SLED: Semantic Label Embedding Dictionary Representation for Multilabel Image Annotation

Xiaochun Cao, Senior Member, IEEE, Hua Zhang, Xiaojie Guo, Member, IEEE, Si Liu, and Dan Meng, Senior Member, IEEE

Abstract—Most existing methods on weakly supervised image annotation rely on jointly unsupervised feature representation, the components of which are not directly correlated with specific labels. In practical cases, however, there is a big gap between the training and the testing data, say the label combination of the testing data is not always consistent with that of the training. To bridge the gap, this paper presents a semantic label embedding dictionary representation that not only achieves the discriminative feature representation for each label in the image, but also mines the semantic co-occurrence between co-occurrence labels for context information. More specifically, to enhance the discriminative representation of labels, the training data is first divided into a set of overlapped groups by graph shift based on the exclusive label graph. Afterward, given a group of exclusive labels, we try to learn multiple label-specific dictionaries to explicitly decorrelate the feature representation of each label. A joint optimization approach is proposed according to the Fisher discrimination criterion for seeking its solution. Then, to discover the context information hidden in the co-occurrence labels, we explore the semantic relationship between visual words in dictionaries and labels in a multitask learning way with respect to the reconstruction coefficients of the training data. In the annotation stage, with the discriminative dictionaries and exclusive label groups as well as a group sparsity constraint, the reconstruction coefficients of a test image can be easily obtained. Finally, we introduce a label propagation scheme to compute the score of each label for the test image based on its reconstruction coefficients. Experimental results on three challenging data sets demonstrate that our proposed method leads to significant performance gains over existing methods.

I. INTRODUCTION

WITH the popularity of photo and video sharing websites (Flickr, Instagram, and YouTube, etc.), a large number of weakly labeled or unlabeled visual data are spread. The growing scale of visual data requires an effective retrieval mechanism to obtain the content of such data. Even though the visual search has been studied for years, the search engines retrieve relevant images still mainly based on textual queries instead of raw images. Consequently, the automatic image annotation that associates images with human-provided keywords or labels has received much research interest. A variety of algorithms [1]–[16] have been proposed for this task with promising results, which can be roughly divided into two groups, i.e. the parametric and non-parametric methods.

The parametric methods [2], [4], [5], [8]–[14], [17] treat the image annotation as a multi-label classification problem. These methods are mostly based on the BoW paradigm [18] that is composed of local feature extraction, dictionary learning, feature encoding, pooling and the classifier training. In general, visual low-level features, e.g. SIFT [19] and HOG [20], are initially extracted from each image. Then the dictionary is constructed by either k-means or sparse coding [9]–[11], [17], [21]. To capture the final representations of the images, a max or mean pooling strategy is employed to cluster the low-level features. Based on the representations, several classifiers are created using machine learning techniques such as SVM [22] for annotating the test images. The performances of created classifiers based image annotation methods are highly depend on the training data. It would be sharply degraded when the testing data has inconsistent label combinations with the training data. An example is shown in Fig. 1. Although both training and testing data contain label “Bear”, they have different color histogram representations as shown in the second row of Fig. 1(a) and (b). The different features between these images are caused by the intra-class variations and the distinct background, which would further influence the performance of label prediction.

The non-parametric methods [1], [3], [6] consider the image annotation as a sparse reconstruction problem, which is based on the reconstruction coefficients. In this scheme, the feature representations of training images are directly selected as the basis term. Then the labels of training data are propagated to the test data according to the reconstruction coefficients. However, the main limitation of such kind of methods is that
In each exclusive group, we train the label-specific dictionaries by introducing the Fisher discriminant criterion [23] as a regularization term. Afterwards, we obtain the discriminative dictionaries and the corresponding label of each dictionary item as shown in Fig. 2(c). Inspired by the observation that co-occurrence labels would provide the context information [24] e.g. “Waterfall” → “Water” and “Whale” → “Water”, we propose a multi-task learning based method to exploit the semantic correlation between the dictionary items and the labels. Furthermore, we add the context information into the original dictionary labels based on the correlations as shown in Fig. 2(e). Finally, with the reconstruction coefficients and dictionary labels, the test image is annotated through a robust label propagation method based on the RankSVM [25] as shown in Fig. 2(f) and (g). To demonstrate the efficacy of our method, extensive experiments are conducted on several image annotation benchmarks, including NUS-WIDE-Lite [26], Corel 5K [12] and IAPR-TC12 [2] datasets.

The main contributions of this paper can be summarized as: (a) We propose a novel semantic label embedding dictionary (SLED) method to get the discriminative image representation in the weakly supervised setting. (b) We propose a label expansion method based on the label correlation learning to integrate the label correlation into our reconstruction based framework. (c) A novel image annotation framework based on the exclusive label groups and label propagation is proposed.

The rest of this paper is organized as follows. In Sec. II, we review the most relevant works on dictionary learning and image annotation. The method to extract exclusive group of labels is discussed in Sec. III-A. In Sec. III-B and Sec. III-C, we present the exclusive dictionary learning algorithm and dictionary label expansion method, respectively. The approach for image annotation based on group regularized dictionary learning is displayed in Sec. III-D, including training the label propagation model. The optimization of our proposed method is discussed in Sec. III-E. After that we introduce the process of applying our model on the test images in Sec. III-F. The experimental setup and results are given in Sec. IV. We conclude our proposed method in Sec. V.

II. RELATED WORK

In this section, we briefly present a review on existing image annotation methods.

A. Image Annotation

Automatic image annotation aims to assign images with human predefined labels, which is usually viewed as a typical multi-label learning problem. The existing methods can be roughly divided into three categories including classification-based [4], [5], probabilistic modeling-based [14], [27], and reconstruction-based methods [3], [6]. The classification based scheme annotates the images by training the semantic

Fig. 1. Motivation of our proposed method. When the training and the testing images do not have the consistent label combinations (the first row of (a) and (b)), their original features (RGB histograms) are significantly different even they contain the same label “bear” (the second row of (a) and (b)). In the proposed method SLED (the third row of (a) and (b)), the visual words are related to a label have a higher response value. For example, the terra-cotta bins in third rows (reconstruction coefficients) that refer to label “Bear”, have a higher response comparing to those of unrelated labels. (a) Training images for Bear. (b) Testing images for Bear.

In each exclusive group, we train the label-specific dictionaries by introducing the Fisher discriminant criterion [23] as a regularization term. Afterwards, we obtain the discriminative dictionaries and the corresponding label of each dictionary item as shown in Fig. 2(c). Inspired by the observation that co-occurrence labels would provide the context information [24] e.g. “Waterfall” → “Water” and “Whale” → “Water”, we propose a multi-task learning based method to exploit the semantic correlation between the dictionary items and the labels. Furthermore, we add the context information into the original dictionary labels based on the correlations as shown in Fig. 2(e). Finally, with the reconstruction coefficients and dictionary labels, the test image is annotated through a robust label propagation method based on the RankSVM [25] as shown in Fig. 2(f) and (g). To demonstrate the efficacy of our method, extensive experiments are conducted on several image annotation benchmarks, including NUS-WIDE-Lite [26], Corel 5K [12] and IAPR-TC12 [2] datasets.

