Generalized Semi-supervised and Structured Subspace Learning for Cross-Modal Retrieval

Liang Zhang, Bingpeng Ma, Guorong Li, Qingming Huang, *Senior Member, IEEE*, and Qi Tian, *Fellow, IEEE*

***Abstract*—Motivated by the fact that unlabeled data can be easily collected and help to exploit the correlations among different modalities, this paper proposes a novel method named generalized semi-supervised structured subspace learning (GSS-SL) for the task of cross-modal retrieval. First, to predict more relevant class labels for unlabeled data, we propose a label graph constraint that ensures the intrinsic geometric structures of different feature spaces consistent with that of label space. Second, considering that class labels directly reveal the semantic information of multimedia data, GSS-SL takes the label space as a linkage to model the correlations among different modalities. Concretely, the label graph constraint, label-linked loss function, and regularization are integrated into a joint minimization formulation to learn a discriminative common subspace. Finally, an efficient optimization algorithm is designed to alternately optimize multiple linear transformations for different modalities and update the class indicator matrices for unlabeled data. Furthermore, an arbitrary number of modalities can be solved in the proposed framework. Extensive experiments on three standard benchmark datasets demonstrate that GSS-SL outperforms previous methods on exploiting the correlations among different modalities.**

***Index Terms*—Cross-modal retrieval, documents and images, semi-supervised learning.**

I. Introduction

With the rapid growth of multimedia data such as text, image, video and audio, cross-modal retrieval has become increasingly important and attracted considerable attention in the multimedia research field [1]–[14]. The task of cross-modal retrieval is to predict whether samples from different modalities represent the same semantic concept. In this paper, we pay more attention to enhance the correlation between image modality and text modality although the fundamental framework is applicable to any two different modalities.

The major problem of cross-modal retrieval is how to evaluate the similarity among the multimedia data since different modalities lie on different feature spaces. Thus, a lot of works have been developed to settle this problem by learning multiple transformations to map the different modalities into a common subspace. In Fig. 1, we show the common subspaces learned by different methods. Based on the unsupervised learning, one kind of common subspace learning method obtains the common subspace by learning the transformations using the paired samples between two modalities [1]–[3]. Paired samples denote that samples from different modalities represent the same semantic concept. Since the unsupervised methods require the training data to be paired, they only unite the paired samples in the common subspace. Besides, two samples from the different modalities can be paired if their underlying semantic are consistent, so they cannot deal with the unlabeled data, as shown in Fig. 1(b). The other kind of common subspace learning method is based on the supervised learning [4], [5], [7]–[14]. These methods use the class information to learn a discriminant latent space, where same-class samples are united and different classes are separated. However, they cannot predict labels for the unlabeled data because they ignore the underlying data distribution, as shown in Fig. 1(c).

The above methods only exploit labeled data to learn the common subspace. Although the labeled data (including single-labeled and multi-labeled data) are accurate for exploiting the correlations among the different modalities, we are often faced with a large amount of unlabeled data since manually annotated data is very expensive in practice. Hence, it is particularly important to design a semi-supervised framework that can jointly use both labeled data and unlabeled data to exploit the correlations among multi-modal data. Although some semi-supervised methods are proposed for matching heterogeneous samples [15], [16], they cannot handle the problem of out-of-sample testing. When new testing samples appear, such methods need to combine those new testing data with the existing data, and then reconstruct the graph model based on the combined data. Hence, such methods are very inefficient for processing the out-of-sample data.

To overcome the aforementioned problems, this paper proposes a novel semi-supervised framework, named **G**eneralized **S**emi-supervised and **S**tructured **S**ubspace **L**earning (GSS-SL),

Manuscript received January 17, 2017; revised May 16, 2017; accepted June 29, 2017. Date of publication July 5, 2017; date of current version December 14, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61332016, Grant 61620106009, Grant 61572465, Grant 61429201, Grant U1636214, Grant 61650202, and Grant 61303153, in part by the National Basic Research Program of China (973 Program) under Grant 2015CB351800, and in part by the Key Research Program of Frontier Sciences, CAS under Grant QYZDJ-SSW-SYS013. The work of Q. Tian was supported in part by the ARO Grant W911NF-15-1-0290 and in part by the Faculty Research Gift Awards by the NEC Laboratories of America and Blippar. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marco Bertini. *(Corresponding author: Bingpeng Ma.)*

L. Zhang, B. Ma, G. Li, and Q. Huang are with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100190, China, and also with the Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhangliang14@mails.ucas.ac.cn; bpma@ucas.ac.cn; liguorong@ucas.ac.cn; qmhuang@ucas.ac.cn).

Q. Tian is with the Department of Computer Sciences, University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: qitian@cs.utsa.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2017.2723841

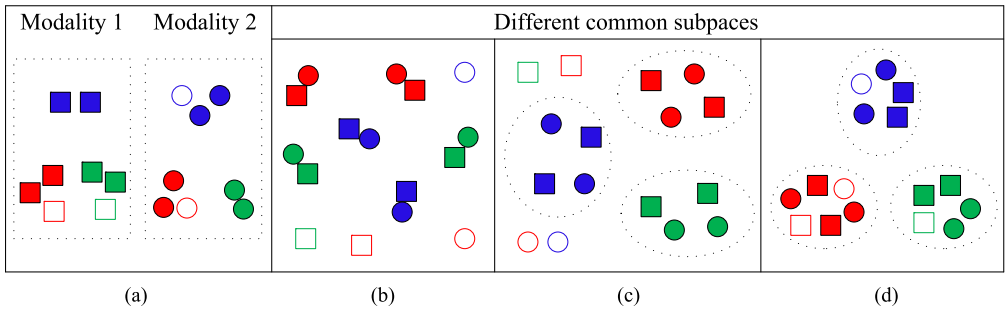


Fig. 1. Simple demonstration of common subspaces learned by various cross-modal methods along with the proposed method. Shapes represent different modalities (i.e., text and image). The same color indicates relevant semantics. Empty circles and squares represent the unlabeled data.

for cross-modal retrieval. First, we propose a label graph to predict the proper class labels for the unlabeled data, which can help to exploit the semantic information for multi-modal data. According to graph theory, if the edge weight between two vertices on affinity matrix of feature space is large, the class labels of these two samples should be relevant. To make the similar features have more relevant labels, we adopt the label information to construct the label graph, and then we constrain the structures of feature graphs of different modalities are close to that of the label graph. By this way, labels of similar features are more relevant and heterogeneous samples within the same class are close to each other in the common subspace.

Second, to alleviate the problem on ignoring data distribution in most of the supervised learning methods, we exploit the intrinsic geometric structure at the common subspace learning stage. Specially, we adopt the graph constraint to ensure that the neighboring data in the original space are close to each other in the common subspace. Furthermore, considering that labels directly reveal the semantic information of the multi-modal data, we argue that the label space can be taken as the common subspace of the different modalities. Then we enhance the correlations between the transformed spaces of the different modalities and the label space by a $\ell_{2,1}$-norm based loss function, by which all the features are close to their true labels and far away from their irrelevant labels. The label graph constraint, label-linked loss function and regularization can be integrated into a generic minimization formulation and then a more discriminative common subspace is learned.

Finally, we design an iterative algorithm to solve the proposed optimization problem. By alternately optimizing the transformations and the predicted labels, GSS-SL achieves more discriminative transformations for the multi-modal data and predicts more accurate labels for the unlabeled data. It is important to note that we use the label space as the linkage among the different modalities in the optimization process. By this way, the arbitrary number of modalities can be handled in our framework.

Fig. 1(d) shows the common subspace learned by the proposed method. In this figure, GSS-SL clusters the same-class samples and separates the different classes in the common subspace. Meanwhile, the structures of different modalities are consistent with that of the label space such that the labels of the unlabeled data can be predicted. The main contributions of this paper are summarized as follows:

1) The label graph is proposed to effectively handle the unlabeled data. By this way, the unlabeled data can achieve more relevant labels such that single-labeled, multi-labeled and unlabeled training data can simultaneously be used to model the correlations among different modalities.
2) A novel semi-supervised framework is developed for the cross-modal retrieval. It can unify the label graph, $\ell_{2,1}$-norm based loss function and regularization into a generic minimization formulation so that label prediction, discriminative feature selection and subspace learning can be performed.
3) A novel optimization strategy is designed to efficiently solve the complex minimization problem. In each iteration, the minimization problem is simplified to solve multiple linear transformations and predicted label matrices such that it can be easily generalized to arbitrary number of modalities.

To validate GSS-SL's effectiveness on handling the unlabeled data, we conduct extensive experiments on three public datasets: Wiki [3], Pascal VOC 2007 [17] and NUS-WIDE [18]. On all three datasets, GSS-SL outperforms the existing cross-modal methods using part of labeled training data. The performance of our approach can be further improved using all the labeled training data.

II. Related Work

Cross-modal retrieval has become increasingly important in many real-world applications. Hence, extensive methods have been developed to improve its performance in the recent past. Taking the class labels of multimedia data into account or not, the traditional methods can be categorized into unsupervised learning, semi-supervised learning and supervised learning.

The unsupervised cross-modal methods use paired training samples between two modalities to learn the common subspace, including Partial Least Squares (PLS) [21], Canonical Correlation Analysis (CCA) [19], Bilinear Model (BLM) [22], Kernel CCA (KCCA) [19], and Deep CCA (DCCA) [2]. PLS [21] linearly maps images from different modalities into a common subspace in which the different modalities are highly correlated [21]. CCA learns a common subspace by maximizing the correlation between the projected vectors of two different modalities. Rasiwasia *et al.* [3] first adopt CCA to solve the

TABLE I
REQUIREMENTS OF POPULAR METHODS FOR CROSS-MODAL RETRIEVAL

Properties \ Methods	CCA[19]	SCM[3]	LCFS[6]	MvDA[20]	LGCFL[4]	ml-CCA[5]	GMLDA & GMMFA [12]	GSS-SL
S		√	√	√	√	√	√	√
SS								√
ML	√		√		√	√		√
FS			√					√

Note: **S**: **S**upervised, **SS**: **S**emi-**S**upervised, **ML**: **M**ulti-**L**abel, **F**eature **S**election ('√' Indicates Presence of Property).

problem of cross-modal multimedia retrieval. As a kernelized extension of CCA, KCCA [19] utilizes kernel function to project the features into a higher-dimensional space, and then maximizes the correlations between two modalities by performing CCA. Based on KCCA, Hwang *et al.* [23] design an unsupervised rank approach to leverage the implicit order information between images and tags, and Ballan *et al.* [24] adopt advanced nearest-neighbor voting algorithm to connect visual and textual modalities. Specifically, DCCA [2] combines the autoencoder model [25] with CCA to learn a set of flexible nonlinear transformations for the multimedia data. However, these methods require the training samples to be paired, so they just focus on the pair-wise closeness in the common subspace.

