

When Dictionary Learning Meets Deep Learning: Deep Dictionary Learning and Coding Network for Image Recognition with Limited Data

Hao Tang, Hong Liu, Wei Xiao and Nicu Sebe, *Senior Member, IEEE*

Abstract—We present a new Deep Dictionary Learning and Coding Network (DDLCN) for image recognition tasks with limited data. The proposed DDLCN has most of the standard deep learning layers (e.g., input/output, pooling, fully connected, etc.), but the fundamental convolutional layers are replaced by our proposed compound dictionary learning and coding layers. The dictionary learning learns an over-complete dictionary for input training data. At the deep coding layer, a locality constraint is added to guarantee that the activated dictionary bases are close to each other. Then the activated dictionary atoms are assembled and passed to the compound dictionary learning and coding layers. In this way, the activated atoms in the first layer can be represented by the deeper atoms in the second dictionary. Intuitively, the second dictionary is designed to learn the fine-grained components shared among the input dictionary atoms, thus a more informative and discriminative low-level representation of the dictionary atoms can be obtained. We empirically compare DDLCN with several leading dictionary learning methods and deep learning models. Experimental results on five popular datasets show that DDLCN achieves competitive results compared with state-of-the-art methods when the training data is limited. Code is available at <https://github.com/Ha0Tang/DDLCN>.

Index Terms—Dictionary Learning, Feature Representation, Deep Learning, Image Recognition, Limited Data.

I. INTRODUCTION

THE key step of classifying images is obtaining feature representations encoding relevant label information. In the last decade, the most popular representation learning methods are dictionary learning (or sparse representation) and deep learning. Dictionary learning is learning a set of atoms so that a given image can be well approximated by a sparse linear combination of these learned atoms, while deep learning methods aim at extracting deep semantic feature representations via a deep network. Scholars from various research fields have realized and promoted the progress of dictionary learning with great efforts, e.g., [1], [2] from the statistics and machine learning community, [3] and [4] from the signal processing community and [5], [6], [7] from the computer vision and image processing communities. However, what is a sparse representation and how can we benefit from it?

Hao Tang and Nicu Sebe are with the Department of Information Engineering and Computer Science (DISI), University of Trento, Trento 38123, Italy. E-mail: hao.tang@unitn.it, sebe@disi.unitn.it.

Hong Liu is with the Shenzhen Graduate School, Peking University, Shenzhen 518055, China. E-mail: hongliu@pku.edu.cn.

Wei Xiao is with the Shenzhen Lingxi Artificial Intelligence Co., Ltd, Shenzhen 518109, China. E-mail: xiaoweithu@163.com.

Corresponding author: Hao Tang and Hong Liu.

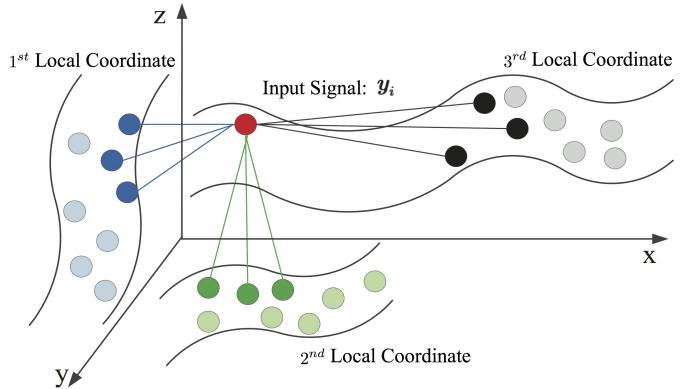


Fig. 1: Multiple local coordinates and ‘fake’ anchor points.

These two questions represent the points we attempt to clarify among the fundamental philosophies of sparse representation. In the following, we start with a brief description of the sparse representation and then put more emphasis on the relationship between sparsity and locality, to elicit the research focus points of the paper.

Dictionary Learning [2], [6], [8], also called sparse representation or sparse coding, represents a signal with a linear combination of a few elements/atoms from a dictionary matrix containing several prototype signal-atoms. The coefficient vector specifies how we select those seemingly ‘useful’ atoms from dictionary to linearly combine the original signal, i.e., each entry of the coefficient vector corresponds to a specific atom, and its value is viewed as a weight that specifies the linear combination proportion of each selected atom. The important asset of sparse representation is that the complex signal can be represented in a concise manner enabling the following classification procedure to adopt a simpler classifier (e.g., linear classifier). This is a fundamental advantage of sparse coding. Besides, it is worth noting that sparse-inducing regularization and sparse representation have been demonstrated to be extremely powerful for representing complex signals in recent works. Signals such as audios, videos and images admit naturally sparse representations, while the key idea of sparse representation is mapping the original chaotic signals to their corresponding concise representations with a regularized or uniform style. These sparse representations have an intimate relationship due to their over-complete dictionaries.

Locality and Sparsity. Recently, Yu et al. [9] extend the form of sparse representation and propose the Local Coordinate Coding (LCC) framework. LCC points out that locality is

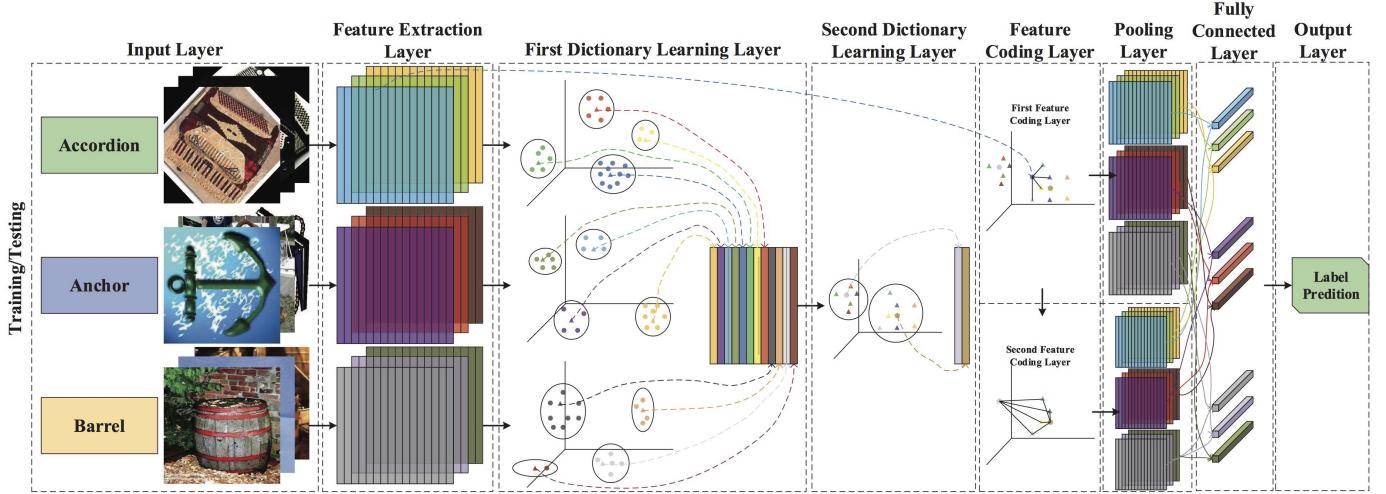


Fig. 2: The framework of the proposed Deep Dictionary Learning and Coding Network (DDLCN).

an intrinsic property rather than the sparsity under certain assumptions since locality will lead to sparsity but not necessarily vice versa. Specifically, LCC learns a non-linear high dimension function by forming an adaptive set of basis functions on the data manifold. To achieve a higher approximation accuracy, coding should be as local as possible to be better incorporated by the subspace spanned by the adaptive set of basis functions. Thereafter, Wang et al. [10] further developed a fast version of LCC, i.e., Locality-constrained Linear Coding (LLC). LLC employs the locality constraint to project each feature into its local-coordinate system. This fast version relaxes the local regularization term using a ℓ_2 norm-based term. That is why it runs faster. As with LLC, two modified versions are proposed in [11] and [12], respectively, where a second layer is introduced to further improve the approximation power using different strategies, such as [11] with local PCA and [12] with deep coding network. In the context of nonlinear function approximation, sparse coding brings an in-depth understanding of its fundamental connotation and success. More importantly, it also obtains a deeper insight into its parentage ties with the essence, i.e., locality.

Limitations. Methods such as LCC [9] and LLC [10] are based on the observation that the sparse representations tend to be ‘local’. However, these methods such as LLC have a major disadvantage, i.e., to achieve higher approximation, we need to employ a large number of so-called ‘anchor points’ for making a better linear approximation of the original signal. Since LLC is a local linear approximation of a complex signal y_i , which means the anchor points need to provide higher approximation power, making some of them not to be necessarily ‘real’ local anchors on the manifold where the original signal y_i resides. Therefore, the goal of the paper is to equip anchors with higher descriptive power to better approximate the input y_i for making more accurate inferences from it (see Fig. 1 for a better understanding). Moreover, we observe that most studies in dictionary learning so far adopt a shallow (single layer) network architecture. For instance, existing leading dictionary learning approaches are K-SVD [4], Discriminative K-SVD (D-KSVD) [13] and Label Consistent K-SVD (LC-KSVD) [14], which aim to decompose the input data into a dense

basis and sparse coefficients. However, shallow network architectures are difficult to fully extract the intrinsic properties of the input data. In our preliminary experiments, we observe that both limitations lead to very poor classification performance when the data are limited.

