Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

Towards Learning a Semantic-Consistent Subspace for Cross-Modal Retrieval

Meixiang Xu · Zhenfeng Zhu* · Yao Zhao

the date of receipt and acceptance should be inserted later

Abstract A great many of approaches have been developed for cross-modal retrieval, among which subspace learning based ones dominate the landscape. Concerning whether using the semantic label information or not, subspace learning based approaches can be categorized into two paradigms, unsupervised and supervised. However, for multi-label cross-modal retrieval, supervised approaches just simply exploit multi-label information towards a discriminative subspace, without considering the correlations between multiple labels shared by multi-modalities, which often leads to an unsatisfactory retrieval performance. To address this issue, in this paper we propose a general framework, which jointly incorporates semantic correlations into subspace learning for multi-label cross-modal retrieval. By introducing the HSIC-based regularization term, the correlation information among multiple labels can be not only leveraged but also the consistency between the modality similarity from each modality is well preserved. Besides, based on the semantic-consistency projection, the semantic gap between the low-level feature space of each modality and the shared high-level semantic space can be balanced by a mid-level consistent one, where multi-label cross-modal retrieval can be performed effectively and efficiently. To solve the optimization problem, an effective iterative algorithm is designed, along with its convergence analysis theoretically and experimentally. Experimental results on real-world datasets have shown the superiority of the proposed method over several existing cross-modal subspace learning methods.

Meixiang Xu · Zhenfeng Zhu · Yao Zhao
Institute of Information Science, Beijing Jiaotong University, Beijing, China
E-mail: xumx0721@bjtu.edu.cn; zhfzhu@bjtu.edu.cn

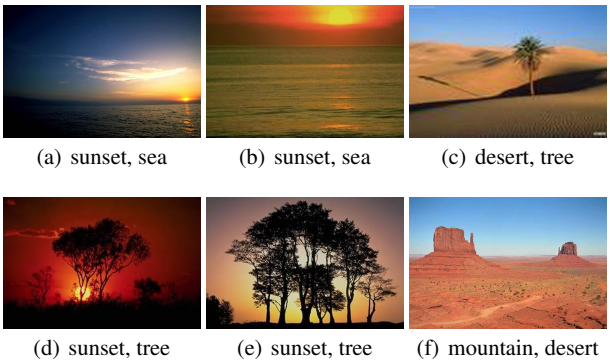
Meixiang Xu · Zhenfeng Zhu · Yao Zhao
Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China
E-mail: yzhao@bjtu.edu.cn

Keywords Cross-modal · Semantic-correlation · Subspace learning · Multi-label

1 Introduction

In numerous real-world applications, data are often presented in diverse forms such as text, image, video, audio and music, etc. These data represent the same semantic content of the objects in different types, which is often referred to as multimedia data. To efficiently manage and manipulate multimedia data, three mainstream tasks, i.e. classification, retrieval and annotation, are involved. Among them, classification aims to categorize multimedia data into a set of predefined semantic concepts and annotation is to designate data objects (e.g. images and videos) a set of labels which describe their content from the semantic level [8, 46]. While retrieval refers to querying data which is relevant to the given data. In this paper, the focus is on cross-modal retrieval.

The goal of cross-modal retrieval is to retrieve the relevant data objects from one modality given one data object from another modality as query. In most cases, multimodal data are high-dimensional and heterogeneous, posing a great challenge for cross-modal retrieval in both efficiency and effectiveness. To address this challenge, one common way is to learn a low-dimensional shared subspace by finding projections for each modality with the paired correspondence across two modalities, where they can be compared. To this end, many subspace learning methods, including unsupervised ones [14, 28, 29, 35] and supervised ones [30, 9, 55, 37], have been studied and proposed to address practical applications such as cross-media retrieval [26, 4], cross-ligual retrieval [36], cross-modal document retrieval [2], cross-modal face recognition [29], etc. However, most of these approaches are specifically designed for single label cross-modal retrieval, which means each cross-modal data



(a) sunset, sea (b) sunset, sea (c) desert, tree

(d) sunset, tree (e) sunset, tree (f) mountain, desert

Fig. 1 Samples with multiple labels from the dataset NUS-WIDE

object only contains one semantic class. Whereas, more often than not, a given data object may be marked with multiple labels at the same time and these labels may be correlated. For example, as shown in Fig. 1, each image contains two semantic concepts, and commonly they are co-occurrent. Usually, 'sunset' and 'tree', 'sunset' and 'sea' are strongly relevant, and also 'tree' and 'sea' is strongly relevant. Thus, multi-label retrieval task is explored. Accordingly, several multi-label datasets [3, 6, 40] are provided.

To learn a common subspace for multi-label cross-modal retrieval, straightforwardly, there are two ways. One is using the unsupervised methods for single-label retrieval directly, ignoring the multi-label information. Since unsupervised ones don't exploit multi-label information, naturally, the learnt subspace will be less discriminative, which often leads to unsatisfactory retrieval results. Motivated by multi-label learning, another way is decomposing multi-label cross-modal retrieval into multiple single-label retrieval by cutting apart one multi-label data object into multiple single-label data objects through the semantic segmentation method. However, this way completely amounts to single-label retrieval in a sense, ignoring the semantic correlations between multiple single label objects that the multi-label data object contains, which is not pleasurable for retrieving the multi-label data object. Consequently, multi-label information and the correlations information are of vital importance for multi-label cross-modal retrieval to learn a more discriminative subspace. Taking the multi-label semantic features as the third view, the proposed three-view CCA (CCA-3V) by Gong et al [9] can be used to multi-label cross-modal subspace learning. To extend CCA for multi-label cross-modal retrieval, Ranjan et al [37] introduce multi-label CCA (ml-CCA), which learns a shared discriminative subspace for two modalities by incorporating multi-label annotations as high level semantic information. Since multi-label retrieval is more complicated than single label retrieval, so far, work on multi-label cross-modal retrieval is still few.

As demonstrated in multi-label learning, correlation among multiple labels belonging to one semantic object is beneficial to handle multi-semantic problems. For better using constraints between multiple labels, it is potential to combine latent common space learning with multi-label classification problems [1][53]. Motivated by this, Zhang and Schenider [49] have proposed an approach by combining CCA with multi-label decoding for multi-label classification. However, few work touch on the exploration of correlation between multiple semantic labels in multi-modal subspace learning. Driven by this and inspired by [50], in this paper we propose a general framework, which jointly incorporates semantic correlation into subspace learning towards learning a semantic-consistent subspace for multi-label cross-modal retrieval.

The main contributions of this paper can be summarized as:

- We propose a semantic-consistent subspace learning approach by incorporating label correlations for multi-label cross-modal retrieval.
- Based on Hilbert-Schmidt Independence Criteria (HSIC), we introduced the HSIC-based regularization. Through the regularization term, the semantic correlation information between multiple labels can be exploited, which is of vital significance to learn a more discriminative subspace for multi-label retrieval. Simultaneously, the consistency between feature-based similarity of each individual modality and semantic-based similarity can be preserved. And also, inter-modality similarity and intra-modality similarity can be well preserved implicitly.
- Under the proposed framework, two problems i.e. high-dimension and heterogeneity of cross-modal data is well addressed by low-dimension embedding and learning a mid-level semantic consistent space jointly.
- To solve the optimization problem, an effective iterative algorithm is provided, along with convergence analysis theoretically and experimentally.

The rest of the paper is structured as follows. In Section 2, a review on some related work is made. In Section 3, formulation of the proposed semantic-consistent subspace learning method with similarity-consistency and semantic-correlation (abbr. SCSL) for multi-label cross-modal retrieval, as well as an effective iterative optimization algorithm with the corresponding theoretical proof, are provided elaborately. Experimental results for evaluating the proposed approach on two multi-modal datasets are reported in Section 4. Finally, Section 5 concludes the paper.

2 Related Work

The recent years have witnessed a surge of interests in cross-modal retrieval and many approaches have been developed [11, 12, 29, 33, 35, 38, 55, 38, 18, 32, 43, 47], such as subspace learning based methods [35, 29, 38, 47], topic model

based methods [17, 52, 32], deep learning based methods [7, 34, 31, 51] and so on and so forth. Among which, the most popular ones are the subspace learning based ones.