The supervised cross-modal methods learn the common subspace by using class information. Considering that the class labels are directly applied to multimedia data, extensive supervised learning methods have been developed [3]–[5], [9], [12], [20], [26]–[29] to model the correlations among multi-modal data. For example, Generalized Multiview Analysis (GMA) [12] is proposed to exploit the label information for the discriminant latent space learning. Multiview Discriminant Analysis (MvDA) [20] obtains the common subspace by learning discriminative information from both intra-view and inter-view scenarios. Wang *et al.* [7] propose a half quadratic optimization to learn the coupled feature spaces for two modalities.

Since a lot of samples may naturally be annotated with multiple labels in practice, it is imperative to develop a framework to handle the multi-labeled data. Ranjan *et al.* [5] propose the multi-label CCA (ml-CCA) to learn the shared subspaces by considering the multi-labeled data. Kang *et al.* [4] use both single-labeled and multi-labeled data to learn the consistent feature representations for the multi-modal data.

Besides, many deep models have been developed to utilize the class information to enhance correlations among the multimedia data. Specifically, multimodal auto-encoders [28] and deep boltzmann machines [29] are proposed to exploit the discriminative information of two modalities by learning deep-based shared representation.

In this paper, we propose a generalized semi-supervised framework for cross-modal retrieval. The semi-supervised methods use both labeled and unlabeled data to learn the common subspace. Some semi-supervised methods are proposed to handle the unlabeled data [15], [16]. However, these methods only use the unlabeled data to increase the diversity of training data, and cannot handle the problem of out-of-sample testing. By contrast, our framework can optimize label prediction and handle multi-labeled data. In Table I, we list the popular cross-modal methods as well as their satiable properties. From this table, we can conclude that only our method satisfies all the mentioned requirements.

III. SEMI-SUPERVISED STRUCTURED SUBSPACE LEARNING

In this section, we first present a novel semi-supervised framework for cross-modal learning, which can be easily generalized to deal with arbitrary number of modalities. Then, an iterative method is designed to solve this framework.

A. Notations

Assume we have a training set consisting of m modalities $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_m\}$, where $\mathbf{X}_r$ denotes the rth modality. In real application, there are amounts of unlabeled data for each modality, so the rth modality $\mathbf{X}_r = [\check{\mathbf{X}}_r, \hat{\mathbf{X}}_r] \in \mathbb{R}^{d_r \times n_r}$ contains both labeled data $\check{\mathbf{X}}_r = [\mathbf{x}_{r1}, \mathbf{x}_{r2}, \ldots, \mathbf{x}_{rn_l}]$ and unlabeled data $\hat{\mathbf{X}}_r = [\mathbf{x}_{r(n_l+1)}, \ldots, \mathbf{x}_{rn_r}]$, where $\mathbf{x}_{ri} \in \mathbb{R}^{d_r}$ denotes the ith sample of the rth modality and n_r is the number of samples of the rth modality.

As for label representation, each sample $\mathbf{x}_{ri}$ is assigned with a c-dimensional binary-valued vector $\mathbf{p}_{ri} \in \mathbb{R}^c$, where c is the class number of a dataset. If $\mathbf{x}_{ri}$ is classified as the kth class, p_{rik} is set to 1, otherwise 0. For the single-labeled data, their label vectors only contain one nonzero value. However, it is difficult to describe the semantic content of a sample using single label in practice. Thus a large amount of multimedia data are usually annotated multiple labels because of its complicated semantics. Their label vectors contain more than one nonzero values. The class indicator matrix for the labeled data is constructed as $\mathbf{P}_l = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{n_l}]^T$, which is identical for all modalities. After obtaining the predicted class indicator matrix $\hat{\mathbf{P}}_r$ for the unlabeled data $\hat{\mathbf{X}}_r$, the class indicator matrix for rth modality is denoted as $\mathbf{P}_r = \left[\begin{smallmatrix}\mathbf{P}_l \\ \hat{\mathbf{P}}_r\end{smallmatrix}\right] \in \mathbb{R}^{n_r \times c}$. In this paper, we can simultaneously deal with unlabeled data, single-labeled data and multi-labeled data.

Obviously, different modalities lie on the different feature spaces, so it is difficult to compute the similarity between two heterogeneous samples. Hence, we focus on learning multiple transformations $\{\mathbf{U}_1, \cdots, \mathbf{U}_m\}$ to link the original feature spaces and the learned common subspace, where $\mathbf{U}_r \in \mathbb{R}^{d_r \times c}$ is the transformation for the rth modality and c is the dimensionality of the common subspace. Then for each sample $\mathbf{x}_{ri}$, the corresponding representation in the common subspace is $\mathbf{h}_{ri} = \mathbf{U}_r^T \mathbf{x}_{ri}$. Finally, we utilize the cosine similarity $\cos(\mathbf{h}_{ri}, \mathbf{h}_{mi}) = \frac{\mathbf{h}_{ri} \cdot \mathbf{h}_{mi}}{\|\mathbf{h}_{ri}\| \|\mathbf{h}_{mi}\|}$ to calculate the similarity between two heterogeneous samples in the common subspace, where

$\mathbf{h}_{ri} \cdot \mathbf{h}_{mi}$ is the inner-product of $\mathbf{h}_{ri}$ and $\mathbf{h}_{mi}$, and $\|\cdot\|$ denotes the Euclidean norm of a vector.

B. Objective Function

We construct a generalized semi-supervised structured subspace learning framework to learn the optimal transformations for the multiple modalities. In this framework, multiple transformations are learned and labels of unlabeled data are predicted simultaneously:

$$\min_{\mathbf{U}_1,\mathbf{P}_1;\cdots;\mathbf{U}_m,\mathbf{P}_m} J(\mathbf{U}_1,\mathbf{P}_1;\cdots;\mathbf{U}_m,\mathbf{P}_m) = \min_{\mathbf{U}_1,\mathbf{P}_1} J_1(\mathbf{U}_1,\mathbf{P}_1) + \cdots + \min_{\mathbf{U}_m,\mathbf{P}_m} J_m(\mathbf{U}_m,\mathbf{P}_m) \quad (1)$$

where $J_r(\mathbf{U}_r,\mathbf{P}_r)$ learns the optimal transformation $\mathbf{U}_r$ and predicts labels for the rth modality

$$J_r(\mathbf{U}_r,\mathbf{P}_r) = \mathcal{L}(\mathbf{U}_r,\mathbf{P}_r) + \lambda\Psi(\mathbf{U}_r,\mathbf{P}_r) + \gamma\Omega(\mathbf{U}_r,\mathbf{P}_r) \quad (2)$$

where $\mathcal{L}(\mathbf{U}_r,\mathbf{P}_r)$ is the label-linked loss function to ensure that samples from the different modalities are close to each other within the same class while far away between the different classes. $\Psi(\mathbf{U}_r,\mathbf{P}_r)$ is a label graph defined to exploit the underlying data structure and ensure labels of similar features close to each other. $\Omega(\mathbf{U}_r,\mathbf{P}_r)$ avoids the overfitting problem and selects discriminative features from original features. $\lambda > 0$, $\gamma > 0$ are the balancing parameters.

Loss constraint: This constraint item is defined to minimize the labeling approximation error such that samples from the different modalities within the same class are close to each other in the label space.

It is well known that class labels directly reveal the semantic information of multimedia data, i.e., semantic content described by multimedia data usually can be summarized as label annotation. Thus, we regard the label space as the common latent subspace, into which the different modalities should be transformed. In the common subspace, samples from different modalities are linked with the labels so that the heterogeneous features from the same semantic class become more similar after transforming.

The labeling error minimized by the least squares loss is very sensitive to outliers [30], [31]. Hence, we use the sum of labeling error based on $\ell_{2,1}$-norm to define this item as

$$\mathcal{L}(\mathbf{U}_r,\mathbf{P}_r) = \frac{1}{2}\sum_{i=1}^{n_r}\|\mathbf{U}_r^T\mathbf{x}_{ri} - \mathbf{p}_{ri}\|_{2,1}. \quad (3)$$

For each sample $\mathbf{x}_{ri}$, we use a class label vector $\mathbf{p}_{ri}$ to explicitly characterize its multi-label information. Then we embed the class information of $\mathbf{x}_{ri}$ into the $\ell_{2,1}$-norm loss function, by which the transformed $\mathbf{x}_{ri}$ will be enforced close to the entries "+1" of $\mathbf{p}_{ri}$ so that $\mathbf{x}_{ri}$ will construct semantic relationship with its ground truth labels. Furthermore, it shall be noted that GSS-SL uses the label space as a linkage to ensure samples from different modalities close to their true labels and far away from their irrelevant labels. Hence, different modalities are connected with each other and their correlations are effectively exploited.

Label graph constraint: We define this item to exploit the geometry structures for both label space and feature space. Furthermore, to predict relevant labels for the unlabeled data, we introduce the label information to the graph structure and ensure that the structures of different feature spaces are consistent with that of the label space.

We assume that the neighboring data in the original feature space should be close to each other in the common subspace. It may lead to undesirable transformations when the underlying data structure and distribution are ignored at the learning stage. Moreover, by constraining the consistent structures between feature space and label space, we guarantee that the nearest features have the closest labels such that label propagation can be conducted from labeled data to unlabeled data.

