Adaptive Semi-Supervised Feature Selection for Cross-Modal Retrieval

En Yu, Jiande Sun, Jing Li, Xiaojun Chang, Xian-Hua Han, *Member, IEEE*, and Alexander G. Hauptmann

Abstract—**In order to exploit the abundant potential information of the unlabeled data and contribute to analyzing the correlation among heterogeneous data, we propose the semi-supervised model named adaptive semi-supervised feature selection for cross-modal retrieval. First, we utilize the semantic regression to strengthen the neighboring relationship between the data with the same semantic. And the correlation between heterogeneous data can be optimized via keeping the pairwise closeness when learning the common latent space. Second, we adopt the graph-based constraint to predict accurate labels for unlabeled data, and it can also keep the geometric structure consistency between the label space and the feature space of heterogeneous data in the common latent space. Finally, an efficient joint optimization algorithm is proposed to update the mapping matrices and the label matrix for unlabeled data simultaneously and iteratively. It makes samples from different classes to be far apart, while the samples from same class lie as close as possible. Meanwhile, the $l_{2,1}$-norm constraint is used for feature selection and outlier reduction when the mapping matrices are learned. In addition, we propose learning different mapping matrices corresponding to different sub-tasks to emphasize the semantic and structural information of query data. Experiment results on three datasets demonstrate that our method performs better than the state-of-the-art methods.**

Index Terms—**Semi-supervised, cross-modal retrieval, feature selection.**

Manuscript received February 2, 2018; revised July 15, 2018 and September 13, 2018; accepted October 3, 2018. Date of publication October 22, 2018; date of current version April 23, 2019. This work was supported in part by the Natural Science Foundation for Distinguished Young Scholars of Shandong Province (JQ201718), in part by the Key Research and Development Foundation of Shandong Province (2016GGX101009), in part by the Natural Science Foundation of China (U1736122), in part by Shandong Provincial Key Research and Development Plan (2017CXGC1504), and in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract no. D17PC00340. *(Corresponding author: Jiande Sun.)*

E. Yu and J. Sun are with the School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China (e-mail: sdnu_enyu@hotmail.com; jiandesun@hotmail.com).

J. Li is with the School of Mechanical and Electrical Engineering, Shandong Management University, Jinan 250014, China, and also with Shandong Normal University, Jinan 250014, China (e-mail: lijingjdsun@hotmail.com).

X. Chang is with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia (e-mail: cxj273@gmail.com).

X.-H. Han is with the Graduate School of Science and Technology for Innovation, Yamaguchi University, Yamaguchi 753-8511, Japan (e-mail: hanxhua@yamaguchi-u.ac.jp).

A. G. Hauptmann is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: alex@cs.cmu.edu).

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

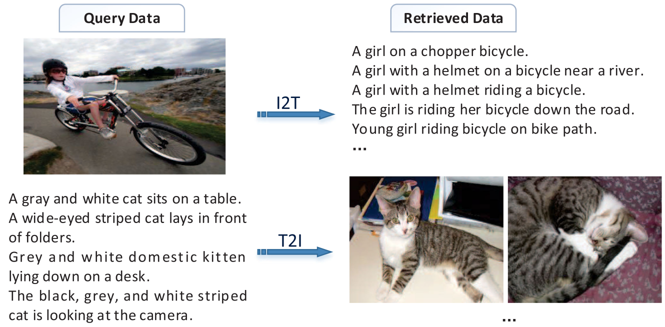
Digital Object Identifier 10.1109/TMM.2018.2877127

I. Introduction

With the explosive growth of multi-modal data just like texts, images, videos and audios, more and more researchers concentrate on analyzing the correlation among various modal information. In the field of multi-modal research, cross-modal retrieval has attracted much attention in recent years [1]–[10]. Given query data from one modality, it aims to retrieve semantically related data from other modalities, such as using an image to retrieve the relevant texts (I2T) or using a text to retrieve the relevant images (T2I), as shown in Fig. 1. In this paper, we investigate a semi-supervised generalized framework for the cross-modal retrieval among multi-modalities and verify it through the retrieval task between images and texts.

The main challenge of cross-modal retrieval is to bridge the semantic-gap between multi-modal data [11]–[16]. In order to address the problem, many classical methods have been proposed, such as dictionary learning, subspace learning, hash-based methods and so on. In this paper, we focus on subspace learning, and it means learning a common latent subspace shared by multi-modal data, in which the similarity across different modalities can be measured directly [4], [14], [17]–[19]. The traditional subspace learning methods can be roughly divided into two categories depending on it is trained by labeled data or unlabeled data. Unsupervised subspace learning methods merely leverage the pairwise information of unlabeled samples from different modalities to learn the common latent subspace [14], [18], [20], [21]. It keeps the closeness of different modalities in common subspace. However, these unsupervised cross-modal retrieval methods only exploit the unlabeled data for training and ignore the explicit high-level semantics. On the other hand, supervised subspace learning methods use the labeled data completely for training and the label information is exploited to enforce the samples from same classes lie as close as possible [22]–[25], so it can learn more discriminative subspace with supervised semantic labels. Fig. 2 shows the differences between unsupervised and supervised methods. However, annotating a great deal of data may consume a lot of labor and time. It is very difficult to obtain the labeled data in real world for supervised learning. Fortunately, semi-supervised cross-modal retrieval methods can utilize both labeled and unlabeled data by exploring the potential information underlying the labeled and unlabeled data to improve the diversity of training data.

Motivated by the above analysis, a semi-supervised method, named Adaptive Semi-supervised Feature Selection (ASFS), is proposed in this paper. Since we can conclude from supervised



A girl on a chopper bicycle.
A girl with a helmet on a bicycle near a river.
A girl with a helmet riding a bicycle.
The girl is riding her bicycle down the road.
Young girl riding bicycle on bike path.

A gray and white cat sits on a table.
A wide-eyed striped cat lays in front of folders.
Grey and white domestic kitten lying down on a desk.
The black, grey, and white striped cat is looking at the camera.

Fig. 1. Cross-modal retrieval tasks between image and text modalities. (The images and texts are selected from Pascal Sentence dataset).

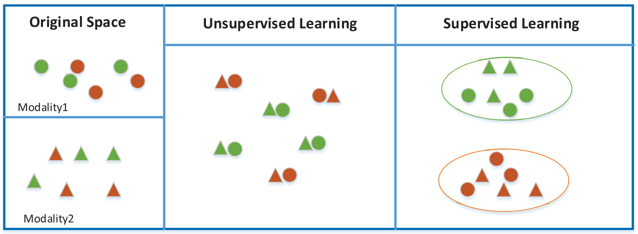


Fig. 2. The difference between unsupervised and supervised subspace learning methods. The same color indicates the same class, and the same shape indicates the same modality.

methods that labels can reflect the abundant high-level semantics of multi-modal data, we firstly propose preserving semantic consistency between query modal feature space and semantic space by the linear regression. It can keep the closeness of the data within the same class in the latent subspace. Besides, the correlation between pairwise data is also considered to keep neighbor relationship of heterogeneous data in the common latent subspace inspired by unsupervised learning. Secondly, we adopt the label graph propagation constraint to predict accurate labels for unlabeled data, which can keep the geometric structural consistency between feature space and label space. Meanwhile, the $l_{2,1}$-norm constrain is utilized when the mapping matrices are learned. It can not only select the informative and discriminative features, but also improve the robustness of the semi-supervised framework via reducing the effects of outliers caused by unlabeled data [26]. Finally, an efficient joint optimization algorithm is proposed to update the mapping matrices and the label matrix for unlabeled data simultaneously and iteratively. Fig. 3 depicts the whole framework of the proposed semi-supervised method, and we can see that the mapping matrices and predicted label matrix interact with each other to update iteratively. Therefore, it can adaptively learn the most accurate labels for unlabeled data and ensure that the raw feature distribution will be as consistent as possible with the semantic distribution in the latent subspace. In addition, because of different modalities have their own semantic and structural distributions, the accurate semantic and structural representation of query data in the latent subspace is helpful to the retrieval performance. If the semantic distribution of the query data is not properly represented, it is difficult to retrieve the relevant data in other modalities. Therefore, the more discriminative mapping matrices are learned depend on the query modality and sub-tasks. The contributions of our paper can be summarized as follows:

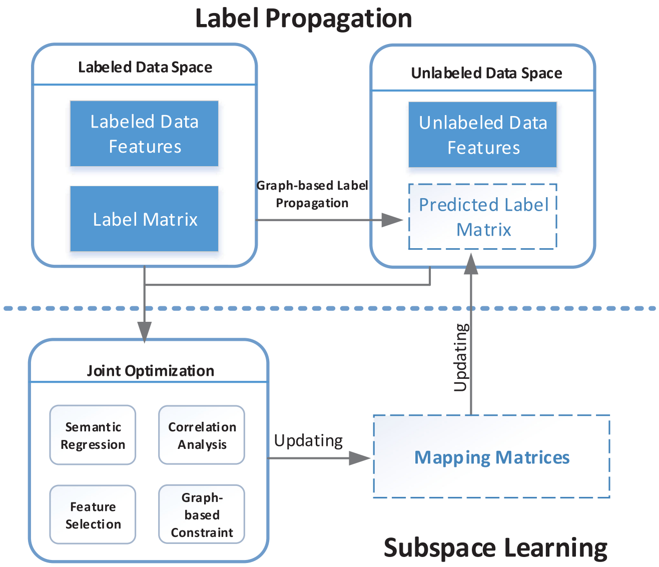


Fig. 3. The semi-supervised framework of ASFS. Firstly, the labels for unlabeled data are predicted by the graph-based label propagation. Then the unlabeled data with the predicted labels are combined with the labeled data to learn the mapping matrices. Meanwhile, the mapping matrices update the predicted label matrices, which can ensure that the raw feature distribution will be as consistent as possible with the semantic distribution in the subspace after several iterations.