Deep Dictionary Learning. To fix both limitations, recent works such as [15], [16], [17], [18], [19], [20], [21], [22], [23], [24] have demonstrated that deeper network architectures can be built from dictionary learning methods. For example, Liu et al. [16] present a new dictionary learning layer to replace both the conventional fully connected layer and the rectified linear unit in a deep learning framework for scene recognition tasks. Song et al. [23] propose MDDL with a locality constraint for image classification tasks. Mahdizadehaghdam et al. [24] introduce a new model, which tries to learn a hierarchy of deep dictionaries for image classification tasks. Different from conventional deep neural networks, deep dictionary learning methods usually first extract feature representations at a patch-level, and then rely on the patch-based learned dictionary to output a global sparse feature representation for the input data.

Our proposed method borrows some useful ideas from CNNs and differs from these multi-layer dictionary learning methods in two aspects. First, our dictionary is directly learned from image features and then the learned dictionary acts as a candidate pool for learning the next layer dictionary. The learned dictionaries from different layers have connections while existing methods use a fixed dictionary in different layers, i.e., there is no message passing between the dictionaries of different layers. Second, to represent an atom in the previous layer, our model picks out a few atoms in the next layer and linearly combine them. The activated atoms have a linear contribution in constructing the atom in the previous layer. This could incorporate more information into the next layer’s codes and alleviate the influence of incorrect atoms.

Contributions. In this paper, we aim to improve the deep representation ability of dictionary learning. To this end, we propose a novel Deep Dictionary Learning and Coding Network (DDLCN), which mainly consists of several layers, i.e., input, feature extraction, dictionary learning, feature coding, pooling, fully connected and output layer, as shown in Fig. 2.

The design motivation of the proposed DDLCN is derived from both Convolutional Neural Networks (CNNs) and dictionary learning approaches. However, the biggest difference being that the convolutional layers in CNNs are replaced by our proposed dictionary learning and coding layers. By doing so, the proposed DDLCN can learn edge, line and corner representations from the shallow dictionary layers. Then additional sophisticated ‘hierarchical’ feature representations can be learned from deeper dictionary layers. The proposed DDLCN has a better approximation capability of the input since the introduction of the proposed dictionary learning and coding layer, which takes advantage of the manifold geometric structure to locally embed points from the underlying data manifold into a lower-dimensional deep structural space. Moreover, it also fully considers each fundamental basis vector adopted in the shallow layer coding, and incorporates additional gradient effects of nonlinear functions on it into the deeper local representation. Thus, the proposed DDLCN can transfer a very difficult nonlinear learning problem into a simpler linear learning one. More importantly, the approximation power is higher than its single-layer counterpart.

Overall, the contributions of the paper are summarized as: 1) We present a novel deep dictionary learning DDLCN framework, which combines the advantages of both dictionary learning and deep learning methods. 2) We introduce a novel compound dictionary learning and coding layer, which can substitute the convolutional layer in the standard deep learning architectures. 3) Extensive experiments on a broader range of datasets with limited training data demonstrate that the proposed method outperforms state-of-the-art dictionary learning approaches and achieves competitive results compared with existing deep learning methods.

Part of this work has been published in [25]. The additional contributions are: 1) We present a more detailed analysis by including recently published works about deep dictionary learning. 2) We generalize the two-layer framework of DDLCN in [25] to a deeper one and validate the effectiveness. 3) We present an in-depth description of the proposed layer and framework, providing all the architectural and implementation details of the method. 4) We also extend the experimental evaluation provided in [25] in several directions. First, we conduct extensive experiments on five popular datasets, demonstrating the wide application scope of our DDLCN framework. Second, we conduct more interpretative experiments to show the superiority of the proposed DDLCN compared with the traditional convolutional layers. Third, we introduce several variants of the proposed DDLCN, i.e., DDLCN-3, DDLCN-4, DDLCN-5 and DDLCN-6, which achieve better results than DDLCN-2 proposed in [25]. Lastly, we have also included new state-of-the-art baselines, e.g., MDDL [23], and we observe that the proposed DDLCN achieves always better results.

II. RELATED WORK

In this section, we briefly review the related work. Some important notations used in our paper are given in Table I.

Sparse coding (or sparse representation) represents an original signal \mathbf{y} with a sparse signal \mathbf{x} based on a dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_p] \in \mathbb{R}^{m \times p}$, where the dictionary \mathbf{D} is

TABLE I: The notations used in this paper.

\mathbf{Y}	$[\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n] = \{\mathbf{y}_i\}_{i=1}^n$
\mathbf{y}	signal (also called data vector or descriptor or feature)
\mathbf{y}'	single-layer physical approximation of \mathbf{y}
\mathbf{y}''	two-layer physical approximation of \mathbf{y}
\mathbf{X}	$[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] = \{\mathbf{x}_i\}_{i=1}^n$
\mathbf{x}	coefficient vector or solution
\mathbf{v}	anchor point of C^1 or first layer atom
\mathbf{v}'	physical approximation of \mathbf{v}
\mathbf{u}	anchor point of $C^{2,v}$ or second layer atom
C^1	set of anchor points to \mathbf{y}
$C^{2,v}$	set of anchor points to \mathbf{v}
γ^1	map of \mathbf{y} to $\gamma^1(\mathbf{y})$
$\gamma^{2,v}$	map of \mathbf{v} to $\gamma^{2,v}(\mathbf{v})$
\mathbf{D}	$[\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_p]$, a dictionary or codebook with p atoms
\mathbf{d}	element or atom or codeword
D_i	dictionary size of the i layer
m	dimension of \mathbf{y}
l	number of signals
n	number of deep layers

learned by dictionary learning algorithms. Usually, it is formulated as the following constrained optimization objective,

$$(P_{0,\varepsilon}) \quad \min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{D}\mathbf{x}, \quad (1)$$

where $\|\cdot\|_0$ is the ℓ_0 pseudo-norm, which aims to count the non-zeros entries of a vector. However, the equality constraint $\mathbf{y} = \mathbf{D}\mathbf{x}$ is too strict for solving the problem. Hence, we relax the optimization problem with a small threshold as follows,

$$(P_{0,\varepsilon}) \quad \min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \leq \varepsilon, \quad (2)$$

or equivalently, its corresponding unconstrained form is as follows using the Lagrange multipliers,

$$(P_{0,\lambda}) \quad \min_{\mathbf{x}} \left[\frac{1}{2} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_0 \right]. \quad (3)$$

Unfortunately, the approaches that provide an approximate solution to the problem are pursuit procedures but are not global optimal because this problem is NP-hard. In the past few decades, various pursuit procedures have been introduced. For instance, both Matching Pursuit (MP) [26] and Orthogonal Matching Pursuit (OMP) [27] are greedy algorithms that select the dictionary atoms sequentially, which involves the computation of inner products between the input signal and dictionary atoms.

Another type of pursuit procedure is a relaxation strategy. For instance, Basis Pursuit (BP) [28] converts the problems of Eq. (1) or Eq. (2) to their convex counterparts by replacing the ℓ_0 pseudo-norm with the ℓ_1 norm. Focal Underdetermined System Solver (FOCUSS) [29] employs the ℓ_q norm with $q \leq 1$ to replace the ℓ_0 pseudo-norm. Since the ℓ_0 optimization problem is non-convex, one of the popular relaxations is Lasso [1], which uses the ℓ_1 norm instead of ℓ_0 pseudo-norm,

$$(P_{1,\varepsilon}) \quad \min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \leq \varepsilon, \quad (4)$$

where $\|\cdot\|_1$ is the ℓ_1 norm and the corresponding unconstrained form is as follows, using the Lagrange multipliers,

$$(P_{1,\lambda}) \quad \min_{\mathbf{x}} \left[\frac{1}{2} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1 \right]. \quad (5)$$

It is well known that, as λ goes larger, \boldsymbol{x} tends to be more sparse, such that only a few dictionary elements are involved.

This paper focuses on learning the sparse representations in a situation where the data have only a few significant patterns. This greatly benefits the applications of classifications and information fusion. There are several popular extensions of traditional sparse coding, i.e., Group Sparse Coding [30], [31], LCC [9], [11] and its fast implementation algorithm LLC [10]. Group Sparse Coding encourages the solutions of sparse regularized problems to have the specific patterns of non-zero coefficients, which benefits higher-level tasks such as image recognition [32] and compressive sensing [33]. LCC and LLC empirically observe that sparse representation results tend to be ‘local’. As indicated in LCC [9], [10], the locality is more essential than sparsity, especially for supervised learning, since the locality will lead to sparsity but not necessarily vice versa.

The corresponding LLC [10] representations can be obtained by solving the convex programming,

$$\min_{\boldsymbol{x}} \left[\sum_{i=1}^n \|\boldsymbol{y}_i - \boldsymbol{D}\boldsymbol{x}_i\|_2^2 + \lambda \|\boldsymbol{b}_i \odot \boldsymbol{x}_i\|_2^2 \right] \quad s.t. \quad \mathbf{1}^\top \boldsymbol{x}_i = 1, \quad (6)$$

where \boldsymbol{y}_i is the i^{th} local descriptor of $\boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_n]$ and $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n]$ is the set of codes for \boldsymbol{Y} . The constraint $\mathbf{1}^\top \boldsymbol{x}_i = 1$ follows the shift-invariant requirements of the LLC code. \odot represents the element-wise multiplication. $\boldsymbol{b}_i = \exp(\frac{\text{dist}(\boldsymbol{y}_i, \boldsymbol{D})}{\psi})$ is the locality adapter, which provides different degrees of freedom to each basis vector \boldsymbol{d}_j proportional to its similarity to the input descriptor \boldsymbol{y}_i , where $\text{dist}(\boldsymbol{y}_i, \boldsymbol{D}) = [\text{dist}(\boldsymbol{y}_i, \boldsymbol{d}_1), \dots, \text{dist}(\boldsymbol{y}_i, \boldsymbol{d}_p)]^\top$, $\text{dist}(\boldsymbol{y}_i, \boldsymbol{d}_j)$ is the Euclidean distance between \boldsymbol{y}_i and \boldsymbol{d}_j , and ψ is used for adjusting the weight decay speed for the locality adapter.