The main contributions of this paper can be summarized as: (a) We propose a novel semantic label embedding dictionary (SLED) method to get the discriminative image representation in the weakly supervised setting. (b) We propose a label expansion method based on the label correlation learning to integrate the label correlation into our reconstruction based framework. (c) A novel image annotation framework based on the exclusive label groups and label propagation is proposed.

The rest of this paper is organized as follows. In Sec. II, we review the most relevant works on dictionary learning and image annotation. The method to extract exclusive group of labels is discussed in Sec. III-A. In Sec. III-B and Sec. III-C, we present the exclusive dictionary learning algorithm and dictionary label expansion method, respectively. The approach for image annotation based on group regularized dictionary learning is displayed in Sec. III-D, including training the label propagation model. The optimization of our proposed method is discussed in Sec. III-E. After that we introduce the process of applying our model on the test images in Sec. III-F. The experimental setup and results are given in Sec. IV. We conclude our proposed method in Sec. V.

II. RELATED WORK

In this section, we briefly present a review on existing image annotation methods.

A. Image Annotation

Automatic image annotation aims to assign images with human predefined labels, which is usually viewed as a typical multi-label learning problem. The existing methods can be roughly divided into three categories including classification-based [4], [5], probabilistic modeling-based [14], [27], and reconstruction-based methods [3], [6]. The classification based scheme annotates the images by training the semantic
Fig. 2. The overview of our proposed approach on image annotation. Exclusive label groups (b) are first discovered based on the ground truth label of training data (a) as described in Sec. III-A. Then the discriminative dictionary (c) is trained for each exclusive group, as discussed in Sec. III-B. Then, label expansion (d) is introduced to fuse correlation between the co-occurrence of labels (Sec. III-C). In the test stage, image reconstruction (e) with group regularization is proposed to compute the reconstruction coefficients of the test image as described in Sec. III-D. Then the label propagation method (f) is used to compute the score of each label for the test image and the results are shown in (g). Note that the dictionary label in (f) is derived from (d).

label classifiers. In [5], Cusano et al. conquer the annotation problem by solving a multi-classification problem. Similarly, Carneiro et al. [4] propose to annotate the images without prior segmentation for each class by estimating the corresponding semantic class probability map. The main limitation of classification based method is that they need the supervised label information of the training images to train the classification model. The probabilistic modeling-based methods focus on modeling the distribution of each label based on the visual features and attempt to infer the correlation or joint probability between images and annotation keywords. Ueda and Saito [14] propose a generative model for multi-label leaning that explicitly incorporates the pairwise correlation between any two class labels. On the other hand in [27], a Bayesian model is introduced to assign labels based on underlying label representations. However, the computation efficient and accuracy are two limitations for modeling the semantic distributions. The third category makes use of the sparse reconstruction framework to accomplish this task. In [3], Wang et al. employ the sparse coding framework to get the coefficients of reconstructions. Based on the non-zero coefficients, the labels of reference images can thus be transferred to the test image. Furthermore, Chen et al. [6] introduce a prior about the exclusive labels to automatically annotate the images. One of the drawback is that the existed methods are based on the raw low-level features, which cause the annotation results to be sensitive to the noisy feature representation.

B. Dictionary Learning

Recent prevailing work on dictionary learning can be generally classified into two categories, say unsupervised and supervised dictionary learning. The former one focuses on simultaneously minimizing the reconstruction error and the residual error. In [28], the K-SVD and K-means clustering techniques are used to learn an over-complete dictionary from image patches. Lee et al. [29] treat the dictionary learning problem as a least squares problem, and solve it by an iterative algorithm to minimize the reconstruction error. Wright et al. [30] directly employ the training samples of the whole training set as the dictionary, and then they compute the least residual errors for face recognition. The dictionaries generated via unsupervised learning is expected to be of low reconstruction error and better generalization. Due to the lack of the label information, these methods have limited discriminative power.

Alternatively, the supervised dictionary learning aims for discriminative representation by embedding the information of labels into the dictionary. The existing supervised dictionary learning approaches can be roughly divided into three main categories regarding the structure of dictionaries. The first category is based on learning multiple dictionaries to promote
the discrimination among different categories [9], [31]–[33]. In [31] and [32], they propose to learn a dictionary for each class, and then the test images are classified based on the reconstruction errors of the corresponding classes. Zhou et al. [9] learns multiple dictionaries for visually correlated object category based on the Fisher discrimination criterion. In [33], they propose to wrap the dictionary learning process inside a boosting procedure for learning multiple dictionaries. One drawback of this kind of category is its computational inefficiency, especially when the number of classes is relatively large. The thought of the second category is to learn a compact and discriminative dictionary by merging or selecting informative visual words from a larger dictionary. In [34], they develop to merge the visual dictionary visual words by considering the trade-off between the intra-class compactness and inter-class discrimination power. In [35] and [36], the dictionary is constructed through merging two visual words by maximizing the mutual information of class distributions. Qiu et al. [37] present an approach for dictionary learning of action attributes via information maximization based on Gaussian Process. The shortcoming of this category is that the results are easily influenced by the noise and the intra-class visual variation. The last kind of category incorporates the label information as a regularization term into the objective function [10], [11], [23]. Yang et al. [23] propose to form a structured dictionary with class labels via Fisher discriminative criterion. Jiang et al. [10], [11] simultaneously integrate the label consistent constraint, the reconstruction error and the classification error into one single objective function to achieve the goal. The numerical evaluations of the supervised dictionary learnings demonstrate the advantage of using the label training data. But it costs a high price to collect the fully supervised data, which is obviously impractical in real cases.

C. Discussion

Our work is fundamentally different from the previous studies on discriminative dictionary learning and image annotation. The key differences lie in the following aspects. First, considering that the supervised methods [3]–[6] need the fully labeled training data, our proposed method needs the weakly labeled training data by introducing the semantic feature representation. Then, to construct the semantic representation, we introduce a novel dictionary learning method, which is also under the weakly supervised setting instead of supervised learning [10], [11], [23]. Moreover, our dictionary learning method harnesses the exclusive label groups to overcome the lack of location information in a weakly supervised manner. In addition, our reconstruction based image annotation method integrates the label correlation information by exploring the semantic meaning of each visual word in the dictionary. Finally, the weighting scheme is introduced into the phase of label propagation to guarantee that the related labels achieve higher responses than the unrelated ones.

III. SLED METHOD

In this section we define some common notations used throughout this paper. Let \( I = \{I_1, \ldots, I_n\} \) be a set of training images, and \( X = [x_1, \ldots, x_n] \in \mathbb{R}^{m \times n} \) be the feature matrix corresponding to \( I \) where \( m \) is the dimension of feature representation and \( n \) is the number of the training images. \( C = [c_1, \ldots, c_m] \in \{0, 1\}^{p \times n} \) is the label indicating matrix, where \( p \) is the number of predefined labels. The elements of the label indicator \( c_i \) are set to 1 when the image \( x_i \) contains the corresponding labels, otherwise 0. Since we consider a multi-label annotation problem, it is possible that more than one elements in \( c_i \) are 1. The learned dictionary \( D = \{D_1, \ldots, D_g, \ldots, D_G\} \in \mathbb{R}^{m \times K} \) is composed by the group dictionary \( D_g \), where \( G \) and \( K \) are the number of exclusive groups and the dictionary size, respectively. Table I lists the main symbols in this paper and their definitions.