1) In our semi-supervised method, the label graph propagation is devised to explore the potential relationship between labeled and unlabeled data for label propagation. Furthermore, it can keep the geometric structure consistency between raw feature space and label space.
2) We integrate the semantic regression, heterogeneous correlation, label graph propagation and feature selection based on $l_{2,1}$-norm constrain into a joint cross-modal learning framework. They interact with each other and embed more semantic and structural information in the shared cross-modal retrieval subspace.
3) An efficient joint optimization algorithm is proposed to update the mapping matrices and the label matrix for unlabeled data simultaneously and iteratively, so that the more relevant labels and more discriminative mapping matrices can be obtained adaptively.

The rest of our paper is organized as follows: Section II introduces the related works about the cross-modal retrieval. In Section III, we describe the proposed method. Section IV presents the experiment performance and analysis. Finally, we make a conclusion in Section V.

II. Related Work

Cross-modal retrieval has drawn more and more attention in relevant research areas and been used in many real-world applications. Thus, various approaches have been proposed to improve the retrieval performance. These existing (subspace-based) cross-modal retrieval methods can be divided into three categories according to training data type: 1) unsupervised learning that only used the unlabeled data for training. 2) supervised learning that only used the labeled data for training. 3) semi-supervised learning that used both the labeled and unlabeled data for training.

The unsupervised subspace learning methods used the pairwise data from different modalities to learn the latent subspace. For instance, Canonical Correlation Analysis (CCA) [27], the most popular unsupervised subspace learning method, aimed at maximizing the correlations between two set of variables to get one couple of projection matrices. Based on CCA, Rasiwasia *et al.* [14] proposed Semantic Matching (SM) and Semantic Correlation Matching (SCM). Zhang *et al.* [28] proposed Kernel Canonical Correlation Analysis (KCCA). Besides, Partial Least Squares (PLS) [29] and Bilinear Model (BLM) [20] were also the popular unsupervised methods for cross-modal retrieval. But, unsupervised methods only considered the pairwise correlations among heterogeneous data and ignored the explicit high-level semantics in subspace learning. Therefore, many supervised subspace learning methods were developed to capture the high-level semantic information by using class information. For example, based on CCA, the Locality Correlation Preserving based Support Vector Machine (LCPSVM) was proposed by Zhang *et al.* [30] to keep local correlation of categories. Three-view CCA proposed in [31] used a third view to capture the high-level semantic information, and Generalized Multi-view Analysis (GMA) was proposed in [20]. Recently, some non-linear multi-view kernel-based learning approaches have been proposed to improve the retrieval performance. For instance, Cao *et al.* [32] proposed the kernel-based Multi-view Modular Discriminant Analysis (MvMDA), which used the kernel-based method for non-linear embedding and considered the inter-view and intra-view covariances when learning the latent subspace. And the Kernel-based Multi-view Nonparametric Discriminant Analysis (KMvNDA) [33] was proposed to exploit the class boundary structure and discrepancy information of the available views.

Recently, semi-supervised subspace learning has achieved promising performance for cross-modal retrieval [34], [35], since semi-supervised learning can not only use the label information but also explore the potential information of unlabeled data. For example, Zhang *et al.* [36] proposed the Generalized Semi-supervised Structured Subspace Learning (GSS-SL) method. It used the graph-based method for label propagation firstly, and then the label graph constraint and label-linked loss function were integrated to learn the common subspace, which can cluster the samples with the same semantic and keep the structured information in the latent subspace. However, Zhang *et al.* ignored correlation among heterogeneous data and the pairwise relationship across different modalities in the latent subspace. Furthermore, in their method, one couple of mapping matrices was learned for both I2T and T2I tasks without considering the importance of the accurate semantic and structural representation of query data. To address these problems, we propose an adaptive semi-supervised method for cross-modal retrieval according to retrieval tasks. In our method, different mapping matrices are learned corresponding to the query modality. Meanwhile, the pairwise correlation among heterogeneous data and semantic information are considered to learn more discriminative common subspace. And we propose to update the mapping matrices and the label matrix of unlabeled data simultaneously and iteratively, so that it can make samples from different classes to be far apart while those from the same class lie as close as possible. Furthermore, we impose the $l_{2,1}$-norm constraint to select the informative and discriminative features and reduce the effects of outliers caused by unlabeled data when learning the mapping matrices.

TABLE I
LIST OF NOTATIONS

Notation	Description
t	Index of different sub-task
q	Index of different query modality
M	The total number of modalities
U_{tq}	The mapping matrices for query modal data
U_{tm}	The mapping matrices for m-th modal data
d, p	Feature dimensions of image and text
$Y = [\hat{Y}, \overset{\vee}{Y}]$	Label matrix for all labeled and unlabeled data
$X_1 = [\hat{X}_1, \overset{\vee}{X}_1]$	Feature matrix for all labeled and unlabeled image
$X_2 = [\hat{X}_2, \overset{\vee}{X}_2]$	Feature matrix for all labeled and unlabeled text
U_{I1}, U_{I2}	Mapping matrices in I2T sub-task
U_{T1}, U_{T2}	Mapping matrices in T2I sub-task

III. Proposed Method

In this section, we give a detailed description about our method, which is called Adaptive Semi-supervised Feature Selection (ASFS) for cross-modal retrieval. Firstly, we define the generalized objective function among multi-modalities for different sub-retrieval tasks. And it includes semantic consistency analysis term, heterogeneous data correlation analysis term, graph model constraint and feature selection constraint. Then, we will introduce the details of the formulated terms in the objective function and give the specific objective for I2T and T2I tasks. Finally, an effective iterative optimization algorithm is designed to solve the formulated optimization problem, and its convergence is proved.

A. Generalized Objective Function

As we have mentioned, in our proposed method, different mapping matrices are learned based on the query modality corresponding to sub-retrieval tasks. Thus, we define the generalized objective function corresponding to the query modality and sub-task as:

$$\begin{aligned} &\min_{U_{tq}, U_{t1}..U_{tm}..,U_{tM-1}, \overset{\vee}{Y}} f_t(U_{tq}, U_{t1}, ..U_{tm}.., U_{tM-1}; Y) \\ &= \beta S_q(U_{tq}, Y) + (1-\beta) \sum_{m \neq q}^{M} C(U_{tq}, U_{tm}) \\ &\quad + \gamma G_q(U_{tq}, Y) + N(U_{tq}, U_{t1}, \ldots U_{tm} ..., U_{tM-1}) \quad (1) \end{aligned}$$

where β is the balance parameters and $0 \leq \beta \leq 1$. γ is the positive hyper-parameter for adjusting the weight of the graph constraint. Y is the label matrix for both labeled and unlabeled data and $Y = [\hat{Y}, \overset{\vee}{Y}]$. More descriptions of notations can be seen in Table I. The first and third terms are the semantic regression and graph regularization term, respectively, which depend on the query modality only. The second term is the correlation analysis term, which indicates to keep the closeness of query modality and any other modalities, and the last term is the feature selection term based on $l_{2,1}$-norm.

In details, we use the semantic regression term to explore the label information, which aims at making the data within the same class as close as possible in the common latent subspace. So the semantic regression term can be formulated as:

$$S_q(U_{tq}, Y) = ||{X_q}^T U_{tq} - Y||_F^2 \tag{2}$$

And in order to obtain the correlation relationship between different modalities, the correlation analysis term of heterogeneous data is proposed to keep the pairwise data as close as possible in the common latent subspace, and it is formulated as:

$$C(U_{tq}, U_{tm}) = \sum_{m \neq q}^{M} ||{X_q}^T U_{tq} - {X_m}^T U_{tm}||_F^2 \tag{3}$$

where $q, m \in (1, 2 \ldots, M)$.

B. Specific Objective Function for I2T and T2I

Firstly, we suppose the dataset of n instances is $D = \{(x_{1i}, x_{2i})\}_{i=1}^{n}$, where $x_{1i} \in \mathbb{R}^d$ denotes the image feature and $x_{2i} \in \mathbb{R}^p$ denotes the text feature. Specifically, the labeled feature matrices of image and text can be denoted as $\hat{X}_1 \in \mathbb{R}^{d \times \hat{n}}$ and $\hat{X}_2 \in \mathbb{R}^{p \times \hat{n}}$. The unlabeled data feature matrices of image and text are denoted as $\overset{\vee}{X}_1 \in \mathbb{R}^{d \times \overset{\vee}{n}}$ and $\overset{\vee}{X}_2 \in \mathbb{R}^{p \times \overset{\vee}{n}}$. As for label representation, each sample is assigned with a c-dimensional binary-valued vector. If the i-th sample is classified as k-th class, y_{ik} is set to 1, otherwise 0. So the label matrix for labeled data can be represented as $\hat{Y} \in \mathbb{R}^{\hat{n} \times c}$. For unlabeled data, we use the graph model to predict the label matrix as the initial value. Therefore, the overall label matrix can be denoted as $Y = [\hat{Y}, \overset{\vee}{Y}] \in \mathbb{R}^{n \times c}$, where c is the total number of classes in the dataset. According to the analysis of semantic regression and correlation analysis terms, we can conclude that the proposed method combines the advantages of unsupervised method and supervised method. It not only considers the label information but also utilizes the pairwise relationship when learning the optimal mapping matrices, as shown in Fig. 4. Besides, in order to emphasize the semantic and structural information of query data, we define different objective functions corresponding to sub-retrieval tasks according to Eq. (1):

a) The objective function for I2T:

$$\begin{aligned} \min_{U_{I1}, U_{I2}, \overset{\vee}{Y}} \; f_I = {} & \beta||X_1^T U_{I1} - Y||_F^2 \\ & + (1-\beta)||X_1^T U_{I1} - X_2^T U_{I2}||_F^2 \\ & + \gamma G_1(U_{I1}, Y) + N(U_{I1}, U_{I2}) \end{aligned} \tag{4}$$

where $q = 1$ represents image query and $t = I$ represents I2T sub-task.