In view of whether the prior semantic label information is available or not, subspace learning based approaches can be categorized into two paradigms, unsupervised and supervised. For unsupervised methods, they usually make use of the paired correspondence information to learn a common subspace shared by multiple modalities, where the similarity between modalities can be compared. Typical unsupervised subspace learning methods for cross-modal retrieval include CCA [14, 26, 33], BLM [35] and PLS [29], etc. These methods mainly focus on finding projection matrices for each modality, by which data from each modality can be projected into a latent common space. And they have been widely applied to various practical applications such as cross-media retrieval [26], cross-ligual retrieval [36], cross-modal document retrieval [2], cross-modal face recognition [29], etc. To mine the nonlinear correlations and to derive the comparable low-dimensional representation for heterogeneous modalities, Song et al [32] propose multimodal Similarity Gaussian Process latent variable model (m-SimGP) to learn non-parametric mapping functions between the intra-modal similarities and latent representation, by which mapping heterogeneous modalities into a common latent space. Albeit these methods have demonstrated their effectiveness in practical applications [12, 29, 33, 35, 38], they only exploit pairwise closeness without using semantic class information and any other prior knowledge. Hence, the learned subspace is less discriminative, which often leads to unsatisfying retrieval performance. In fact, using semantic class information and some prior knowledge is of vital significance to learning a more discriminant subspace for cross-modal retrieval [5, 19, 30]. Therefore, supervised approaches [44, 48, 45, 38] are investigated. By incorporating the semantic labels as a third view, Gong [9] proposed the three-view CCA method. Encouraged by structured sparsity, Zhuang et al. [55] proposed the supervised coupled dictionary learning method for multi-modal retrieval. To extend CCA to the supervised case, multi-view discriminant analysis methods are proposed [5, 19, 30]. Integrating high-level semantic information in the form of multi-label annotations into subspace learning, Ranjan et al [37] introduced multi-label Canonical Correlation Analysis (ml-CCA) to learn a discriminative subspace for multi-label cross-modal retrieval. Different from the standard CCA, ml-CCA does not require pairwise information of two modalities, rather it can build the correspondences between two modalities based on the shared multi-label information. To construct a joint cross-modal probabilistic graphical model to mine the mutual consistent semantic topics by interactions between model factors, Wang et al [41] proposed a supervised multi-modal mutual topic reinforce modelling approach for cross-media retrieval. Using a compound nonparametric Bayesian multi-modal prior to depict the correlation structure of data of each modality and between two modalities, Liao et al [22] present the nonparametric Bayesian upstream supervised multi-modal topic model to analyze multi-modal data, which is an extension of the hierarchical Dirichlet process by embedding upstream supervised response variables and values of latent functions under Gaussian Process. Although supervised subspace learning approaches have demonstrated their superiority over unsupervised ones, most of them just purely incorporate the label information to learn a unique discriminant subspace where single label retrieval can be effectively implemented. While for retrieving multi-label semantic objects, it is usually not easy to expectation of satisfactory retrieval precision. More more work can be found in [39].

Supervised subspace learning method for single label cross-modal retrieval can be directly used to address multi-label cross-modal retrieval by incorporating multi-label information. However, as shown in multi-label learning, correlations often exist between multiple labels belonging to one object and generally such correlation information is beneficial to handle multi-semantic problems. To this end, in this paper we jointly integrate semantic correlation into subspace learning for multi-label cross-modal retrieval, which will be discussed subsequently.

3 Semantic-Consistent Subspace Learning (SCSL) for Multi-label Cross-Modal Retrieval

3.1 Notations and HSIC

Notations Let's begin with introducing some notations utilized in this paper. For any matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{A}^{\cdot i}$ and $\mathbf{A}^{:j}$ are used to represent its i-th row and j-th column, respectively. The Frobenius norm of $\mathbf{A}$ is defined as $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{n} \left\|\mathbf{A}^{:i}\right\|_2^2}$. Besides, $tr(\cdot)$ denotes the trace operator and $\mathbf{I}$ is an identity matrix with an appropriate size. Throughout this paper, matrices are represented in capital boldface and vectors in lower boldface, respectively.

Hilbert-Schmidt Independence Criteria (HSIC) Given N independent observations from $\mathscr{Z} := \{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\} \subseteq \mathscr{X} \times \mathscr{Y}$ with joint distribution P_{xy}, HSIC is to compute the square of the norm of the cross-covariance operator over the domain $\mathscr{X} \times \mathscr{Y}$. An empirical estimate of HSIC, expressed as HSIC$(\mathscr{Z}, \mathscr{Y}, P_{xy})$, is defined as:

$$HSIC(\mathscr{Z}, \mathscr{Y}, P_{xy}) = (N-1)^{-2} tr(\mathbf{H}\mathbf{K}_1\mathbf{H}\mathbf{K}_2) \tag{1}$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are two Gram matrices with $k_{1,ij} = k_1(x_i, x_j)$, $k_{2,ij} = k_2(y_i, y_j)$. $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$, is a centering matrix, and $\mathbf{1}_n \in \mathbb{R}^n$ is a full-one column vector. For more details about HSIC, please see the literature [10].

3.2 Formulation

This paper addresses the multi-label cross-modal task, where the involved retrieval objects are with multiple labels, as shown in Fig. 2. Suppose that there is a data set of n training samples with c classes from M modalities, denoted as $\Omega = \{\mathbf{x}_i^1, \mathbf{x}_i^2, \cdots, \mathbf{x}_i^M\}_{i=1}^n$. Let $\mathbf{X}_v = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d_v}$ be the feature matrix of the v-th modality $(v = 1, \cdots, M)$ and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n]^T \in \mathbb{R}^{n \times c}$ represent the semantic matrix with the i-th row being the semantic vector corresponding to $\mathbf{X}_v$. $y_{ij} = 1$ if $\mathbf{x}_i^v$ belongs to the j-th class, and $y_{ij} = 0$ otherwise $(j = 1, \cdots, c)$. Where d_v is the dimension of the v-th modality and M is the number of modality. Our method proposes to learn a semantic-consistent common subspace, where the consistency between feature-based similarity of each modality and semantic-based similarity can be preserved, simultaneously the correlation among multiple semantic labels is taken into consideration. To target this goal, our proposed SCSL is formulated as:

$$\min O = S(\cdot) + R(\cdot), \tag{2}$$

where $S(\cdot)$ is the semantic-consistent projection term, which is used to learn projection matrices for projecting multiple modality data into a mid-level semantic-consistent subspace. $R(\cdot)$ is the HSIC-based regularization term, which is introduced to incorporate semantic correlation information among the shared multi-label by multiple modalities and preserve the consistence between the modality similarity of each modality.

3.3 The Semantic-Consistent Projection Term

In many real-world applications, multi-modality data are usually with high-dimensional features and heterogeneity, which poses a great challenge for cross-modal retrieval. On one hand, high-dimensionality causes highly computational complexity and affects the retrieval efficiency. On the other hand, heterogeneity makes multiple modalities incomparable which means that the similarity between them can not be directly measured. To address high-dimensionality, we first learn a low-dimensional embedding for each modality, i.e. $\mathbf{Z}_v = \mathbf{X}_v\mathbf{P}_v (v = 1, \cdots, M)$, here $\mathbf{P}_v$ plays the role of removing redundant features for reducing dimensionality of each modality. To handle heterogeneity, further we seek for projection matrices $\mathbf{Q}_v$ by which each modality can be projected into a common space where multiple modalities are comparable. Since multiple modalities describe the same semantic object in different forms and each individual modality is considered to be from a specific feature space where it describes the semantic objects, they share the same semantic space (As illustrated in Fig. 2). Therefore, more often than not, many subspace-based learning methods assume that different modalities are projected into the shared semantic space.

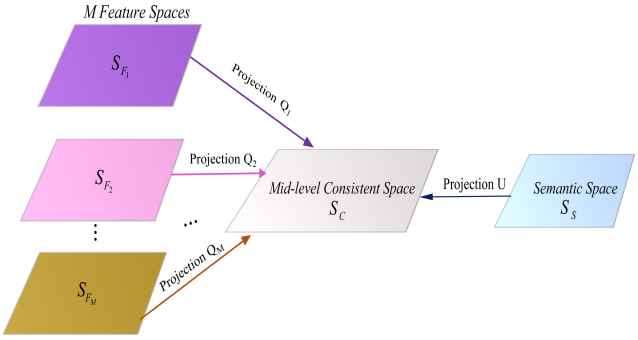


Fig. 3 Framework of the proposed method towards learning a hidden consistent discriminant subspace S_C. Given M modalities together with their semantic labels, they are corresponding to M feature spaces $S_{F_v} (v = 1, \cdots, M)$ and the shared semantic space S_S. Since multi-modal data are often high-dimensional, a low-dimension embedding with the projection matrix $\mathbf{P}_v$ for each modality, denoted as $\mathbf{Z}_v = \mathbf{X}_v\mathbf{P}_v$, is first performed. This step is not displayed in the framework. To reduce the semantic gap and guarantee that each modality is close to each other as much as possible in S_C, each individual modality with a separate projection matrix $\mathbf{Q}_v$ and their shared semantic space are simultaneously projected into the consistent space S_C.

