Neurocomputing 461 (2021) 404–418

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Learning sufficient scene representation for unsupervised cross-modal retrieval

Jieting Luo, Yan Wo *, Bicheng Wu, Guoqiang Han

[a] School of Computer Science and Engineering, South China University of Technology, Canton, China

ARTICLE INFO

Article history:
Received 24 January 2021
Revised 8 May 2021
Accepted 28 July 2021
Available online 30 July 2021
Communicated by Zidong Wang

Keywords:
Unsupervised cross-modal retrieval
Common representation
Statistical manifold
Gaussian Mixture Model
Geodesic distance

ABSTRACT

In this paper, a novel unsupervised **C**ross-**M**odal retrieval method via **S**ufficient **S**cene **R**epresentation (CMSSR) is proposed. Distinguished from the existing methods which mainly focus on simultaneously preserving the mutually-constrained intra- and inter-modal similarity relation, CMSSR considers data of different modalities as the descriptions of a scene from different views and accordingly integrates information of different modalities to learn a complete common representation containing sufficient information of the corresponding scene. To obtain such common representation, Gaussian Mixture Model (GMM) is firstly utilized to generate statistic representation of each uni-modal data, while the uni-modal spaces are accordingly abstracted as uni-modal statistical manifolds. In addition, the common space is assumed to be a high-dimensional statistical manifold with different uni-modal statistical manifolds as its sub-manifolds. In order to generate sufficient scene representation from uni-modal data, a representation completion strategy based on logistic regression is proposed to effectively complete the missing representation of another modality. Then, the similarity between different multi-modal data can be more accurately reflected by the distance metric in common statistical manifold. Based on the distance metric in common statistical manifold, Iterative Quantization is utilized to further generate binary code for fast cross-modal retrieval. Extensive experiments on three standard benchmark datasets fully demonstrate the superiority of CMSSR compared with several state-of-the-art methods.

1. Introduction

With the rapid growth of multimedia data including text, image, audio, and video, cross-modal retrieval task, which takes query data from one modality and retrieve relevant data from any other modality, has attracted a great deal of research attention in recent years. In cross-modal retrieval, the most challenging problem is how to narrow the 'heterogeneity gap' which means the representation forms of different modalities are inconsistent, and thus the cross-modal similarity cannot be directly calculated. For the sake of bridging the 'heterogeneity gap', the mainstream solution is learning a common space where the cross-modal similarity between different multi-modal data can be directly calculated for retrieval. Based on this idea, many methods have been proposed in both supervised [1–7] and unsupervised [8–28] learning. As the unsupervised method does not require any label information, it is suitable for the retrieval problems with limited or scarce label information. Furthermore, considering the basic idea of unsupervised cross-modal retrieval can be applied to any combination of different modalities while visual and textual contents are two major forms of data in our world, we focus on the unsupervised image-text bi-directional retrieval in this work. Unsupervised methods utilize the one-to-one correspondence provided by the multi-modal dataset to learn common representation which is complete and can fully represent each multi-modal data. Generally, researchers attempt to generate such an ideal common representation through preserving the similarity information of each multi-modal data in the common space. In addition to the inter-modal similarity provided by the one-to-one correspondence, most of the methods also construct intra-modal similarity information such as neighbor information [8,9,16], cluster information [17,18] or graph information [12,22,23] in the uni-modal space. However, Huang et al. [25] analyze the process of cross-modal learning from the perspective of statistic. With the help of variational inference, they proved that maximizing the intra- and inter-modal similarity are mutually constrained. It is impossible to simultaneously maximize intra-modal similarity and inter-modal similarity. This indicates that preserving inter-modal similarity will destroy the intra-modal similarity relation, vice versa. Therefore, the constant

* Corresponding author.
E-mail address: woyan@scut.edu.cn (Y. Wo).

https://doi.org/10.1016/j.neucom.2021.07.078
0925-2312/

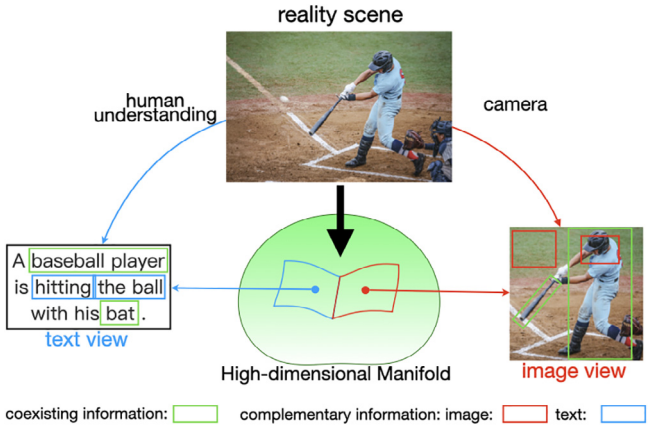


Fig. 1. The illustration showing the relation between data from different modalities.

competition between this two kind of similarities will make data sample in chaotic distribution after being projected into common space, which affects the cross-modal retrieval performance.