The LLC solution in Eq. (6) is not sparse in the sense of ℓ_0 pseudo-norm or any other sparse inducing norms, while it is really sparse in the sense that the solution only has few non-zero values. This idea enriches the connotation of ‘sparse’, as it makes possible various applications such as signal representation with simultaneous sparse and discriminative properties. This property of sparse is mainly induced by locality regularization term $\lambda \|\boldsymbol{b}_i \odot \boldsymbol{x}_i\|_2^2$. More specifically, each entry in \boldsymbol{b}_i constrains the corresponding entry in \boldsymbol{x}_i , i.e., the more non-zero entries in \boldsymbol{b}_i , the fewer values in the counterpart \boldsymbol{x}_i . This effect is amplified exponentially by \boldsymbol{b}_i .

Compared with these methods, the newly introduced coding strategy captures more accurate correlations. Traditional sparse coding methods only pursue the solo goal, i.e., to be as sparse as possible in the final representation. While LLC aims to catch the delicate atoms structure of the manifold where the input signals reside, and it further uses these activated atoms for signal representation. However, existing methods have a major limitation, i.e., to achieve higher approximation, one has to use a large number of so-called ‘anchor points’ to achieve a better linear approximation of the input signal. To fix this limitation, in this paper, we aim to equip anchors with more descriptive power to better approximate the input data \boldsymbol{y}_i for making more accurate inferences from it. We also provide an illustrative example in Fig. 1 for better understanding.

III. THE PROPOSED DDLCN FRAMEWORK

We sequentially introduce each layer of the proposed DDLCN in this section. Note that we only illustrate details of two-layer DDLCN for simplicity. Extension of the proposed DDLCN to multiple layers is straight forward.

Feature Extraction Layer. We first adopt a feature extractor F to extract a set of m -dimensional local descriptors $\boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_l] \in \mathbb{R}^{m \times l}$ from the input image I , where l is the total number of local descriptors. To highlight the effectiveness of the proposed method, we only use a single feature extractor in our experiment, i.e., Scale-Invariant Feature Transform (SIFT) [34]. The SIFT descriptor has been widely used in dictionary learning [35], [36], [37], [38], [39]. However, one can always use multiple feature extractor to further improve performance. Specifically, for the input image I , we extract the SIFT feature \boldsymbol{y}_i by using the feature extractor F , this process can be formulated as $\boldsymbol{y}_i = F(I), i \in [1, \dots, l]$.

First Dictionary Learning Layer. Let r denote the total number of classes in the dataset. Then we randomly select p images in each class to train the dictionary of the corresponding class, and the number of the first layer dictionary per category is denoted as q . Thus, the number of the dictionary for the first layer can be calculated by $D_1 = r * q$. Next, we adopt the following dictionary learning algorithm,

$$\min_{\boldsymbol{V}_i} \left[\frac{1}{2} \|\boldsymbol{y}_i - \boldsymbol{V}_i \alpha_i\|_2^2 \right] \quad s.t. \quad \|\alpha_i\|_1 \leq \lambda \quad (7)$$

where $\boldsymbol{V}_i = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_q]$ is the dictionary for i^{th} class in the first-layer dictionary, which contains q atoms, i.e., \boldsymbol{v}_i . We then group all of them to form the first-layer dictionary \boldsymbol{V} after separately learning the dictionary of each class. Thus $\boldsymbol{V} = [\boldsymbol{V}_1, \boldsymbol{V}_2, \dots, \boldsymbol{V}_r] = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_{D_1}] \in \mathbb{R}^{m \times D_1}$. α_i is a sparse coefficient introduced in [40]. In this way, the dictionary \boldsymbol{V}_i and the coefficients α_i can be learned jointly.

First Feature Coding Layer. After learning \boldsymbol{V} , each local feature is then encoded by \boldsymbol{V} through several nearest atoms for generating the first coding. By doing so, the first feature coding layer transfers each local descriptor \boldsymbol{y}_i into a D_1 dimensional code $\boldsymbol{\gamma}^1 = [\gamma_1^1, \gamma_2^1, \dots, \gamma_l^1] \in \mathbb{R}^{D_1 \times l}$. Specifically, each code can be obtained using the following optimization,

$$\min_{\boldsymbol{\gamma}_i^1} \left[\sum_{i=1}^l \frac{1}{2} \|\boldsymbol{y}_i - \boldsymbol{V} \boldsymbol{\gamma}_i^1\|_2^2 + \beta \|\boldsymbol{\gamma}_i^1 \odot \boldsymbol{\zeta}_i^1\|_1 \right] \quad s.t. \quad \mathbf{1}^\top \boldsymbol{\gamma}_i^1 = 1, \quad (8)$$

where $\boldsymbol{\zeta}_i^1 \in \mathbb{R}^{D_1}$ is a distance vector to measure the distance between \boldsymbol{y}_i and \boldsymbol{v}_i . \odot denotes the element-wise multiplication. Typically, $\boldsymbol{\zeta}_i^1$ can be obtained by reducing a reconstruction loss in the corresponding layer. We note that [12] adopts a simple sparse coding model at the first layer, which overlooks the importance of quantity distributions of each item in the code $\boldsymbol{\gamma}_i^1$, thus it is prone to a rough approximation at the first layer. Therefore, the physical approximation of \boldsymbol{y} in the first layer can be expressed as,

$$\boldsymbol{y}' = \sum_{\boldsymbol{v} \in \boldsymbol{C}^1} \boldsymbol{\gamma}^1(\boldsymbol{y}) \boldsymbol{v}, \quad (9)$$

where \boldsymbol{C}^1 is the set of anchor points to \boldsymbol{y} . An illustrative example is shown in Fig. 3.

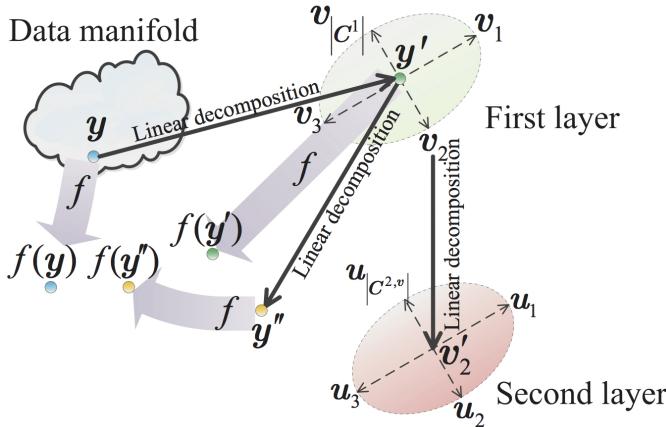


Fig. 3: Multi layers coding strategy. The first layer is mainly used to partition the space, while the main approximation power is achieved within the second layer, which embodies a ‘divide and conquer’ strategy.

Second Dictionary Learning Layer. As discussed in the introduction, most existing dictionary learning frameworks only use a single layer, which significantly limits the discriminative ability of the feature coding. Meanwhile, we observe that better representation will be obtained by using deeper layers in most computer vision tasks. Thus, we borrow some idea from deep CNNs and present a new deeper dictionary learning and coding layer. Then the second layer dictionary $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{s_2}]$ can be learned from the first layer dictionary \mathbf{V} ,

$$\min_{\mathbf{U}} \left[\frac{1}{2} \|\mathbf{v}_i - \mathbf{U}\alpha_i\|_2^2 \right] \quad s.t. \quad \|\alpha_i\|_1 \leq \lambda \quad (10)$$

where $\mathbf{v}_i \in \mathbf{V}$ is one of the basis vectors in the first activated dictionary. At the second layer, we put more emphasis on the representation of each \mathbf{v}_i or each group of \mathbf{v}_i to further refine each basis \mathbf{v}_i . Specifically, after coding at the first layer, we try to map a nonlinear function f to a simplified local coordinate space with low intrinsic dimensionality. However, from the viewpoint of Lipschitz smoothness [9], [12], this solo layer mapping only incorporates limited information about f with its derivative on \mathbf{y} , such that it is incapable of guaranteeing better approximation quality. That is why we would move deeper into the second layer to seek more information about f for further improving the approximation. By doing so, the first layer can capture the fine low-level structures from the input image, then the second coherently captures more complex structures from the first layer.