![Fig. 3. The training flowchart of our proposed method.](image)

**TABLE I DEFINITION OF MAIN SYMBOLS**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( I )</td>
<td>The set of training images</td>
</tr>
<tr>
<td>( X \in \mathbb{R}^{m \times n} )</td>
<td>The feature matrix of training set</td>
</tr>
<tr>
<td>( C \in {0, 1}^{p \times n} )</td>
<td>The label indicating matrix for training images</td>
</tr>
<tr>
<td>( D \in \mathbb{R}^{m \times K} )</td>
<td>The learned dictionary by training images</td>
</tr>
<tr>
<td>( S_f )</td>
<td>The exclusive label group</td>
</tr>
<tr>
<td>( A_f )</td>
<td>The reconstruction coefficients</td>
</tr>
<tr>
<td>( S_w )</td>
<td>The intra-class scatter matrix</td>
</tr>
<tr>
<td>( S_0 )</td>
<td>The inter-class scatter matrix</td>
</tr>
<tr>
<td>( L_f )</td>
<td>The labels of the visual words in the ( f )th dictionary</td>
</tr>
<tr>
<td>( L_{exp} )</td>
<td>The expansion label matrix</td>
</tr>
<tr>
<td>( \Theta \in \mathbb{R}^{K \times m} )</td>
<td>The correlation parameter</td>
</tr>
<tr>
<td>( W \in \mathbb{R}^{p \times K} )</td>
<td>The weighting parameter computed by the RankSVM</td>
</tr>
<tr>
<td>( I_f )</td>
<td>The index for one of the labels in ( c_i )</td>
</tr>
<tr>
<td>( a_{f, t} )</td>
<td>The reconstruction coefficients of one training image</td>
</tr>
<tr>
<td>( a_{f, st} )</td>
<td>The reconstruction coefficients of the test image</td>
</tr>
<tr>
<td>( y_f )</td>
<td>The set of images which contain label ( t )</td>
</tr>
<tr>
<td>( y_t )</td>
<td>The set of images which do not contain label ( t )</td>
</tr>
</tbody>
</table>
between the visual words and their corresponding labels by using concept of entropy [39]. Further, a multi-task learning framework [40] is used to expand the labels of visual words (Fig. 3(d)). We introduce the group sparsity based reconstruction framework to get the representation for the images, which is robust to the noise in the feature representation (Fig. 3(e)). Finally, to obtain the robust label propagation, the Rank SVM [25] is used to compute the weighting of different reconstruction coefficients for image annotation (Fig. 3(f)).

A. Exclusive Label Group Discovery

In the proposed method the detection of exclusive label groups using the training data is considered as a subgraph seeking problem [6]. We first construct the graph \( O = \{V, E\} \), where \( V = \{1, 2, ..., p\} \) is the node set of graph, which contains all the predefined labels and \( E \subseteq V \times V \) is the set of edges between labels, which requires that labels should not appear together in the same image. The exclusive label weighted matrix \( T \) associated with \( O \) is defined as: \( T_{i,j} = 1 \) if label \( i \) and label \( j \) never exist together in the same training image, and \( T_{i,j} = 0 \) otherwise. Afterwards, the graph shift algorithm [38] is employed to get the subgraphs to achieve the goal of exclusive label grouping.

Suppose that we extract \( G \) exclusive groups encoded as \( \{S_1, ..., S_g, ..., S_G\} \). Each \( S_g = \{p_1, p_2, p_3, ...\} \) contains the indices of the specific labels in that group. We impose that the exclusive groups can be overlapped. Please notice that if we use all the extracted groups to develop the discriminative dictionaries, two problem will be arisen, i.e. 1) the imbalance number of visual words distribution and 2) the computational efficiency of training the label specific dictionaries. To avoid these problems, we add two additional constrains to filter the redundant groups from the initial ones. Firstly, we propose the inter-group similarity restriction to decrease the number of groups. Specifically, if two groups have a significant amount of overlap labels (more than 60% of the smaller group), we merge them into a single bigger group. This process is iterated until no further merging occurs. Furthermore, if there exists co-occurrence labels between groups, we would delete the smaller group. Secondly, the computational efficiency of dictionary learning drives us to restrict the number of labels in each group to be less than a certain threshold, e.g. 40. This constraint is used in the step of group merging to prevent the group from growing too large.

B. Semantic Label Embedding Dictionary (SLED) Learning

In this part, we discuss how to embed the semantic label information into the dictionary representation to achieve the label-specific discriminative power for dictionaries. Since the exclusive groups are independent with each other, we could train the dictionary of each group in parallel. Given a group of exclusive labels \( S_q \), by following the traditional dictionary learning framework [23], we have:

\[
\begin{align*}
\{D_g, A_g\} = \arg\min_{i=1}^m \sum ||X_i^g - D_i^g A_i^g||_2^2 + \lambda ||A_i^g||_1 \\
+ \beta \Omega(A_g^1, A_g^2, ..., A_g^m),
\end{align*}
\]

where \( X_i^g \) is the low-level feature representation of training images for label \( i \) in the group \( g \). \( A_i^g \) is the reconstruction coefficient with respect to the dictionary of the \( i^{th} \) label \( D_i^g \). \( \Omega(\cdot) \) is the discriminative regularization term, and \( ||\cdot||_2 \) and \( ||\cdot||_1 \) indicates the Frobenius norm and the \( \ell_1 \) norm, respectively. \( \lambda \) is a scalar weight to balance the importance of the coefficient sparsity. \( \beta \) controls the trade-off between the reconstruction and discrimination.

For a discriminative label-specific dictionary, the discrimination regularization term \( \Omega(A^1_g, A^2_g, ..., A^m_g) \) should be designed to be not only discriminative between inter-categories but also compact within intra-categories. According to the Fisher discriminative criterion [9], [23], we can obtain the discriminative label-special dictionaries by simultaneously minimizing the intra-class scatter matrix as well as maximizing the inter-class scatter matrix. In detail, the definition of the intra-class matrix is given by:

\[
S_W = \sum_{j=1}^m \sum_{a_i \in A_i^g} (a_i - \mu_j)(a_i - \mu_j)^T,
\]

where \( \mu_j \) is the mean vector of the reconstruction coefficients of \( A_i^g \) and \( a_i \) is \( i^{th} \) column of the reconstruction coefficients \( A_i^g \). This term enforce the compact of the dictionary. While the inter-class matrix is defined as:

\[
S_B = \sum_{j=1}^m n_j (\mu_j - \mu)(\mu_j - \mu)^T,
\]

where \( n_j \) is the number of samples in label \( j \), \( \mu \) is the mean of \( A_g \) in the group. This term is proposed to guarantee the discrimination between different labels. So far, we can define the discriminative regularization term as:

\[
\Omega(A^1_g, A^2_g, ..., A^m_g) = tr(S_W) - tr(S_B),
\]

By plugging Eq. 4 into Eq. 2, we have the objective function for our dictionary learning model as:

\[
\{D_g, A_g, L_g\} = \arg\min_{i=1}^m [||X_i^g - D_i^g A_i^g||_F^2 + \lambda ||A_i^g||_1 \\
+ \beta (tr(S_W) - tr(S_B))],
\]

where \( L_g \) reflects the label of each dictionary item. From the discussion above, we can see that the dictionary learning with Fisher discriminative regularization has several advantages: 1) the sparsity constraint can make the created dictionary robust to the noise; 2) Fisher regularizer can help extract the discriminative and compact label-special dictionary by computing the intra-label and inter-label scatter matrix; and 3) the objective function Eq. 5 is convex and easy to optimize.