We first construct the k-nn graph for the rth modality $\mathbf{X}_r$. All training samples including the labeled and unlabeled ones are regarded as the vertices in the graph. $\mathbf{x}_{ri}$ and $\mathbf{x}_{rj}$ can be connected if one of them belongs to the other's k-nearest neighbors, and the corresponding edge weight is calculated as

$$S_r^{ij} = \begin{cases} \exp(-\frac{\|\mathbf{x}_{ri}-\mathbf{x}_{rj}\|^2}{\sigma_r^2}) & \mathbf{x}_{ri} \in \mathcal{N}_k(\mathbf{x}_{rj}) \text{ or} \mathbf{x}_{rj} \in \mathcal{N}_k(\mathbf{x}_{ri}) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $\mathcal{N}_k(\mathbf{x}_{ri})$ (or $\mathcal{N}_k(\mathbf{x}_{ri})$) denotes the k-nearest neighbors of $\mathbf{x}_{ri}$ (or $\mathbf{x}_{rj}$), and σ_r is the bandwidth parameter for the rth modality. $\mathbf{S}_r$ is a $n_r \times n_r$ symmetric undirected graph and its edge weights are non-negative values. Let $\mathbf{E}_r$ be the diagonal matrix with $E_r^{ii} = \sum_{j=1}^{n_r} S_r^{ij}$. The normalized graph Laplacian matrix $\mathbf{L}_r$ is formulated as

$$\mathbf{L}_r = \mathbf{E}_r^{-\frac{1}{2}}(\mathbf{E}_r - \mathbf{S}_r)\mathbf{E}_r^{-\frac{1}{2}}. \quad (5)$$

We use the graph $\mathbf{L}_r$ to achieve two goals. First, $\mathbf{L}_r$ can realize that the neighboring data points in the original feature space are close to each other in the common subspace. Second, according to graph theory, the class labels of two samples are similar if the edge weight between the two vertices is large [32]–[34]. Thus, $\mathbf{L}_r$ is used to explore the underlying structures of feature space and label space and propagate the label information from labeled data to unlabeled data in GSS-SL. Based on the above analyses, we design $\Psi(\mathbf{U}_r,\mathbf{P}_r)$ for the rth modality as

$$\Psi(\mathbf{U}_r,\mathbf{P}_r) = \frac{1}{2}\sum_{i,j=1}^{n_r} S_r^{ij}\left\|\frac{\mathbf{U}_r^T\mathbf{x}_{ri}}{\sqrt{E_r^{ii}}} - \frac{\mathbf{U}_r^T\mathbf{x}_{rj}}{\sqrt{E_r^{jj}}}\right\|_2^2 - \frac{1}{2}\sum_{i,j=1}^{n_r} S_r^{ij}\left\|\frac{\mathbf{p}_{ri}}{\sqrt{E_r^{ii}}} - \frac{\mathbf{p}_{rj}}{\sqrt{E_r^{jj}}}\right\|_2^2. \quad (6)$$

Although we construct a local graph for each modality, all modalities are also connected with each other because the graphs of different modalities are constrained to be consistent with the same label graph. Besides, by building a local graph for each modality, GSS-SL reduces the computational cost compared with the global graph, which is constructed by using all training samples and testing samples [9], [15], [16]. Furthermore, the global graph methods are inefficient in realizing the out-of-sample testing. This is because they must utilize all labeled and unlabeled data to construct graph model, and then the labels of

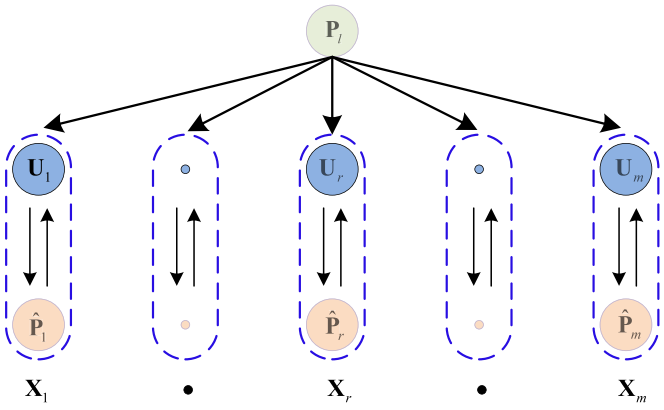


Fig. 2. Correlation of multiple modalities in the optimization procedure.

unlabeled data are obtained by propagating the label information from labeled data to unlabeled data. To make it more efficient, we constrain the structure of feature space consistent with that of label space by the unified graph. With this supervised instruction, the multi-modal data will further construct semantic correlation with class labels. When a new testing sample arrives, the proposed loss constraint ensures that features from the same semantic class become similar after transforming. Then the label graph will help the nearest features to find the most relevant labels. Hence, although the new data is not combined with the previous data to construct the graph model, the proposed method effectively solve the problem of out-of-sample testing.

Regularization constraint: This constraint item aims to avoid the overfitting problem induced by the sparse features and select discriminative features from the original feature. The $\ell_{2,1}$-norm based regularization first obtains a vector by computing 2-norm for each row of a matrix, and then 1-norm is acted on the obtained vector. These operations guarantee that matrix is sparse in rows, by which feature selection can be conducted from feature space to the common subspace [30], [31]. We utilize $\ell_{2,1}$-norm to formulate this item as follows:

$$\Omega(\mathbf{U}_r, \mathbf{P}_r) = \frac{1}{2}\|\mathbf{U}_r\|_{2,1} + \frac{1}{2}\|\mathbf{P}_r\|_{2,1}. \tag{7}$$

For any modality $\mathbf{X}_r, r = 1, \cdots, m$, we can model it by using the above semi-supervised learning framework. Then we obtain the overall objective function in (1).

C. Optimization

To exploit the correlations among different modalities, it is important to exchange the information of different modalities during the optimization procedure. Fig. 2 shows the illustration of our optimizing principles. For each modality, we alternately optimize the transformations and update the predicted labels for the unlabeled data. As for the correlations among the multi-modal data, we use the label space as the linkage to simultaneously optimize multiple transformations, and their correlations are strengthened with each other.

From (3) and (6), it is very difficult to take the derivative with respect to $\mathbf{U}_r$ or $\mathbf{P}_r$ in the optimization process. The computational process would be iterate over n_r, which is very slow for the large-scale dataset. Based on the definition of $\ell_{2,1}$-norm and the graph theory, we rewrite the compact matrix representation of the objective function as

$$\begin{aligned} J_r(\mathbf{U}_r, \mathbf{P}_r) = {} & \mathrm{Tr}((\mathbf{X}_r^T\mathbf{U}_r - \mathbf{P}_r)^T\mathbf{D}_r(\mathbf{X}_r^T\mathbf{U}_r - \mathbf{P}_r)) \\ & + \lambda\mathrm{Tr}(\mathbf{U}_r^T\mathbf{X}_r\mathbf{L}_r\mathbf{X}_r^T\mathbf{U}_r - \mathbf{P}_r^T\mathbf{L}_r\mathbf{P}_r) \\ & + \gamma\mathrm{Tr}(\mathbf{U}_r^T\mathbf{Q}_r\mathbf{U}_r) + \gamma\mathrm{Tr}(\mathbf{P}_r^T\mathbf{S}_r\mathbf{P}_r) \end{aligned} \tag{8}$$

where $\mathbf{D}_r$, $\mathbf{Q}_r$ and $\mathbf{S}_r$ are the diagonal matrices with $D_r^{ii} = \frac{1}{2\|(\mathbf{X}_r^T\mathbf{U}_r - \mathbf{P}_r)^i\|_2}, i = 1, \ldots, n_r$, $Q_r^{ii} = \frac{1}{2\|\mathbf{U}_r^i\|_2}, i = 1, \ldots, d_r$ and $S_r^{ii} = \frac{1}{2\|\mathbf{S}_r^i\|_2}, i = 1, \ldots, n_r$.

Convexity analysis: We observe that it is difficult to solve the objective function because of its non-smoothness. To solve it, we first prove that the optimization problem in (8) is jointly convex with respect to $\mathbf{U}_r$ and $\mathbf{P}_r$.

Theorem 1: $J_r(\mathbf{U}_r, \mathbf{P}_r)$ is jointly convex with respect to $\mathbf{U}_r$ and $\mathbf{P}_r$.

Proof: $J_r(\mathbf{U}_r, \mathbf{P}_r)$ can be rewritten as the matrix form

$$J_r(\mathbf{U}_r, \mathbf{P}_r) = \mathrm{Tr}\left(\begin{bmatrix}\mathbf{U}_r \\ \mathbf{P}_r\end{bmatrix}^T \left[\mathbf{B}_{\mathcal{L}} + \lambda B_{\Psi} + \gamma\mathbf{B}_{\Omega}\right] \begin{bmatrix}\mathbf{U}_r \\ \mathbf{P}_r\end{bmatrix}\right) \tag{9}$$

where $\mathbf{B}_{\mathcal{L}} = \begin{bmatrix}\mathbf{X}_r\mathbf{D}_r\mathbf{X}_r^T & -\mathbf{X}_r\mathbf{D}_r \\ -\mathbf{D}_r\mathbf{X}_r^T & \mathbf{D}_r\end{bmatrix}$, $\mathbf{B}_{\Psi} = \begin{bmatrix}\mathbf{X}_r\mathbf{L}_r\mathbf{X}_r^T & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_r\end{bmatrix}$ and $\mathbf{B}_{\Omega} = \begin{bmatrix}\mathbf{Q}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_r\end{bmatrix}$.

Based on the second-order conditions of convex functions [35], $\mathbf{B}_{\mathcal{L}} + \lambda\mathbf{B}_{\Psi} + \gamma\mathbf{B}_{\Omega}$ is required to be positive semi-definite if $J_r(\mathbf{U}_r, \mathbf{P}_r)$ is jointly convex with respect to $\mathbf{U}_r$ and $\mathbf{P}_r$. Since $\mathbf{B}_{\Omega}$ is positive semi-definite according to (8), we only need to prove that $\mathbf{B}_{\mathcal{L}}$ and $\mathbf{B}_{\Psi}$ are positive semi-definite.

Given arbitrary vector $\mathbf{z} = \begin{bmatrix}\mathbf{z}_1 \\ \mathbf{z}_2\end{bmatrix}^T \in \mathbb{R}^{d_r + n_r}$, where $\mathbf{z}_1 \in \mathbb{R}^{d_r \times 1}$ and $\mathbf{z}_2 \in \mathbb{R}^{n_r \times 1}$, we have

$$\begin{aligned} & \mathrm{Tr}\left(\begin{bmatrix}\mathbf{z}_1 \\ \mathbf{z}_2\end{bmatrix}^T \mathbf{B}_{\mathcal{L}} \begin{bmatrix}\mathbf{z}_1 \\ \mathbf{z}_2\end{bmatrix}\right) \\ & \quad = \mathrm{Tr}\left(\begin{bmatrix}\mathbf{z}_1 \\ \mathbf{z}_2\end{bmatrix}^T \begin{bmatrix}\mathbf{X}_r\mathbf{D}_r\mathbf{X}_r^T & -\mathbf{X}_r\mathbf{D}_r \\ -\mathbf{D}_r\mathbf{X}_r^T & \mathbf{D}_r\end{bmatrix} \begin{bmatrix}\mathbf{z}_1 \\ \mathbf{z}_2\end{bmatrix}\right) \\ & \quad = (\mathbf{X}_r^T\mathbf{z}_1 - \mathbf{z}_2)^T\mathbf{D}_r(\mathbf{X}_r^T\mathbf{z}_1 - \mathbf{z}_2) \\ & \quad \geq 0. \end{aligned} \tag{10}$$

$\mathbf{D}_r$ is a diagonal matrix with positive values and the above equation holds up. Hence, $\mathbf{B}_{\mathcal{L}}$ is positive semi-definite.

Similarly, $\mathbf{B}_{\Psi}$ is also positive semi-definite. In the proofs, one thing to note is that L_r is a symmetric undirected matrix with non-negative edge weight and the diagonal entries of $\mathbf{L}_r$ are no less than the non-diagonal entries.