Second Feature Coding Layer. We can obtain the code of the second layer by using the following optimization,

$$\min_{\gamma_i^2} \left[\sum_{i=1}^{D_1} \frac{1}{2} \|\mathbf{v}_i - \mathbf{U}\gamma_i^2\|_2^2 + \beta \|\gamma_i^2 \odot \zeta_i^2\|_1 \right] \quad (11)$$

$$s.t. \quad 1^\top \gamma_i^2 = 1,$$

where $\gamma_i^2 = [\gamma_i^2(\mathbf{u}_1), \gamma_i^2(\mathbf{u}_2), \dots, \gamma_i^2(\mathbf{u}_{D_2})]^\top \in \mathbb{R}^{D_2}$ is the second coding and D_2 is the number dictionary of the second layer. $\zeta_i^2 \in \mathbb{R}^{D_2}$ is used to measure the distance between \mathbf{v}_i and each atom in \mathbf{U} . $\mathbf{v}_i \in \mathbf{V}$ is one of the basis vectors adopted in the representation of \mathbf{y}_i at the first layer.

By doing so, the activated atoms \mathbf{v}_i in the first layer can be further decomposed to obtain the second layer coding using \mathbf{U} . Thus, the approximation of \mathbf{y} in the second layer can be defined as,

$$\mathbf{y}'' = \sum_{v \in C^1} \left[\gamma^1(\mathbf{y}) \sum_{u \in C^{2,v}} \gamma^{2,v}(\mathbf{v}) \mathbf{u} \right], \quad (12)$$

where $C^{2,v}$ is the set of anchor points to v . We also provide an illustrative example in Fig. 3 for better understanding. The core idea of the two-layer coordinate coding is that if both coordinate codings, i.e., \mathbf{y}' and $\mathbf{v}' = \sum_{u \in C^{2,v}} \gamma^{2,v}(\mathbf{v}) \mathbf{u}$, are sufficiently localized, then a point \mathbf{y} lies on a manifold, which would be locally embedded into a lower-dimensional two-layer structure space. More importantly, not only the data point \mathbf{y} is locally linearly represented, but also the function $f(\mathbf{y})$. This significant observation lays the foundation for our approach.

The n^{th} Dictionary Learning Layer. Similarly, we can learn the n^{th} dictionary $\mathbf{D}^n = [d_1^n, d_2^n, \dots, d_{D_n}^n]$ from the previous layer dictionary \mathbf{D}^{n-1} ,

$$\min_{\mathbf{D}^n} \left[\frac{1}{2} \|d_i^{n-1} - \mathbf{D}^n \alpha_i\|_2^2 \right] \quad s.t. \quad \|\alpha_i\|_1 \leq \lambda, \quad (13)$$

where $d_i^{n-1} \in \mathbf{D}^{n-1}$ is one of the activated basis vectors in the previous $(n-1)^{th}$ dictionary layer.

The n^{th} Feature Coding Layer. Therefore, we can generalize the two-layer framework of DDLCN to a deeper one,

$$\min_{\gamma_i^n} \left[\frac{1}{2} \|d_i^{n-1} - \mathbf{D}^n \gamma_i^n\|_2^2 + \beta \|\gamma_i^n \odot \zeta_i^n\|_1 \right] \quad (14)$$

$$s.t. \quad 1^\top \gamma_i^n = 1,$$

where γ_i^n is the n^{th} layer coding and ζ_i^n is employed to measure the distance between d_i^{n-1} and each atom in \mathbf{D}^n . $d_i^{n-1} \in \mathbf{D}^{n-1}$ is one of the basis vectors adopted in the feature representation of \mathbf{y}_i at the $(n-1)^{th}$ coding layer. Through the proposed multi-layer learning and coding strategy, the proposed DDLCN can output a robust feature representation to accurately represent the input image. Moreover, DDLCN increases and boosts the separability of feature representations from different semantic classes. Lastly, DDLCN preserves the locality information of the input local features, avoiding very large values in the coding representation and reducing the error caused by over-fitting.

Pooling Layer. After the last dictionary learning and feature coding layer, we use a pooling layer for removing the fixed-size constraint of the input images [41]. Specifically, for each input image, we adopt 1×1 , 2×2 and 4×4 spatial pyramids with max-pooling.

Fully Connected Layer. The final feature representations of \mathbf{y}_i can be obtained by integrating feature representation from each layer. Task two-layer framework for an example, each item (such as the j^{th} item) in the first layer’s codes γ_i^1 can be augmented into the form of $[\gamma_i^1(\mathbf{v}_j), \gamma_i^1(\mathbf{v}_j)[\gamma_j^2(\mathbf{u}_1), \gamma_j^2(\mathbf{u}_2), \dots, \gamma_j^2(\mathbf{u}_{s_2})]]^\top$. Then we concatenate the first layer coding and the second layer coding to form the final coding representation, which is a $D_1 \times (1+D_2)$ dimensional vector. We also provide the two-layer framework of our DDLCN in Algorithm 1.

Algorithm 1 The two-layer framework of our DDLCN.

Require: : $\mathbf{Y} \in \mathbb{R}^{m \times l}$
Ensure: : γ_i

- 1: First dictionary learning: $\mathbf{V} \leftarrow \mathbf{V}_{Dictionary}$
- 2: First locality constraint calculating:

$$\zeta_i^1 = [\|\mathbf{y}_i - \mathbf{v}_1\|_2, \|\mathbf{y}_i - \mathbf{v}_2\|_2, \dots, \|\mathbf{y}_i - \mathbf{v}_{D_1}\|_2]^T$$
- 3: First feature coding:

$$\begin{aligned} \text{for } i = 1 \text{ to } l \\ \gamma_i^1 \leftarrow \min_{\gamma_i^1} \left[\frac{1}{2} \|\mathbf{y}_i - \mathbf{V} \gamma_i^1\|_2^2 + \beta \|\gamma_i^1 \odot \zeta_i^1\|_1 \right] \\ \text{s.t. } \mathbf{1}^T \gamma_i^1 = 1 \\ \text{end} \end{aligned}$$
- 4: Second dictionary learning: $\mathbf{U} \leftarrow \mathbf{U}_{Dictionary}$
- 5: Second locality constraint calculating:

$$\zeta_i^2 = [\|\mathbf{v}_i - \mathbf{u}_1\|_2, \|\mathbf{v}_i - \mathbf{u}_2\|_2, \dots, \|\mathbf{v}_i - \mathbf{u}_{D_2}\|_2]^T$$
- 6: Second feature coding:

$$\begin{aligned} \text{for } i = 1 \text{ to } D_1 \\ \gamma_i^2 \leftarrow \min_{\gamma_i^2} \left[\frac{1}{2} \|\mathbf{v}_i - \mathbf{U} \gamma_i^2\|_2^2 + \beta \|\gamma_i^2 \odot \zeta_i^2\|_1 \right] \\ \text{s.t. } \mathbf{1}^T \gamma_i^2 = 1 \\ \text{end} \end{aligned}$$
- 7: Coding augmentation:

$$\begin{aligned} \text{for } i = 1 \text{ to } l \\ \text{for } j = 1 \text{ to } D_1 \\ \gamma_i^1(\mathbf{v}_j) \leftarrow [\gamma_i^1(\mathbf{v}_j), \gamma_i^1(\mathbf{v}_j)[\gamma_j^2(\mathbf{u}_1), \dots, \gamma_j^2(\mathbf{u}_{D_2})]]^T \\ \text{end} \\ \text{end} \end{aligned}$$

Output Layer. We adopt the Support Vector Machine (SVM) [42] as our classifier, which has been validated in many classification tasks such as [10], [43], [23], [44], [45]. Specifically, we employ LIBSVM [46] to implement our multi-class SVM.

The classification of the input image is ultimately carried out by assembling deep dictionaries from different layers and assessing their contribution. Moreover, through jointly minimizing both the classification errors and the reconstruction errors of all different layers, the proposed DDLCN iteratively adapts the deep dictionaries to help to build better feature representations for image recognition tasks.

IV. EXPERIMENTS

We conduct extensive experiments (including face recognition, object recognition and hand-written digits recognition) to evaluate the effectiveness of the proposed DDLCN.

A. Experimental Setting

Datasets. We follow [38], [47], [25], [48], [20], [18], [19], [49] and evaluate the effectiveness of the proposed DDLCN on five widely-used datasets, i.e., Extended YaleB [50], AR Face [51], Caltech 101 [52], Caltech 256 [53] and MNIST [54], which are all standard datasets for dictionary learning evaluation. Note that we follow the same evaluation procedure with the previous works on each dataset for a fair comparison.

Parameter Setting. Compared with existing CNN-based methods, which have lots of hyper-parameters, while the proposed DDLCN only has three parameters needed to be tuned. The three parameters of the proposed DDLCN are:

- p , the number of training dictionary samples per category.
- q , the number of the first layer dictionary per category.
- t , the number of training samples per category.

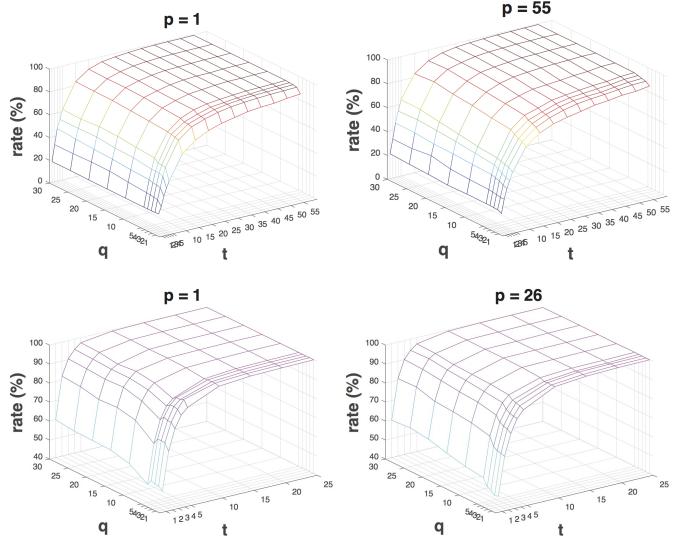


Fig. 4: Classification accuracy with different parameter p on Extended YaleB (top two) and AR Face (bottom two).