C. Feature Mining and Label Correlation Learning

Inspired by the observation that not all the visual words are equally discriminative for their corresponding labels, we attempt to mine the most discriminative visual words for each label based on the mutual information between the labels and the visual words. The dictionary \( D_g \) for the \( g^{th} \) group, its corresponding label \( L_g \), and reconstruction coefficients \( A_g \)
of each image based on the dictionary are determined by Eq. 5. The discriminative score \( S(t) \) is computed to measure relevance between the visual words and their corresponding labels, where \( t \) represent the visual word in the \( t^{th} \) dictionary. By using the concept of entropy [39], we define \( S(t) \) as:

\[
S(t) = 1 + \frac{\sum_c p(c|t) \cdot \log p(c|t)}{\log C},
\]

\[
p(c|t) = \frac{\sum_{j=1}^{N_f} F(t|I_j) \cdot p(c|I_j)}{\sum_{j=1}^{N_f} F(t|I_j)}.
\]

where \( I_j \) is the \( j^{th} \) image and \( N_f \) is the total number of images in the \( g^{th} \) group. \( F(t|I_j) \) represents the frequency of \( t^{th} \) dictionary visual words appearing in the image \( I_j \). We set \( p(c|I_j) = 1 \) if the label of \( I_j \) including \( c \) and 0 otherwise. \( p(c|t) \) means the probability of belonging to the class \( c \) given the dictionary items \( t \). A higher value of \( S(t) \) implies that the dictionary items \( t \) are more related with the its corresponding label. Based on the discriminative score, we select these visual words having higher relevance to the labels.

Even so, we may still lose the semantic correlations between the co-occurrence labels, e.g. “Bear” and “Grass”, “Bear” and “Water” that exist in different exclusive groups. Since each training image would have different label combinations as shown in Fig. 1, the correlation between co-occurrence labels should related to all the labels that exist in the query images. To compute the correlation, a trace regularization based multi-task learning framework [40] is adopted based on the reconstruction coefficients.

Given the dictionary \( D = \{D_1, ..., D_G\} \), the exclusive label groups and the low-level features of training data, we first compute the reconstruction coefficients \( U \) based on the group sparse reconstruction framework (discussed in Sec. III-D). Then \( U = [u_1, u_2, ..., u_N] \in \mathbb{R}^{K \times N} \) is the reconstruction coefficient of all the training samples. With the ground truth labels of training images denoted as \( C \), the correlation parameter \( \Theta \in \mathbb{R}^{K \times m} \) can be calculated by:

\[
\Theta = \arg\min_{\Theta} ||U - \Theta C||_F^2 + \eta||\Theta||_s,
\]

where \( \eta \) is the weight associated with \( ||\Theta||_s \), which is the trace norm (the sum of eigenvalues of \( \Theta \)). This term is to avoid the over-fitting on the expansion parameters. After we obtain the solution of Eq. 8, we set a threshold \( \delta \) for the robustness of expansion (\( \delta \) is determined by the validation dataset). The expansion label matrix \( L_{\text{exp}} \in \mathbb{R}^{m \times K} \) is obtained as follows:

\[
L_{\text{exp}}^{i,j} = \begin{cases} 
0, & |\Theta_i^{j}| < \delta \& L^{i,j} = 0, \\
1, & |\Theta_i^{j}| \geq \delta \& L^{i,j} = 1,
\end{cases}
\]

where \( L_{\text{exp}}^{i,j} \) is the value in the \( i^{th} \) row and \( j^{th} \) column of \( L_{\text{exp}} \). \( L = [L_1, ..., L_G] \) is the label of a dictionary visual word, which is computed by Eq. 5. \( \Theta_i^{j} \) represents the element of correlation matrix \( \Theta \). The visualization of proposed label expansion is shown in Fig. 4. From the visualization, we can find that different labels refer to the distinct number of visual words.

### D. Image Annotation Based on Sparse Coding

1) Image Feature Reconstruction Based on Group Regularization: Once the discriminative dictionaries have been selected by our proposed feature mining method. An intuitive way for annotating the image could be done effectively by making use of the non-zero reconstruction coefficients provided by different group dictionaries. The reasons of choosing reconstruction instead of training the classifiers is that reconstruction based method is more robust to the image feature variations especially caused by the different label combinations. In our setting, the nonzero reconstruction coefficients are considered semantically related to the query image. Based on the reconstruction coefficients, we could find that partial similarity image would share the similar nonzero coefficients. Therefore it could reconstruct the query image based on their partial similarity.

Given a test image \( x_q \in \mathbb{R}^{d \times 1} \), the objective function is:

\[
a_q = \arg\min_{a_q} ||x_q - Da_q||_2^2 + \alpha_1||a_q||_1 + \alpha_2 \sum_{g=1}^{G} ||a_g||_2, \quad (10)
\]

where \( a_q \) is the reconstruction coefficients of the input image and \( a_g \) denotes the reconstruction coefficient of the \( S_g \) group. \( \alpha_1 \) and \( \alpha_2 \) are the weighting parameters to control the sparsity of reconstruction coefficients.

The advantages of using group sparsity to reconstruct the image can be summarized into two aspects. Firstly, for minimizing the reconstruction error, the reconstruction coefficients should related to all the labels that exist in the query images. Secondly, only the most relevant groups will give the non-zero reconstruction coefficients, which leads to the sparsity solutions.

2) Robust Label Propagation: When the reconstruction coefficient \( a_q^{\text{test}} \) of the test image is obtained, we can transfer the dictionary label to the query. There exist two ways to
transfer the labels: an intuitive way is to sum the reconstruction coefficients for each label, then sort the labels based on the scores. However, since the discovered exclusive groups are overlapped, the number of dictionary items for each label is not even. This would generate the bias ranking for the labels. Furthermore, in this way each dictionary item is assumed to have the same contribution to its labels, which would further make the results sensitive to noise.