In the shared semantic space, if each $\mathbf{Z}_v\mathbf{Q}_v = \mathbf{X}_v\mathbf{P}_v\mathbf{Q}_v$ is close enough to $\mathbf{Y}$, each $\mathbf{Z}_v\mathbf{Q}_v$ will be close to one another, in which case cross-modal retrieval on $\mathbf{X}_v$ would be quite accurate. However, it is not as expected that each modality $\mathbf{Z}_v\mathbf{Q}_v$ is close to $\mathbf{Y}$ in most cases, due to the semantic gap between low-level features and high-level semantic concepts. What's worse, $\mathbf{Z}_v\mathbf{Q}_v$ may be even far away from each other at times. Consequently, cross-modal retrieval on $\mathbf{X}_v$ will be not as accurate as expected. Towards the goal that each modality $\mathbf{Z}_v\mathbf{Q}_v$ in the projected space is semantically as close as possible, assuming that there exists a projection matrix $\mathbf{U}$, which projects the shared semantic space into a mid-level consistent space where multiple modalities are maximally semantic-relevant (as illustrated in Fig. 3). Adopting the F-norm as the metric, we take $S(.)$ as:

$$S(.) = \sum_{v=1}^{M} \|\mathbf{Z}_v\mathbf{Q}_v - \mathbf{YU}\|_F^2 = \sum_{v=1}^{M} \|\mathbf{X}_v\mathbf{P}_v\mathbf{Q}_v - \mathbf{YU}\|_F^2 \tag{3}$$

where $\mathbf{P}_v \in \mathbb{R}^{d_v \times D_v}$, $\mathbf{Q}_v \in \mathbb{R}^{D_v \times d}$ and $\mathbf{U} \in \mathbb{R}^{c \times d}$.

3.4 The HSIC-based Regularization Term

Embedding the semantic correlation information between multiple labels shared by M modalities and the similarity-consistency constraint of each modality is jointly incorporated into the HSIC-based regularized term $R(.)$. In the following, we shall explain the derivation of $R(.)$. Since the semantic-correlation embedding is built on the similarity-consistency, we first introduce the similarity-consistency constraint.

Similarity-Consistency Constraint This constraint depends on the assumption that two identical examples should

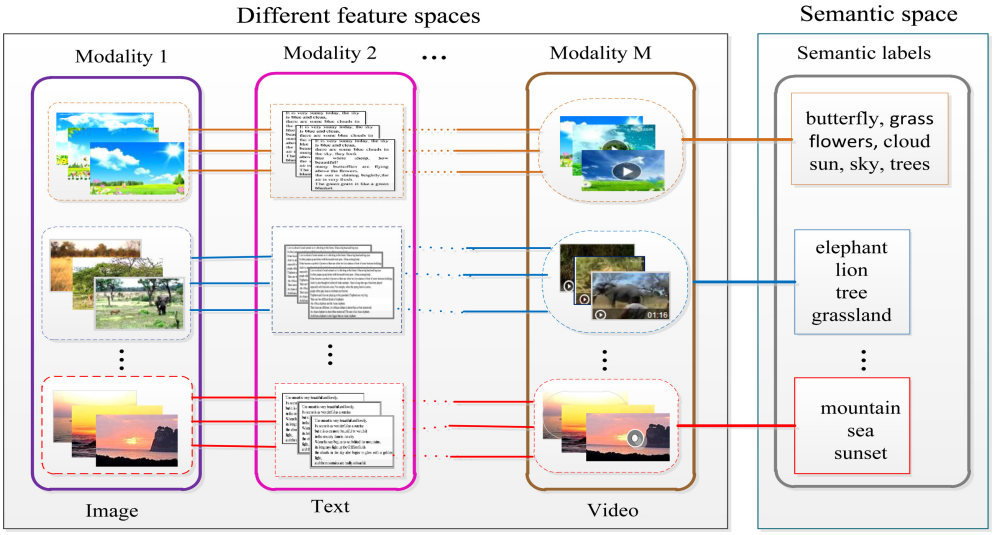


Fig. 2 Illustration of multiple modalities. For every semantic object, it has multiple labels and M modalities, each of which can describe the object itself. Different modalities represent the semantic content from different levels, corresponding to different feature space, but sharing with the same semantic space.

have identical semantic labels, or not strictly speaking, two examples sharing high similarity in features are apt to have overlap in their semantic labels [23]. More specifically, for each modality, consider two examples $\mathbf{x}_i^v$ and $\mathbf{x}_j^v$ with their labels $\mathbf{y}_i$ and $\mathbf{y}_j$, correspondingly. To evaluate the similarity between two examples, two different ways can be used. One is feature-based, which measures sample similarity in each modality feature space. The other is semantic-based, which measures sample similarity in the shared label space. Adopting the kernel function as measuring similarity of two types of similarity, feature-based similarity and semantic-based similarity can be denoted as $k_{x_v}(\mathbf{x}_i^v, \mathbf{x}_j^v)$ and $k_y(\mathbf{y}_i, \mathbf{y}_j)$, respectively. Since multiple modalities share with the same semantic labels, two similarity measurements should be consistent, namely $k_{x_v}(\mathbf{x}_i^v, \mathbf{x}_j^v) \approx k_y(\mathbf{y}_i, \mathbf{y}_j)$. For all the examples of each modality, we denote two styles of similarities as $\mathbf{K}_{\mathrm{X}_v}(\mathbf{X}_v, \mathbf{X}_v)$ and $\mathbf{K}_{\mathrm{Y}}(\mathbf{Y}, \mathbf{Y})$, respectively. To preserve the consistency between feature-based similarity of each modality and semantic-based similarity, using the HISC criterion [10] we expect:

$$\max \sum_{v=1}^{M} tr\left(\mathbf{H}\mathbf{K}_{\mathrm{X}_v}\mathbf{H}\mathbf{K}_{\mathrm{Y}}\right) \tag{4}$$

Here, $R(.) = -\sum_{v=1}^{M} tr\left(\mathbf{H}\mathbf{K}_{\mathrm{X}_v}\mathbf{H}\mathbf{K}_{\mathrm{Y}}\right)$.

For simplicity, here we take the linear kernel to measure the similarity between two examples, then $\mathbf{K}_{\mathrm{X}_v}$ and $\mathbf{K}_{\mathrm{Y}}$ can be defines as $\mathbf{K}_{\mathrm{X}_v} = \langle \mathbf{X}_v\mathbf{P}_v\mathbf{Q}_v, \mathbf{X}_v\mathbf{P}_v\mathbf{Q}_v \rangle = \mathbf{X}_v\mathbf{P}_v\mathbf{Q}_v\mathbf{Q}_v^T\mathbf{P}_v^T\mathbf{X}_v^T$ and $\mathbf{K}_{\mathrm{Y}} = \langle \mathbf{Y}, \mathbf{Y} \rangle = \mathbf{Y}\mathbf{Y}^T$, respectively.

Embedding Semantic-Correlation Correlation relationships between multiple labels as useful information are beneficial to multi-label learning, and label correlation has been studied in multi-label learning [23, 1, 50, 20] to improve the performance of the learning model. Inspired by this, here we incorporate multi-label correlation information into subspace learning towards a more discriminative subspace for retrieving multiple semantic objects across different modalities. To this end, we introduce a matrix $\mathbf{C} = [C_{k,l}]_{c\times c}(k = 1, \cdots, c; l = 1, \cdots, c)$ to capture the correlations between multiple semantic labels. For the correlation matrix $\mathbf{C}$, each element $C_{k,l} \geq 0$ indicates the score of label correlation between two semantic labels.

Embedding correlation information, we redefine $\mathbf{K}_{\mathrm{Y}}$ as $\tilde{\mathbf{K}}_{\mathrm{Y}} = \mathbf{Y}\mathbf{C}\mathbf{Y}^T$. Then, Eq. (4) is re-expressed as:

$$\max \sum_{v=1}^{M} tr\left(\mathbf{H}\mathbf{K}_{\mathrm{X}_v}\mathbf{H}\tilde{\mathbf{K}}_{\mathrm{Y}}\right) \tag{5}$$

From the above elaborations, finally the proposed general model is formulated as:

$$\begin{aligned} &\min_{\mathbf{P}_v,\mathbf{Q}_v,\mathbf{U}} \sum_{v=1}^{M} \|\mathbf{X}_v\mathbf{P}_v\mathbf{Q}_v - \mathbf{Y}\mathbf{U}\|_F^2 - \alpha \sum_{v=1}^{M} tr\left(\mathbf{H}\mathbf{K}_{\mathrm{X}_v}\mathbf{H}\tilde{\mathbf{K}}_{\mathrm{Y}}\right) \\ &s.t.\ \ \mathbf{P}_v^T\mathbf{P}_v = \mathbf{I}_{d_v} \end{aligned} \tag{6}$$

where $\mathbf{P}_v \in \mathbb{R}^{d_v\times D_v}$,$\mathbf{Q}_v \in \mathbb{R}^{D_v\times d}$ and $\mathbf{U} \in \mathbb{R}^{c\times d}(d < c)$. The second term also preserves the intra-modality similarity relationship among samples within each individual modality and inter-modality similarity implicitly.

3.5 Optimization

Three varibles $\mathbf{P}_v$, $\mathbf{Q}_v$ and $\mathbf{U}$ need to be optimized in Eq. (6). Obviously, it is is not a trivial task to simultaneously optimize them in a direct way. Therefore, in the following we develop an efficient iterative algorithm to solve them.