It is true that the sum of three convex function is convex, so $J(\mathbf{U}_r, \mathbf{P}_r)$ is jointly convex with respect to $\mathbf{U}_r$ and $\mathbf{P}_r$. ■

Initialization: We initialize labels for unlabeled data based on the graph theory. The class labels of two samples should be relevant if the edge weight between the two vertices on graph matrix is large. According to the above assumption, we

propagate the class label information from labeled data to unlabeled data by the following way:

$$\min_{\mathbf{P}} \mathbf{P}_r^T \mathbf{L}_r \mathbf{P}_r, r = 1, \cdots, m \tag{11}$$

where $\mathbf{L}_r$ is the normalized Laplacian matrix defined in (5). Hence, (11) can be rewritten as following:

$$\min_{\hat{\mathbf{P}}_r} = \text{Tr}\left(\begin{bmatrix} \mathbf{P}_l \\ \hat{\mathbf{P}}_r \end{bmatrix}^T \begin{bmatrix} \mathbf{L}_r^{ll} & \mathbf{L}_r^{lu} \\ \mathbf{L}_r^{ul} & \mathbf{L}_r^{uu} \end{bmatrix} \begin{bmatrix} \mathbf{P}_l \\ \hat{\mathbf{P}}_r \end{bmatrix} \right) \tag{12}$$

since we know the label matrix $\mathbf{P}_l$ for the first n_l labeled samples, the above formulation has the unique solution

$$\hat{\mathbf{P}}_r = -\mathbf{L}_{uu}^{-1} \mathbf{L}_{ul} \mathbf{P}_l. \tag{13}$$

Once the initial value $\hat{\mathbf{P}}_r$ is given, in each iteration, we first update $\mathbf{U}_r$ given the total label matrix $\mathbf{P}_r = \left[\begin{smallmatrix} \mathbf{P}_l \\ \hat{\mathbf{P}}_r \end{smallmatrix}\right]$. Then we update the predicted labels $\hat{\mathbf{P}}_r$ for the unlabeled data. The two alternating steps are described as below.

Update $\mathbf{U}_r$, $r = 1, 2, \cdots, m$*:* Differentiate $J_r(\mathbf{U}_r, \mathbf{P}_r)$ with respect to $\mathbf{U}_r$ and then set the obtained equation to zero, we can achieve the equation relating to U_r as

$$\begin{aligned} \frac{\partial J_r(\mathbf{U}_r, \mathbf{P}_r)}{\partial \mathbf{U}_r} &= \mathbf{X}_r \mathbf{D}_r (\mathbf{X}_r^T \mathbf{U}_r - \mathbf{P}_r) + \lambda \mathbf{X}_r \mathbf{L}_r \mathbf{X}_r^T \mathbf{U}_r \\ &\quad + \gamma \mathbf{Q}_r \mathbf{U}_r \\ &= 0. \end{aligned} \tag{14}$$

After some algebraic manipulations, we could obtain the following analytical solution:

$$\mathbf{U}_r = (\mathbf{X}_r \mathbf{D}_r \mathbf{X}_r^T + \lambda \mathbf{X}_r \mathbf{L}_r \mathbf{X}_r^T + \gamma \mathbf{Q}_r)^{-1} \mathbf{X}_r \mathbf{D}_r \begin{bmatrix} \mathbf{P}_l \\ \hat{\mathbf{P}}_r \end{bmatrix}. \tag{15}$$

Update $\hat{\mathbf{P}}_r$*,* $r = 1, 2, \cdots, m$*:* To effectively propagate the class label information from labeled data to unlabeled data, we rewrite the objective function as follows:

$$\begin{aligned} J_r(\mathbf{U}_r, \mathbf{P}_r) = \text{Tr} &\left(\begin{bmatrix} \check{\mathbf{X}}_r^T \mathbf{U}_r - \mathbf{P}_l \\ \hat{\mathbf{X}}_r^T \mathbf{U}_r - \hat{\mathbf{P}}_r \end{bmatrix}^T \begin{bmatrix} \check{\mathbf{D}}_r & 0 \\ 0 & \hat{\mathbf{D}}_r \end{bmatrix} \right. \\ &\left. \times \begin{bmatrix} \check{\mathbf{X}}_r^T \mathbf{U}_r - \mathbf{P}_l \\ \hat{\mathbf{X}}_r^T \mathbf{U}_r - \hat{\mathbf{P}}_r \end{bmatrix} \right) \\ &+ \lambda \text{Tr} \left(\begin{bmatrix} \check{\mathbf{X}}_r^T \mathbf{U}_r \\ \hat{\mathbf{X}}_r^T \mathbf{U}_r \end{bmatrix}^T \begin{bmatrix} \mathbf{L}_r^{ll} & \mathbf{L}_r^{lu} \\ \mathbf{L}_r^{ul} & \mathbf{L}_r^{uu} \end{bmatrix} \begin{bmatrix} \check{\mathbf{X}}_r^T \mathbf{U}_r \\ \hat{\mathbf{X}}_r^T \mathbf{U}_r \end{bmatrix} \right) \\ &- \lambda \text{Tr} \left(\begin{bmatrix} \mathbf{P}_l \\ \hat{\mathbf{P}}_r \end{bmatrix}^T \begin{bmatrix} \mathbf{L}_r^{ll} & \mathbf{L}_r^{lu} \\ \mathbf{L}_r^{ul} & \mathbf{L}_r^{uu} \end{bmatrix} \begin{bmatrix} \mathbf{P}_l \\ \hat{\mathbf{P}}_r \end{bmatrix} \right) \\ &+ \gamma \text{Tr}(\mathbf{U}_r^T \mathbf{Q}_r \mathbf{U}_r) \\ &+ \gamma \text{Tr} \left(\begin{bmatrix} \mathbf{P}_l \\ \hat{\mathbf{P}}_r \end{bmatrix}^T \begin{bmatrix} \mathbf{S}_r^{ll} & \mathbf{S}_r^{lu} \\ \mathbf{S}_r^{ul} & \mathbf{S}_r^{uu} \end{bmatrix} \begin{bmatrix} \mathbf{P}_l \\ \hat{\mathbf{P}}_r \end{bmatrix} \right). \end{aligned} \tag{16}$$

Algorithm 1: GSS-SL

Input: Data sets: $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_m\}$, each modality contains labeled and unlabeled data: $\mathbf{X}_r = [\check{\mathbf{X}}_r, \hat{\mathbf{X}}_r]$, label sets: $\mathbf{P} = \{\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_m\}$, each modality contains labeled and unlabeled class indicator matrices: $\mathbf{P}_r = \left[\begin{smallmatrix} \mathbf{P}_l \\ \hat{\mathbf{P}}_r \end{smallmatrix}\right]$

Output: Optimal $\mathbf{U}_r$ and $\hat{\mathbf{P}}_r$, $r = 1, 2, \cdots, m$

1: Calculate normalized graph Laplacian matrices $\mathbf{L}_r$ ($r = 1, 2, \cdots, m$) for the r-th modality according to (5)
2: Initialize the labels $\hat{\mathbf{P}}_r$ ($r = 1, 2, \cdots, m$) for the unlabeled data of the r-th modality according to (13)
3: Construct objective function for the r-th modality: $J_r(\mathbf{U}_r, \mathbf{P}_r)$, $r = 1, 2, \cdots, m$
4: **repeat**
5: Update $\mathbf{U}_r$ ($r = 1, 2, \cdots, m$) by (15)
6: Update $\hat{\mathbf{P}}_r$ ($r = 1, 2, \cdots, m$) by (19)
7: Calculate $J = \sum_{i=1}^{m} J_r$
8: **until** $|J_{t-1} - J_t| \leq 0.001$

We remove the constraints which are irrelevant to $\hat{\mathbf{P}}_r$ in (16), and obtain the new formula $J_r(\hat{\mathbf{P}}_r)$ as follow:

$$\begin{aligned} J_r(\hat{\mathbf{P}}_r) = &\text{Tr}((\hat{\mathbf{X}}_r^T \mathbf{U}_r - \hat{\mathbf{P}}_r)^T \hat{\mathbf{D}}_r (\hat{\mathbf{X}}_r^T \mathbf{U}_r - \hat{\mathbf{P}}_r)) \\ &- \lambda \text{Tr}(2\hat{\mathbf{P}}_r^T \mathbf{L}_r^{ul} \mathbf{P}_l + \hat{\mathbf{P}}_r^T \mathbf{L}_r^{uu} \hat{\mathbf{P}}_r) \\ &+ \gamma \text{Tr}(2\hat{\mathbf{P}}_r^T \mathbf{S}_r^{ul} \mathbf{P}_l + \hat{\mathbf{P}}_r^T \mathbf{S}_r^{uu} \hat{\mathbf{P}}_r). \end{aligned} \tag{17}$$

Similarly, we also differentiate $J_r(\hat{\mathbf{P}}_r)$ with respect to $\hat{\mathbf{P}}_r$ and set the obtained equation to zero, we get

$$\begin{aligned} \frac{\partial J_r(\hat{\mathbf{P}}_r)}{\partial \hat{\mathbf{P}}_r} &= -\hat{\mathbf{D}}_r (\hat{\mathbf{X}}_r^T \mathbf{U}_r - \hat{\mathbf{P}}_r) - \lambda (\mathbf{L}_r^{ul} \mathbf{P}_l + \mathbf{L}_r^{uu} \hat{\mathbf{P}}_r) \\ &\quad + \gamma (\mathbf{S}_r^{ul} \mathbf{P}_l + \mathbf{S}_r^{uu} \hat{\mathbf{P}}_r) \\ &= 0. \end{aligned} \tag{18}$$

Then, $\hat{\mathbf{P}}_r$ can be computed as

$$\hat{\mathbf{P}}_r = (\hat{\mathbf{D}}_r - \lambda \mathbf{L}_r^{uu} + \gamma \mathbf{S}_r^{uu})^{-1} (\hat{\mathbf{D}}_r \hat{\mathbf{X}}_r^T \mathbf{U}_r + \lambda \mathbf{L}_r^{ul} \mathbf{P}_l - \gamma \mathbf{S}_r^{ul} \mathbf{P}_l). \tag{19}$$

To search an optimal solution, we alternately optimize the transformation $\mathbf{U}_r$ and the class indicator matrix $\hat{\mathbf{P}}_r$ of unlabeled data, $r = 1, \cdots, m$. We summarize the optimization procedure in Algorithm 1. From it, we observe that the induction of transformation $\mathbf{U}_r$ depends on the class indicator matrix $\hat{\mathbf{P}}_r$ of unlabeled data, and the predicted labels for unlabeled data rely on the transformation $\mathbf{U}_r$. Both the transformations and the predicted labels are expected to benefit from this interdependent procedure. To our best knowledge, it is the first time to use the class indicator matrix of unlabeled data doing inductive learning in cross-modal learning. The convergence criterion in our experiments is that the number of iterations is more than 10 or $|J_{t-1} - J_t| \leq 0.001$,

where J_t is the value of the objective function in the tth iteration.