Considering the above-mentioned issues, we introduce the weighting scheme into the label propagation process. The structure SVM [25] is introduced to make the relevant label have a higher response score than others. Given the reconstruction coefficients of training images \( \mathbf{A} \) computed by Eq. 10, the dictionary label \( \mathbf{L}_{e_{x_p}} \) and the image labels \( \mathcal{C} \), we extract the index of dictionary visual words for label \( t \) denoted as \( \mathbf{id}_t \) by exploring the non-zero elements of the \( t \)-th row in \( \mathbf{L}_{e_{x_p}} \). Then the feature mapping for label \( t \) on the training data is represented as: 
\[
\Phi(\mathbf{A}^t, \mathbf{y}) = (0, ..., 0, \mathbf{A}_{\mathbf{id}_t}^t, 0, ..., 0),
\]
where \( \mathbf{y} \) represent the image and \( \mathbf{0} \) is a zero vector. We denote \( \mathbf{A}_{\mathbf{id}_t}^t \) the subvector of \( \mathbf{A} \) indexing by \( \mathbf{id}_t \). This subvector contains all the visual words related to the label \( t \). Under this feature map setting, each image \( I_i \) is divided into several parts based on its label \( c_i \). Finally, the multi-label weighting scheme is constructed as:
\[
\argmin_{\mathbf{w}, \mathbf{z}} \mathbf{w}^T \mathbf{w} + \rho \sum_t \zeta_t,
\]
\[
\forall t \quad \mathbf{w}^T (\Phi(\mathbf{A}^t, \mathbf{y}_t^i) - \Phi(\mathbf{A}^t, \mathbf{y}_t^i)) \geq \Delta(\mathbf{y}_t^i, \mathbf{y}_t^i) - \zeta_t,
\]
where \( \mathbf{w} \) is the weighting parameters and \( \mathbf{y}_t^i \) is the set of images which contains the specified label \( t \). While \( \mathbf{y}_t \) represents the other set of images that does not include label \( t \). \( \Delta(\mathbf{y}_t^i, \mathbf{y}_t^i) \) is the loss function which is set as Hamming loss in all the datasets. The reason of using hamming loss is that the objective function can be efficiently solved as described in [25]. \( \zeta_t \) indicates the slack variables. \( \rho \) is a constant that controls the trade-off between the training error and the max-margin.

### E. Optimization

In our approach, there exist two objective functions Eq. 5 and Eq. 10 need to optimize.

1) Optimization of Dictionary Learning: The procedure of the dictionary learning could be divided into two iteratively sub-procedures: updating the coefficients \( \mathbf{A}_{\mathbf{x}} \) by fixing dictionary \( \mathbf{D}_{\mathbf{x}} \), and updating the dictionary \( \mathbf{D}_{\mathbf{x}} \) by fixing coefficients \( \mathbf{A}_{\mathbf{x}} \). When the dictionary is fixed, the objective function becomes:
\[
E(\mathbf{A}_{\mathbf{x}}^t) = \arg\min_{\mathbf{A}_{\mathbf{x}}^t} ||\mathbf{X}_g - \mathbf{D}_{\mathbf{x}} \mathbf{A}_{\mathbf{x}}^t||_F^2 + \lambda ||\mathbf{A}_{\mathbf{x}}^t||_1 + \beta (tr(\mathbf{S}_W) - tr(\mathbf{S}_B)),
\]
which can be efficiently solved by TwLST algorithm [41].

2) Optimization of Reconstruction: When the coefficients \( \mathbf{A}_{\mathbf{x}} \) are fixed, the objective function turns to be:
\[
E(\mathbf{D}_{\mathbf{x}}^t) = \arg\min_{\mathbf{D}_{\mathbf{x}}^t} ||\mathbf{X}_g - \mathbf{D}_{\mathbf{x}}^t||_F^2
\quad s.t. ||d_j||_2^2 = 1 \quad \forall j = 1, 2, ..., K^t_g.
\]

### Algorithm 1 The Main Dictionary Learning Steps of Our Proposed Method

1. **Input:** The training data \( \mathbf{X} \in \mathbb{R}^{m \times n} \), \( \lambda, \beta \)
2. **Initialize** \( \mathbf{D}_{\mathbf{x}} \).
3. **repeat**
   4. **Update** reconstruction coefficients \( \mathbf{A}_{\mathbf{x}}^t \) according to Eq. 12 with \( \mathbf{D}_{\mathbf{x}} \) fixed.
   5. **Update** dictionary \( \mathbf{D}_{\mathbf{x}} \) according to Eq. 13 with \( \mathbf{A}_{\mathbf{x}} \) fixed.
   6. **until** convergence: the values of \( E(A_{\mathbf{x}}^t) \) in adjacent iterations are close enough, or reached the maximum number of iterations.
7. **Output:** Reconstruction coefficients \( \mathbf{A}_{\mathbf{x}} \), discriminative dictionary \( \mathbf{D}_{\mathbf{x}} \).

Since Eq. 13 is the convex function, it can be solved by the Lagrangian method. For clarity, the optimization procedure of dictionary learning is summarized in Algorithm 1.

2) Optimization of Reconstruction: The objective function is to reconstruct the query image, which is a classic group lasso problem [42] with the overlapped group. To optimize this objective, we adopt the method proposed in [43]. The effectiveness of this method has been demonstrated.

### F. Image Label Prediction

When the test image comes, the low-level features are first extracted, and then we obtain the reconstruction coefficients \( \mathbf{a}_q^{test} \) by Eq. 10. Based on the reconstruction coefficients and the weighting \( \mathbf{W} \), the score of prediction results \( \mathbf{S}_{pred} \) can be estimated by: 
\[
\mathbf{S}_{pred} = < \mathbf{W}, \Phi(\mathbf{a}_q^{test}, \mathbf{I}^{test}) > ,
\]
and the predict score \( \mathbf{S}_{pred} \) is sorted to obtain the final prediction label list.

### IV. EXPERIMENTS

We evaluate our proposed approach on three widely used image annotation datasets, including NUS-WIDE-LITE image set [26], Corel 5K [12], and IAPR-TC12 dataset [2] whose summarization is displayed in Table II.

**NUS-WIDE-LITE** [26]: This dataset consists of 55,615 images with each image being annotated by at most 15 labels downloaded from Flicker. We use the same setting as presented in [26] to split the dataset into the training part and testing part (27,807 for training and 27,808 for testing). There are, in total, 81 unique predefined labels in this dataset. Moreover, each label associates with about 580 images on average. The labels in this dataset have a wide range from scenes like "airport" and "night time" to animals e.g. "Whales" and "Zebra".

**Corel5K** [12]: It contains of 4,999 images collected from the larger Corel CD set. Moreover, following the same experimental setup, we select 4,500 images for training and the rest for testing. Each image is manually annotated by at most 5 labels and each label averagely exist in about 58 images. There are 260 different labels in this dataset.

**IAPR-TC12** [2]: This dataset consists of 19,627 images of natural scenes. Following the setting in [2], 17,665 images are selected for training and 1,962 images are chosen for testing.
The content of images includes different sports and actions, photographs of people, animals, cities, landscapes and many other aspects of contemporary life. The number of labels in this dataset is 291 with an average of 5 labels per image. Besides, each label averagely relates with 347 images.

A. Experimental Setup

1) Evaluation Metric: For fairness of comparison, we adopt the same evaluation metric as [12]. All the test images are annotated by the top 5 relevant labels, i.e. the top 5 labels with highest predicted score. Specially, for NUS-WIDE-LITE dataset, Average Precision (AP) for each label and Mean Average Precision (MAP) over all the labels are also selected as the criteria to evaluate the performance. For Corel5K and IAPR-TC12 datasets, we use the average precision (AP) and average recall (AR) over each label as the evaluation measurement.

2) Low-Level Feature: For the NUS-WIDE-LITE data, we use the same set of features as [26], including color histogram (64 D), color correlogram (144 D), edge direction histogram (73 D), wavelet texture (128 D), block-wise color moments (225 D) and bag of words based on SIFT (500 D). We simply concatenate all these low-level features to form a feature vector with 1,134 dimensions. For the Corel5K and IAPR-TC12, we adopt the feature representation from [2]. It provides 8 different visual descriptors which contains Gist descriptor (512 D), six global color histograms (4,196 D) and one local bag of visual words features (1,000 D) based on Dense SIFT. Each image is described by a 5,708 D feature vector.

