1697–1704.

[26] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based Fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 209–232, Sep. 2014.

[27] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16, 2003, pp. 153–160.

[28] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognit.*, vol. 43, no. 1, pp. 331–341, 2010.

[29] J. Gou, L. Du, K. Cheng, and Y. Cai, "Discriminative sparsity preserving graph embedding," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2016, pp. 4250–4257.

[30] J. Gou, Y. Yang, Z. Yi, J. Lv, Q. Mao, and Y. Zhan, "Discriminative globality and locality preserving graph embedding for dimensionality reduction," *Expert Syst. Appl.*, vol. 144, Apr. 2020, Art. no. 113079.

[31] M. Zheng *et al.*, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, May 2011.

[32] H. Chang, H. Tang, F. Zhang, Y. Chen, and H. Zheng, "Graph-regularized discriminative analysis-synthesis dictionary pair learning for image classification," *IEEE Access*, vol. 7, pp. 55398–55406, 2019.

[33] L. Zhang, Z. Liu, J. Pu, and B. Song, "Adaptive graph regularized nonnegative matrix factorization for data representation," *Appl. Intell.*, vol. 50, no. 2, pp. 438–447, 2020.

[34] D. Ding, F. Xia, X. Yang, and C. Tang, "Joint dictionary and graph learning for unsupervised feature selection," *Int. J. Speech Technol.*, vol. 50, no. 5, pp. 1379–1397, May 2020.

[35] Y. Rong, S. Xiong, and Y. Gao, "Double graph regularized double dictionary learning for image classification," *IEEE Trans. Image Process.*, vol. 29, pp. 7707–7721, 2020.

[36] A. S. Charles, N. Cermak, R. O. Affan, B. B. Scott, J. Schiller, and G. Mishne, "GraFT: Graph filtered temporal dictionary learning for functional neural imaging," *IEEE Trans. Image Process.*, vol. 31, pp. 3509–3524, 2022.

[37] C. Vincent-Cuaz, T. Vayer, R. Flamary, M. Corneli, and N. Courty, "Online graph dictionary learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 10564–10574.

[38] A. Baltoiu, A. Patrascu, and P. Irofti, "Graph anomaly detection using dictionary learning," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 3551–3558, 2020.

[39] S. Tariyal, A. Majumdar, R. Singh, and M. Vatsa, "Deep dictionary learning," *IEEE Access*, vol. 4, pp. 10096–10109, 2016.

[40] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, Feb. 2015.

[41] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[42] V. Singhal and A. Majumdar, "Age and gender estimation via deep dictionary learning regression," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[43] V. Singhal and A. Majumdar, "A domain adaptation approach to solve inverse problems in imaging via coupled deep dictionary learning," *Pattern Recognit.*, vol. 100, p. 100, Apr. 2020.

[44] H. Tang, H. Liu, W. Xiao, and N. Sebe, "When dictionary learning meets deep learning: Deep dictionary learning and coding network for image recognition with limited data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2129–2141, May 2021.

[45] A. Montazeri, M. Shamsi, and R. Dianat, "MLK-SVD, the new approach in deep dictionary learning," *Vis. Comput.*, vol. 37, no. 4, pp. 707–715, Oct. 2020.