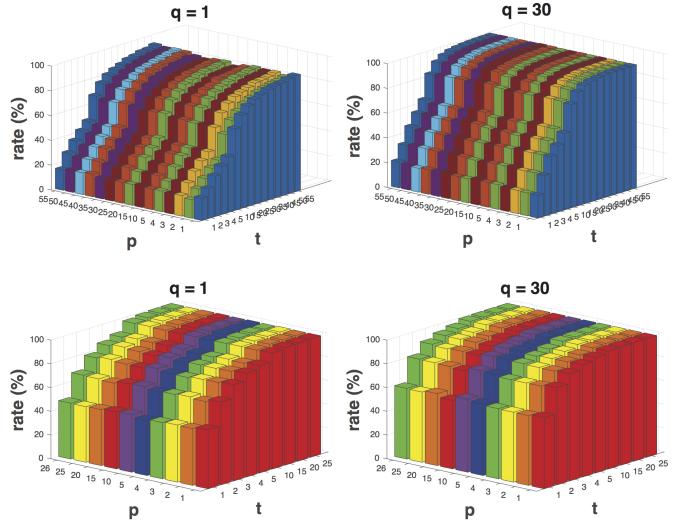


Fig. 5: Classification accuracy with varying q on Extended YaleB (top two) and AR Face (bottom two).

For simplicity, we use $p-q$ to represents that p images are randomly selected per category for training dictionary and q dictionary bases are learned per category in the first dictionary. For example, ‘5-5’ means that $p=5$ and $q=5$. We first conduct extensive experiments to evaluate the performance of different values of p , q and t . For all experiments, we repeat 10 times to achieve reliable results and then average them to obtain the final results.

(i) Parameter p . To demonstrate the superiority of the proposed DDLCN, we set the number of the first dictionary to $p=[1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$ and $p=[1, 2, 3, 4, 5, 10, 15, 20, 25, 26]$ on Extended YaleB and AR Face, respectively. The results of two extreme cases on both datasets are shown in Fig. 4. We can see that the proposed method achieves good results when $p=1$, validating our design motivation. Moreover, we observe that the classification performance achieves a peak with 10 training samples and then

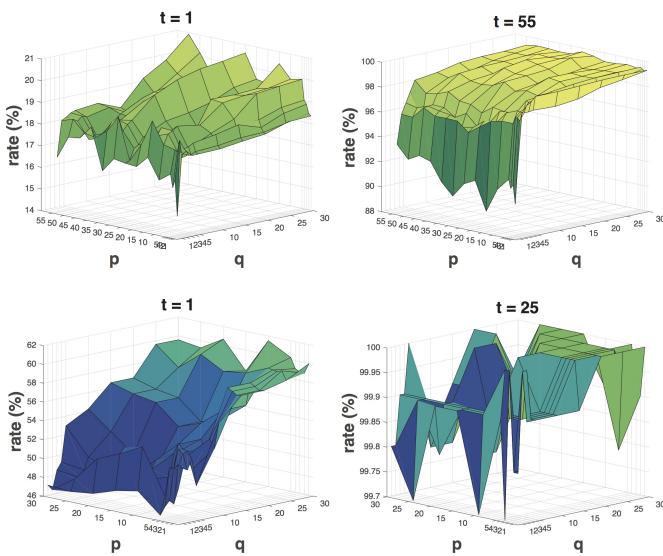


Fig. 6: Classification accuracy with different t on Extended YaleB (top two) and AR Face (bottom two).

TABLE II: Comparison results between the proposed dictionary layer and conventional convolutional layer.

Method	Extended YaleB	AR Face
DDLCN-2 (Dictionary Layer)	99.18 \pm 0.46	99.87 \pm 0.19
DDLCN-2 (Convolutional Layer)	98.94 \pm 0.57	99.59 \pm 0.38
DDLCN-3 (Dictionary Layer)	99.32 \pm 0.41	99.92 \pm 0.15
DDLCN-3 (Convolutional Layer)	99.07 \pm 0.53	99.67 \pm 0.36

tends to be stable.

(ii) Parameter q . We set $q=[1, 2, 3, 4, 5, 10, 15, 20, 25, 30]$ on both Extended YaleB and AR Face datasets. The results of two extreme cases on both datasets are reported in Fig 5. We can see that the gaps between the two cases are marginal due to the introduction of the proposed layer on Extended YaleB. Moreover, the proposed DDLCN achieves nearly 100% classification accuracy when only using 1 atom per person and 20 images per class for training on AR Face. This means that the proposed DDLCN can exploit the intrinsic structure of the manifold where features reside, leading to better classification performance with limited dictionaries and training data.

(iii) Parameter t . We set the number of the training images to $t=[1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$ and $t=[1, 2, 3, 4, 5, 10, 15, 20, 25]$ on Extended YaleB and AR Face, respectively. The results of two extreme cases on both datasets are illustrated in Fig. 6. We can draw two conclusions: 1) the classification accuracy first rises to the peak rapidly and then tends to be stable as t increasing. 2) there is a small impact to classification accuracy when changing p .

B. The Proposed Layer vs. Convolutional Layer

We then conduct experiments to validate the effectiveness of the proposed compound dictionary learning and coding layer. Specifically, we employ the proposed DDLCN as our backbone and replace the proposed compound layers with conventional convolutional layers keeping all other details the same. Comparison results of both Extended YaleB and AR Face datasets are shown in Table II. We can see that the proposed compound dictionary learning and coding layer achieves better results than the convolutional layer, meaning

TABLE III: Classification accuracy (%) on Extended YaleB.

Method	Included (%)	Excluded* (%)	Time (ms)
SRC (15 per person) [6]	80.50	86.70	11.22
LLC (30 local bases) [10]	82.20	92.10	-
DL-COPAR [55]	86.47 \pm 0.69	-	31.11
FDDL [56]	90.01 \pm 0.69	-	42.48
LLC (70 local bases) [10]	90.70	96.70	-
DBDL [57]	91.09 \pm 0.59	-	1.07
JBDC [47]	92.14 \pm 0.52	-	1.02
K-SVD (15 per person) [4]	93.10	98.00	-
SupGraphDL-L [58]	93.44	-	-
D-KSVD (15 per person) [13]	94.10	98.00	-
LC-KSVD1 (15-15) [14]	94.50	98.30	0.52
LC-KSVD2 (15-15) [14]	95.00	98.80	0.49
MBAP [38]	95.12	-	-
VAE+GAN [59]	96.4	-	-
EasyDL [60]	96.22	-	-
LC-KSVD2 (A-15) [14]	96.70	99.00	-
SRC (all training samples) [6]	97.20	99.00	20.78
MDDL-2 [23]	98.2	-	-
MDDL-3 [23]	98.3	-	-
DDL [24]	99.1	-	-
RRC_L1 (300) [61]	99.80	-	-
PCANet-1 [62]	97.77	-	-
PCANet-2 [62]	99.85	-	-
DDLCN-2 (I-1)	87.42 \pm 1.33	89.54 \pm 1.02	0.18
DDLCN-2 (15-15)	97.38 \pm 0.54	98.48 \pm 0.48	0.71
DDLCN-2 (55-15)	97.68 \pm 0.60	98.64 \pm 0.52	0.92
DDLCN-2 (A-15)	98.34 \pm 0.56	99.18 \pm 0.46	0.98
DDLCN-3 (A-15)	98.76 \pm 0.42	99.32 \pm 0.41	1.32

that the proposed layer indeed obtains a more informative and discriminative representation, and confirming our design motivation.

C. Comparison Against State-of-the-Art Methods

We compare the proposed DDLCN (i.e., DDLCN-2 and DDLCN-3) with both dictionary learning and deep learning methods on five public datasets, i.e., Extended YaleB, AR Face, Caltech 101, Caltech 256 and MNIST datasets.

Specifically, DDLCN-2 uses two compound dictionary learning and coding layers, the first layer aims to learn a dictionary to represent the input image and the second layer target to learn a dictionary to represent the activated atoms in the first layer. For a given input image, we use both the first and second dictionaries to learn the coding representation and then concatenate both of them to obtain the final coding representation. For a deeper version, i.e., DDLCN-3, which uses three compound dictionary learning and coding layers. The first layer aims to learn a dictionary to represent input image, then the second layer targets to learn a dictionary to represent the activated atoms of the first layer, finally, the third layer learns another dictionary to represent the activated atoms of the second layer. We use the first, second and third dictionaries to learn the corresponding coding representation and then concatenate all three to form the final coding representation. The only difference between DDLCN-2 and DDLCN-3 consists in the number of the proposed layer, while all other layers such as fully connected and pooling layers, and the training details are all the same.

Extended YaleB. We adopt both state-of-the-art dictionary learning methods and deep learning methods as our baselines. Specifically, we compare the proposed DDLCN with dictionary learning methods such as D-KSVD [13], LC-KSVD [14] and MDDL [23]. We also compare DDLCN with deep learning models including PCANet [62] and VAE+GAN [59]. Results are shown in the second column of Table III, we can see

TABLE IV: Classification accuracy (%) on AR Face.