B. Component-Wise Model Evaluation

To gain the insight into our proposed method, we conduct experiments to analyze the effect of each component. A validation dataset is collected from the NUS-WIDE-LITE by randomly selecting 1,000 images from the testing part. In general, four parts (dictionary size, label correlation learning, query image reconstruction, and label propagation) mainly influence the performance of our approach are evaluated.

1) Dictionary Size: In this part, we investigate how sensitive our method is to the dictionary size. Intuitively, increasing the dictionary size might give better results at the expense of increasing the computational cost. In our experimental setup, each label is set to include at most \( k \) visual words selected based on the discriminative score \( S(i) \) given by Eq. 7. In particular, 5 candidate dictionary sizes \{300, 500, 1,000, 3,000, 5,000\} are tested.

For better revealing the efficacy of different components, in the step of label propagation, the reconstruction coefficient \( A_q \) computed by Eq. 10 is directly combined with the dictionary label \( L \) obtained through Eq. 5 to calculate the score of each label for the test. We plot the overall label annotation performance (AP and AR) across different dictionary sizes in Fig. 5(a) and (b). We can observe that the performance does not always increase as the dictionary size grows. This demonstrates the effectiveness of the discriminative scores on visual words.

2) Label Correlation Learning: Here, we discuss the efficacy of embedding the semantic correlations into the dictionary labels, the impact of parameters \( \eta \) in Eq. 8 and \( \delta \) in Eq. 9. In detail, the value of \( \eta \) ranges from \( 10^{-3} \) to \( 10^{2} \), and \( \delta \) from 0.001 to 1. The dictionary size is selected as 300. In the step of label propagation, reconstruction coefficients \( A_q \) computed by Eq. 10 are directly combined with the dictionary label \( L_{cap} \) obtained through Eq. 9 to calculate the score of each label for the test image. The experimental results (AP and AR) are shown in Fig. 6(a) and (b). From the results, we observe that the best performance is achieved by \( \eta = 0.1 \) and \( \delta = 0.3 \). Furthermore, comparing the results between Fig. 6(a) and Fig. 5(a) on 300 visual words,
are two parameters of weighting scheme in the step of label propagation. There are fewer labels (strong sparsity, $\alpha_t$ introduced. In contrast, when the test image is related with more labels, weak sparsity, $\alpha_1 = 0.0001$ and $\alpha_2 = 0.0001$), the noisy labels are very likely to be introduced. In contrast, when the test image is related with fewer labels (strong sparsity, $\alpha_1 = 0.7$ and $\alpha_2 = 0.7$), because the less relevant labels having non-zero reconstruction coefficients and the precision and recall are both decreasing. According to these results, we find that the best performance appears at $\alpha_1 = 0.01$ and $\alpha_2 = 0.1$. Thus, we need to find proper reconstruction parameters to associate with the right number of labels based on validation dataset.

3) Reconstruction Parameters: In this part, we examine the influence of the parameters $\alpha_1$ and $\alpha_2$, which control the sparsity of reconstruction coefficients and further affect the number of labels relating to the test images, respectively. The dictionary size for this evaluation uses 500. We plot the results (AP and AR) of different combinations in Fig. 7(a) and (b). The results demonstrate that when the test image is related with more labels (weak sparsity, $\alpha_1 = 0.0001$ and $\alpha_2 = 0.0001$), the noisy labels are very likely to be introduced. In contrast, when the test image is related with fewer labels (strong sparsity, $\alpha_1 = 0.7$ and $\alpha_2 = 0.7$), because the less relevant labels having non-zero reconstruction coefficients and the precision and recall are both decreasing. According to these results, we find that the best performance appears at $\alpha_1 = 0.01$ and $\alpha_2 = 0.1$. Thus, we need to find proper reconstruction parameters to associate with the right number of labels based on validation dataset.

4) Weighting Parameters: We further examine the efficacy of weighting scheme in the step of label propagation. There are two parameters $\rho$, which controls the trade-off between the training error and the max-margin and the stop iteration error $e$ related with the weighting scheme. For the experimental setup $\rho$ ranges from 0.001 to 1,000 and $e$ is 3 values ($10^{-3}, 10^{-4}, 10^{-5}$). The dictionary size for this evaluation is set to 500 and the reconstruction and label expansion parameters are determined by the validation set. The experimental results are shown in Fig. 8(a) and (b). The best parameter combinations is $\rho = 100$ and $e = 0.001$. Comparing with the results in Fig. 6(a) and (b), we can observe that the weighting scheme can boost the performance.

5) Dictionary Visualization: Fig. 9 visualizes the dictionary basis via the training example images. Each row in Fig. 9 shows 5 example images whose reconstruction coefficients are closest to a basis (a column vector in the learned dictionary $D$). In detail, we set the dictionary size for each label equal to 50 for IAPR-TC12 dataset, and then the reconstruction coefficients computed by Eq. 10 are sorted based on their values. The examples suggest that the learned basis capture certain common properties across different training images. However, the limitation of our model is that it can not distinguish the visual similarity labels. For example, the “sea water” and the “sky” are easily confused based on their similarity as shown in the bottom row of Fig. 9.

Overall, the contributions of each component in our proposed framework can be observed from Figs. 5–8. Each term has different influences to the final results, therefore in all the following experiments, we determine the parameters of each component by using a 3-fold validation dataset.

C. Evaluation on NUS-WIDE-LITE

In this experiment, the dataset is first divided into training and testing part as proposed in [26]. And then 1,000 images are extracted from the training part to construct the validation dataset, which are used for determining the parameters of our method. Second, to train the discriminative dictionary, 64 exclusive label groups are extracted with average 12 labels and maximum 16 labels per group. Finally, the reconstruction parameters are set to $\alpha_1 = 0.01$ and $\alpha_2 = 0.01$, the label expansion parameters are $\eta = 0.01$ and $\delta = 10^{-4}$, and the weighting parameters $\rho = 100$. We test 3 dictionary sizes {300, 500, 1000}. Fig. 10 presents the comparison between our method and some state-of-the-art methods using the AP for each label. We extract six state-of-the-art algorithms: KNN [44] which finds the 5 nearest images, and then scores the labels based on their occurrence. SVM [22] by following one-vs-all method for each label and ranking the label score for each image. Label Co-occurrence Linear Representation (LCLR) and Label Exclusive Linear Representation (LELR) [6], LCLR based on the label co-occurrence to propagate the label to the test images. While LELR proposes to make the exclusive
Fig. 10. The comparison of APs for the 81 labels with six methods on NUS-WIDE-LITE is shown on the above two images. (a) We display the AP for the top 40 labels and (b) we show the AP for the rest labels. Best view in color and please zoom in for the clear comparisons.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.09</td>
<td>0.095</td>
<td>0.175</td>
<td>0.278</td>
<td>0.332</td>
<td>0.358</td>
</tr>
<tr>
<td>Methods</td>
<td>WODL(D=300)</td>
<td>WODL(D=500)</td>
<td>WODL(D=1,000)</td>
<td>WODL+SVML(D=300)</td>
<td>WODL+SVML(D=500)</td>
<td>WODL+SVML(D=1,000)</td>
</tr>
<tr>
<td>MAP</td>
<td>0.095</td>
<td>0.120</td>
<td>0.108</td>
<td>0.1114</td>
<td>0.1594</td>
<td>0.2133</td>
</tr>
<tr>
<td>Methods</td>
<td>WEDL(D=300)</td>
<td>WEDL(D=500)</td>
<td>WEDL(D=1,000)</td>
<td>SLED (D=300)</td>
<td>SLED (D=500)</td>
<td>SLED (D=1,000)</td>
</tr>
<tr>
<td>MAP</td>
<td>0.2775</td>
<td>0.1565</td>
<td>0.122</td>
<td>0.3026</td>
<td>0.2655</td>
<td>0.2034</td>
</tr>
</tbody>
</table>