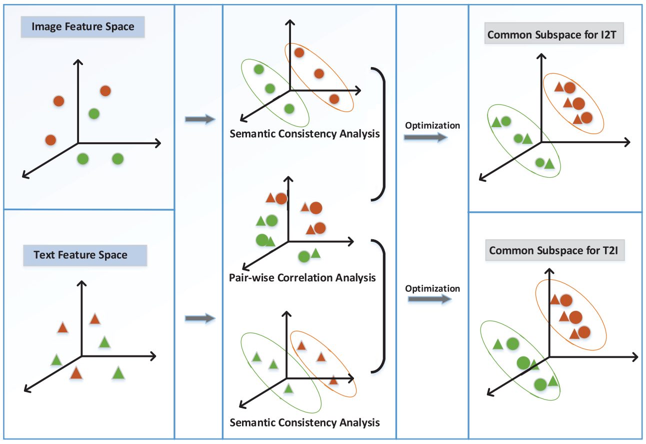


Fig. 4. The correlation analysis among different modalities in this method. The same color indicates the same class, and the same shape indicates the same modality.

b) The objective function for T2I:

$$\begin{aligned} \min_{U_{T1}, U_{T2}, \overset{\vee}{Y}} \; f_T = {} & \beta||X_2^T U_{T2} - Y||_F^2 \\ & + (1-\beta)||X_1^T U_{T1} - X_2^T U_{T2}||_F^2 \\ & + \gamma G_2(U_{T2}, Y) + N(U_{T1}, U_{T2}) \end{aligned} \tag{5}$$

where $q = 2$ represents the text query and $t = T$ represents T2I sub-task.

C. Graph Constraint

In this paper, we define a graph constraint term for two goals: 1) To explore the geometric structural neighborhood relationship of data samples in the common latent subspace. 2) To keep the structural consistency between feature space and label space so that the label information can be propagated from labeled data to unlabeled data accurately.

Specifically, we construct different graph models corresponding to the subtasks. In other words, we regard the query modal data as the vertices in graph so that the structure information of query modal data can be emphasized. Thus, we regard all labeled and unlabeled data from query modality as the vertices in the graph, and the edge weights can be formulated as:

$$W_q^{i,j} = \begin{cases} \exp\left(-\frac{||x_{qi} - x_{qj}||_2^2}{2\sigma^2}\right), & x_{qi} \in N_k(x_{qj}) \; or \; x_{qj} \in N_k(x_{qi}) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $N_k(x_{qi})$ (or $N_k(x_{qj})$) denotes k nearest neighbors of x_{qi} (or x_{qj}). $W_q^{i,j}$ is a $n \times n$ symmetric graph for t-th sub-task and its edge weights are non-negative values. Then we denote the $D_q = diag(d_{q1}, d_{q2}, \ldots d_{qn})$ and $d_{qi} = \sum_{j=1}^{n} W_q^{i,j}$, so the normalized graph Laplacain matrix L_q can be formulated as:

$$L_q = {D_q}^{-1/2}(D_q - W_q){D_q}^{-1/2} \tag{7}$$

As mentioned above, we construct different undirected weighted graphs corresponding to sub-tasks according to the query modal data, the weights of which represent the pairwise relationships. We argue that the graph of query modality must be consistent with the label graphs if the more accurate results are desired. According to the purpose and analysis [36], the graph term for t-th sub-task can be formulated as:

$$G_q(U_{tq}, Y) = \frac{1}{2}\sum_{i,j=1}^{n} W_q^{i,j}\left(\left\|\frac{U_{tq}^T x_{qi}}{\sqrt{D_q^{ii}}} - \frac{U_{tq}^T x_{qj}}{\sqrt{D_q^{jj}}}\right\|_2^2\right) - \frac{1}{2}\sum_{i,j=1}^{n} W_q^{i,j}\left(\left\|\frac{y_i}{\sqrt{D_q^{ii}}} - \frac{y_j}{\sqrt{D_q^{jj}}}\right\|_2^2\right) \tag{8}$$

where the first term ensures data points with similar semantics in the feature space can keep the geometric structural neighborhood relationship in the latent common subspace, and the second term can ensure the labels of pairwise data with large weights should be similar [36], [37]. Therefore, the structural consistency between feature space and label space can be kept and the label information can be propagated from labeled data to unlabeled data by the graph model.

D. Feature Selection Constraint

In this paper, we propose using the $l_{2,1}$-norm for feature selection, which can ensure that the optimal mapping matrices can be learned by selecting the informative and discriminative features from heterogeneous feature spaces. And we formulate the feature selection constraint corresponding to retrieval sub-tasks as:

$$N(U_{tq}, U_{tm}) = \lambda_q ||U_{tq}||_{2,1} + \sum_{m\neq q}^{M-1} \lambda_m ||U_{tm}||_{2,1} \tag{9}$$

where λ_q and λ_m are positive parameters for balancing the regularization terms.

In details, we generally solve the constrained $l_{2,1}$-norm by using the half-quadratic minimization. For example, assuming the problem is

$$\min_U ||U||_{2,1} \tag{10}$$

where U is the projection matrix, as shown in [38], the following objective is often used

$$\min_U \sum \sqrt{\varepsilon + ||u^i||_2^2} \tag{11}$$

where ε is a smoothing term, and it is usually a small constant value. Now, we define $\phi(x) = \sqrt{\varepsilon + x^2}$ and we can simplify Eq. (11) as:

$$\min_U \sum \phi(||u^i||_2) \tag{12}$$

It has been proved in [38] that for a fixed $||u^i||_2$, there is a dual potential function $\varphi(\cdot)$:

$$\phi(||u^i||_2) = \inf_{s\in R}\{s||u^i||_2^2 + \varphi(s)\} \tag{13}$$

where s is determined by $\varphi(\cdot)$ with respect to $\phi(\cdot)$. Based on the half-quadratic minimization in [39], Eq. (10) can be replaced by:

$$||U||_{2,1} = Tr\left(U^T R U\right) \tag{14}$$

where $R = diag(r)$. r can be formulated:

$$r = \frac{1}{2\sqrt{||u^i||_2^2 + \varepsilon}} \tag{15}$$

E. Joint Optimization

Based on the definition of $l_{2,1}$- norm in Eq. (14) and the graph theory [36], [40], [41], the objective functions of I2T and T2I sub-tasks in Eq. (4) and Eq. (5) can be rewritten as:

a) For I2T sub-task:

$$\begin{aligned}\min_{U_{I1}, U_{I2}, \overset{\vee}{Y}} f_I &= \beta||X_1^T U_{I1} - Y||_F^2 \\ &+ (1-\beta)||X_1^T U_{I1} - X_2^T U_{I2}||_F^2 \\ &+ \gamma Tr\left({U_{I1}}^T X_1 L_1 X_1^T U_{I1} - Y^T L_1 Y\right) \\ &+ \lambda_1 Tr\left({U_{I1}}^T R_{I1} U_{I1}\right) \\ &+ \lambda_2 Tr\left({U_{I2}}^T R_{I2} U_{I2}\right)\end{aligned} \tag{16}$$

b) For T2I sub-task:

$$\begin{aligned}\min_{U_{T1}, U_{T2}, \overset{\vee}{Y}} f_T &= \beta||X_2^T U_{T2} - Y||_F^2 \\ &+ (1-\beta)||X_1^T U_{T1} - X_2^T U_{T2}||_F^2 \\ &+ \gamma Tr\left({U_{T2}}^T X_2 L_2 X_2^T U_{T2} - Y^T L_2 Y\right) \\ &+ \lambda_1 Tr\left({U_{T1}}^T R_{T1} U_{T1}\right) \\ &+ \lambda_2 Tr\left({U_{T2}}^T R_{T2} U_{T2}\right)\end{aligned} \tag{17}$$

Label Propagation: According the graph theory, vertices with larger edge weight should have more similar semantic information. Therefore, we firstly propose the graph-based label propagation to predict more accurate labels for unlabeled data as the initialized label matrix. The graph Laplacian matrix can explore the geometric structures of heterogeneous feature space and propagate the labels. Here, we formulate the propagation function:

$$\min_{\overset{\vee}{Y}} \left(Y^T L_q Y\right) \tag{18}$$

where L_q is the normalized Laplacian matrix constructed by query data. Now, we rewrite the Eq. (18) as:

$$\min_{\overset{\vee}{Y}} Tr\left(\begin{bmatrix}\hat{Y} \\ \overset{\vee}{Y}\end{bmatrix}^T \begin{bmatrix}L_q^{ll} & L_q^{lu} \\ L_q^{ul} & L_q^{uu}\end{bmatrix} \begin{bmatrix}\hat{Y} \\ \overset{\vee}{Y}\end{bmatrix}\right) \tag{19}$$