For convenience, let us first introduce intermediate variable $\mathbf{W}_v = \mathbf{P}_v\mathbf{Q}_v (v = 1, \cdots, M)$, then $\mathbf{K}_{\mathrm{X}_v}$ can be expressed as $\tilde{\mathbf{K}}_{\mathrm{X}_v} = \mathbf{X}_v\mathbf{W}_v\mathbf{W}_v^T\mathbf{X}_v^T$. Thereby, the objective function can be reformulated as:

$$\begin{aligned}
&\min_{\mathbf{P}_v,\mathbf{Q}_v,\mathbf{W}_v,\mathbf{U}} \sum_{v=1}^{M} \|\mathbf{X}_v\mathbf{W}_v - \mathbf{YU}\|_F^2 - \alpha \sum_{v=1}^{M} tr\left(\mathbf{H}\,\tilde{\mathbf{K}}_{\mathrm{X}_v}\mathbf{H}\tilde{\mathbf{K}}_Y\right) \\
&\qquad + \gamma \sum_{v=1}^{M} \|\mathbf{W}_v - \mathbf{P}_v\mathbf{Q}_v\|_F^2 \\
\Leftrightarrow &\min_{\mathbf{P}_v,\mathbf{Q}_v,\mathbf{W}_v,\mathbf{U}} \sum_{v=1}^{M} tr\left((\mathbf{X}_v\mathbf{W}_v - \mathbf{YU})^T(\mathbf{X}_v\mathbf{W}_v - \mathbf{YU})\right) \\
&\qquad - \alpha \sum_{v=1}^{M} tr\left(\mathbf{HX}_v\mathbf{W}_v\mathbf{W}_v^T\mathbf{X}_v^T\mathbf{H}\,\tilde{\mathbf{K}}_Y\right) \\
&\qquad + \gamma \sum_{v=1}^{M} tr\left((\mathbf{W}_v - \mathbf{P}_v\mathbf{Q}_v)^T(\mathbf{W}_v - \mathbf{P}_v\mathbf{Q}_v)\right) \\
&s.t.\ \ \mathbf{P}_v^T\mathbf{P}_v = \mathbf{I}_{D_v}
\end{aligned} \tag{7}$$

The optimization problem formulated by Eq. (6) is equivalent to that by Eq. (7). To solve it, we optimize the following four sub-minimization problems under alternative rules.

3.5.1 Solve $\boldsymbol{Q}_v$, fixing $\boldsymbol{W}_v$, $\boldsymbol{P}_v$ and $\boldsymbol{U}$

For any given $\mathbf{P}_v$, $\mathbf{W}_v$ and $\mathbf{U}$, the optimizing problem in Eq. (7) turns to an unconstrained one:

$$\begin{aligned}
&\min_{\mathbf{Q}_v} O = \sum_{v=1}^{M} tr\left((\mathbf{W}_v - \mathbf{P}_v\mathbf{Q}_v)^T(\mathbf{W}_v - \mathbf{P}_v\mathbf{Q}_v)\right) \\
\Leftrightarrow &\min_{\mathbf{Q}_v} O = \sum_{v=1}^{M} tr\left(\mathbf{P}_v^T\mathbf{X}_v^T\mathbf{X}_v\mathbf{P}_v - 2\mathbf{P}_v^T\mathbf{X}_v^T\mathbf{YQ}_v + \mathbf{Q}_v^T\mathbf{Y}^T\mathbf{YQ}_v\right)
\end{aligned} \tag{8}$$

Setting the derivative $\frac{\partial O}{\partial \mathbf{Q}_v}$ of O w.r.t. $\mathbf{Q}_v$ to 0 yields:

$$\mathbf{Q}_v = \mathbf{P}_v^T\mathbf{W}_v. \tag{9}$$

3.5.2 Solve $\boldsymbol{W}_v$, fixing $\boldsymbol{Q}_v$, $\boldsymbol{P}_v$ and $\boldsymbol{U}$

Substituting $\mathbf{Q}_v = \mathbf{P}_v^T\mathbf{W}_v$ for the expression $\mathbf{Q}_v$ in Eq. (7), we will obtain the equivalent optimization problem w.r.t $\mathbf{W}_v$:

$$\begin{aligned}
\min_{\mathbf{W}_v} O = &\sum_{v=1}^{M} tr\left((\mathbf{X}_v\mathbf{W}_v - \mathbf{YU})^T(\mathbf{X}_v\mathbf{W}_v - \mathbf{YU})\right) \\
&- \alpha \sum_{v=1}^{M} tr\left(\mathbf{HX}_v\mathbf{W}_v\mathbf{W}_v^T\mathbf{X}_v^T\mathbf{H}\,\tilde{\mathbf{K}}_Y\right) \\
&+ \gamma \sum_{v=1}^{M} tr\left(\left(\mathbf{W}_v - \mathbf{P}_v\mathbf{P}_v^T\mathbf{W}_v\right)^T\left(\mathbf{W}_v - \mathbf{P}_v\mathbf{P}_v^T\mathbf{W}_v\right)\right)
\end{aligned} \tag{10}$$

With $\left(\mathbf{I}_{d_v} - \mathbf{P}_v\mathbf{P}_v^T\right)^T\left(\mathbf{I}_{d_v} - \mathbf{P}_v\mathbf{P}_v^T\right) = \mathbf{I}_{d_v} - \mathbf{P}_v\mathbf{P}_v^T$, the objective function is reexpressed as:

$$\begin{aligned}
O = &\sum_{v=1}^{M} tr\left(\mathbf{W}_v^T\mathbf{X}_v^T\mathbf{X}_v\mathbf{W}_v - 2\mathbf{W}_v^T\mathbf{X}_v^T\mathbf{YU} + \mathbf{U}^T\mathbf{Y}^T\mathbf{YU}\right) \\
&- \alpha \sum_{v=1}^{M} tr\left(\mathbf{W}_v^T\mathbf{X}_v^T\mathbf{H}\tilde{K}_Y\mathbf{HX}_v\mathbf{W}_v\right) \\
&+ \gamma \sum_{v=1}^{M} tr\left(\mathbf{W}_v^T\left(\mathbf{I}_{d_v} - \mathbf{P}_v\mathbf{P}_v^T\right)\mathbf{W}_v\right)
\end{aligned} \tag{11}$$

Then, the derivative $\frac{\partial O}{\partial \mathbf{Q}_v}$ of O w.r.t. $\mathbf{W}_v$ is derived as:

$$\begin{aligned}
\frac{\partial O}{\partial \mathbf{W}_v} = &2\mathbf{X}_v^T\mathbf{X}_v\mathbf{W}_v - 2\mathbf{X}_v^T\mathbf{YU} - 2\alpha\mathbf{X}_v^T\mathbf{H}\,\tilde{\mathbf{K}}_Y\mathbf{HX}_v\mathbf{W}_v \\
&+ 2\gamma\left(\mathbf{I}_{d_v} - \mathbf{P}_v\mathbf{P}_v^T\right)\mathbf{W}_v.
\end{aligned}$$

Setting $\frac{\partial O}{\partial \mathbf{W}_v} = 0$ yields:

$$\begin{aligned}
&\left(\mathbf{X}_v^T\mathbf{X}_v - \alpha\mathbf{X}_v^T\mathbf{H}\,\tilde{\mathbf{K}}_Y\mathbf{HX}_v + \gamma\left(\mathbf{I}_{d_v} - \mathbf{P}_v\mathbf{P}_v^T\right)\right)\mathbf{W}_v = \mathbf{X}_v^T\mathbf{YU} \\
\Leftrightarrow\ \ &\mathbf{W}_v = {\mathbf{F}_v}^{-1}\mathbf{X}_v^T\mathbf{YU}
\end{aligned} \tag{12}$$

where $\mathbf{E}_v = \mathbf{X}_v^T\mathbf{X}_v - \alpha\mathbf{X}_v^T\mathbf{H}\,\tilde{\mathbf{K}}_Y\mathbf{HX}_v + \gamma\mathbf{I}_{d_v}$, $\mathbf{F}_v = \mathbf{E}_v - \gamma\mathbf{P}_v\mathbf{P}_v^T$.