In consideration of the above problems, we rethink the relation between different modalities. Essentially, corresponding image and text are the descriptions of the same scene from different views. For any reality scene, an image is taken by camera to depict it, and a text is written down to describe it after understanding this scene by human being. From the image view, information of the scene is represented through corner and texture constructed from pixel points, which has higher fidelity but lacks the process of semantic abstraction. As for the text view, much semantically independent detail of the scene is filtered by human being, but some semantic information is also added after the understanding of human being. Therefore, a scene is described by image and text data in a complementary manner. And there should be co-existed information and complementary information between different views of the scene. For example, as is shown in Fig. 1, the baseball player and his bat, mentioned in the text description and displayed in the image, describe the co-existed information, while textual concept of "hitting" and the visual information of black helmet are complementary information. Cross-modal retrieval aims to retrieve and return data of another modal which describes similar scene with the query data. To accurately describe the scene, common representation should contain sufficient complementary information and co-existed information. Complementary information is exclusive information of each modal while co-existed information are shared information between different modal. Most of the previous methods attempt to preserve both co-existed and complementary information into their common representation through maximizing the intra- and inter-modal similarity simultaneously. However, maximizing inter-modal similarity will inhibit the complementary information while maximizing intra-modal similarity will intensify the complementary information. It is precisely because of this two mutual constraint process which makes these methods hard to generate a common representation containing complete information. In light of the above observation and analysis, a new unsupervised cross-modal retrieval method via sufficient scene representation (CMSSR) is proposed in this paper. In CMSSR, each scene is modeled as a consistency concept for aligning multi-modal data and abstracted as a data point lying in a high-dimensional manifold, which is shown in Fig. 1. And the representation of each scene contains information from both the image and text view. Accordingly, those uni-modal manifolds can be seen as the sub-manifolds of this high-dimensional manifold. Unlike those previous methods which mainly focus on the similarity relation in the common space, CMSSR learns a sufficient scene representation for each multi-modal data and measure their similarity on the scene level.

In order to effectively extract information from raw image and text data and embed them into the same semantic level, CMSSR firstly utilizes Gaussian Mixture Model (GMM) to generate statistic representation for them. Let $\mathbb{G}^I$ and $\mathbb{G}^T$ denote the Gaussian density function set estimated from the training image and text set, respectively. Uni-modal data manifold spanned by $\mathbb{G}^I$ and $\mathbb{G}^T$ are defined as image statistical manifold $\mathscr{M}^I$ and text statistical manifold $\mathscr{M}^T$, respectively. Accordingly, statistical manifold spanned by the concatenation of $\mathbb{G}^I$ and $\mathbb{G}^T$ is defined as common statistical manifold $\mathscr{M}^C$. In this way, $\mathscr{M}^I$ and $\mathscr{M}^T$ are sub-manifolds of $\mathscr{M}^C$. And each data point in uni-modal statistical manifold ($\mathbb{G}^I$ or $\mathbb{G}^T$) represents a uni-modal data (image or text) while each data point in $\mathscr{M}^C$ represents a scene containing sufficient information from both the image and text view. Fig. 2 shows the flowchart of the proposed CMSSR. Given any multi-modal data (image or text), CMSSR firstly embeds it into its corresponding uni-modal data manifold ($\mathscr{M}^I$ or $\mathscr{M}^T$). In order to generate a sufficient scene representation for the given uni-modal data, it is further projected from its corresponding uni-modal statistical manifold into statistical manifold of another modal to obtain the information estimation and representation in another modality. Then, through concatenating this two representations, CMSSR can generates a sufficient scene representation in the common statistical manifold $\mathscr{M}^C$ in