D. Convergence Analysis

The objective function will be converged by the proposed iterative method, which is proved by the following theorem.

Theorem 2: In each iteration, the value of the objective function value will be monotonically decreased until convergence by adopting the proposed optimization method.

Proof: In the t-th iteration, we denote the transformation and the predicted label matrix as $\mathbf{U}_r^{(t)}$ and $\mathbf{P}_r^{(t)}$. Then, we solve $\mathbf{U}_r^{(t+1)}$ while fix $\mathbf{P}_r$ as $\mathbf{P}_r^{(t)}$. Based on the analysis in [31], we can get the following formulation:

$$\mathbf{U}_r^{(t+1)} = \arg\min_{\mathbf{U}_r} \mathrm{Tr}\left(\begin{bmatrix}\mathbf{U}_r\\ \mathbf{P}_r\end{bmatrix}^T \left[\mathbf{B}_{\mathcal{L}} + \lambda\mathbf{B}_{\Psi} + \gamma\mathbf{B}_{\Omega}\right] \begin{bmatrix}\mathbf{U}_r\\ \mathbf{P}_r\end{bmatrix}\right). \tag{20}$$

Since optimization problem is jointly convex with respect to $\mathbf{U}_r$ and $\mathbf{P}_r$, we can have the following formulation:

$$\mathrm{Tr}\left(\begin{bmatrix}\mathbf{U}_r^{(t+1)}\\ \mathbf{P}_r^{(t)}\end{bmatrix}^T \left[\mathbf{B}_{\mathcal{L}} + \lambda\mathbf{B}_{\Psi} + \gamma\mathbf{B}_{\Omega}\right] \begin{bmatrix}\mathbf{U}_r^{(t+1)}\\ \mathbf{P}_r^{(t)}\end{bmatrix}\right) \leq \mathrm{Tr}\left(\begin{bmatrix}\mathbf{U}_r^{(t)}\\ \mathbf{P}_r^{(t)}\end{bmatrix}^T \left[\mathbf{B}_{\mathcal{L}} + \lambda\mathbf{B}_{\Psi} + \gamma\mathbf{B}_{\Omega}\right] \begin{bmatrix}\mathbf{U}_r^{(t)}\\ \mathbf{P}_r^{(t)}\end{bmatrix}\right). \tag{21}$$

Similarly, we can obtain the following equation by fixing $\mathbf{U}_r$ as $\mathbf{U}_r^{(r)}$:

$$\mathrm{Tr}\left(\begin{bmatrix}\mathbf{U}_r^{(t)}\\ \mathbf{P}_r^{(t+1)}\end{bmatrix}^T \left[\mathbf{B}_{\mathcal{L}} + \lambda\mathbf{B}_{\Psi} + \gamma\mathbf{B}_{\Omega}\right] \begin{bmatrix}\mathbf{U}_r^{(t)}\\ \mathbf{P}_r^{(t+1)}\end{bmatrix}\right) \leq \mathrm{Tr}\left(\begin{bmatrix}\mathbf{U}_r^{(t)}\\ \mathbf{P}_r^{(t)}\end{bmatrix}^T \left[\mathbf{B}_{\mathcal{L}} + \lambda\mathbf{B}_{\Psi} + \gamma\mathbf{B}_{\Omega}\right] \begin{bmatrix}\mathbf{U}_r^{(t)}\\ \mathbf{P}_r^{(t)}\end{bmatrix}\right). \tag{22}$$

After integrating (21) and (22), we achieve the final inequation as follows:

$$\mathrm{Tr}\left(\begin{bmatrix}\mathbf{U}_r^{(t+1)}\\ \mathbf{P}_r^{(t+1)}\end{bmatrix}^T \left[\mathbf{B}_{\mathcal{L}} + \lambda\mathbf{B}_{\Psi} + \gamma B_{\Omega}\right] \begin{bmatrix}U_r^{(t+1)}\\ \mathbf{P}_r^{(t+1)}\end{bmatrix}\right) \leq \mathrm{Tr}\left(\begin{bmatrix}\mathbf{U}_r^{(t)}\\ \mathbf{P}_r^{(t)}\end{bmatrix}^T \left[\mathbf{B}_{\mathcal{L}} + \lambda\mathbf{B}_{\Psi} + \gamma\mathbf{B}_{\Omega}\right] \begin{bmatrix}\mathbf{U}_r^{(t)}\\ \mathbf{P}_r^{(t)}\end{bmatrix}\right). \tag{23}$$

Equation (23) validates that the proposed optimization method decreases the objective function value after each iteration. Since all modalities $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_m\}$ satisfy this property, our proposed method will monotonically decrease the objective in (1), which will be converged to the global optimum eventually. ■

E. Computational Complexity Analysis

In this subsection, we discuss the computational cost of the proposed method. As stated above, for the rth modality, the k-nn graph is first constructed based on the Euclidean distance in the original space, and the corresponding cost is $O(d_r n_r^2)$. Then, the proposed optimization problem is solved iteratively. In each iteration, it costs $O(d_r^3 + d_r^2 n_r + d_r n_r^2 + d_r n_r c)$ to calculate $\mathbf{U}_r$. Since each modality has n_u unlabeled data points and n_l labeled data points, $\hat{\mathbf{P}}_r$ is obtained with the time complexity of $O(n_u^3 + n_u^2 d_r + n_u d_r c + n_u n_l c + n_u^2 c)$. Since $n_r \geq c$, the total cost of calculating m modalities is $O(m \times (d_r n_r^2 + N(d_r^3 + d_r^2 n_r + d_r n_r^2 + n_u^3 + n_u^2 d_r)))$, where N is the number of iterations.

IV. Experimental Results

To validate the effectiveness of the proposed GSS-SL method, extensive experiments of GSS-SL and the compared methods are conducted for the task of text-image retrieval, i.e., image-query-texts and text-query-images. We test all the methods on three standard benchmark datasets: Wiki [3], Pascal VOC 2007 [17] and NUS-WIDE [18].

A. Experimental Settings

We compare GSS-SL with several related methods, including CCA [19], SCM [3], GMLDA and GMMFA [12], LCFS [7], MvDA [20], LGCFL [4], and ml-CCA [5]. For fairness, we ensure that the total number of the training samples is equivalent for all the methods. Then all the compared methods directly use the labeled training set, while 20 percent samples in the training set are set to the unlabeled data for GSS-SL. Besides, we also report the performance of the proposed GSS1-SL, which uses all labeled training data. GSS2-SL not only uses all labeled training data but also the additionally available 20 percent unlabeled data.[1] At the testing stage, we adopt the cosine distance to measure the similarity of features.

The mean average precision (MAP) [3] is used to evaluate the performance. Especially, we adopt MAP@R [8] to measure the retrieval performance at fixed number of the retrieved samples. R is set to 50 for the top 50 retrieved instances and to all for all the retrieved instances. On Wiki, we define that a retrieved sample is relevant to the query if they belong to the same semantic class. On Pascal and NUS-WIDE, a retrieved sample is relevant to the query if they share at least one concept. Besides, the precision-recall curve [3] and precision-scope curve [36] are also displayed for all methods.

Considering that CCA, SCM, GMLDA, GMMFA and MvDA focus on learning the common subspace, we conduct Principal Component Analysis (PCA) on the original features to remove the redundant features. For all these methods except for MvDA, PCA preserves 95% information energy. For MvDA, since it calculates the intra-modality and inter-modality distances for its solution, we adopt PCA to ensure that the dimensions of image and text feature are equal. The dimensionality of the common subspace is set to 10, 20 and 10 for Wiki, Pascal and NUS-WIDE, respectively. For all compared methods, parameters are set by 5-fold cross validation on the training set. For our method, λ, γ and k are respectively set to 0.01, 10 and 30 in all experiments.

[1] On all three datasets, these samples are selected from the NUS-WIDE dataset.

TABLE II
PERFORMANCE COMPARISON IN TERMS OF MAP@R SCORES ON WIKI DATASET

Methods \ Tasks	R = 50			R = all			Training set
	Text query	Image query	Average MAP	Text query	Image query	Average MAP	(Labeled/Unlabeled)
CCA[19]	0.3129	0.2853	0.2991	0.1872	0.2160	0.2016	100%/0%
SCM[3]	0.3556	0.3015	0.3285	0.2336	0.2759	0.2548	
LCFS[6]	0.3687	0.2712	0.3199	0.2043	0.2711	0.2377	
MvDA[20]	0.3917	0.3091	0.3504	0.2319	0.2971	0.2645	
LGCFL[4]	0.4956	0.3868	0.4412	0.3160	0.3775	0.3467	
ml-CCA[5]	0.4521	0.3683	0.4102	0.2873	0.3527	0.3120	
GMLDA[12]	0.4697	0.3270	0.3983	0.2885	0.3159	0.3022	
GMMFA[12]	0.4750	0.3310	0.4030	0.2964	0.3155	0.3060	
GSS1-SL	0.5312	0.4122	0.4717	0.3383	0.4060	0.3721	
GSS-SL	0.5195	0.3966	0.4580	0.3263	0.3897	0.3580	80%/20%
GSS2-SL	**0.5531**	**0.4244**	**0.4888**	**0.3532**	**0.4175**	**0.3853**	100%/20%
JRL	0.5469	0.4409	0.4939	0.3468	0.4390	0.3929	training/testing
S^2UPG	0.5547	0.4572	0.5060	0.3582	0.4473	0.4027	
GSS3-SL	**0.5713**	**0.4791**	**0.5252**	**0.3646**	**0.4552**	**0.4099**	

Note: The '20%' of GSS2-SL denotes additionally unlabeled samples but not from the mentioned dataset. The last three methods use both training and testing samples as training set.

B. Results on the Wiki Dataset

The Wiki[2] dataset [3] is collected from Wikipedia feature articles. It totally contains 2,866 image-text pairs which belongs to 10 semantic classes, and each paired samples belong to a unique semantic class. Our image features are represented by the 4,096 dimensional output from the fc7 layer of Convolutional Neural Network (CNN) [37]. For text features, we first adopt *word2vec* model to learn the 100 dimensional skip-gram word vectors [38]. Then we calculate a mean vector of the word vectors of the words appearing in each text document. On this dataset, we randomly select 2,000 pairs for training and 866 pairs for testing. The MAP scores of all methods are shown in Table II. From the table, we draw the following conclusions:

First, GSS-SL outperforms all compared methods except for GSS1-SL and GSS2-SL, although the training set of GSS-SL includes some unlabeled data. For example, in $R = all$, the Average MAP[3] score of GSS-SL is 0.3580, which is higher than the best result from the compared method, LGCFL (0.3467). Since LGCFL also uses the label space to link the image space and text space, this improvement is possibly due to exploiting the underlying data structures and selecting discriminative features of the proposed method.