3.5.3 Solve $\boldsymbol{P}_v$, fixing $\boldsymbol{W}_v$, $\boldsymbol{Q}_v$ and $\boldsymbol{U}$

The objective function in Eq. (10) can be rewritten as:

$$\begin{aligned}
O = &\sum_{v=1}^{M} tr\left(\mathbf{W}_v^T\mathbf{X}_v^T\mathbf{X}_v\mathbf{W}_v - 2\mathbf{W}_v^T\mathbf{X}_v^T\mathbf{YU} + \mathbf{U}^T\mathbf{Y}^T\mathbf{YU}\right) \\
&- \alpha \sum_{v=1}^{M} tr\left(\mathbf{W}_v^T\mathbf{X}_v^T\mathbf{H}\,\tilde{\mathbf{K}}_Y\mathbf{HX}_v\mathbf{W}_v\right) \\
&+ \gamma \sum_{v=1}^{M} tr\left(\mathbf{W}_v^T\left(\mathbf{I}_{d_v} - \mathbf{P}_v\mathbf{P}_v^T\right)\mathbf{W}_v\right) \\
= &\sum_{v=1}^{M} tr\left(\mathbf{W}_v^T\left(\mathbf{F}_v\mathbf{W}_v - 2\mathbf{X}_v^T\mathbf{YU}\right)\right) + Mtr\left(\mathbf{U}^T\mathbf{Y}^T\mathbf{YU}\right)
\end{aligned} \tag{13}$$

Plugging $\mathbf{W}_v = {\mathbf{F}_v}^{-1}\mathbf{X}_v^T\mathbf{YU}$ into Eq. (13) gives the following expression:

$$\begin{aligned}
O = &\sum_{v=1}^{M} tr\left(\mathbf{U}^T\mathbf{Y}^T\mathbf{X}_v\mathbf{F}_v^{-1}\left(\mathbf{F}_v{\mathbf{F}_v}^{-1}\mathbf{X}_v^T\mathbf{YU} - 2\mathbf{X}_v^T\mathbf{YU}\right)\right) \\
&+ Mtr\left(\mathbf{U}^T\mathbf{Y}^T\mathbf{YU}\right) \\
= &-\sum_{v=1}^{M} tr\left(\mathbf{U}^T\mathbf{Y}^T\mathbf{X}_v\mathbf{F}_v^{-1}\mathbf{X}_v^T\mathbf{YU}\right) + Mtr\left(\mathbf{U}^T\mathbf{Y}^T\mathbf{YU}\right)
\end{aligned} \tag{14}$$

The optimization problem w.r.t $\mathbf{P}_v$ is reduced to:

$$\begin{aligned} &\max_{\mathbf{P}_v} O = \sum_{v=1}^{M} tr\left(\mathbf{U}^T\mathbf{Y}^T\mathbf{X}_v\mathbf{F}_v^{-1}\mathbf{X}_v^T\mathbf{Y}\mathbf{U}\right) \\ &s.t.\ \ \mathbf{P}_v^T\mathbf{P}_v = \mathbf{I} \end{aligned} \tag{15}$$

where $\mathbf{F}^{-1}$ can be calculated by resorting to the Sherman-Morrison-Woodbury formula in [13] as:

$$\begin{aligned} \mathbf{F}_v^{-1} &= \left(\mathbf{E}_v - \gamma\mathbf{P}_v\mathbf{P}_v^T\right)^{-1} \\ &= \mathbf{E}_v^{-1} + \gamma\mathbf{E}_v^{-1}\mathbf{P}_v\left(\mathbf{I}_{D_v} - \gamma\mathbf{P}_v^T\mathbf{E}_v^{-1}\mathbf{P}_v\right)^{-1}\mathbf{P}_v^T\mathbf{E}_v^{-1} \end{aligned} \tag{16}$$

Accordingly, the objective function in Eq. (15) can be equivalently expressed as:

$$\begin{aligned} O &= \gamma\sum_{v=1}^{M} tr\left(\mathbf{U}^T\mathbf{Y}^T\mathbf{X}_v\mathbf{E}_v^{-1}\mathbf{P}_v\left(\mathbf{I}_{D_v} - \gamma\mathbf{P}_v^T\mathbf{E}_v^{-1}\mathbf{P}_v\right)^{-1}\mathbf{P}_v^T\mathbf{E}_v^{-1}\mathbf{X}_v^T\mathbf{Y}\mathbf{U}\right) \\ &= \gamma\sum_{v=1}^{M} tr\left(\left(\mathbf{I}_{D_v} - \gamma\mathbf{P}_v^T\mathbf{E}_v^{-1}\mathbf{P}_v\right)^{-1}\mathbf{P}_v^T\mathbf{E}_v^{-1}\mathbf{X}_v^T\mathbf{Y}\mathbf{U}\mathbf{U}^T\mathbf{Y}^T\mathbf{X}_v\mathbf{E}_v^{-1}\mathbf{P}_v\right) \\ &= \gamma\sum_{v=1}^{M} tr\left(\left(\mathbf{P}_v^T\mathbf{G}_v\mathbf{P}_v\right)^{-1}\mathbf{P}_v^T\mathbf{N}_v\mathbf{P}_v\right) \end{aligned}$$

where $\mathbf{G}_v = \mathbf{I}_{d_v} - \gamma\mathbf{E}_v^{-1}$ and $\mathbf{N}_v = \mathbf{E}_v^{-1}\mathbf{X}_v^T\mathbf{Y}\mathbf{U}\mathbf{U}^T\mathbf{Y}^T\mathbf{X}_v\mathbf{E}_v^{-1}$. The optimization problem formulated by Eq. (15) is equivalent to:

$$\begin{aligned} &\max_{\mathbf{P}_v} \sum_{v=1}^{M} tr\left(\left(\mathbf{P}_v^T\mathbf{G}_v\mathbf{P}_v\right)^{-1}\mathbf{P}_v^T\mathbf{N}_v\mathbf{P}_v\right) \\ &s.t.\ \mathbf{P}_v^T\mathbf{P}_v = \mathbf{I}_{d_v} \\ \Leftrightarrow &\max_{\mathbf{P}_v} \sum_{v=1}^{M} tr\left(\mathbf{P}_v^T\mathbf{G}_v^{-1}\mathbf{N}_v\mathbf{P}_v\right) \\ &s.t.\ \mathbf{P}_v^T\mathbf{P}_v = \mathbf{I}_{d_v} \end{aligned} \tag{17}$$

Observe that $\mathbf{G}_v$ is positive definite [16, 21], thus $\mathbf{P}_v$ can be derived by eigen-decomposition of $\mathbf{G}_v^{-1}\mathbf{N}_v$.

3.5.4 Solve $\mathbf{U}$, fixing $\mathbf{Q}_v$, $\mathbf{W}_v$ and $\mathbf{P}_v$

With $\mathbf{Q}_v$, $\mathbf{W}_v$ and $\mathbf{P}_v$ fixed, the optimizing problem also turns to an unconstrained one:

$$\min_{\mathbf{U}} O = \sum_{v=1}^{M} tr\left(\mathbf{W}_v^T\mathbf{X}_v^T\mathbf{X}_v\mathbf{W}_v - 2\mathbf{W}_v^T\mathbf{X}_v^T\mathbf{Y}\mathbf{U} + \mathbf{U}^T\mathbf{Y}^T\mathbf{Y}\mathbf{U}\right) \tag{18}$$

Setting $\frac{\partial O}{\partial \mathbf{U}} = -2(\sum_{v=1}^{M}\mathbf{W}_v^T\mathbf{X}_v^T\mathbf{Y})^T + 2M\mathbf{Y}^T\mathbf{Y}\mathbf{U}$ to 0, we will obtain:

$$\mathbf{U} = \left(M\mathbf{Y}^T\mathbf{Y}\right)^{\dagger}(\sum_{v=1}^{M}\mathbf{W}_v^T\mathbf{X}_v^T\mathbf{Y})^T \tag{19}$$

where '†' refers to the pseudo inverse of a matrix.

To better understand the procedure for solving the proposed method, we summarize in detail the solver for solving the optimization problem in Eq. (7) as Algorithm 1.

Algorithm 1 : Semantic-Consistent subspace learning for cross-modal retrieval

Input:
Multi-modality data $\mathbf{X}_v \in \mathbb{R}^{N\times d_v}$, $v = 1,...,M$; the trade-off parameters α and γ; the semantic label matrix $\mathbf{Y}$.

Output:
$\mathbf{P}_v$, $\mathbf{Q}_v$ and $\mathbf{U}$

Construct the kernel matrix $\tilde{\mathbf{K}}_Y$ of $\mathbf{Y}$;
Initializing: Initialize $\mathbf{P}_v(v = 1,...,M)$ and $\mathbf{U}$ randomly;
for v =1:M **do**
  Compute $\mathbf{E}_v = \mathbf{X}_v^T\mathbf{X}_v - \alpha\mathbf{X}_v^T\mathbf{H}\,\tilde{\mathbf{K}}_Y\mathbf{H}\mathbf{X}_v + \gamma\mathbf{I}_{d_v}$;
  Compute $\mathbf{G}_v = \mathbf{I}_{d_v} - \gamma\mathbf{E}_v^{-1}$;
end for
Repeat
for v=1:M **do**
  Compute $\mathbf{F}_v = \mathbf{E}_v - \gamma\mathbf{P}_v{\mathbf{P}_v}^T$;
  Compute $\mathbf{N}_v = \mathbf{E}_v^{-1}\mathbf{X}_v^T\mathbf{Y}\mathbf{U}\mathbf{U}^T\mathbf{Y}^T\mathbf{X}_v\mathbf{E}_v^{-1}$
  Solve $\mathbf{P}_v$ by eigen-value decomposition of $\mathbf{G}_v^{-1}\mathbf{N}_v$;
  Compute $\mathbf{W}_v = {\mathbf{F}_v}^{-1}\mathbf{X}_v^T\mathbf{Y}\mathbf{U}$
  Compute $\mathbf{Q}_v = {\mathbf{P}_v}^T\mathbf{W}_v$;
end for
Compute $\mathbf{U} = \left(M\mathbf{Y}^T\mathbf{Y}\right)^{\dagger}(\sum_{v=1}^{M}\mathbf{W}_v^T\mathbf{X}_v^T\mathbf{Y})^T$.
Until satisfying convergence criterion.
Return $\mathbf{P}_v$, $\mathbf{Q}_v$ and $\mathbf{U}$;

3.6 Convergence and Computational Complexity

The convergence behavior of the proposed iterative optimization algorithm in Algorithm 1 is summarized by the following Theorem 1.