Then, calculating the derivative of $\overset{\vee}{Y}$ and setting it to zero, we can get the solution:

$$\overset{\vee}{Y} = -\left(L_q^{uu}\right)^{-1} L_q^{ul} \hat{Y} \tag{20}$$

Alternately Updating: We use the graph-based label propagation to ensure the initial labels are more accurate. Therefore, it ensures predicted labels as accurate as possible in the process of iteration, and the feature distribution to be as consistent as possible with the semantic distribution in the latent subspace. Specifically, once the initial label matrix $Y = [\hat{Y}, \overset{\vee}{Y}]$ is given, then we can update the predicted label matrix and mapping matrices iteratively. Obviously, we can find the objective function in Eq. (16) is non-convex, but it convex with respect to either U_{t1}, U_{t2} or $\overset{\vee}{Y}$ while the other two are treated as constants. So we can get the optimal solutions of mapping matrices and label matrix by calculating the partial derivatives of U_{t1}, U_{t2} and $\overset{\vee}{Y}$ in Eq. (16) and Eq. (17) and setting them to zero.

a) Updating mapping matrices: In this part, we will focus on the I2T retrieval task to explain the optimization strategy. Thus, we calculate the partial derivatives of U_{I1} and U_{I2} for I2T task and set them to zero respectively. Then the follow results can be obtained:

$$\begin{aligned}\frac{\partial f_I}{\partial U_{I1}} &= X_1X_1^T + \gamma X_1L_1X_1^TU_{I1} + \lambda_1R_{I1}U_{I1} - \beta X_1Y\\ &\quad - X_1X_2^TU_{I2} + \beta X_1X_2^TU_{I2} = 0 \qquad (21)\end{aligned}$$

$$\begin{aligned}\frac{\partial f_I}{\partial U_{I2}} &= (1-\beta)X_2X_2^TU_{I2} - (1-\beta)X_2X_1^TU_{I1}\\ &\quad + \lambda_2R_{I2}U_{I2} = 0 \qquad (22)\end{aligned}$$

From the above formulas, we can get:

$$\begin{aligned}U_{I1} &= \left(X_1X_1^T + \gamma X_1L_1X_1^T + \lambda_1R_{I1}\right)^{-1}[\beta X_1Y\\ &\quad + (1-\beta)X_1X_2^TU_{I2}] \qquad (23)\end{aligned}$$

$$U_{I2} = \left[(1-\beta)X_2X_2^T + \lambda_2R_{I2}\right]^{-1}(1-\beta)X_2X_1^TU_{I1} \qquad (24)$$

b) Updating label matrix: According to the definition of label matrix $Y = [\hat{Y}, \overset{\vee}{Y}]$, our goal is to update the unknown matrix $\overset{\vee}{Y}$. We remove the constraints which are irrelevant to $\overset{\vee}{Y}$ in Eq. (16), and it can be rewritten as:

$$\begin{aligned}f_I &= \beta Tr\left[\left(\overset{\vee}{X}{}_1^T U_{I1} - \overset{\vee}{Y}\right)^T\left(\overset{\vee}{X}{}_1^T U_{I1} - \overset{\vee}{Y}\right)\right]\\ &\quad - \gamma\left(\overset{\vee}{Y}{}^T L_1^{ul}\hat{Y} + \hat{Y}{}^T L_1^{ul}\overset{\vee}{Y} + \overset{\vee}{Y}{}^T L_1^{uu}\overset{\vee}{Y}\right) \qquad (25)\end{aligned}$$

Then, we calculate the partial derivatives of $\overset{\vee}{Y}$ and set to zero. We can get:

$$\frac{\partial f_I}{\partial \overset{\vee}{Y}} = \beta\left(\overset{\vee}{Y} - \overset{\vee}{X}{}_1^T U_{I1}\right) - \gamma\left(L_1^{ul}\hat{Y} + L_1^{uu}\overset{\vee}{Y}\right) = 0 \qquad (26)$$

$$\overset{\vee}{Y} = (\beta - \gamma L_1^{uu})^{-1}\left(\beta\overset{\vee}{X}{}_1^T U_{I1} + \gamma L_1^{ul}\hat{Y}\right) \qquad (27)$$

Similarly, we can get the optimal solutions of mapping matrices and label matrix in T2I task according to the above optimization strategy:

$$U_{T1} = [(1-\beta)X_1X_1^T + \lambda_1R_{T1}]^{-1}(1-\beta)X_1X_2^TU_{T2} \qquad (28)$$

$$\begin{aligned}U_{T2} &= (X_2X_2^T + \lambda_2R_{T2} + \gamma X_2L_2X_2^T)^{-1}[\beta X_2Y\\ &\quad + (1-\beta)X_2X_1^TU_{T1}] \qquad (29)\end{aligned}$$

$$\overset{\vee}{Y} = (\beta - \gamma L_2^{uu})^{-1}\left(\beta\overset{\vee}{X}{}_2^T U_{T2} + \gamma L_2^{ul}\hat{Y}\right) \qquad (30)$$

The optimization procedure of our method is summarized in Algorithm 1. We can find that the joint optimization strategy uses the initialized label matrix Y to update the mapping matrices U_{tm} firstly. Then, the learned mapping matrices update the predicted label matrix iteratively. Therefore, the label matrix and mapping matrices rely on each other to learn the optimal results.

```
Algorithm 1: ASFS-I2T(or T2I)
Input: Image feature matrix $X_1 = [\overset{\wedge}{X}_1, \overset{\vee}{X}_1]$ and text feature
    matrix $X_2 = [\overset{\wedge}{X}_2, \overset{\vee}{X}_2]$, the label matrix $Y = [\overset{\wedge}{Y}, \overset{\vee}{Y}]$,
    parameters: $\beta$, $\gamma$, $\lambda_1$, $\lambda_2$.
Initialize:
  Calculate normalized graph Laplacian matrix $L_1$(or $L_2$)
    according to Eq. (7). And initialize the predicted label
    matrix $\overset{\vee}{Y}$ for unlabeled data according to Eq. (20)
  Initialize $U_{t1}$ and $U_{t2}$ as identity matrixes.
  while not converge do
  Fixing the $U_{I2}$(or $U_{T2}$) and $Y$, update the $U_{I1}$(or $U_{T1}$)
    according to Eq. (23) (or Eq. (28)).
  Fixing the $U_{I1}$(or $U_{T1}$) and $Y$, update the $U_{I2}$(or $U_{T2}$)
    according to Eq. (24) (or Eq. (29)).
  Fixing the $U_{I1}$(or $U_{T1}$) and $U_{I2}$ (or $U_{T2}$), update the $\overset{\vee}{Y}$
    according to Eq. (27) (or Eq. (30)).
end
Output: The projection matrices $U_{I1}$ (or $U_{T1}$), $U_{I2}$ (or
    $U_{T2}$) and label matrix $Y$.
```

F. Convergence Analysis

In order to prove that the proposed optimize algorithm in Algorithm 1 will converge, we present the following lemmas.

Lemma 1: Supposing the u_i^t donates the i-th row of the updated matrix U_i^t in previous iteration and the u_i^{t+1} donates the i-th row of the variable U_i^{t+1} in the next iteration, then we can obtain the conclusion:

$$||u_i^{t+1}||_2 - \frac{||u_i^{t+1}||_2^2}{2||u_i^t||_2} \leq ||u_i^t||_2 - \frac{||u_i^t||_2^2}{2||u_i^t||_2} \qquad (31)$$

Proof: We consider the following function

$$f(x) = px^2 - 2x^p + (2-p) \qquad (32)$$

where $p \in (0, 2)$. We expect to show that when $x > 0$, $f(x) \geq 0$. Now, we calculate the first and second order derivatives of the function in Eq. (32), $f'(x) = 2px - 2px^{p-1}$ and $f''(x) =$

$2p - 2p(p-1)x^{p-2}$. We can find that $x = 1$ is the only point that satisfies $f'(x) = 0$, and when $0 < x < 1$, $f'(x) < 0$ and when $1 < x$, $f'(x) > 0$. It means that $f(x)$ is monotonically decreasing when $0 < x < 1$ and monotonically increasing when $x > 1$. Furthermore, we have $f''(1) = 2p(2-p) > 0$. Therefore, when $x > 0$, $f(x) \geq g(1) = 0$. ■

Then, by substituting $x = \frac{||u_i^{t+1}||_2}{||u_i^t||_2}$ into Eq. (32) and set $p = 1$, we obtain the conclusion:

$$\begin{aligned}
&\frac{||u_i^{t+1}||_2^2}{||u_i^t||_2^2} - 2\frac{||u_i^{t+1}||_2}{||u_i^t||_2} + 1 \geq 0\\
&\Leftrightarrow ||u_i^{t+1}||_2^2 - 2||u_i^{t+1}||_2||u_i^t||_2 + ||u_i^t||_2^2 \geq 0\\
&\Leftrightarrow \frac{||u_i^{t+1}||_2^2}{||u_i^t||_2} - 2||u_i^{t+1}||_2 + ||u_i^t||_2 \geq 0\\
&\Leftrightarrow 2||u_i^{t+1}||_2 - \frac{||u_i^{t+1}||_2^2}{||u_i^t||_2} \leq ||u_i^t||_2\\
&\Leftrightarrow ||u_i^{t+1}||_2 - \frac{||u_i^{t+1}||_2^2}{2||u_i^t||_2} \leq ||u_i^t||_2 - \frac{||u_i^t||_2^2}{2||u_i^t||_2} \qquad (33)
\end{aligned}$$

According to the Lemma 1, if we sum up all the rows of U_i^t, we can easily get:

Lemma 2: Given the $U_i^t = [u_1^t, u_2^t, \ldots, u_d^t]$, where u_i^t donates the i-th row of the updated matrix U_i^t, then the inequality can be formulates as:

$$\sum_{i=1}^{d} ||u_i^{t+1}||_2 - \sum_{i=1}^{d} \frac{||u_i^{t+1}||_2^2}{2||u_i^t||_2} \leq \sum_{i=1}^{d} ||u_i^t||_2 - \sum_{i=1}^{d} \frac{||u_i^t||_2^2}{2||u_i^t||_2} \qquad (34)$$