labels should not appear in the same image. Entropic Graph Semi-supervised classification (EGSSC) [13] which propose to embed the image similarity by constructing the graph. Large scale multi-label propagation (LSMP) [1] is to construct a probability model to model the correlation between labels. Table III shows the comparison results of our method with above state-of-the-art algorithms including the performance of each important component in the step of label propagation. Specifically, the label prediction score is computed by combining the reconstruction coefficients with the Original Dictionary Label (WODL). The experimental results are obtained by combining the reconstruction coefficients with the Expansion Dictionary Label (WEDL). Moreover, WODL + SVM indicates that we add the weighting scheme
based on rank SVM to compute the score of each label. Finally, SLED is used to represent the whole framework.

Several conclusions can be drawn from the results. 1) It is clear that our proposed method outperforms the other algorithms that directly use the jointly feature representations instead of label embedding dictionary representation. LCLR and LE LR achieve 23.2% and 25.8% in terms of MAP, while our method obtains 30.26%. 2) Comparing the results between WODL and WEDL, there exists a significant improvement especially on D = 300. This can demonstrate the effectiveness of our proposed label expansion on the image annotation task. While the results on introducing the weighting scheme in label propagation give about 3% improvement on unexpanded dictionary label and average 7% improvement on label expansion one. 3) we also can find that different dictionary sizes generate distinct precisions. The best performance exists for D = 300. This also tell us that a suitable size dictionary is important to achieve a better performance.

There may be three reasons to explain our superior performance. Firstly, the dictionary representation we use is able to decorrelate the representation of each label, and is more robust to the noise and less rely on the training data. Secondly, the semantic correlations between co-occurrence labels are explored by the multi-task learning to expand the dictionary labels. The expansion dictionary label can bring more information in the step of label propagation. Thirdly, the weighting scheme is introduced based on the degree of correlation between dictionary items and labels, which further helps determine the contribution of each visual word for its corresponding labels.

D. Evaluation on Corel5K

Compared with the NUS-WIDE-LITE, Corel5K contains more labels (260) and fewer images (4,999), which can generate a strong correlation between labels. In this part, we want to check if our method is able to adopt to the large scale labels situation. Following the standard evaluation procedure, the dataset is divided into training and testing part and we randomly extract 300 images from training data to develop a validation dataset. The dictionary sizes in this experiment are selected as \{100, 300, 500, 700\}. While $\alpha_1 = 0.01$ and $\alpha_2 = 0.05$ are used in the reconstruction objective function, and 100 exclusive groups are extracted from the training part. The label expansion parameters adopt $\delta = 0.0008$ and $\eta = 10$.

We evaluate our approach under different conditions compared with state-of-art approaches. Several competitors participate in the comparison, including Continuous-space Relevance Model (CRM) [45], an approach using Inference Network (InfNet) [46], a method based on Non-Parametric Density Estimation (NPDE) [47], Supervised Multi-class Labeling (SML) [4], Two-phrases Graph Learning Method (TGLM) [48], Multiple Bernoulli Relevance Models (MBRM) [53], Joint Equal Contribution (JEC) [49], L1-Penalized Logistic Regression (LSR) [49], Multi-label Sparse Coding (MSC) [3], TagProp [2], image annotation using Group Sparsity (GS) [42], Large Scale Max-Margin Multi-Label Classification (LM3L) [51], Learning Social Tag Relevance (TagRel) [52], Fast Image Tagging (FastTag) [54], Linear Sparse Reconstructions (LSR(n=2)) [50], and Linear Sparse Reconstructions using all the images ($\pi$ LSR(n=2)) [50]. Moreover, we also can test the different combinations of the components. In addition, we adopt average precision and average recall over all the test images instead of each label as the evaluation metrics.

The comparison results are shown in Table IV. Several observations can be drawn from the comparisons. First, Our proposed method consistently outperforms almost related approaches. Second, with merging the label correlation into the dictionary label, the performance is boosted from 15% to 32% which demonstrates the effectiveness of correlation in learning. Third, comparing with LASSO and GS that use the sparse reconstruction framework to annotate the images, our proposed method uses the label embedding feature representations to obtain a better result, which further proves the effectiveness of our proposed dictionary representation. Fourth, our method does not achieve a better results comparing with LSR(n=2), $\pi$ LSR(n=2) [50], the reason is that [50] only uses the top 2 labels instead of the top 5.

E. Evaluation on IAPR-TC12

Different from the above-mentioned datasets, the IAPR-TC12 dataset contains images covering a wide range of scenarios and viewpoints. Following the standard evaluation procedure, several low-level features are concatenated to describe the images. 201 exclusive groups are discovered and each group averagely contains 20 labels. The dictionary sizes in this data are selected as \{50, 100\}. The parameters of reconstruction are set as $\alpha_1 = 0.01$ and $\alpha_2 = 0.05$. Moreover, the label expansion parameters are $\delta = 0.017$ and $\eta = 10$.

We compare the performance of our method against several state-of-the-art approaches, i.e. MBRM [53], JEC [49], LASSO [49], TagProp [2], GS [42], FastTag [54]. The results of related work are copied from their papers. However, the results of MBRM are obtained from [49]. We further evaluate

<table>
<thead>
<tr>
<th>Methods</th>
<th>AP</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRM [45]</td>
<td>0.16</td>
<td>0.19</td>
</tr>
<tr>
<td>NPDE [47]</td>
<td>0.18</td>
<td>0.21</td>
</tr>
<tr>
<td>TGLM [50]</td>
<td>0.25</td>
<td>0.29</td>
</tr>
<tr>
<td>JEC [50]</td>
<td>0.27</td>
<td>0.32</td>
</tr>
<tr>
<td>LSR (n=2) [51]</td>
<td>0.37</td>
<td>0.39</td>
</tr>
<tr>
<td>LASSO [49]</td>
<td>0.34</td>
<td>0.29</td>
</tr>
<tr>
<td>TagProp [2]</td>
<td>0.33</td>
<td>0.42</td>
</tr>
<tr>
<td>LML [52]</td>
<td>0.32</td>
<td>0.51</td>
</tr>
<tr>
<td>WODL(D = 100)</td>
<td>0.14</td>
<td>0.20</td>
</tr>
<tr>
<td>WODL(D = 300)</td>
<td>0.15</td>
<td>0.21</td>
</tr>
<tr>
<td>WODL(D = 500)</td>
<td>0.14</td>
<td>0.20</td>
</tr>
<tr>
<td>WODL(D = 700)</td>
<td>0.14</td>
<td>0.20</td>
</tr>
<tr>
<td>WODL(D = 100) + SVM</td>
<td>0.23</td>
<td>0.36</td>
</tr>
<tr>
<td>WODL(D = 300) + SVM</td>
<td>0.20</td>
<td>0.24</td>
</tr>
<tr>
<td>WODL(D = 500) + SVM</td>
<td>0.18</td>
<td>0.25</td>
</tr>
<tr>
<td>WODL(D = 700) + SVM</td>
<td>0.17</td>
<td>0.23</td>
</tr>
</tbody>
</table>
how well our method is performed by calculating the MAP of individual components of our model. We further show the quantitative analysis by randomly selecting several images in this dataset as shown in Fig. 11.