Second, the MAP scores of GSS1-SL are further improved by using all labeled training data, i.e., GSS1-SL adopts the same training set of compared methods to model the correlations between image modality and text modality. In $R = all$, GSS1-SL achieves the best Average MAP scores 0.3721, which is superior than all compared methods and GSS-SL. This is because using all labeled data helps to exploit the correlation among different modalities under our framework. Hence, GSS-SL still works well in the supervised setting.

Finally, GSS2-SL improves the performance of GSS1-SL. This phenomenon validates that the unlabeled data can help to model the correlations among the different modalities in our framework. This is mainly because the label graph ensures the consistent structures of feature space and label space, and label-linked loss function enforces similar labels among the nearest features. Then our framework can effectively select features from labeled and unlabeled data.

The precision-recall curves and precision-scope curves of the image-query and text-query are plotted in Fig. 3(a) and 3(b). The scope (i.e., the top K retrieved samples) of the precision-scope varies from $K = 100$ to 800. We observe that compared with the other compared methods, our method achieves better results on all curves.

C. Results on the Pascal Dataset

The Pascal[4] dataset [17] consists of 5,011/4,952(testing/training) images-tag pairs. All pairs belong to one or more of 20 semantic classes. We use the publicly available 512-dimensional GIST features for images. For texts, we use the 399-dimensional word frequency features. We use the original training-test split and remove some pairs since their text features are all zeros. Finally, 5,000 pairs are used for training and 4,919 pairs for testing. The MAP scores are shown in Table III.

GSS-SL is compared with CCA, LCFS, ml-CCA and LGCFL because they can handle the multi-labeled data. From Table III, we can see that results on this dataset are generally higher than those on Wiki dataset. These results validate that the multi-label information helps to learn a discriminative semantic space which is more suitable for exploiting correlation between two modalities. Compared with the other methods, the improvement of GSS-SL is significant as that on the Wiki dataset. Recall that in the Pascal dataset, one image or text is associated with multiple labels. GSS-SL exploits the structure information of the label space by constructing the label graph, which is suitable to reveal semantic information of all samples. For example, in $R = all$, the Average MAP score of GSS-SL are 0.4233, which is about 8.5% higher than LGCFL. Furthermore, the MAP scores of GSS1-SL and GSS2-SL are also improved further than GSS-SL.

[2][Online]. Available: http://www.svcl.ucsd.edu/projects/crossmodal/

[3]"Average MAP" denotes the average of MAP scores of image-query-texts and text-query-images

[4][Online]. Available: http://www.cs.utexas.edu/%7egrauman/research/datasets.html

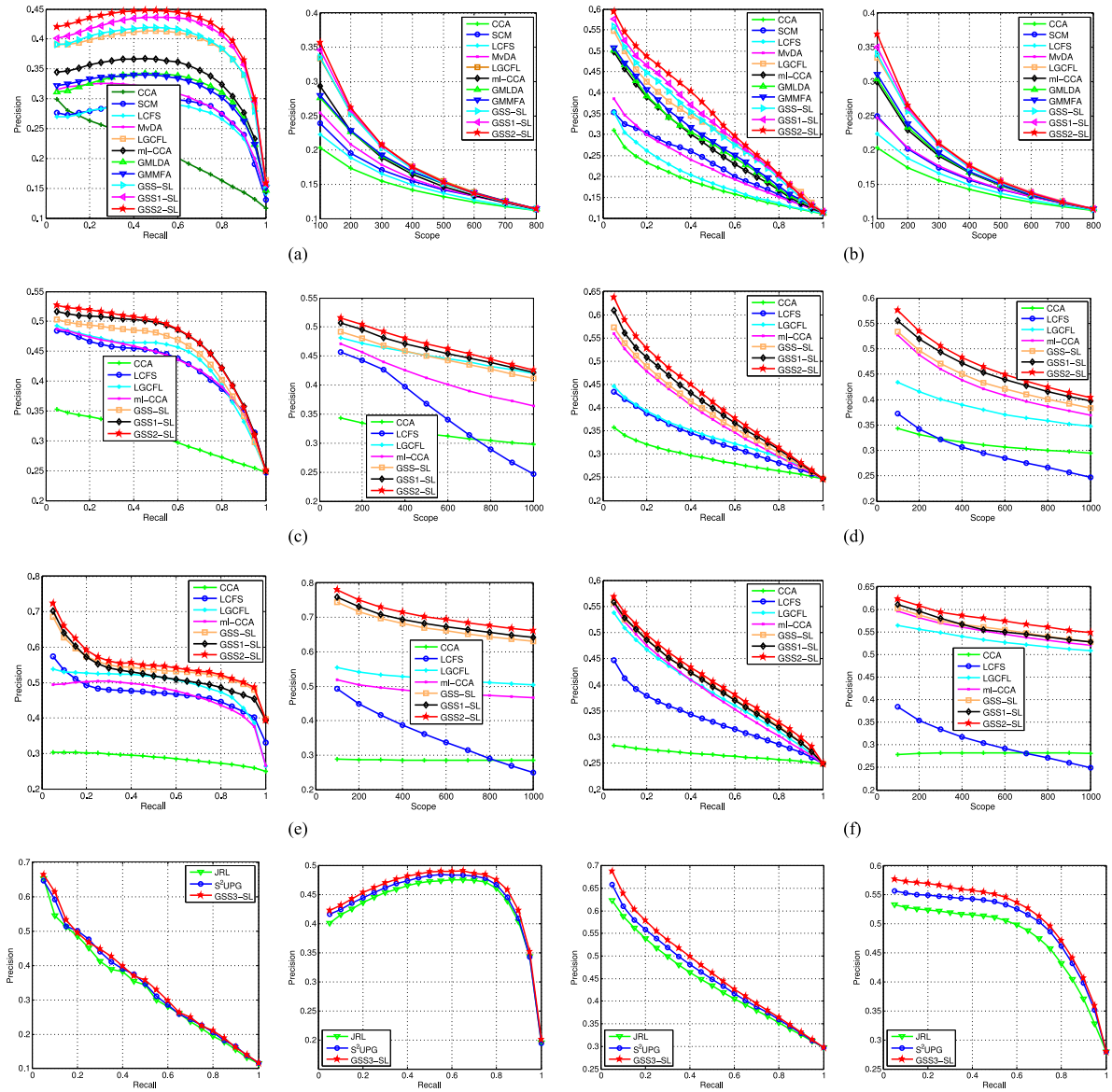


Fig. 3. Precision-recall curves and precision-scope curves for the image-query-tests and text-query-images experiments on Wiki, Pascal, and NUS datasets. (a) Wiki Image query. (b) Wiki Text query. (c) Pascal Image query. (d) Pascal Text query. (e) NUS Image query. (f) NUS Text query. (g) Wiki Text query. (h) Wiki Image query. (i) Pascal Text Query. (j) Pascal Image query.

The precision-recall curves and precision-scope curves are also displayed in Fig. 3(c) and 3(d). From the precision-recall curves, we observe that under the same recall, GSS-SL obtains the higher precision than all compared methods. These results are consistent with the precision-scope curves, which show the precision by varying the scope from 0 to 1,000.

D. Results on the NUS-WIDE Dataset

The NUS-WIDE[5] dataset consists of 40,834/27,159 (training/testing) image-tag pairs, which are pruned from the original train-test split of the NUS dataset [18] by keeping the pairs that belong to one or more of the 10 largest classes. Each text is represented by a 1,000-dimensional word frequency vector based tag features, and each image is represented as a 500-dimensional SIFT feature. Since NUS-WIDE is a large scale dataset, the experiments on this dataset can validate the scalability of the different methods. The experimental results are shown in Table IV, Fig. 3(e) as well as 3(f).

From Table IV, we can see that results on this dataset are generally high, which may be partly due to the large number of training samples (40 k). Therefore, the large number of training samples are sufficient for reliable learning. We also see that GSS-SL outperforms the compared methods. For example, in

[5][Online]. Available: http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

TABLE III
PERFORMANCE COMPARISON IN TERMS OF MAP@R SCORES ON PASCAL DATASET

Methods \ Tasks	R = 50			R = all			Training set
	Text query	Image query	Average MAP	Text query	Image query	Average MAP	(Labeled/Unlabeled)
CCA[19]	0.4119	0.3955	0.4037	0.2945	0.3073	0.3009	
LCFS[6]	0.4501	0.4864	0.4682	0.3355	0.4278	0.3816	
LGCFL[4]	0.5026	0.5292	0.5159	0.3440	0.4362	0.3901	100%/0%
ml-CCA[5]	0.6095	0.5204	0.5649	0.3885	0.4303	0.4094	
GSS1-SL	0.6552	0.5560	0.6056	0.4120	0.4660	0.4390	
GSS-SL	0.6125	0.5381	0.5753	0.3970	0.4496	0.4233	80%/20%
GSS2-SL	**0.6819**	**0.5581**	**0.6200**	**0.4252**	**0.4697**	**0.4474**	100%/20%
JRL	0.6854	0.5594	0.6224	0.4277	0.4690	0.4484	
S^2UPG	0.6871	0.5634	0.6253	0.4315	0.4715	0.4515	training/testing
GSS3-SL	**0.6976**	**0.5672**	**0.6324**	**0.4437**	**0.4805**	**0.4621**	

Note: The "20%" of GSS2-SL denotes additionally unlabeled samples but not from the mentioned dataset. The last three methods use both training and testing samples as training set.

TABLE IV
PERFORMANCE COMPARISON IN TERMS OF MAP@R SCORES ON NUS-WIDE DATASET

Methods \ Tasks	R = 50			R = all			Training set
	Text query	Image query	Average MAP	Text query	Image query	Average MAP	(Labeled/Unlabeled)
CCA[19]	0.3211	0.3312	0.3262	0.2667	0.2869	0.2768	
LCFS[6]	0.5458	0.6742	0.6100	0.3363	0.4742	0.4053	
LGCFL[4]	0.5989	0.5900	0.5945	0.3907	0.4972	0.4440	100%/0%
ml-CCA[5]	0.6473	0.5684	0.6078	0.3908	0.4689	0.4299	
GSS1-SL	0.6828	0.8690	0.7759	0.4050	0.5501	0.4776	
GSS-SL	0.6821	0.8354	0.7587	0.4040	0.5364	0.4702	80%/20%
GSS2-SL	**0.7162**	**0.8819**	**0.7991**	**0.4367**	**0.5938**	**0.5153**	100%/20%

Note: The "20%" of GSS2-SL denotes additionally unlabeled samples but not from the mentioned dataset.