[46] C. Qiao, L. Yang, V. D. Calhoun, Z.-B. Xu, and Y.-P. Wang, "Sparse deep dictionary learning identifies differences of time-varying functional connectivity in brain neuro-developmental study," *Neural Netw.*, vol. 135, pp. 91–104, Mar. 2021.
[47] U. Rodriguez-Dominguez and O. Dalmau, "Hierarchical discriminative deep dictionary learning," *IEEE Access*, vol. 8, pp. 142680–142690, 2020.
[48] G. Yang, H.-C. Li, W.-Y. Wang, W. Yang, and W. J. Emery, "Unsupervised change detection based on a unified framework for weighted collaborative representation with RDDL and fuzzy clustering," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8890–8903, Nov. 2019.
[49] S. Xu, C. Zhang, and J. Zhang, "Adaptive quantile low-rank matrix factorization," *Pattern Recognit.*, vol. 103, Jul. 2020, Art. no. 107310.
[50] Y. Zuo, N. Wang, L. Jia, H. Zhang, Z. Wang, and Y. Qin, "Fully decomposed singular value and fixed dictionary extreme learning machine for bogie fault diagnosis," *IEEE Trans. Intell. Transp. Syst.*, early access, Jun. 24, 2021, doi: 10.1109/TITS.2021.3089181.
[51] X. Zhu, J. Guo, W. Nejdl, X. Liao, and S. Dietze, "Multi-view image clustering based on sparse coding and manifold consensus," *Neurocomputing*, vol. 403, pp. 53–62, Aug. 2020.
[52] G. E. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.
[53] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[54] A. Martinez and R. Benavente, "The AR face database: CVC technical report 24," Dept. Comput. Sci., Universitat Autonoma de Barcelona, Barcelona, Spain, Tech. Rep. 24, 1998.
[55] C. E. Thomaz and G. A. Giraldi, "A new ranking method for principal components analysis and its application to face image analysis," *Image Vis. Comput.*, vol. 28, no. 6, pp. 902–913, Jun. 2010.
[56] 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.
[57] Q. Zhang, J. Zhou, J. He, X. Cun, S. Zeng, and B. Zhang, "A shell dataset, for shell features extraction and recognition," *Sci. Data*, vol. 6, no. 1, pp. 1–9, Dec. 2019.
[58] O. Soderkvist, "Computer vision classification of leaves from Swedish trees," M.S. thesis, 2001.
[59] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-100)," Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-006-96, 1996.

Jianping Gou (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2012. He was a Post-Doctoral Research Fellow with The University of Sydney. He is currently an Associate Professor with the School of Computer Science and Communication Engineering, Jiangsu University, China. He has published over 100 papers in international journals or conferences, such as in *IJCV*, IEEE TRANSACTIONS ON CYBERNETICS, and *TKDD*. His current research interests include pattern classification and machine learning. He is an Editorial Board Member of *Mathematics* and a Senior Member of CCF and CSIG. He is an Academic Editor of *Scientific Programming*.

Xia Yuan is currently pursuing the master's degree with the School of Computer Science and Communication Engineering, Jiangsu University, China. Her research interests include pattern recognition and machine learning.

Lan Du received the Ph.D. degree in computer science from Australian National University in 2012. He is currently a Senior Lecturer with the Faculty of Information Technology, Monash University, Australia. Before he joined Monash University, he was a Post-Doctoral Research Fellow associated with the Computational Linguistic Group, Macquarie University. His research interests include machine learning, natural language understanding, and health analytics.

Shuyin Xia (Member, IEEE) received the Ph.D. degree from the College of Computer Science, Chongqing University, China. He is a Professor and a Ph.D. Supervisor with the Chongqing University of Posts and Telecommunications, Chongqing, China, where he is also the Executive Deputy Director of the Big Data and Network Security Joint Laboratory. He has published more than 30 papers in prestigious journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, and *IS*. His research interests include classifiers and granular computing.

Zhang Yi (Fellow, IEEE) received the Ph.D. degree in mathematics from the Institute of Mathematics, Chinese Academy of Sciences, Beijing, China, in 1994. He is currently a Professor with the College of Computer Science, Sichuan University, Chengdu, China. He has coauthored three books entitled *Convergence Analysis of Recurrent Neural Networks* (Kluwer, 2004), *Neural Networks: Computational Models and Applications* (Springer, 2007), and *Subspace Learning of Neural Networks* (CRC, 2010). His current research interests include neural networks and intelligent medicine. He was an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS from 2009 to 2012 and is an Associate Editor of IEEE TRANSACTIONS ON CYBERNETICS.