Theorem 1 *Using the iterative optimizing rules in Algorithm 1, the objective function defined by Eq. (7) monotonically decreases until convergence after certain iterations.*

Proof From the optimization procedure in Section 3, we can express $\mathbf{Q}_v$ and $\mathbf{U}$ by $\mathbf{W}_v$ and $\mathbf{P}_v$. Therefore, we can only consider the updating rules of $\mathbf{W}_v$ and $\mathbf{P}_v$. Assume that in the t-th iteration, we got $\mathbf{W}_v^{(t)}$ and $\mathbf{P}_v^{(t)}(v = 1,\cdots,M)$. Fix $\mathbf{W}_v$ and update $\mathbf{P}_v$ by solving the optimization problem in Eq. (16), and we can obtain $\mathbf{P}_v^{(t+1)}$ by eigen-decomposition. Consequently, we have:

$$O(\mathbf{P}_v^{(t+1)}, \mathbf{W}_v^{(t)}) \leq O(\mathbf{P}_v^{(t)}, \mathbf{W}_v^{(t)})(v = 1,\cdots,M) \tag{20}$$

Likewise, fixing $\mathbf{P}_v^{(t+1)}(v = 1,\cdots,M)$ and updating $\mathbf{W}_v^{(t)}(v = 1,\cdots,M)$ by solving the problem in Eq. (9), we have:

$$O(\mathbf{P}_v^{(t+1)}, \mathbf{W}_v^{(t+1)}) \leq O(\mathbf{P}_v^{(t+1)}, \mathbf{W}_v^{(t)})(v = 1,\cdots,M) \tag{21}$$

Combining Eq. (20) with Eq. (21), we arrive at:

$$O(\mathbf{P}_v^{(t+1)}, \mathbf{W}_v^{(t+1)}) \leq O(\mathbf{P}_v^{(t)}, \mathbf{W}_v^{(t)})(v = 1,\cdots,M) \tag{22}$$

The above equation conveys that the objective function value monotonically decreases in each iteration under the iterative optimization rules in Algorithms 1, which completes the proof of Theorem 1.

In the following, we would like to roughly analyze the computational complexity of the proposed algorithm. In our case, $c \ll n$ and $c \ll d_v (v = 1, \cdots, M)$. The complexity for calculating the inverse of a few matrices $\mathbf{E}_v$, $\mathbf{G}_v$ and $\mathbf{F}_v$ $(v = 1, \cdots, M)$ is $o(d_v^3)$ and the eigen-decomposition of $\mathbf{N}_v^{-1}\mathbf{T}_v$ also needs $o(d_v^3)$ in complexity. In each iteration, it takes $o(d_v^3 + nd_v^2)$ to update $\mathbf{Q}_v$, $o(d_v^3)$ to update $\mathbf{W}_v$ and $o(c^3)$ to update $\mathbf{U}$. Thus, after t times iterations, the total cost for solving SCSL is $t\left(\sum_{v=1}^{M} o\left(d_v^3 + nd_v^2\right) + o(c^3)\right)$ approximately.

4 Experiments

To test the performance of the proposed SCSL for cross-modality retrieval, experiments were conducted on real-world multi-label cross-modal datasets. Given a cross-modal problem, using the iterative algorithm in Algorithm 1, we can learn projection matrices $\mathbf{P}_v$ and $\mathbf{Q}_v$ for low-dimension embedding and learning semantic-consistency space on the training set. After that, data from different modalities can be projected into a common subspace, where we can measure the relevance of projected data from each modality. In the testing phase, taking data in one modality as a query set, we can retrieve the relevant data from another modality. Without loss of generality, we mainly consider the two modalities case in the following experiments.

4.1 Datasets

Two multi-label cross-modal datasets, i.e. NUS-WIDE and VOC2007, which have been commonly used for retrieval, are used in the experiments. Brief descriptions of them are as follows:

- **NUS-WIDE:** This dataset is originally from [3], including 190420 image examples totally, each with 21 possible labels. For each image-text pair, 500-dimensional SIFT BoVW features are extracted for image and 1000-dimensional text annotations are used for text. Statistically, there are 86670 single-label image-text pairs and 103750 multi-label image-text pairs in NUS-WIDE. In our experiments, we utilize the 103750 image-text pairs (half for training and half for testing) with multiple labels for evaluation.

- **VOC2007:** This dataset is a subset from [25], containing 9963 image-text pairs that are labeled with 20 semantic classes altogether. The whole dataset is originally divided into two parts, one for training (5011 pairs) and the other for testing (4952 pairs). For each image-text pair, 4096-dimensional CNN features are extracted to represent the modality image and 798-dimensional tag ranking features are used to represent the modality text [42][15]. According to statistics, there are 2808 single-label pairs and 2203 multi-label pairs in the training set, 2841 pairs single-label pairs and 2111 pairs multi-label pairs in the testing set. In our experiments, we use multi-label pairs (2203 pairs for training and 2111 pairs for testing) for evaluation.

4.2 Compared Approaches and Experimental Setup

We compare the proposed SCSL with the following representative methods:

- **CCA**[14]: As a classical multivariate data analysis method, CCA aims to find pairs of vectors that can maximize the correlation between a set of paired variables. For multi-view learning, it amounts to finding a common subspace where the low dimensional embeddings of data from two views are maximally correlated.
- **KCCA** [28]: Kernel CCA (KCCA) is the kernel extension of the classic CCA.
- **PLS**[29]: To avoid information loss during correlating different modals, PLS correlates the subspaces of CCA by virtue of the least square method.
- **KPLS** [27]: kernel PLA (KPLS) is the kernel extension of PLS.
- **BLM**[35]: BLM tries to learn a shared subspace where data with the same content and different views can be projected onto the same coordinates. It enables the style and content to be separated through singular value decomposition (SVD).
- **GMLDA**[30]: GMLDA is to find a set of projection directions for each view which can separate different content's class means and unite samples from the same class of different views in the projected feature subspace. It is an extension of linear discriminant analysis (LDA).
- **CCA-3V**[9]: CCA-3V is a three view extension of CCA for cross-modal retrieval in the multi-label setting by taking the multi-label semantic information as the third view.
- **ML-CCA**[37]: ML-CCA is an extension of CCA for multi-label cross-modal retrieval, which learns a shared discriminative subspace for two modalities by incorporating multi-label annotations as high level semantic information.
- **SCM**[24]: SCM is based on two hypotheses i.e. correlation and abstraction. It uses Logistic regression in the space of CCA projected coefficients.
- **JFSSL**[38]: JFSSL is a graph-regularized subspace learning method. It imposes $\ell_{2,1}$-norm on each projection matrix specified for each modality to perform coupled feature selection and simultaneously it introduces a multi-modal graph regularization term to preserve the inter-

modality and intra-modality similarity of the multi-modal data.

Among all the aforementioned approaches, CCA, KCCA, BLM, PLS and KPLS are unsupervised, while GMLDA, CCA-3V, SCM and ML-CCA are supervised ones that utilize label information. KCCA and KPLS are also kernel-based methods. ML-CCA is specified for multi-label multi-modal retrieval.

For CCA, BLM, PLS and GMLDA, we use the implementations of them from the GMA package [30] and the involved parameters are determined by five-fold cross validation. For SCM, we follow the procedure provided by literature [24]. Equally important, seeing that our SCSL performs low-dimensional embedding for each modality to deal with the high-dimensional issue, for fairness, we execute Principal Component Analysis (PCA) on the original features of each modality for all compared approaches. During the training phase of SCSL, to compute the label correlation matrix $\mathbf{C}$, first we represent each class c by a binary vector whose elements are set to be one if the corresponding training multi-modal examples belong to the class c and zeros otherwise. Then we calculate the pairwise class similarity based on their vector representation adopting the normalized RBF kernel which has been used in [23]. Besides, we set $\alpha = 1$ on NUS-WIDE and $\alpha = 10^{-2}$ on VOC20007, and $\gamma = 100$, the dimension of the common subspace is limited to 20.