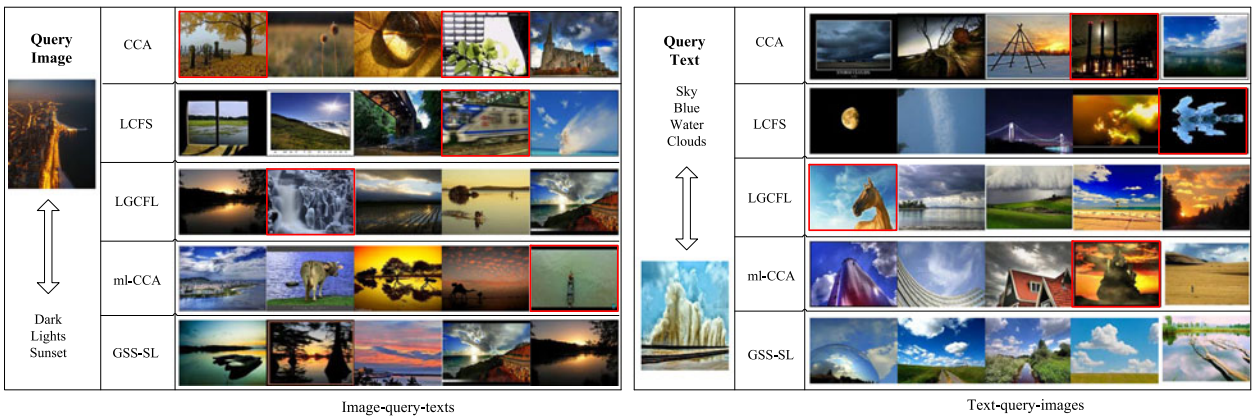


Fig. 4. Two examples of image-query-texts (in left half) and text-query-images (in right half) on the NUS-WIDE dataset. For each example, we show the query and its corresponding top retrieved results.

$R = all$, the Average MAP score of GSS-SL is about 5.9% higher than LGCFL. The precision-recall and the precision-scope curves are shown in Fig. 3(e) and 3(f). The curves also clearly show that the proposed method has the best performance on both directions of retrieval task. Fig. 4 shows two retrieved examples from two retrieved directions. Based on intuitive judgement, we draw the conclusion that GSS-SL achieves the best results comparing with its four counterparts.

E. Comparison With Semi-supervised Methods

In this part, we compare with two semi-supervised cross-modal methods on Wiki, Pascal datasets. Both JRL of [16] and S^2 UPG of [15] construct the graph model by jointly using training samples and testing samples, which is different from our experimental setting. For fairness, we also run our code with this experimental setting, denotes as GSS3-SL. On wiki and Pascal datasets, we tune the parameter values until the best performances are achieved. The MAP scores of the three semi-supervised methods are shown in Tables II and III. To give a clear comparison of three semi-supervised methods, their precision-recall curves are specifically plotted in Fig. 3(g)–3(j).

From these results, we conclude that the retrieval performance of GSS3-SL is higher than that of JRL and S^2 UPG. JRL and S^2 UPG jointly utilize training data and testing data to construct

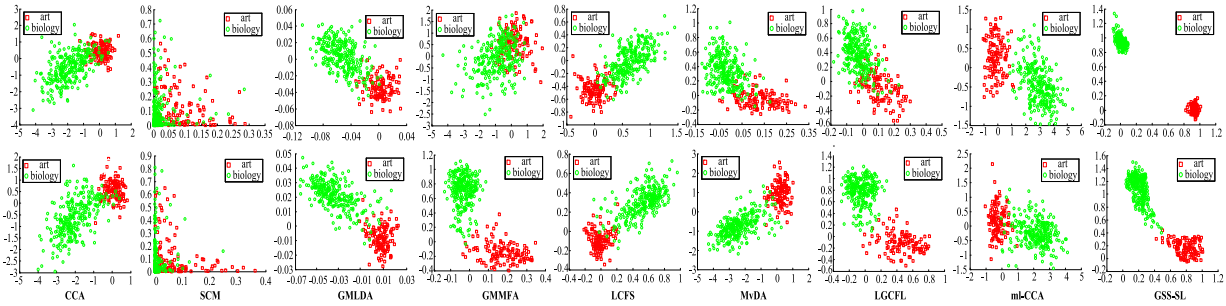


Fig. 5. Low-dimensional transforming of images and texts from "art" and "biology" classes of Wiki dataset. The top row shows the transforming for image modality and the bottom shows the transforming for text modality.

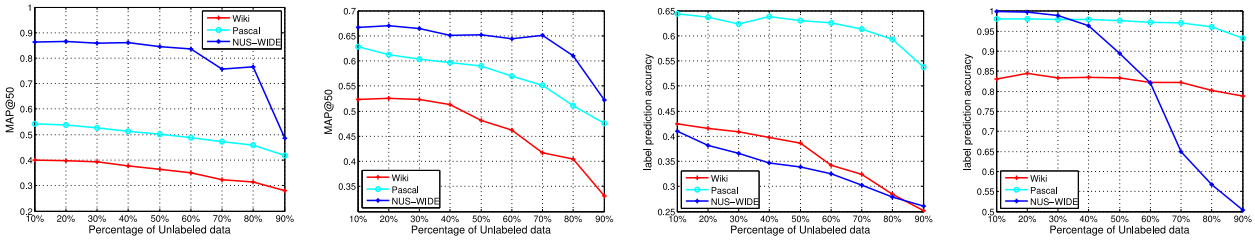


Fig. 6. Results on handling unlabeled data. (a) and (b) show the MAP scores with increasing percentage of unlabeled data. (c) and (d) show the label prediction accuracy for the unlabeled data with increasing percentage of unlabeled data.

graph to propagate the label information from labeled data to unlabeled data. Although the graph ensures that the nearest features have similar class labels, it cannot guarantee that the similar class labels can exactly annotate their corresponding features because labeled data cannot construct semantic association with their ground truth labels in the process of label propagation. GSS3-SL enhances the semantic association in the proposed label graph, thus it outperforms both semi-supervised algorithms. Besides, on Wiki and Pascal datasets, we observe that text query of GSS3-SL only improves 3.2% and 4.3% than GSS2-SL. The training set of GSS2-SL consists of all training samples and extra 20% unlabeled data. Hence, the performance of semi-supervised method does not always improve with the increase of unlabeled data. This is possibly due to the fact that part of unlabeled data express confused semantic information, which cannot help to explore the semantic association between different modalities.

F. Results on Transformations

To analyze the discriminant of transformations learned by the different methods, we present the low-dimensional embedding of the different modalities. We adopt the first two classes of the Wiki dataset ('art' and 'biology') to construct a toy dataset. For image and text modalities, the first and second most correlated entries of all the methods are shown in Fig. 5. We use the red squares to denote the distribution of the 'art' class, and green circles represents the 'biology' class. It is clear from the figure that GSS-SL units the same-class samples and separates the different classes for both directions of retrieval, but the second best results (LGCFL and GMMFA) only unit the same-class samples and separate the different classes for text query. Moreover, both image and text distributions of GSS-SL are in the same coordinate range, while the other methods' coordinate ranges are quite different. These results validate that GSS-SL ensures the consistent structures between the image and text spaces such that the low-dimensional embedding is effective with the stronger discriminant.

G. Results on Various Percentages of Unlabeled Data

In this part, we show the performance with various percentages of unlabeled data on the three datasets. Fig. 6(a) and 6(b) show the MAP@50 scores on different percentages of unlabeled data. Compared to Tables II–IV, we conclude that GSS-SL achieves comparable results when the percentage of unlabeled data is no larger than 20%. These results validate that GSS-SL can effectively deal with the unlabeled data.

Fig. 6(c) and 6(d) report the label prediction accuracies on the three datasets. For each unlabeled sample, we select the class assigned with the maximum predicted value as its label. On Wiki, it is a correct prediction if the predicted label is the same as its true label. For the other datasets, a correct prediction is defined as that this predicted label shares one concept with its multi-label annotations. From Fig. 6(c) and 6(d), it is clear that the predicted precisions for text query are much larger than those of image query. This is likely due to the fact that class labels apply more directly to texts than images. In fact, class labels are usually treated as textual description on some datasets. For example, most of the concepts (used as class labels) are equivalent to the taglist (used as textual feature) on NUS-WIDE dataset.

Fig. 7. Some examples of predicted labels for the unlabeled data on the Wiki dataset. The incorrect predicted results are marked in a red frame, and the irrelevant predicted labels are in red font.

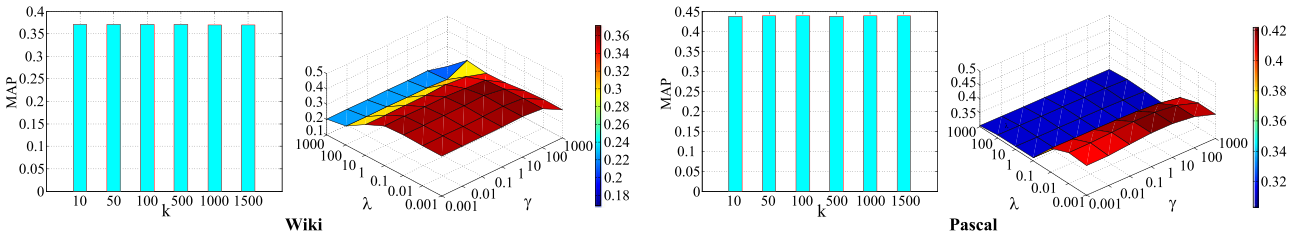


Fig. 8. Sensitivity analysis on the parameters (λ, γ and k) with respect to average MAP scores of image-query-texts and text-query-images experiments.

TABLE V
PROCESSING TIME COMPARISON (SECONDS)

Tasks \ Methods	CCA	SCM	LCFS	MvDA	GMLDA	GMMFA	LGCFL	ml-CCA	GSS-SL
Train	0.059	0.112	32.175	0.436	54.098	55.497	3.002	521.878	8.919
Test	0.227	0.212	0.185	0.201	0.200	0.201	0.199	0.205	0.198

In Fig. 7, we show some predicted results for the unlabeled data. We can get the intuitive judgement that the proposed method can achieve the reasonable labels for the unlabeled samples. We also note that there are some inconsistent predictions, but this misclassification is reasonable. For example, the image of a sculpture could also fit into the "art" class. "Royalty" misclassified into "warfare" can be accepted since the two classes share the similar imagery.

H. Parameter Sensitivity Analysis

There are mainly three parameters in the proposed GSS-SL method: λ, γ and k. Without loss of generality, we use the training sets of Wiki and Pascal to conduct the parameter sensitivity analysis to test how they impact the performance.