Similarly, for the given U_{tm}^t and Y^t, we can similarly get the same conclusions according to Lemma 1 and Lemma 2:

$$||u_{tm(i)}^{t+1}||_2 - \frac{||u_{tm(i)}^{t+1}||_2^2}{2||u_{tm(i)}^t||_2} \leq ||u_{tm(i)}^t||_2 - \frac{||u_{tm(i)}^t||_2^2}{2||u_{tm(i)}^t||_2} \qquad (35)$$

$$\begin{aligned}
\sum_{i=1}^{p} ||u_{tm(i)}^{t+1}||_2 - \sum_{i=1}^{p} \frac{||u_{tm(i)}^{t+1}||_2^2}{2||u_{tm(i)}^t||_2} &\leq \sum_{i=1}^{p} ||u_{tm(i)}^t||_2\\
&\quad - \sum_{i=1}^{p} \frac{||u_{tm(i)}^t||_2^2}{2||u_{tm(i)}^t||_2} \qquad (36)
\end{aligned}$$

$$||y_i^{t+1}||_2 - \frac{||y_i^{t+1}||_2^2}{2||y_i^t||_2} \leq ||y_i^t||_2 - \frac{||y_i^t||_2^2}{2||y_i^t||_2} \qquad (37)$$

$$\sum_{i=1}^{c} ||y_i^{t+1}||_2 - \sum_{i=1}^{c} \frac{||y_i^{t+1}||_2^2}{2||y_i^t||_2} \leq \sum_{i=1}^{c} ||y_i^t||_2 - \sum_{i=1}^{c} \frac{||y_i^t||_2^2}{2||y_i^t||_2} \qquad (38)$$

Theorem 1: At each iteration of Algorithm 1, the value of the objective function monotonically decreases until convergence by adopting the proposed optimization method.

Proof: We suppose $\Gamma(U_{t1}^t, U_{t2}^t, Y^t)$ denoting the first three terms on the objective function in Eq. (1), and U_{t1}^t U_{t2}^t and Y^t are the optimized solution, then we obtain the following conclusion:

$$\begin{aligned}
&\Gamma(U_{t1}^{t+1}, U_{t2}^{t+1}, Y^{t+1}) + \lambda_1 Tr((U_{t1}^{t+1})^T R_{t1}(U_{t1}^{t+1}))\\
&\quad + \lambda_2 Tr((U_{t2}^{t+1})^T R_{t2}(U_{t2}^{t+1}))\\
&\leq \Gamma(U_{t1}^t, U_{t2}^t, Y^t) + \lambda_1 Tr((U_{t1}^t)^T R_{t1}(U_{t1}^t))\\
&\quad + \lambda_2 Tr((U_{t2}^t)^T R_{t2}(U_{t2}^t))\\
&\Rightarrow \Gamma(U_{t1}^{t+1}, U_{t2}^{t+1}, Y^{t+1}) + \lambda_1 \sum_{i=1}^{d} \frac{||u_{t1(i)}^{t+1}||_2^2}{2||u_{t1(i)}^{t+1}||_2}\\
&\quad + \lambda_2 \sum_{i=1}^{p} \frac{||u_{t2(i)}^{t+1}||_2^2}{2||u_{t2(i)}^{t+1}||_2}\\
&\leq \Gamma(U_{t1}^t, U_{t2}^t, Y^t) + \lambda_1 \sum_{i=1}^{d} \frac{||u_{t1(i)}^t||_2^2}{2||u_{t1(i)}^t||_2} + \lambda_2 \sum_{i=1}^{p} \frac{||u_{t2(i)}^t||_2^2}{2||u_{t2(i)}^t||_2}\\
&\Rightarrow \Gamma(U_{t1}^{t+1}, U_{t2}^{t+1}, Y^{t+1}) + \lambda_1 \sum_{i=1}^{d} ||u_{t1(i)}^{t+1}||_2\\
&\quad - \lambda_1 \left(\sum_{i=1}^{d} ||u_{t1(i)}^{t+1}||_2 - \sum_{i=1}^{d} \frac{||u_{t1(i)}^{t+1}||_2^2}{2||u_{t1(i)}^{t+1}||_2} \right)\\
&\quad + \lambda_2 \sum_{i=1}^{p} ||u_{t2(i)}^{t+1}||_2 - \lambda_2 \left(\sum_{i=1}^{p} ||u_{t2(i)}^{t+1}||_2 - \sum_{i=1}^{p} \frac{||u_{t2(i)}^{t+1}||_2^2}{2||u_{t2(i)}^{t+1}||_2} \right)\\
&\leq \Gamma(U_{t1}^t, U_{t2}^t, Y^t) + \lambda_1 \sum_{i=1}^{d} ||u_{t1(i)}^t||_2\\
&\quad - \lambda_1 \left(\sum_{i=1}^{d} ||u_{t1(i)}^t||_2 - \sum_{i=1}^{d} \frac{||u_{t1(i)}^t||_2^2}{2||u_{t1(i)}^t||_2} \right)\\
&\quad + \lambda_2 \sum_{i=1}^{p} ||u_{t2(i)}^t||_2 - \lambda_2 \left(\sum_{i=1}^{p} ||u_{t2(i)}^t||_2 - \sum_{i=1}^{p} \frac{||u_{t2(i)}^t||_2^2}{2||u_{t2(i)}^t||_2} \right) \qquad (39)
\end{aligned}$$

Given the conclusion of Lemma 2, we finally arrive at:

$$\begin{aligned}
&\Gamma(U_{t1}^{t+1}, U_{t2}^{t+1}, Y^{t+1}) + \lambda_1 \sum_{i=1}^{d} ||u_{t1(i)}^{t+1}||_2 + \lambda_2 \sum_{i=1}^{p} ||u_{t2(i)}^{t+1}||_2\\
&\quad \leq \Gamma(U_{t1}^t, U_{t2}^t, Y^t) + \lambda_1 \sum_{i=1}^{d} ||u_{t1(i)}^t||_2 + \lambda_2 \sum_{i=1}^{p} ||u_{t2(i)}^t||_2 \qquad (40)
\end{aligned}$$

■

Hence, the value of the objective function monotonically decrease in each iteration.

IV. Experiment

A. Datasets

Wikipedia: The dataset has 10 semantic categories, and includes totally 2866 image-text pairs. In experiments, we select 2173 image-text pairs for training and the other 693 pairs for testing randomly. As for the features, we use the 4096-dimensional

TABLE II
mAP OF ALL COMPARED APPROACHES ON THREE DATASETS. THE BEST RESULT IN EACH COLUMN IS MARKED WITH BOLD

Methods	***Wikipedia-CNN***			***Pascal Sentence***			***INRIA-Websearch***		
	I2T	T2I	Average	I2T	T2I	Average	I2T	T2I	Average
PLS [29]	35.95%	35.10%	35.53%	36.53%	37.63%	37.08%	19.38%	26.03%	22.71%
CCA [27]	33.16%	31.66%	32.41%	37.99%	37.20%	37.59%	26.03%	27.95%	26.99%
SM [14]	36.85%	38.67%	37.76%	44.98%	43.39%	44.19%	37.83%	35.31%	36.57%
SCM [14]	37.48%	39.26%	38.37%	40.71%	39.35%	40.03%	35.44%	30.87%	33.16%
GMMFA [20]	28.41%	24.87%	26.64%	37.32%	34.70%	36.01%	28.09%	30.37%	29.23%
GMLDA [20]	30.03%	28.06%	29.05%	40.80%	38.77%	39.79%	47.59%	54.07%	50.83%
MDCR [42]	41.07%	37.75%	39.41%	43.22%	46.22%	44.72%	47.09%	45.99%	46.54%
JLSLR [43]	39.38%	36.91%	38.15%	45.43%	45.53%	45.48%	52.50%	54.48%	53.49%
ASFS	**41.80%**	**44.84%**	**43.32%**	**48.90%**	**59.61%**	**54.26%**	**54.48%**	**70.61%**	**62.55%**

CNN [44] features for images and 100-dimensional LDA [45] features for texts.

Pascal Sentence: This dataset has 20 semantic categories, and each category includes 50 image-text pairs. We select 30 pairs in each category for training and the rest for testing. In this dataset, images are represented with 4096-dimensional CNN features, and texts are represented with 100-dimensional LDA features.

INRIA-Websearch: This dataset has 353 semantic categories, and includes totally 71478 image-text pairs. In experiments, we use largest 100 semantic classes to construct a experimental dataset, and select 70% of each category for training (10332 pairs) and the rest for testing (4366 pairs). For features, we use the 4096-dimensional CNN features for images and 1000-dimensional LDA features for texts.

B. Experiment Setting

1) Compared Methods: In this paper, we use the cosine distance for retrieval task and compare our ASFS with several state-of-the-art methods: **a)** Unsupervised methods: Partial Least Squares (PLS) [29] and Canonical Correlation Analysis (CCA) [27]. **b)** Supervised methods: Semantic Matching (SM) [14], Semantic Correlation Matching (SCM) [14], Generalized Multiview Marginal Fisher Analysis (GMMFA) [20], Generalized Multiview Linear Discriminant Analysis (GMLDA) [20], modality-dependent cross-media retrieval (MDCR) model [42] and Joint Latent Subspace Learning and Regression (JLSLR) [43]. **c)** Semi-supervised methods: Joint Representation Learning (JRL) [41] and Generalized Semi-supervised and Structured Subspace Learning (GSS-SL) [36].