Method	Accuracy (%)	Time (ms)
SRC (5 per person) [6]	66.50	17.76
LLC (30 local bases) [10]	69.50	-
DL-COPAR [55]	83.29 ± 1.23	36.49
FDDL [56]	85.97 ± 1.23	50.03
DBDL [57]	86.15 ± 1.19	1.20
K-SVD (5 per person) [4]	86.50	-
JBDC [47]	87.17 ± 0.99	1.18
LLC (70 local bases) [10]	88.70	-
D-KSVD (5 per person) [13]	88.80	-
LC-KSVD1 (5-5) [14]	92.50	0.541
LC-KSVD2 (5-5) [14]	93.70	0.479
MBAP [38]	93.88	-
MDDL-2 [23]	94.9	-
MDDL-3 [23]	95.0	-
RRC_L1[61]	96.30	-
ADDL (5 items, 20 labels) [20]	97.00	-
SRC (all training samples) [6]	97.50	83.79
LC-KSVD2 (A-5) [14]	97.80	-
LGII [63]	99.00	-
PCANet-1 [62]	98.00	-
PCANet-2 [62]	99.50	-
DDLCN-2 (1-1)	99.56 ± 0.21	0.73
DDLCN-2 (5-5)	99.84 ± 0.36	1.26
DDLCN-2 (A-5)	99.87 ± 0.19	1.63
DDLCN-3 (A-5)	99.92 ± 0.15	1.82

that the proposed DDLCN achieves better results than all the dictionary learning methods. Moreover, we observe that the proposed DDLCN achieves slightly worse results than deep learning based models such as [62] and [61]. However, the proposed DDLCN still outperforms the PCANet-1 version [62] (97.77%), which proves the effectiveness of the proposed deep dictionary learning framework.

Moreover, we follow the evaluation metric of [14] and conduct another experiment with the bad images excluded to further validate the effectiveness of the proposed method. The comparison results are shown in Table III (the third column). We observe that our DDLCN obtains the best classification accuracy compared to the other methods. Lastly, we list the computation time of predicting a single image during the testing stage. Results compare with SRC [6], LC-KSVD [14], DBDL [57] and JBDC [47] are shown in Table III (the fourth column). We can see that the proposed DDLCN costs remarkably less time than other baselines.

AR Face. We adopt several advanced methods such as LC-KSVD [14] and MDDL [23] as our baselines. The comparison results are reported in Table IV. We can see that the proposed DDLCN achieves better results than other baselines including [62] and [61] when only using a 1-1 strategy, which strongly validates the effectiveness of the proposed deep dictionary learning framework.

Moreover, we show the computation time of predicting an image during the testing stage to simultaneously highlight the efficiency of the proposed DDLCN. Results are shown in the third column of Table IV, we see that DDLCN only spends more time than LC-KSVD, but costs significantly less time than SRC, DL-COPAR, JBDC, DBDL and FDDL. It means that the proposed DDLCN is not only a robust method but also an efficient solution for image nonrecognition tasks.

Caltech 101. We follow the standard experimental settings and

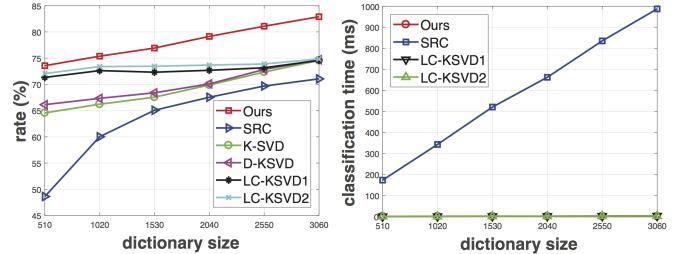


Fig. 7: Performance comparisons on Caltech 101. (left) Performance comparisons with varying dictionary size. (right) Computation time (ms) for classifying an image during testing.

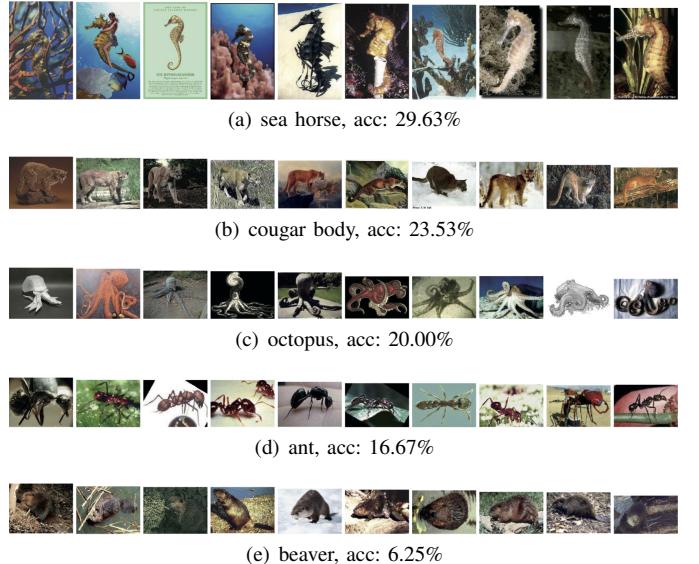


Fig. 8: The five categories with the lowest classification accuracy on Caltech 101.

randomly select 5, 10, 15, 20, 25, and 30 images per category for training and the remaining for testing. In Table V, we observe that the proposed method achieves better classification performance in all the situations except when using 15 and 30 training samples. The classification accuracy of DDLCN-2 are slightly worse than [66], while [66] adopts a three-layer model. In this case, [66] achieves a higher recognition accuracy since more stacked layers can enrich the ‘level’ of features as indicated in [67]. We also note that such CNN-based methods are supervised while our DDLCN is an unsupervised method. Moreover, DDLCN-3 further achieves competitive results compared with [66]. Currently, our method only uses one hand-crafted feature, one can always use more powerful image-level features such as CNN-based features for further improving the performance. Besides, we observe that DDLCN outperforms other methods when using only a few training samples such as 5 and 10 for each category, which is of great benefit when the training data is limited.

Next, we follow [14] and evaluate the performance of our method with different dictionary sizes. Specifically, we set $K=[510, 1020, 1530, 2040, 2550, 3060]$, respectively. Fig. 7(left) shows that the proposed DDLCN achieves better classification accuracy than state-of-the-art dictionary learning methods such as K-SVD, D-KSVD, SRC and LC-KSVD. In

TABLE V: Classification accuracy (%) on Caltech 101.

Number of Train. Samp.	5	10	15	20	25	30
KC [45]	-	-	-	-	-	64.14 ± 1.18
Griffin [64]	44.20	54.50	59.00	63.30	65.80	67.60
SRC [6]	48.80	60.10	64.90	67.70	69.20	70.70
D-KSVD [13]	49.60	59.50	65.10	68.60	71.10	73.00
K-SVD [4]	49.80	59.80	65.20	68.70	71.00	73.20
ScSPM [65]	-	-	67.00 ± 0.45	-	-	73.20 ± 0.54
LC-KSVD1 [14]	53.50	61.90	66.80	70.30	72.10	73.40
LLC [10]	51.15	59.77	65.43	67.74	70.16	73.44
LC-KSVD2 [14]	54.00	63.10	67.70	70.50	72.30	73.60
MBAP [38]	54.8	63.6	68.3	72.2	72.7	73.9
EasyDL [60]	-	-	68.40	-	-	-
MDDL-2 [23]	-	-	-	-	-	77.4
MDDL-3 [23]	-	-	-	-	-	77.6
Deep Convolutional Learning [66]	-	-	75.24	-	-	82.78
DDLCN-2 (1-1)	48.39 ± 1.29	56.77 ± 0.62	60.90 ± 0.61	64.34 ± 0.55	66.96 ± 0.40	68.75 ± 0.37
DDLCN-2 (31-30)	58.46 ± 0.84	67.26 ± 0.98	72.40 ± 0.59	76.53 ± 0.51	78.90 ± 0.42	80.16 ± 0.36
DDLCN-3 (31-30)	60.21 ± 0.62	70.12 ± 0.65	74.62 ± 0.48	77.92 ± 0.39	80.13 ± 0.26	81.98 ± 0.27

TABLE VI: Classification accuracy (%) on Caltech 256.