For each parameter, we perform the empirical analysis by changing its value and fixing the other parameters. Fig. 8 shows the Average MAP scores of image-text retrieval with the different tradeoff parameters. On all four tasks, GSS-SL can achieve superior performance under a wide range of parameter values, i.e., $\lambda \in [0.01\ 1]$, $\gamma \in [1\ 10]$, $k \in [10\ 1000]$. We also test how two constraints impact the retrieval performance on the Wiki dataset. When λ and γ are all set to 0, the Average MAP is 0.3378. When λ is set to 0 and γ is set to the optimal value 5, Average MAP 0.3563 is obtained. When λ is set to the optimal value 0.01 and γ is set to 0, the Average MAP score is 0.3656. These results validate that every constraint proposed in this paper is important to enhance the correlations between different modalities.

I. Convergency and Computational Time

To compare the computational complexity of all methods, we use the training set (2000 paired samples) and the testing set (866 paired samples) from the Wiki dataset to evaluate the computational time. We show the training and testing time of all methods with Matlab R2013a in Table V.

From Table V, we observe that the training time of all the supervised methods are longer than the unsupervised method like CCA. The reason is that the supervised methods use the class information to construct more complex framework for exploiting the correlations between two modalities. The proposed GSS-SL is only longer than SCM and LGCFL but shorter than the other supervised methods. That is because GSS-SL computes Laplacian graphs for each modality, which requires more algebraic operation. Most of the other supervised methods need to solve eigenvalues and eigenvectors which lead to a higher computational complexity. With regard to the test time, GSS-SL spends about 0.198 seconds to handle the 866 paired samples, which is much faster than other compared methods except for LCFS. It is may be that LCFS adopts the trace norm constraining the transformations to obtain the sparse structures, which leads to a more efficient low-dimensional embedding. Since the training process is conducted offline, the time spent on the training process is not as significant as that of the testing process.

In Fig. 9, we analyze the convergence of the proposed GSS-SL method on the Wiki dataset. The value of objective function is reported by varying the number of iterations. From this figure, we can conclude that the performance of GSS-SL becomes stable after about 30 iterations.

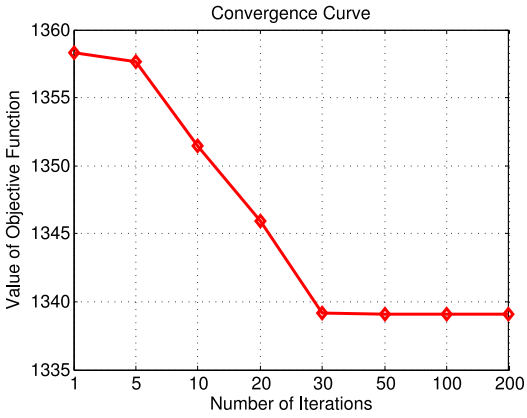


Fig. 9. Value of objective function of our approach versus different number of iterations on the Wiki dataset.

V. Conclusion

In this paper, we propose a generalized semi-supervised learning for cross-modal retrieval. By combining the label graph with $\ell_{2,1}$-norm based loss function and regularization, the discriminative subspace learning, feature selection and label prediction can be performed simultaneously. We use the label space as the linkage to optimize multiple transformations, so arbitrary number of modalities can be solved under this framework. Extensive experiments demonstrate that the proposed method outperforms the state-of-the-art methods on three cross-modal datasets.

In the future, we will focus on making our method more efficient and exploiting the relations among the class labels. On one hand, it is well known that the computational cost of constructing the Laplacian graph is relatively high. Thus, we will propose an efficient method to realize the same goal as graph such that our method can be applied to deal with more practical problems. On the other hand, in the real-world, the important degree of each associated label from the multi-label data is generally different, so we will exploit the correlations among the multiple class labels.

References

[1] J. Pereira *et al.*, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 521–535, Mar. 2014.

[2] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013.

[3] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM Int. Conf. MultiMedia*, 2010, pp. 251–260.

[4] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 370–381, Mar. 2015.

[5] V. Ranjan, N. Rasiwasia, and C. Jawahar, "Multi-label cross-modal retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4094–4102.

[6] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Cross-modal retrieval using multi-ordered discriminative structured subspace learning," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1220–1233, Jun. 2017.

[7] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2088–2095.

[8] F. Wu *et al.*, "Cross-media semantic representation via bi-directional learning to rank," in *Proc. ACM Int. Conf. MultiMedia*, 2013, pp. 877–886.

[9] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 1198–1204.

[10] M. Katsurai, T. Ogawa, and M. Haseyama, "A cross-modal approach for extracting semantic relationships between concepts using tagged images," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1059–1074, Jun. 2014.

[11] V. Mahadevan *et al.*, "Maximum covariance unfolding: Manifold learning for bimodal data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 918–926.

[12] A. Sharma, A. Kumar, D. Hal, and D. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2160–2167.

[13] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "PL-ranking: A novel ranking method for cross-modal retrieval," in *Proc. ACM Int. Conf. MultiMedia*, 2016, pp. 1355–1364.

[14] Y. Hua, S. Wang, S. Liu, Q. Huang, and A. Cai, "Cross-modal correlation learning by adaptive hierarchical semantic aggregation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1201–1216, Jun. 2016.

[15] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, "Semi-supervised cross-media feature learning with unified patch graph regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 583–596, Mar. 2016.

[16] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semi-supervised regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 965–978, Jun. 2014.

[17] M. Everingham, V. Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2010.

[18] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, pp. 368–375.

[19] D. Hardoon, S. Szedmark, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[20] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 1, 2016.

[21] A. Sharma and D. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 593–600.

[22] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Comput.*, vol. 12, no. 6, pp. 1247–1283, 2000.

[23] S. Hwang and K. Grauman, "Learning the relative importance of objects from tagged images for retrieval and cross-modal search," *Int. J. Comput. Vis.*, vol. 100, pp. 134–153, 2012.

[24] L. Ballan, T. Uricchio, L. Seidenari, and A. Bimbo, "A cross-media model for automatic image annotation," in *Proc. Int. Conf. Multimedia Retrieval.*, 2014.

[25] G. hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5768, pp. 504–507, 2006.

[26] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.

[27] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2014, pp. 823–831.

[28] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Lear.*, 2011.

[29] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv Neural Inf. Process. Syst.*, 2012, pp. 2424–2432.

[30] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2085–2098, Oct. 2015.

[31] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.

[32] S. Yan *et al.*, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

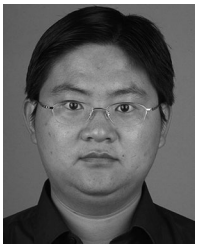
[33] D. Zhou, O. Bousquet, T. Navin, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004.

[34] X. Cai, F. Nie, and W. Cai, "Heterogeneous image features integration via multi-modal semi-supervised learning model," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1737–1744.

[35] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[36] N. Rasiwasia, P. Moreno, and N. Vascondelos, "Bridging the gap: Query by semantic example," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 923–938, Aug. 2007.
[37] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd Int. Conf. Multimedia*, 2014, pp. 675–678.
[38] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

Liang Zhang received the M.S. degree in technology of computer application from the University of Jinan, Jinan, China, in 2014, and is currently working toward the Ph.D. degree at the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China.

His research interests include image and text retrieval, metric learning, and deep learning.



Bingpeng Ma received the B.S. degree in mechanics and the M.S. degree in mathematics from the Huazhong University of Science and Technology, Beijing, China, in 1998 and 2003, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2009.

He was a Postdoctoral Researcher with the University of Caen, Caen, France, from 2011 to 2012. In March 2013, he joined the School of Computer and Control Engineering, University of Chinese Academy of Sciences, where he is currently an Associate Professor. His research interests include computer vision, pattern recognition, and machine learning. His research especially focuses on face recognition, person reidentification, and the related research topics.

Guorong Li received the B.S. degree in technology of computer application from the Renmin University of China, Beijing, China, in 2006, and the Ph.D. degree in technology of computer application from the Graduate University of the Chinese Academy of Sciences, Beijing, China, in 2012.

She is currently an Associate Professor with the University of the Chinese Academy of Sciences. Her research interests include object tracking, video analysis, pattern recognition, and cross-media analysis.

Qingming Huang (A'04–M'04–SM'08) received the B.S. degree in computer science and the Ph.D. degree in computer engineering, both from the Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively.

He is currently a Professor with the University of Chinese Academy of Sciences, Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He was a Postdoctoral Fellow with the National University of Singapore, Singapore, from 1995 to 1996, and served as a member and research staff with the Institute for Infocomm Research, Singapore, from 1996 to 2002. He joined the University of Chinese Academy of Sciences as a Professor under the Science100 Talent Plan in 2003, and has been granted by the China National Funds for Distinguished Young Scientists in 2010. He was chosen in the National Hundreds and Thousands Talents Project in 2014. He has authored or coauthored more than 300 academic papers in prestigious international journals including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and top-level conferences, such as ACM Multimedia, International Conference on Computer Vision (ICCV), Conference on Computer Vision and Pattern Recognition (CVPR), International Joint Conference on Artificial Intelligence, and VLDB. His research interests include multimedia video analysis, image processing, computer vision, and pattern recognition.

Prof. Huang is an Associate Editor of *Acta Automatica Sinica*, and the reviewer of various international journals including the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTION ON IMAGE PROCESSING. He has served as the Program Chair, Track Chair, and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICMR, and PSIVT.

Qi Tian (S'95–M'96–SM'03–F'16) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 1992, the M.S. degree in electronics and computer engineering from Drexel University, Philadelphia, PA, USA, in 1996, and the Ph.D. degree in electronics and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Champaign, IL, USA, in 2002.

He is currently a Full Professor with the Department of Computer Science, University of Texas at San Antonio (UTSA), San Antonio, TX, USA. He was a Tenured Associate Professor from 2008 and 2012 and a Tenure-Track Assistant Professor from 2002 to 2008. During 2008 and 2009, he was on one-year faculty leave at Microsoft Research Asia, Beijing, China, as a Lead Researcher with the Media Computing Group. He has authored more than 340 refereed journal and conference papers. His research interests include multimedia information retrieval, computer vision, pattern recognition, and bioinformatics.

Prof. Tian is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the *Multimedia System Journal*, and is on the editorial board of the *Journal of Multimedia*, and the *Journal of Machine Vision and Applications*. He is the Guest Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, and the *Journal of Computer Vision and Image Understanding*. He was the coauthor of the recipient of the Best Paper Award in ACM ICMR 2015, the Best Paper Award in PCM 2013, the Best Paper Award in MMM 2013, and the Best Paper Award in ACM ICIMCS 2012. He was the recipient of the Top 10% Paper Award in MMSP 2011, and the Best Student Paper Award in ICASSP 2006, and was the coauthor of a Best Student Paper Candidate in ICME 2015 and a Best Paper Candidate in PCM 2007. He was the recipient of the 2014 Research Achievement Award from the College of Science, UTSA and the 2010 ACM Service Award. His research projects are funded by ARO, NSF, DHS, Google, FXPAL, NEC, SALSI, CIAS, Akiira Media Systems, HP, Blippar, and UTSA.