2) Evaluation Metrics: In this paper, we use the mean Average Precision (mAP) of all categories and each category as evaluation standard. In briefly, we can formulate the AP of each query as:

$$AP = \frac{\sum_{i=1}^{N} P(i)rel(i)}{\sum_{i=1}^{N} rel(i)}$$

where N is the test data size. $rel(i) = 1$ when the item at rank i is relevant, otherwise $rel(i) = 0$. $P(i)$ is the precision of the result ranked at i. The mAP scores is calculated by averaging AP scores of all queries. Besides, the Precision-Recall is also used for evaluation.

3) Parameter Setting: On Wikipedia dataset, we set $\beta = 0.6$, $\gamma = 2$, $\lambda_1 = 0.6$, $\lambda_2 = 15$ for I2T task, and $\beta = 0.8$, $\gamma = 2$, $\lambda_1 = 0.01$, $\lambda_2 = 0.1$ for T2I task. On Pascal Sentence dataset, we set $\beta = 0.7$, $\gamma = 3$, $\lambda_1 = 0.5$, $\lambda_2 = 9$ for I2T task, and $\beta = 0.8$, $\gamma = 1$, $\lambda_1 = 0.01$, $\lambda_2 = 0.5$ for T2I task. And on INRIA-Websearch dataset, we set $\beta = 0.7$, $\gamma = 3$, $\lambda_1 = 0.5$, $\lambda_2 = 9$ for I2T task, and $\beta = 0.8$, $\gamma = 1$, $\lambda_1 = 0.01$, $\lambda_2 = 0.5$ for T2I task.

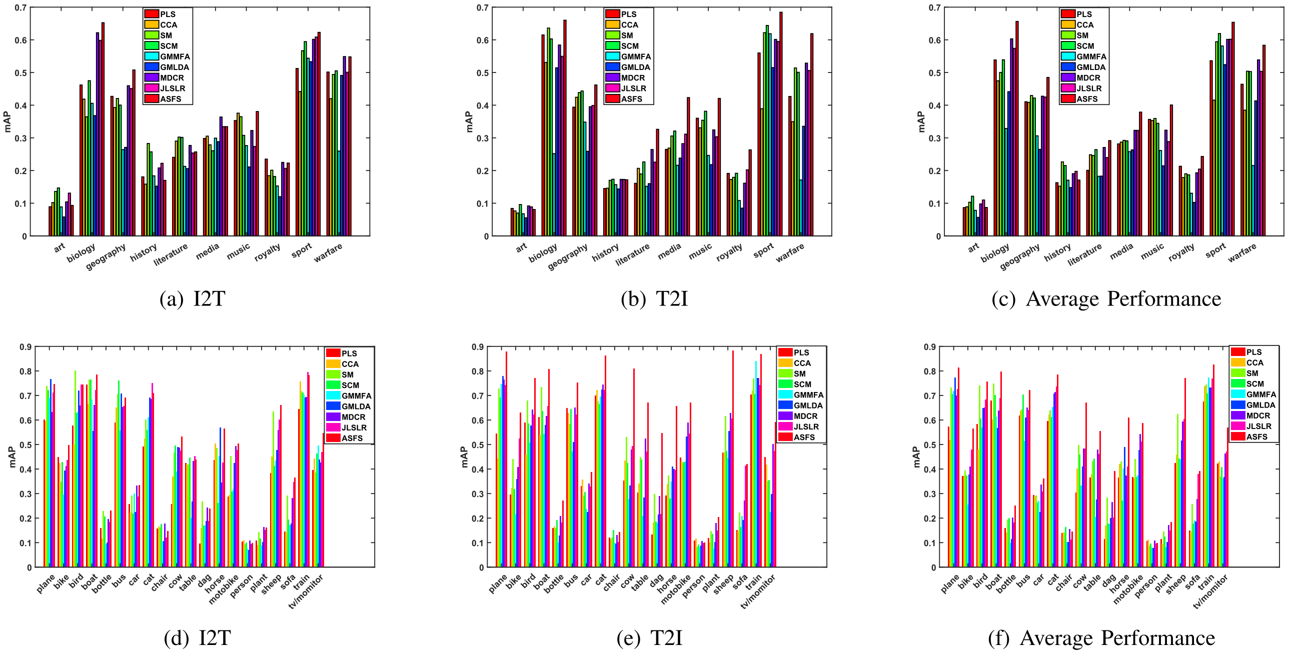
C. Experiment Performance and Analysis

1) Comparison With Unsupervised and Supervised Methods: The Table II shows the mAP scores of all compared approaches on three different datasets. On the Wikipedia dataset, we can find the ASFS method has a better performance on I2T and T2I tasks, especially the mAP score increases about 7% in T2I task. Fig. 5 shows the mAP score of each category, and we can find that our method has better performance on most of the categories than that of the other methods. And the Recall-Precision can be seen in Fig. 6. From the performance on Wikipedia dataset, we can conclude the ASFS method can reach the best performance compared with the other method. In details, compared with the supervised or unsupervised method, our method utilizes all labeled and unlabeled data for training, which can increase the diversity of training data. And the ASFS method not only uses the label information but also considers the pairwise relationship when learning the common subspaces. Besides, the $l_{2,1}$-norm constraint is used for informative and discriminative feature selection to ensure the optimal mapping matrices can be learned.

On Pascal Sentence, we can see from Table II that the improvement of retrieval performance is more obvious than that on Wikipedia dataset. And the mAP scores of our method are the best on most of the classes just as shown in Fig. 5. The precision-recall curves in Fig. 6 also show that our method is the best. The reason is that Pascal Sentence dataset has more categories and the label information is more abundant. Therefore, our semi-supervised model can use discriminative and abundant label information to propagate more accurate labels from labeled data to unlabeled data adaptively. Besides, we also compare the performance between our AFSF and its variant FS (supervised feature selection). From Table III we can find the semi-supervised method has superior performance.

In order to further demonstrate the effectiveness of our method, we test the ASFS method on a larger dataset (INRIA-Websearch dataset) with complex semantic and its performance is the best as shown in Table II and Fig. 6.

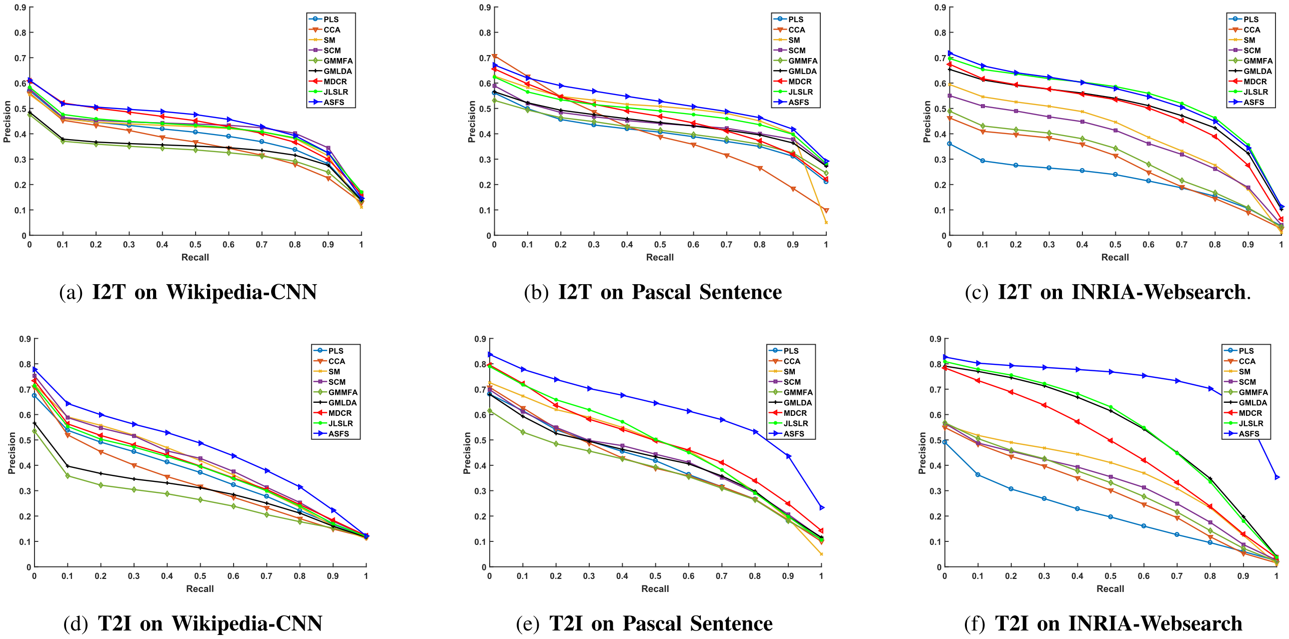
2) Comparison With Semi-Supervised Methods: In this part, we compare the ASFS method with two semi-supervised



(a) I2T (b) T2I (c) Average Performance

(d) I2T (e) T2I (f) Average Performance

Fig. 5. mAP performance of each class on **Wikipedia-CNN** and **Pascal Sentence**.



(a) **I2T on Wikipedia-CNN** (b) **I2T on Pascal Sentence** (c) **I2T on INRIA-Websearch**.

(d) **T2I on Wikipedia-CNN** (e) **T2I on Pascal Sentence** (f) **T2I on INRIA-Websearch**

Fig. 6. *Precision-Recall* curves of all compared approaches on three datasets.