Table 1 MAP Comparison on NUS-WIDE

Methods	Image as query	Text as query	Average
CCA	0.2436	0.2375	0.2406
KCCA	0.2594	0.2421	0.2508
CCA-3V	0.2911	0.2436	0.2674
BLM	0.2413	0.2271	0.2342
PLS	0.2642	0.2537	0.2590
KPLS	0.2752	0.2588	0.2670
GMLDA	0.2794	0.2598	0.2696
SCM	0.3826	0.3579	0.3703
JFSSL	0.3958	0.3523	0.3742
ML-CCA	0.3896	0.3612	0.3754
SCSL	**0.4113**	**0.3787**	**0.3950**

Table 2 MAP Comparison on VOC2007

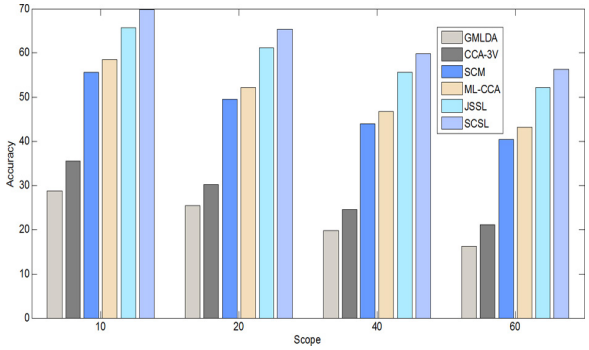
Methods	Image as query	Text as query	Average
CCA	0.2651	0.2579	0.2615
KCCA	0.2728	0.2682	0.2705
CCA-3V	0.2935	0.2754	0.2845
BLM	0.2698	0.2661	0.2680
PLS	0.2773	0.2582	0.2678
KPLS	0.2895	0.2706	0.2801
GMLDA	0.3072	0.2598	0.2835
SCM	0.3259	0.3073	0.3166
JFSSL	0.3458	0.3123	0.3291
ML-CCA	0.3399	0.3068	0.3234
SCSL	**0.3626**	**0.3287**	**0.3457**

4.3 Results on Cross-Modality Retrieval

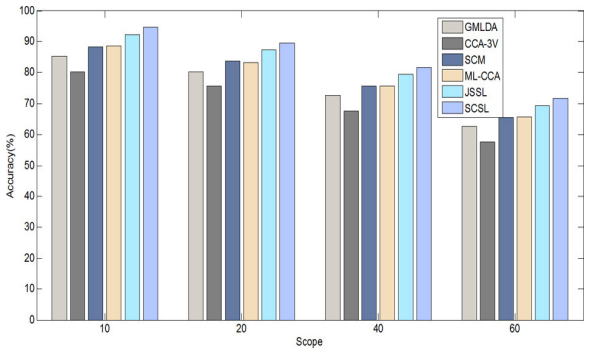
Given a cross-modal problem, using the iterative algorithm in Algorithm 1, we can learn the projections $\mathbf{P}_v$ and $\mathbf{Q}_v$ for low-dimension embedding and the semantic-consistency subspace on the training set. After that, data from different modalities can be projected into a common subspace, where we can measure the relevance of projected data from each modality. In the testing phase, taking data in one modality as a query set, we can retrieve the relevant data from another modality. In the experiment, we use the mean average precision (MAP) [26] as the evaluation metric. As demonstrated in [24], normalized correlation (NC) shows the best average performance for cross-modal retrieval, therefore, here we adopt it as the distance metric to measure similarity. Table 1 and Table 2 provide the MAP scores of all compared methods on NUS-WIDE and VOC2007, respectively. As can be learnt from Table 1 and Table 2, all the subspace learning methods can avail to cross-modal retrieval task. By contrast, supervised ones such as JSSL, GMLDA, CCA-3V, SCM and the proposed SCSL are more effective than unsupervised ones such as CCA, BLM and PLS, which demonstrates that using supervised information (e.g. semantic label information) can facilitate finding a more discriminant common space for cross-modal retrieval. Meanwhile, we can observe that although JSSL, GMLDA, CCA-3V, SCM, SM and SCSL are supervised methods, SCSL and JSSL performs better than GMLDA, CCA-3V and SCM. The reason lies in that SCSL and JSSL can preserve the inter-modality similarity apart from using the semantic information. Moreover, comparing SCSL with JSSL, we can get that the former achieves more pleasurable results than the latter. The reason is that SCSL additionally incorporates semantic correlation between multiple labels and preserves the consistency between feature-based similarity of each modality and semantic-based similarity simultaneously.

4.4 Resulsts on Image-to-Image Retrieval

This subsection is to evaluate the performance of SCSL in terms of the image retrieval task, compared with the supervised methods i.e. GLDA, CCA-3V, SCM, JSSL and ML-CCA. We adopt $Accuracy = N_r/Scope$ as the evaluation metric [54], where N_r and $Scope$ is the number of the returned relevant samples and the total number of all the returned samples, respectively. Setting $Scope = \{10, 20, 40, 60\}$, Fig.

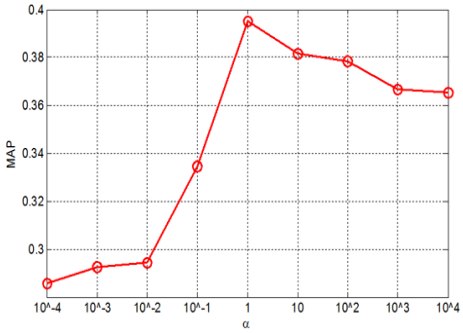


(a) NUS-WIDE

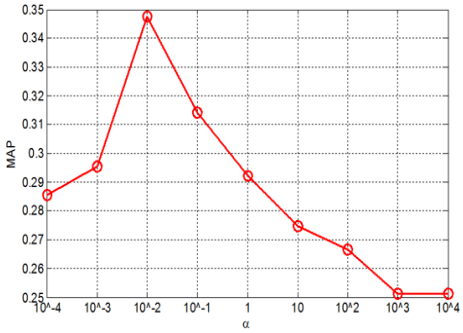


(b) VOC2007

Fig. 4 Accuracy comparison results on image-to-image retrieval



(a) NUS-WIDE



(b) VOC2007

Fig. 5 MAP vs. varying λ on NUS-WIDE and VOC2007

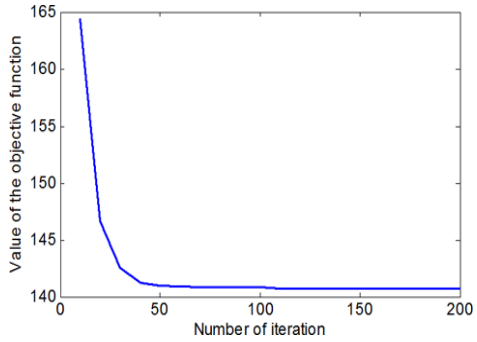
4 shows the retrieval results of the proposed approach on two datasets. As can be seem from Fig. 4, our proposed SCSL, JSSL and ML-CCA achieve better results than the other approaches. Although all the compared methods exploit the semantic label information, our performed SCSL performs best. The main reason behind this is that SCSL not only use the semantic label information but also it incorporates semantic correlation between multiple labels, which is of great benefit to multi-label image retrieval. Moreover, both SCSL and JSSL preserve the intra-modality similarity and inter-modality, while the former performs better than the latter. The reason is that SCSL preserves consistency between the feature-based similarity of each modality and semantic-similarity besides considering intra-modality similarity and inter-modality and semantic correlation. Although both SCSL and ML-CCA exploit the multi-label information to learn a discriminant subspace for retrieval, SCSL outperforms ML-CCA for the retrieval task. It attributes to that SCSL additionally incorporates the label correlation of multi-label images into subspace learning rather than purely using multi-label information like ML-CCA.

4.5 Parameter Sensitivity Analysis

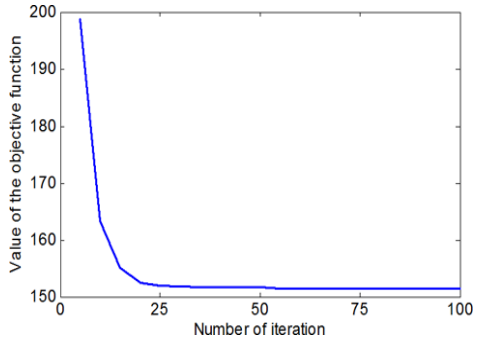
Two parameters α and γ are involved in the optimization model formulated by Eq. (7). Therein, γ is a relaxation one for optimization, which has little influence on the performance of SCSL. Therefore, the follow-up experiments are to show the impact of α on SCSL. Specifically, we tune α in the range of $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$ and execute experiments on NUS-WIDE and VOC2007, respectively. Fig. 5 displays the results of MAP versus different values of α. As shown in Fig. 5, the performance of the proposed SCSL varies when α. It performs best when α is set to 1 and 10^{-2} on NUS-WIDE and VOC2007, respectively. Moreover, α has different effects on SCSL regarding two different datasets. Thus, for different datasets, the best α needs to be explored beforehand to ensure the optimum performance of SCSL.