Num. of Train. Samp.	15	30	45	60
KC [45]	-	27.17 ± 0.46	-	-
LLC [10]	25.61	30.43	-	-
K-SVD [4]	25.33	30.62	-	-
D-KSVD [13]	27.79	32.67	-	-
LC-KSVD1 [14]	28.10	32.95	-	-
SRC [6]	27.86	33.33	-	-
Griffin [64]	28.30	34.10 ± 0.20	-	-
LC-KSVD2 [14]	28.90	34.32	-	-
ScSPM [65]	27.73 ± 0.51	34.02 ± 0.35	37.46 ± 0.55	40.14 ± 0.91
NDL [18]	29.30 ± 0.29	36.80 ± 0.45	-	-
SNDL [18]	31.10 ± 0.35	38.25 ± 0.43	-	-
MLCW [68]	34.10	39.90	42.40	45.60
LP- β [69]	-	45.8	-	-
M-HMP [70]	42.7	50.7	54.8	58.0
Convolutional Networks [71]	-	-	-	74.2 ± 0.3
VGG19 [72]	-	-	-	84.10
DDLCN-2 (1-1)	26.30 ± 0.40	31.45 ± 0.21	34.69 ± 0.31	37.76 ± 0.25
DDLCN-2 (15-15)	35.06 ± 0.26	41.26 ± 0.22	44.17 ± 0.35	47.48 ± 0.26
DDLCN-2 (30-30)	45.25 ± 0.31	51.64 ± 0.51	55.11 ± 0.26	59.66 ± 0.45
DDLCN-3 (30-30)	47.65 ± 0.22	54.28 ± 0.42	57.89 ± 0.32	62.42 ± 0.34

Fig. 7(right), we show the computational speed of classifying one test image using different dictionary sizes. We can see that our method is remarkably faster than SRC, and is very close to LC-KSVD. Lastly, when using 30 images per category for training, our method achieves 100% classification accuracy on 9 classes, i.e., accordion, car side, garfield, inline skate, metronome, minaret, okapi, snoopy and trilobite. We also present the five categories with the lowest classification accuracy in Fig. 8, which are sea horse, cougar body, octopus, ant and beaver. We can observe that the five categories are all moving animals with significant differences in shape, pose, color and background.

Caltech 256. We conduct extensive experiments using 15, 30, 45 and 60 training images per class and compare with state-of-the-art methods. Table VI shows the comparison results. We can see that the proposed DDLCN outperforms existing leading dictionary-based methods such as K-SVD, D-KSVD, LC-KSVD and LLC, which significantly validates the advantages of the proposed DDLCN.

Moreover, we observe that the proposed method achieves slightly worse results than both VGGNet [72] and convolutional network [71] when using 60 training samples. However, 1) [72] uses a very deep convolutional network, i.e., VGG19 [73], which consists of 16 convolutional layers and 3 fully connected layers. 2) Both [72] and [71] have limited practical applicability than our DDLCN since they rely on careful hyper-parameter selection. 3) The proposed approach

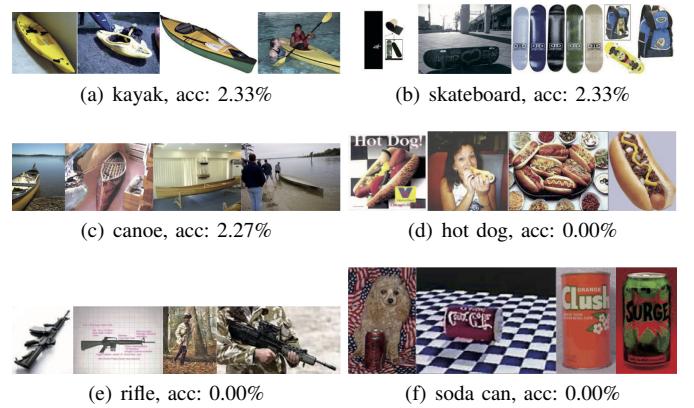


Fig. 9: The six categories with the lowest classification accuracy on Caltech 256.

has much fewer hyper-parameters that need to be tuned. 4) Compared with [72] and [71], the feature learner and encoder of the proposed DDLCN are fixed after extracting features, and only the linear SVM classifier on top is needed to update during training. Thus, the training of DDLCN is offline and its testing is pretty fast. All these represent the big advantages of the proposed DDLCN. Lastly, we show the six categories with the lowest classification accuracy in Fig. 9 when using 60 training images per category, which are kayak, skateboard, canoe, hot dog, rifle and soda can.

MNIST. The results on MNIST are shown in Table VII. We observe that the proposed DDLCN-2 consistently outperforms all the baselines except [66] when using the 500-500 strategy. Pu et al. [66] employs a three layers model to achieve a slightly better result (+1.03%) than us. The reason is that such CNN-based methods are jointly optimized between forward and backward propagation, while our DDLCN has no end-to-end tuning. Therefore, the training of our DDLCN is more efficient than such CNN-based models.

Moreover, when setting $p=1$ and $q=2$, the proposed model achieves 96.56% classification accuracy, which proves again that the proposed DDLCN can also achieve good performance when the number of training samples is limited and the size of the dictionary is small. This advantage is of great benefit

TABLE VII: Classification accuracy (%) on MNIST.

Method	Accuracy
Deep Representation Learning [74]	85.47
D-KSVD [13]	90.33
LC-KSVD [75]	92.58
SRC [6]	95.69
DCN [12]	98.15
TLCC [19]	98.57
Embed CNN [76]	98.50
Convolutional Clustering [77]	98.60
Deep Convolutional Learning [66]	99.58
DDLCN-2 (1-2)	96.56
DDLCN-2 (100-100)	98.55
DDLCN-2 (500-500)	99.02
DDLCN-3 (500-500)	99.35
DDLCN-4 (500-500)	99.47
DDLCN-5 (500-500)	99.54
DDLCN-6 (500-500)	99.57

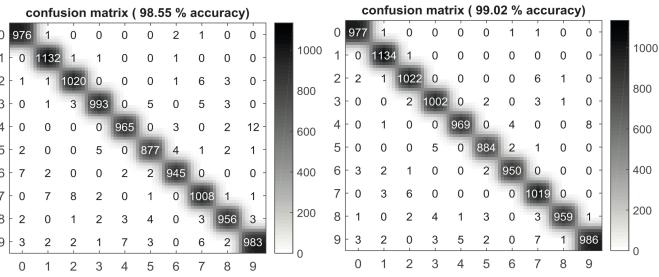


Fig. 10: Confusion matrix on MNIST.

in practical applications when the training data is limited. Also, we observe that when $p=100$ and $q=100$, the proposed method achieves 98.55% classification accuracy. We also show the confusion matrix of each category in Fig. 10(left)). When setting $p=500$ and $q=500$, the performance is further boosted. Specifically, we achieve 99.02% classification accuracy on this dataset. Fig. 10(right) lists the confusion matrix of each category under this experimental setting, and we can see that the most confusing pairs are (2, 7), (4, 9) and (3, 5).

Finally, we note that the classification accuracy of our hierarchical DDLCN increases by adding more layers. For example, the performance of DDLCN-3 achieves better results than the shallow one, i.e., DDLCN-2. To further explore the upper bound of our method, we conduct experiments on MNIST using DDLCN-4, DDLCN-5 and DDLCN-6. Specifically, DDLCN-4, DDLCN-5 and DDLCN-6 adopting 4, 5 and 6 the proposed compound dictionary learning and coding layers, respectively, and keeping all other details the same. Results are reported in Table VII. We observe that as the number of proposed dictionary learning and coding layers increases, the classification accuracy also gradually improves. However, when the number of layers increases to a certain number ('6' in this case), the classification performance saturates, which can be also observed in deep CNN models [72], [71].

V. CONCLUSION

In this paper, we aim to improve the deep representation capability of dictionary learning. To this end, we propose a novel deep dictionary learning method, i.e., DDLCN, to learn multi-layer deep dictionaries, which combines the advantages of both deep learning and dictionary learning approaches,

and achieves promising performance. Moreover, we propose a novel dictionary learning and coding layer and use it to substitute traditional convolutional layers in CNNs. Extensive experimental results on five public datasets with limited training data show that our DDLCN outperforms leading dictionary learning methods and achieves competitive results compared with state-of-the-art CNN-based models.

Acknowledgments. This work is partially supported by National Natural Science Foundation of China (No.U1613209, 61673030), National Natural Science Foundation of Shenzhen (No.JCYJ20190808182209321), and by the Italy-China collaboration project TALENT.