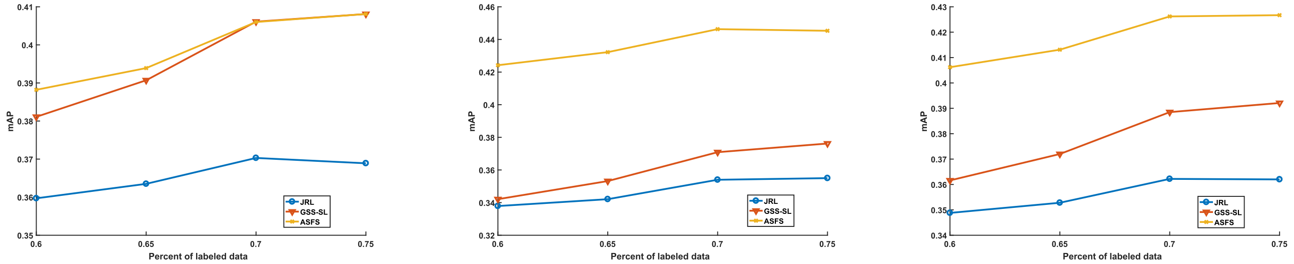
TABLE III
SUPERVISED AND SEMI-SUPERVISED mAP SORES ON PASCAL SENTENCE DATASET (NUMBERS IN BOLDFACE ARE THE BEST)

Method	mAP		
	Image Query	Text Query	Average
FS	46.96%	46.22%	46.59%
ASFS	**48.90%**	**59.61%**	**54.26%**

methods and the mAP scores are shown in Table IV. From this table, we conclude that the retrieval performance of ASFS is better than that of JRL and GSS-SL. These two methods also construct the graph model to propagate the label information from labeled data to unlabeled data and consider the semantic information. But in ASFS method, we adopt the correlation analysis term to enhance the neighborhood relationship of the pairwise data. Meanwhile, the $l_{2,1}$-norm constraint is used to reduce the effects of outliers caused by unlabeled data. Unlike these two methods, we learn different mapping matrices corresponding to different sub-tasks (I2T and T2I) to guarantee the effective semantic and structural representation of the query data in the shared subspace. Hence, our method can achieve better performance.

TABLE IV
mAP of All Compared Semi-Supervised Approaches on Three Datasets. The Best Result in Each Column Is Marked With Bold

Methods	*Wikipedia-CNN*			*Pascal Sentence*			*INRIA-Websearch*		
	I2T	T2I	Average	I2T	T2I	Average	I2T	T2I	Average
JRL [41]	37.71%	36.28%	37.00%	42.32%	47.04%	44.68%	45.06%	50.47%	47.77%
GSS-SL [36]	**41.86%**	38.31%	40.08%	43.52%	43.23%	43.38%	50.47%	53.74%	52.10%
ASFS	41.80%	**44.84%**	**43.32%**	**48.90%**	**59.61%**	**54.26%**	**54.48%**	**70.61%**	**62.55%**



(a) **Image Query Performance** (b) **Text Query Performance** (c) **Average Performance**.

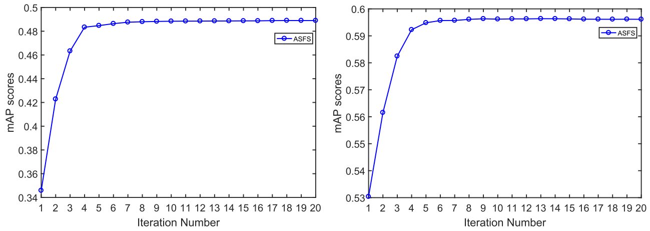
Fig. 7. Performances for varying percentages of labeled data on Wikipedia dataset.

TABLE V
mAP of All Compared Non-Linear Approaches on Three Datasets. The Best Result in Each Column Is Marked With Bold

Methods	*Wikipedia-CNN*			*Pascal Sentence*			*INRIA-Websearch*		
	I2T	T2I	Average	I2T	T2I	Average	I2T	T2I	Average
KMvMDA [32]	39.71%	39.48 %	39.56%	46.65%	47.85%	47.25%	53.27%	56.26%	54.77%
KMvNDA [33]	**42.79%**	38.83%	40.81%	47.44%	47.67%	47.56%	54.02%	56.57 %	55.30%
ASFS	41.80%	**44.84%**	**43.32%**	**48.90%**	**59.61%**	**54.26%**	**54.48%**	**70.61%**	**62.55%**

To further validate the effectiveness of our approach, we compare our AFSF with these two semi-supervised methods based on varying percentages of labeled data. In the experiment, 60%, 65%, 70% and 75% labeled data on Wikipedia dataset are used to test the performance respectively. In Fig. 7, we can see that only the GSS-SL method can achieve the same or even slightly better performance in the I2T task compared with our ASFS. But our algorithm has significantly improved in T2I task on different conditions and it also achieves the best average performance compared with these two semi-supervised methods.

3) Comparison With Non-Linear Methods: As we mentioned above, non-linear methods have been proposed for cross-modal retrieval and reach better performance than the linear methods. Therefore, we compare our semi-supervised linear method with two classical non-linear methods, Kernel-based Multi-view Nonparametric Discriminant Analysis (KMvNDA) [33] and kernel-based Multi-view Modular Discriminant Analysis (MvMDA) [32], to verify the effectiveness of our method. Specifically, we simulate the two-views KMvNDA and KMvMDA, and test them on the three datasets. Comparing two non-linear methods, we find that the KMvNDA is slightly better than KMvMDA in Table V. It is because that the KMvNDA not only considers the inter-view and intra-view covariances when learning the latent subspace but also uses the view-specific penalty graph to push apart the marginal samples from different classes. Table V shows that our method is inferior to KMvNDA in I2T retrieval task on Wikipedia dataset. However, our method has significant preponderance in T2I task on different datasets. This is because ASFS fully explores the potential information of unlabeled data to extend the diversity of training data and uses the graph-based constrain to keep the structural consistency between feature space and semantic space in the latent space. Therefore, our method is competitive with the two nonlinear methods.

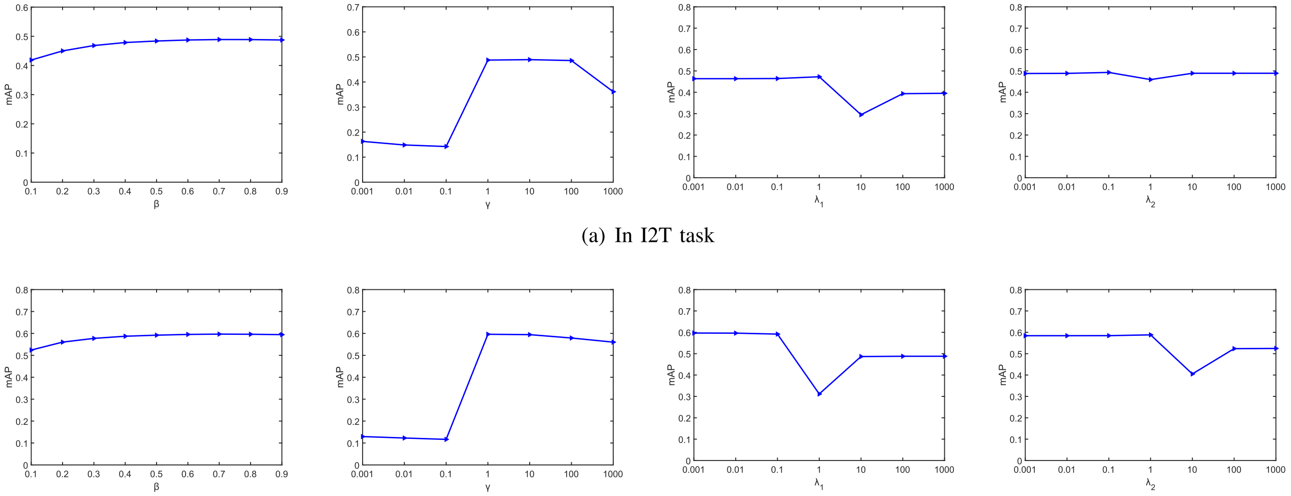


(a) Image Query (b) Text Query

Fig. 8. Convergence Curve on Pascal Sentence Dataset.

4) Convergence Analysis: In this paper, we propose an iterative strategy to optimize the objective function for different sub-tasks. Therefore, the analysis of the convergence is one of most important evaluation indexes. According to the proposed iterative algorithm, we plot the convergence curves on Pascal Sentence dataset, as shown in Fig. 8. From the convergence curves, we can find the mAP scores of our method monotonically increase with the number of iterations increasing, and tend to be stable within about five iterations on all sub-tasks. This experiment results show that the convergence of ASFS can be guaranteed.

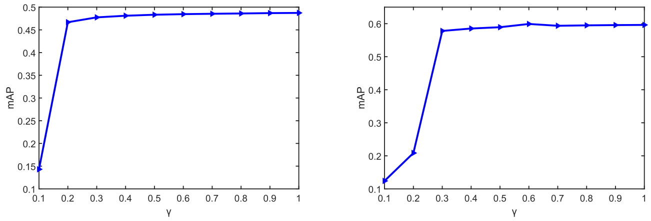
5) Parameters Sensitivity Analysis: We can see from the objective functions in Eq. (16) and Eq. (17), there are four



(a) In I2T task

(b) In T2I task

Fig. 9. Parameters tuning on the Pascal Sentence dataset.