4.6 Convergence and Computational Time Analysis

To solve the optimization problem involved in the proposed formulation, we have designed an iterative optimizing algorithm, as shown in Algorithm 1. And also, we show that the proposed algorithm is convergent under the designed updating rules and provide the detailed proof theoretically. In the following, we will study the convergence behavior of the proposed algorithm experimentally. Fig. 6 displays the relationship between the objective function and the number of iteration on two datasets. As can be seen from From Fig. 6, we can observe that the objective function value decreases monotonically. In particular, it converges rapidly to a stable



(a) NUS-WIDE



(b) VOC2007

Fig. 6 Value of the objective function vs. the number of iteration

Table 3 Training/Testing Time on the Tested Datasets

Dataset	Training (min)	Testing (s)
NUS-WIDE	20.3	56.5
VOC2007	14.1	29.6

state within about fifty iterations, which demonstrates the efficiency of the proposed iterative optimization algorithm. Moreover, Table 3 reports the total computational time for training and testing on NUS-WIDE and VOC2007 respectively (Using matlab2015b on a 64-bit PC with 3.6 GHZ CPU and 16GB RAM). As can be seen from Table 3, it takes less time to train and test by our proposed method, which also shows the efficiency of the optimization algorithm.

5 Conclusion

In this paper, we have proposed a general framework to learn a semantic-consistent subspace for multi-label cross-modal retrieval. The proposed framework consists of low-dimension embedding for each modality, semantic-consistency projection, and HSIC-based regularization. Under the proposed framework, two problems i.e. high-dimension and heterogeneity of multi-modal data can be well addressed by low-dimension embedding and learning a mid-level semantic consistent space jointly. Through the HSIC-based regularization, the semantic correlation information between multiple labels is well incorporated to learn a more discriminative subspace for multi-label retrieval. Simultaneously, the consistency between feature-based similarity of each modality and semantic-based similarity, the inter-modality similarity and intra-modality similarity, can be jointly preserved. To optimize the proposed model, an effective iterative algorithm is well designed, accompanied by its convergence analysis theoretically and experimentally. Experimental results on two benchmark datasets have shown the superiority of the proposed SCSL over several representative subspace learning approaches.

Acknowledgements This work was jointly supported by National Natural Science Foundation of China (NO.61572068, NO.61532005), National Key Research and Development of China (NO.2016YFB0800404).

References

1. X. Chen, X. Yuan, Q. Chen, S. Yan, and T.-S. Chua. Multi-label visual classification with label exclusive context. In *IEEE Interantional Conference on Computer Vision (ICCV)*, 2011.
2. Y. Chen, L. Wang, W. Wang, and Z. Zhang. Continuum regression for cross-modal multimedia retrieval. In *The IEEE International Conference on Image Processing (ICIP)*, 2012.
3. T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zhang. Nus-wide: a real-world web image database from national university of singapore. In *ACM International Conference on Image and Video*, 2009.
4. C. Cui, P. Lin, X. Nie, Y. Yin, and Q. Zhu. Hybrid textual-visual relevance learning for content-based image retrieval. *Journal of Visual Communication and Image Representation*, 48:367–374, 2017.
5. T. Diethe, D. R. Hardoon, and J. Shawe-Taylor. Multiview fisher discriminative analysis. In *NIPS workshop on Learning from Mutlple Sources*, 2008.
6. M. Everingham, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2010.
7. A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems(NIPS)*, 2013.
8. C. G., C. A. B., M. P. .J, and V. N. Supervised learninig of semantic classess for image annotation and retrieval. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 29(3), 2007.
9. Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, pages 1–24, 2013.

10. A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on Algorithmic Learning Theory. Springer Berlin Heidelberg*, 2005.
11. D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12): 2639–2664, 2004.
12. R. He, M. Zhang, L. Wang, Y. Ji, and Q. Yin. Cross-modal subspace learning via pairwise constraints. *IEEE Transaction on Image Processing*, 24(12):5543–5556, 2015.
13. N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, 2002.
14. H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.
15. S. J. Hwang and K. Grauman. Accounting for the relative importance of objects in image retrieval. In *BMVC*, 2010.
16. S. Ji, S. Yu, and J. Ye. A shared-subspace learning framework for multi-label classification. *ACM Transanctions on Knowledge Discovery from Data(TKDD)*, 4(2):1–29, 2010.
17. Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.
18. S. Jiang, X. Song, and Q. Huang. Relative image similarity learning with contextual information for internet cross-media retrieval. *Multimedia System*, 20(6):645–657, 2014.
19. M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discrinative analysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 38(1):188–194, 2016.
20. F. Kang, R. Jin, and R. Sukthankar. Correlated label propogation with application to multi-label learning. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2006.
21. Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu. Clustering-guided sparse structural learning for unsupervised feature selction. *IEEE Transanctions on Knowledge and Data Engineering(TKDE)*, 26(9):2138–2150, 2014.
22. R. Liao, J. Zhu, and Z. Qin. Nonparametric bayesian upstream supervised multi-modal topic models. In *ACM International Conference on Web Search and Data Mining*, 2014.
23. Y. Liu, R. Jin, and L. Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *AAAI*, 2006.
24. J. C. Pereira, G. Doyle, N. Rasiwasia, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 36(3):521–535, 2014.
25. C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *The NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.
26. N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, and R. L. ahd Nuno Vasconcelos. A new appraoch to cross-modal multimedia retrieval. In *International Confeence on MultiMedia*, 2010.
27. R. Rosipal and L. J.Trejo. Kernel partial least square regression in reproducing kernel hilbert space. *Pattern Recognition*, 36(9):1961–1971, 2003.
28. A. S. A kernel method for canonical correlation analysis. In *The International Meeting of the Psychometric Society(IMPS2001)*, 2007.
29. A. Sharma and D. W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, lowresolution and sketch. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2011.
30. A. Sharma, A. Kumar, and H. D. III. Generalized multi-view analysis: A discriminative latent space. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2012.
31. X. Shu, G. Qi, J. Tang, and J. Wang. Weakly-shared deep transfer newworks for heterogeneous-domain knowledge propagation. In *ACM International Conference on Multimedia*, 2015.
32. G. Song, S. Wang, Q. Huang, and Q. Tian. Multimodal similarity gaussian process latent variable model. *IEEE Transanctions on Image Prosessing*, 26(9):4168–4181, 2017.
33. K. Tae-Kyun, J. Kittler, and R. Cipolla. Discrinative learning and recognition of image set classess using canonical correlation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 29(6):1005–1018, 2007.
34. J. Tang, X. Shu, Z. Li, G. Qi, and J. Wang. Generalized deep transfer networks for knowledge propagation in heterogeneous domains. *ACM Transactions on Multimedia Computing, Communications and Applications*, 12(4s):1–22, 2016.
35. J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.
36. R. Udupa and M. Khapra. Improving the multilingual user experience of wikipedia using cross-language name search. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

37. N. R. Viresh Ranjan and C. Jawahar. Multi-label cross-modal retrieval. In *IEEE Interantional Conference on Computer Vision (ICCV)*, 2015.
38. K. Wang, R. He, L. Wang, W. Wang, and T. Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Transanctions on Pattern Analysis and Machine Intelligence*, 38(10):2010–2023, 2016.
39. K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang. A comprehenseice survey on cross-modal retrieval. *arXiv:1607.06215[cs.MM]*, 2016.
40. S. Wang and S. Jiang. Instre:a new benchmark for instance-leverl object retrieval and recognition. *TOMCCAP*, 11(3):1–37, 2015.
41. Y. Wang, F. Wu, J. Song, X. Li, and Y. Zhuang. Multi-modal mutual topic reinforce modeling for cross-media retrieval. In *ACM International Conference on Multimedia*, 2014.
42. Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE Transactions on Cybernetics*, 47 (2):449–460, 2017.
43. Y. Wu, S. Wang, and Q. Huang. Online asymmetric similarity learning for cross-modal retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017.
44. D. Xu and S. Yan. Semi-supervised bilinear subspace learning. *IEEE Transactions on Image Processing*, 18 (7):1671–1676, 2009.
45. J. Yang, S. Yan, and T. S. Huang. Ubiquitously supervised subspace learning. *IEEE Transactions on Image Processing*, 18(2):241–249, 2008.
46. D. Zhang, M. M. Islam, and G. Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 45:346–362, 2012.
47. L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian. Generalized semi-supervised and structured subspace learning for cross-modal retrieval. *IEEE Transactions on Multimedia*, 19(6):1220–1233, 2017.
48. X. Zhang, Y. Yu, M. White, R. Huang, and D. Schuurmans. Convex sparse coding, subspace learning and semi-supervised extensions. In *AAAI*, 2011.
49. Y. Zhang and J. G. Schneider. Multi-label output codes using canonical correaltion analysis. In *The 14th International Conference on Artificial Intelligence and Statistics*, 2011.
50. Y. Zhang and Z. Zhou. Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):14, 2010.
51. F. Zhao, Y. Huang, L. Wang, and T. Tan. Deep semantic ranking based hashing for multi-label image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2015.
52. Y. Zheng, Y. Zhang, and H. Larochelle. Topic modeling of multimodal data: an autoregressive approach. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2014.
53. S. Zhu, X. Ji, W. Xu, and Y. Gong. Multi-labelled classification using maximum entropy method. In *The 28th annual international ACM SIGIR Conferenece on Research and Development in Information Retrieval*, 2005.
54. Z. Zhu, J. Cheng, Y. Zhao, and J. Ye. Lsslp-local structure sensitive label propagation. *Information Sciences*, 332:19–32, 2016.
55. Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, 2013.