REFERENCES

- [1] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *JRSS*, vol. 58, no. 1, pp. 267–288, 1996. [1, 3](#)
- [2] D. L. Donoho, “Compressed sensing,” *IEEE TIT*, vol. 52, no. 4, pp. 1289–1306, 2006. [1](#)
- [3] K. Engan, S. O. Aase, and J. Hakon Husoy, “Method of optimal directions for frame design,” in *ICASSP*, 1999. [1](#)
- [4] M. Aharon, M. Elad, and A. Bruckstein, “K-svd: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE TSP*, vol. 54, no. 11, pp. 4311–4322, 2006. [1, 2, 7, 8, 9](#)
- [5] H. Liu, H. Tang, W. Xiao, Z. Guo, L. Tian, and Y. Gao, “Sequential bag-of-words model for human action classification,” *CAAI TIT*, vol. 1, no. 2, pp. 125–136, 2016. [1](#)
- [6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE TPAMI*, vol. 31, no. 2, pp. 210–227, 2009. [1, 7, 8, 9, 10](#)
- [7] H. Tang and H. Liu, “A novel feature matching strategy for large scale image retrieval,” in *IJCAI*, 2016. [1](#)
- [8] J. Mairal, “Sparse coding for machine learning, image processing and computer vision,” Ph.D. dissertation, École normale supérieure de Cachan-ENS Cachan, 2010. [1](#)
- [9] K. Yu, T. Zhang, and Y. Gong, “Nonlinear learning using local coordinate coding,” in *NeurIPS*, 2009. [1, 2, 4, 5](#)
- [10] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained linear coding for image classification,” in *CVPR*, 2010. [2, 4, 6, 7, 8, 9](#)
- [11] K. Yu and T. Zhang, “Improved local coordinate coding using local tangents,” in *ICML*, 2010. [2, 4](#)
- [12] Y. Lin, Z. Tong, S. Zhu, and K. Yu, “Deep coding network,” in *NeurIPS*, 2010. [2, 4, 5, 10](#)
- [13] Q. Zhang and B. Li, “Discriminative k-svd for dictionary learning in face recognition,” in *CVPR*, 2010. [2, 7, 8, 9, 10](#)
- [14] Z. Jiang, Z. Lin, and L. S. Davis, “Learning a discriminative dictionary for sparse coding via label consistent k-svd,” in *CVPR*, 2011. [2, 7, 8, 9](#)
- [15] S. Tariyal, A. Majumdar, R. Singh, and M. Vatsa, “Greedy deep dictionary learning,” *arXiv preprint arXiv:1602.00203*, 2016. [2](#)
- [16] Y. Liu, Q. Chen, W. Chen, and I. J. Wassell, “Dictionary learning inspired deep network for scene recognition,” in *AAAI*, 2018. [2](#)
- [17] I. Y. Chun and J. A. Fessler, “Convolutional dictionary learning: Acceleration and convergence,” *IEEE TIP*, vol. 27, no. 4, pp. 1697–1712, 2018. [2](#)
- [18] J. Hu and Y.-P. Tan, “Nonlinear dictionary learning with application to image classification,” *PR*, vol. 75, pp. 282–291, 2018. [2, 6, 9](#)
- [19] W. Xiao, H. Liu, H. Tang, and H. Liu, “Two-layers local coordinate coding,” in *CCCV*, 2015. [2, 6, 10](#)
- [20] Z. Zhang, W. Jiang, J. Qin, L. Zhang, F. Li, M. Zhang, and S. Yan, “Jointly learning structured analysis discriminative dictionary and analysis multiclass classifier,” *IEEE TNNLS*, vol. 29, no. 8, pp. 3798–3814, 2017. [2, 6, 8](#)
- [21] H. V. Nguyen, H. T. Ho, V. M. Patel, and R. Chellappa, “Dash-n: Joint hierarchical domain adaptation and feature learning,” *IEEE TIP*, vol. 24, no. 12, pp. 5479–5491, 2015. [2](#)
- [22] W. Dong, Z. Yan, X. Li, and G. Shi, “Learning hybrid sparsity prior for image restoration: Where deep learning meets sparse coding,” *arXiv preprint arXiv:1807.06920*, 2018. [2](#)
- [23] J. Song, X. Xie, G. Shi, and W. Dong, “Multi-layer discriminative dictionary learning with locality constraint for image classification,” *PR*, vol. 91, pp. 135–146, 2019. [2, 3, 6, 7, 8, 9](#)

- [24] S. Mahdizadehaghdam, A. Panahi, H. Krim, and L. Dai, "Deep dictionary learning: A parametric network approach," *IEEE TIP*, 2019. 2, 7
- [25] H. Tang, H. Wei, W. Xiao, W. Wang, D. Xu, Y. Yan, and N. Sebe, "Deep micro-dictionary learning and coding network," in *WACV*, 2019. 3, 6
- [26] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE TSP*, vol. 41, no. 12, pp. 3397–3415, 1993. 3
- [27] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Taylor & Francis IJC*, vol. 50, no. 5, pp. 1873–1896, 1989. 3
- [28] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001. 3
- [29] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm," *IEEE TSP*, vol. 45, no. 3, pp. 600–616, 1997. 3
- [30] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *JRSS*, vol. 68, no. 1, pp. 49–67, 2006. 4
- [31] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *JMLR*, vol. 9, pp. 1179–1225, 2008. 4
- [32] V. Roth and B. Fischer, "The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms," in *ICML*, 2008. 4
- [33] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," *JMLR*, vol. 12, pp. 3371–3412, 2011. 4
- [34] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004. 4
- [35] L. Shen, G. Sun, Q. Huang, S. Wang, Z. Lin, and E. Wu, "Multi-level discriminative dictionary learning with application to large scale image classification," *IEEE TIP*, vol. 24, no. 10, pp. 3109–3123, 2015. 4
- [36] J. Yang and M.-H. Yang, "Top-down visual saliency via joint crf and dictionary learning," *IEEE TPAMI*, vol. 39, no. 3, pp. 576–588, 2017. 4
- [37] S. Kim, R. Cai, K. Park, S. Kim, and K. Sohn, "Modality-invariant image classification based on modality uniqueness and dictionary learning," *IEEE TIP*, vol. 26, no. 2, pp. 884–899, 2017. 4
- [38] C. Bao, H. Ji, Y. Quan, and Z. Shen, "Dictionary learning for sparse coding: Algorithms and convergence analysis," *IEEE TPAMI*, vol. 38, no. 7, pp. 1356–1369, 2016. 4, 6, 7, 8, 9
- [39] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "Stat: spatial-temporal attention mechanism for video captioning," *IEEE TMM*, 2019. 4
- [40] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *ICML*, 2009. 4
- [41] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE TPAMI*, vol. 37, no. 9, pp. 1904–1916, 2015. 5
- [42] C. Cortes and V. Vapnik, "Support-vector networks," *ML*, vol. 20, no. 3, pp. 273–297, 1995. 6
- [43] H. Tang, H. Liu, and W. Xiao, "Gender classification using pyramid segmentation for unconstrained back-facing video sequences," in *ACM MM*, 2015. 6
- [44] H. Tang, H. Liu, W. Xiao, and N. Sebe, "Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion," *Neurocomputing*, vol. 331, pp. 424–433, 2019. 6
- [45] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders, "Kernel codebooks for scene categorization," in *ECCV*, 2008. 6, 9
- [46] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM TIST*, vol. 2, no. 3, p. 27, 2011. 6
- [47] N. Akhtar, A. Mian, and F. Porikli, "Joint discriminative bayesian dictionary and classifier learning," in *CVPR*, 2017. 6, 7, 8
- [48] S. Wu, Y. Yan, H. Tang, J. Qian, J. Zhang, and X.-Y. Jing, "Structured discriminative tensor dictionary learning for unsupervised domain adaptation," *arXiv preprint arXiv:1905.04424*, 2019. 6
- [49] C. Yan, L. Li, C. Zhang, B. Liu, Y. Zhang, and Q. Dai, "Cross-modality bridging and knowledge transferring for image understanding," *IEEE TMM*, 2019. 6
- [50] A. S. Georgiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE TPAMI*, vol. 23, no. 6, pp. 643–660, 2001. 6
- [51] A. M. Martinez, "The ar face database," *CVC TR*, vol. 24, 1998. 6
- [52] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *CVIU*, vol. 106, no. 1, pp. 59–70, 2007. 6
- [53] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," *CIT Technical Report*, 2007. 6
- [54] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998. 6
- [55] D. Wang and S. Kong, "A classification-oriented dictionary learning model: Explicitly learning the particularity and commonality across categories," *PR*, vol. 47, no. 2, pp. 885–898, 2014. 7, 8
- [56] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *IJCV*, vol. 109, no. 3, pp. 209–232, 2014. 7, 8
- [57] N. Akhtar, F. Shafait, and A. Mian, "Discriminative bayesian dictionary learning for classification," *IEEE TPAMI*, vol. 38, no. 12, pp. 2374–2388, 2016. 7, 8
- [58] Y. Yankelevsky and M. Elad, "Structure-aware classification using supervised dictionary learning," in *ICASSP*, 2017. 7
- [59] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representation using adversarial training," in *NeurIPS*, 2016. 7
- [60] Y. Quan, Y. Xu, Y. Sun, Y. Huang, and H. Ji, "Sparse coding for classification via discrimination ensemble," in *CVPR*, 2016. 7, 9
- [61] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized robust coding for face recognition," *IEEE TIP*, vol. 22, no. 5, pp. 1753–1766, 2013. 7, 8
- [62] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcanet: A simple deep learning baseline for image classification?" *IEEE TIP*, vol. 24, no. 12, pp. 5017–5032, 2015. 7, 8
- [63] S. Nikan and M. Ahmadi, "Local gradient-based illumination invariant face recognition using local phase quantisation and multi-resolution local binary pattern fusion," *IET IP*, vol. 9, no. 1, pp. 12–21, 2015. 8
- [64] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," *California Institute of Technology*, 2007. 9
- [65] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR*, 2009. 9
- [66] Y. Pu, X. Yuan, and L. Carin, "A generative model for deep convolutional learning," in *ICLR Workshop*, 2015. 8, 9, 10
- [67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 8
- [68] S. R. Fanello, N. Noceti, C. Ciliberto, G. Metta, and F. Odone, "Ask the image: supervised pooling to preserve feature locality," in *CVPR*, 2014. 9
- [69] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *ICCV*, 2009. 9
- [70] L. Bo, X. Ren, and D. Fox, "Multipath sparse coding using hierarchical matching pursuit," in *CVPR*, 2013. 9
- [71] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014. 9, 10
- [72] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *ICCV*, 2015. 9, 10
- [73] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015. 9
- [74] S. Yang, P. Luo, C. C. Loy, K. W. Shum, X. Tang *et al.*, "Deep representation learning with target coding," in *AAAI*, 2015. 10
- [75] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *IEEE TPAMI*, vol. 35, no. 11, pp. 2651–2664, 2013. 10
- [76] J. Weston, F. Ratle, H. Mobahi, and R. Collobert, "Deep learning via semi-supervised embedding," *Neural Networks: Tricks of the Trade*, pp. 639–655, 2012. 10
- [77] A. Dundar, J. Jin, and E. Culurciello, "Convolutional clustering for unsupervised learning," in *ICLR Workshop*, 2016. 10