(a) In I2T task

(b) In T2I task

Fig. 10. Complementary parameter tuning between (0.1, 1) for γ on Pascal Sentence dataset.

parameters: β, γ, λ_1 and λ_2. And we test the parameters sensitivity on Pascal Sentence dataset. For each parameter, we analyze the sensitivity by tuning its value and fixing the other parameters. From Fig. 9, we can conclude that the balance parameter β is non-sensitive under the range of $(0.1, 0.9)$ in all sub-tasks, and γ has a better performance under the range of (1,100). λ_1 and λ_2 have a stable performance in range of (0.001,1) and (0.001,1000) for I2T task, and for T2I task they have a stable performance in range of (0.001,0.1) and (0.001,1), respectively. Actually, we test more detailed values in experiments for all parameters. In this paper, we just select some representative results for illustration. For example, we also explore more detailed values for γ in range of (0.1, 1) to further analyze the influence of parameters on experimental results. From Fig. 10, we can find that the experiment results have a stable good performance in range of (0.3,1). These results validate that all the parameters in our method can have a stable performance over a wide range.

V. Conclusion

In this paper, we proposed a semi-supervised model, named Adaptive Semi-supervised Feature Selection (ASFS), for cross-modal retrieval tasks. This method aimed at exploring potential information of unlabeled data to improve the retrieval accuracy. The superiority of our method was that it combined the advantages of supervised and unsupervised methods, and the label information and pairwise relationship was considered to get the optimal mapping matrices. Specifically, our method used label graph to explore the geometric structure information between heterogeneous feature space and label space and propagate the labels from labeled data to unlabeled data. In addition, an efficient joint optimization algorithm was proposed to update the mapping matrices and the label matrix for unlabeled data simultaneously and iteratively. Meanwhile, the $l_{2,1}$-norm constraint was used for informative and discriminative feature selection to ensure the optimal mapping matrices can be learned. Experimental results on three datasets demonstrated the superiority of the proposed ASFS compared with several state-of-the-art methods.

Acknowledgment

The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the TITAN X GPU used for this research.

References

[1] Y. Wu, S. Wang, and Q. Huang, “Online asymmetric similarity learning for cross-modal retrieval,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4269–4278.

[2] D. Mandal, K. N. Chaudhury, and S. Biswas, “Generalized semantic preserving hashing for n-label cross-modal retrieval,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2633–2641.

[3] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, “Pl-ranking: A novel ranking method for cross-modal retrieval,” in *Proc. ACM Multimedia Conf.*, 2016, pp. 1355–1364.

[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] X. Mao, B. Lin, D. Cai, X. He, and J. Pei, “Parallel field alignment for cross media retrieval,” in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 897–906.

[6] 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.*, 2013, pp. 2088–2095.

[7] L. Xie, P. Pan, and Y. Lu, “A semantic model for cross-modal and multi-modal retrieval,” in *Proc. ACM Conf. Int. Conf. Multimedia Retrieval*, 2013, pp. 175–182.

[8] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 1070–1076.

[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] J. Shao, Z. Zhao, F. Su, and T. Yue, "Towards improving canonical correlation analysis for cross-modal retrieval," in *Proc. Thematic Workshops ACM Multimedia*, 2017, pp. 332–339.

[11] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4594–4602.

[12] F. Wu, H. Zhang, and Y. Zhuang, "Learning semantic correlations for cross-media retrieval," in *Proc. IEEE Int. Conf. Image Process.*, 2007, pp. 1465–1468.

[13] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, Dec. 2013.

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

[15] J. Wang, S. Kumar, and S. F. Chang, "Semi-supervised hashing for large-scale search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2393–2406, Dec. 2012.

[16] X. Chang, Y. L. Yu, Y. Yang, and E. P. Xing, "Semantic pooling for complex event analysis in untrimmed videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1617–1632, Aug. 2017.

[17] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2407–2414.

[18] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Comput.*, vol. 12, no. 6, pp. 1247–1283, 2014.

[19] J. He, B. Ma, S. Wang, Y. Liu, and Q. Huang, "Cross-modal retrieval by real label partial least squares," in *Proc. ACM Multimedia Conf.*, 2016, pp. 227–231.

[20] A. Sharma, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2160–2167.

[21] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 604–611.

[22] Z. Yu *et al.*, "Discriminative coupled dictionary hashing for fast cross-media retrieval," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 395–404.

[23] D. Zhang and W. J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2177–2183.

[24] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3864–3872.

[25] Y. Cao, M. Long, J. Wang, and H. Zhu, "Correlation autoencoder hashing for supervised cross-modal search," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 197–204.

[26] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint 2,1-norms minimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.

[27] D. R. Hardoon, S. Szedmak, and J. Shawetaylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[28] H. Zhang, Y. Liu, and Z. Ma, "Fusing inherent and external knowledge with nonlinear learning for cross-media retrieval," *Neurocomputing*, vol. 119, no. 16, pp. 10–16, 2013.

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

[30] H. Zhang, L. Cao, and S. Gao, "A locality correlation preserving support vector machine," *Pattern Recognit.*, vol. 47, no. 9, pp. 3168–3178, 2014.

[31] 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.

[32] G. Cao, A. Iosifidis, K. Chen, and M. Gabbouj, "Generalized multi-view embedding for visual recognition and cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2542–2555, Sep. 2018.

[33] G. Cao, A. Iosifidis, and M. Gabbouj, "Multi-view nonparametric discriminant analysis for image retrieval and recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 10, pp. 1537–1541, Oct. 2017.

[34] X. Chang and Y. Yang, "Semisupervised feature analysis by mining correlations among multiple tasks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2294–2305, Oct. 2017.

[35] M. Luo *et al.*, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 648–660, Feb. 2018.

[36] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 128–141, Jan. 2018.

[37] D. Zhou, O. Bousquet, T. N. Lal, and J. Weston, "Learning with local and global consistency," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2003, pp. 321–328.

[38] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.

[39] T. N. Tan, "L21 regularized correntropy for robust feature selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2504–2511.

[40] 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.

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

[42] Y. Wei *et al.*, "Modality-dependent cross-media retrieval," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, 2016, Art. no. 57.

[43] J. Wu, Z. Lin, and H. Zha, "Joint latent subspace learning and regression for cross-modal retrieval," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2017, pp. 917–920.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

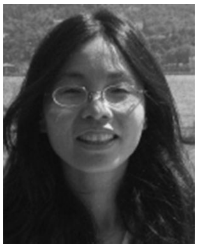
[45] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

En Yu received the bachelor's degree in information science and engineering in 2016 from Shandong Normal University, Jinan, China, where he is currently working toward the master's degree in communication and information systems. His research interests include multimedia processing and analysis, machine learning, and deep learning. Mr. Yu is a student member of the CCF.

Jiande Sun received the Ph.D. degree in communication and information systems from Shandong University, Jinan, China, in 2000 and 2005, respectively. From September 2008 to August 2009, he was a Visiting Researcher with the Institute of Telecommunications System, Technical University of Berlin, Berlin, Germany. From October 2010 to December 2012, he was a Postdoctoral Researcher with the Institute of Digital Media, Peking University, Beijing, China, and with the State Key Laboratory of Digital-Media Technology, Hisense Group. From July 2014 to August 2015, he was a DAAD Visiting Researcher with the Technical University of Berlin and the University of Konstanz, Germany. From October 2015 to November 2016, he was a Visiting Researcher with the Language Technology Institute, School of Computer Science, Carnegie Mellon University, USA. He is currently a Professor with the School of Information Science and Engineering, Shandong Normal University, Jinan. He has authored or coauthored more than 60 journal and conference papers. He has coauthored two books. His current research interests include multimedia content analysis, video hashing, gaze tracking, image/video watermarking, 2-D-to-3-D conversion, and so on.

Jing Li is working toward the Ph.D. degree with Shandong Normal University, Jinan, China. She is also a Lecturer with the School of Mechanical and Electrical Engineering, Shandong Management University, Jinan. Her research interests include machine learning, multimedia processing, and retrieval.

Xiaojun Chang received the Ph.D. degree in artificial intelligence from the University of Technology Sydney, Ultimo NSW, Australia, in 2016. He is currently a Lecturer (a.k.a. Assistant Professor) with the Faculty of Information Technology, Monash University, Clayton, VIC, Australia. He is also an Adjunct Professor with the School of Information Science and Engineering, Shandong Normal University. Before that, he was a Postdoctoral Research Associate with the School of Computer Science, Carnegie Mellon University, in August 2016. He has served as an Associate Editor or a PC member for several prestigious journals and conferences in related fields. His research interests include machine learning and its applications to computer vision and multimedia.



Xian-Hua Han (M'11) received the B.E. degree from Chongqing University, Chongqing, China, the M.E. degree from Shandong University, Jinan, China, and the D.E. degree in 2005 from the University of the Ryukyus, Okinawa, Japan. From April 2007 to March 2013, she was a Postdoctoral Fellow and an Associate Professor with the College of Information Science and Engineering, Ritsumeikan University, Japan. From April 2016 to February 2017, she was a Senior Researcher with the Artificial Intelligence Researcher Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan. She is currently an Associate Professor with the Graduate School of Science and Technology for Innovation, Yamaguchi University, Yamaguchi, Japan. Her current research interests include image processing and analysis, feature extraction, machine learning, computer vision, and pattern recognition. Prof. Han is a member of the Institute of Electronics, Information and Communication Engineers.

Alexander G. Hauptmann received the B.A. and M.A. degrees in psychology from Johns Hopkins University, Baltimore, MD, USA, the degree in computer science from the Technische Universität Berlin, Berlin, Germany, in 1984, and the Ph.D. degree in computer science from Carnegie Mellon University (CMU), Pittsburgh, PA, USA, in 1991. He is currently a Professor with the Department of Computer Science and the Language Technologies Institute, CMU. From 1984 to 1994, he worked on speech and machine translation, when he joined the Informedia project for digital video analysis and retrieval, and led the development and evaluation of news-on-demand applications. His research interests include several different areas: man–machine communication, natural language processing, speech understanding and synthesis, video analysis, and machine learning.