The comparison results are shown in Table V, it can be found that our proposed method performs better than the related methods. Comparing these methods using the reconstruction framework i.e. GS and LASSO, our method achieves significant improvement, which demonstrates the effective of dictionary representation and label correlation learning. Comparing with these methods using label classifiers i.e. FastTag, we still obtain a better results by introducing the weighting scheme. The experimental results between with and without label expansion indicate the effect of this term in the step of label propagation. Moreover, weighting scheme plays an important role in determining the weighting of each visual words for different labels.

### F. Discussion

Our proposed image annotation method **SLED** achieves better results comparing with the state-of-the art algorithms on three datasets. Based on the comparison between our proposed method and [6] on the NUS-WISE-LITE dataset in Table III, our model achieves a significant improvement on mean average precision. This is because we introduce the semantic dictionary representation instead of directly using the raw features as [6]. From the comparison results between our approach and [51], [56] on Corel5K, the advantage of using semantic dictionary representation can be further proved. Moreover, comparing with the traditional label propagation methods [2], [52] using the image label directly, our proposed method is based on the labels of visual word and the RankSVM [25] to propagate the labels to the test images. We can find that our approach can achieve a better results as shown in Tables IV and V.

### G. Computational Complexity & Limitations

Our model is composed by two parts: off-line part and on-line part. For the off-line part, the computational complexity is caused by the dictionary learning, which is the alternative optimizations problem. The total time complexity $\nu(O(q^2 p^4) + \sum_i p_i O(2nq))$ as discussed in [56], where $\nu$ is the number of iterations, $O(q^2 p^4)$ denotes the complexity of updating the coefficients, and $\sum_i p_i O(2nq)$ is the time complexity of updating dictionary atoms. Moreover, $q$ is the feature dimensionality, $p$ is the number of dictionary atoms, and $n$ is the total number of training samples. For the online part, the computational complexity is caused by the group sparse reconstruction step, which is determined by the number of outer iterations (the gradient steps) and the total number of inner iterations (the steps required for computing the proximal operators) as discussed in [43]. And the complexity of inner iterations is $O(tg)$ where $t$ is the size of each group and $g$ is the number of groups.

However, our work has two limitations. First, the imbalance distribution of training images for each label would lead biased performance. Second, the dictionary size can not be automatically determined which impacts the performance and efficient of the whole framework.

### V. Conclusion

In this paper, we proposed a novel label embedding dictionary representation method, named **SLED**, to solve the problem of image annotation under a weakly supervised setting. Our main contribution lies in explicitly fusing the label information into dictionary representation and exploring the semantic correlations between co-occurrence labels. Additionally, based on the semantic dictionary representation and the exclusive label group, our model can solve the problem of inconsistent label combinations between training and testing data. Furthermore, the solution to our proposed method can be efficiently achieved by using the alternating optimization algorithm. The experimental results show that our approach yields very good image annotation results on three well-known public datasets. Our method outperforms recently
proposed methods including [6], [50], [54] especially when
the number of labels are very large. However, our proposed
method also contains three main limitations. One of the main
limitations of our proposed method is the incomplete labeled
training data. The incomplete labeled images can decrease the
discriminative of the dictionary representation. The second
disadvantage is the size of dictionary for each label. In the
experimental part, we determine the size of dictionary by
using validation dataset. The last one would be the unbalanced
training data distribution, which influence the generalization
and discrimination of our proposed dictionary representation.
Future work includes extending our approach to learn the
common-structure aware dictionaries, which contains the com-
mon visual patterns between different categories. It could help
for selecting the discriminative dictionary items for different
labels, which may make the representation be more robust
to the noisy input features. Another direction is to apply our
proposed dictionary learning method to other computer vision
task, such as image classification or attribute classification.

REFERENCES

annotation by probabilistic collaborative multi-label propagation,” in
Discriminative metric learning in nearest neighbor models for image
pp. 309–316.
learning of semantic classes for image annotation and retrieval,”
Mar. 2007.
label visual classification with label exclusive context,” in Proc. ICCV,
Nov. 2011, pp. 834–841.
“Fast semantic diffusion for large-scale context-based image and video
annotation,” IEEE Trans. Image Process., vol. 21, no. 6, pp. 3080–3091,
models and applications,” IEEE Trans. Image Process., vol. 19, no. 7,
dictionary for object recognition,” in Proc. IEEE Conf. CVPR, Jun. 2012,
pp. 3490–3497.
for sparse coding via label consistent K-SVD,” in Proc. IEEE Conf.
[12] P. Duvgulu, K. Barnard, J. F. De Freitas, and D. A. Forsyth,
“Object recognition as machine translation: Learning a lexicon for a fixed image
non-parametric semi-supervised classification,” in Proc. NIPS, 2009,
pp. 1803–1811.
Cross-scenario clothing retrieval via parts alignment and auxiliary set,” in


Xiaochun Cao (SM’14) received the B.E. and M.E. degrees from Beihang University, China, and the Ph.D. degree from the University of Central Florida, USA, all in computer science. His Ph.D. dissertation nominated for the University Level Outstanding Dissertation Award. After graduation, he spent about three years with ObjectVideo Inc., as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University. He is currently a Professor with the Institute of Information Engineering, Chinese Academy of Sciences. He has authored or co-authored over 80 journals and conference papers. He was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition in 2004 and 2010.

Hua Zhang received the B.E. degree from Yantai University, and the M.S. degree from Tianjin University, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Technology. His research interests include computer vision, machine learning, data mining, and image retrieval.

Xiaojie Guo (M’13) received the B.E. degree in software engineering from the School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China, in 2008, and the M.S. and Ph.D. degrees in computer science from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2010 and 2013, respectively. He is currently an Assistant Professor with the Institute of Information Engineering, Chinese Academy of Sciences. He was a recipient of the Piero Zamperoni Best Student Paper Award with the International Conference on Pattern Recognition (International Association on Pattern Recognition), in 2010.

Si Liu received the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2012. She is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences. She used to be a Research Fellow with the Learning and Vision Group, National University of Singapore. Her research interests include computer vision and multimedia.

Dan Meng (SM’02) received the Ph.D. degree in computer science from the Harbin Institute of Technology, in 1995. He is currently a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing. His research interests include high-performance computing and computer architecture. He is a Senior Member of the China Computer Federation.