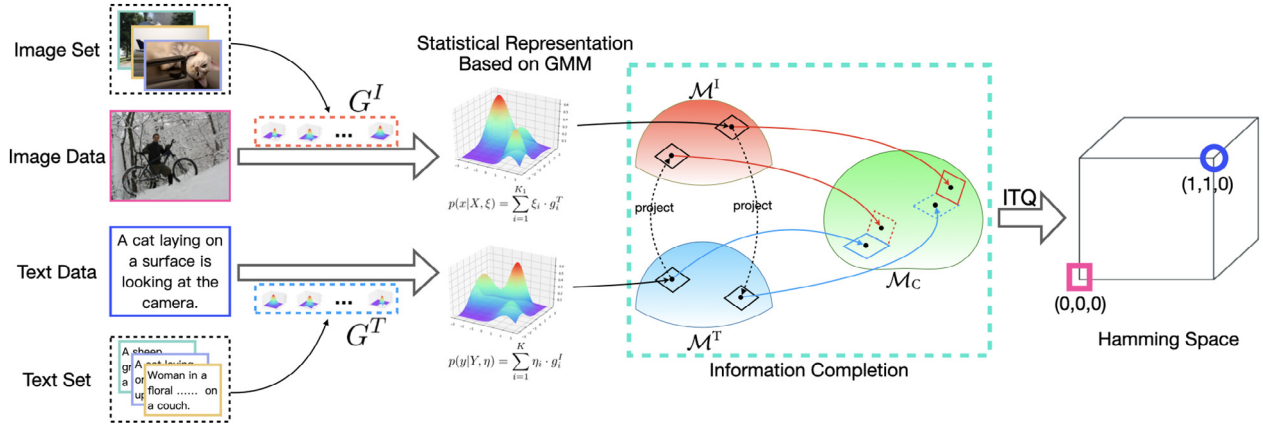


Fig. 2. The flowchart of the proposed unsupervised cross-modal retrieval method via sufficient scene representation. Image statistical manifold, text statistical manifold and common statistical manifold are denoted as $\mathscr{M}^I$, $\mathscr{M}^T$ and $\mathscr{M}^C$, respectively.

which geodesic distance can be calculated to more accurately measure the similarity between different data. To perform efficient cross-modal retrieval, CMSSR utilizes the distance metric in common statistical manifold and implements Iterative Quantization (ITQ) [29] to generate binary code for multi-modal data. Compared with the existing works, the main contributions of this paper are listed as follows:

- This paper proposes to perform cross-modal retrieval between image and text modality from the perspective of scene. Through modeling common space as a high-dimensional manifold with image and text space as its two sub-manifolds, we abstract SSR as the common representation for aligning multi-modal data without the consideration of simultaneously preserving inter- and intra-modal similarity. The sufficient scene representation can be generated from uni-modal data via representation completion strategy, which can estimate the missing scene information and representation in another modality. And the generated scene representation contains more complete information, which helps to enhance the performance of cross-modal retrieval.
- We give the definition of uni-modal statistical manifolds and common statistical manifold. Accordingly, CMSSR generates statistic representation for each image and text data based on GMM and studies unsupervised cross-modal retrieval in the statistical manifold. Unlike traditional linear metric space (such as Euclidean space), statistical manifold can model more complex non-linear relation between data of different modalities and perform more accurate similarity measurement through calculating the geodesic distance. And based on the distance metric in the common statistical manifold, binary code is further generated for each multi-modal data to perform more effective and efficient cross-modal retrieval.
- Through a series of experiments conducted on three widely-used multi-modal datasets, the superior performance of CMSSR in cross-modal retrieval task has been sufficiently demonstrated compared with several state-of-the-art methods.

The rest of this paper is organized as follows. In Section 2, some unsupervised cross-modal representation learning methods will be reviewed. And the details of the CMSSR will be introduced in Section 3. After that, extensive experiment results will be presented and discussed in Section 4. Finally, we briefly summarize our work in Section 5.

2. Related work

The primary target of unsupervised cross-modal retrieval method is to learn a common representation, so that the heterogeneity gap between different modalities can be effectively narrowed. According to different assumptions about common space, common representation can be categorized into coordinated representation and joint representation [30].

Coordinated representation learns separated representation for each modality through the similarity constraint. Through maximizing the pairwise correlation between data from different modalities, canonical correlation analysis (CCA) [31] can separately learn the coordinated representation for data of different modalities. Thus, the early applications of coordinated representation in unsupervised cross-modal retrieval are CCA-based methods [26–28]. After that, a great deal of methods based on similarity preservation have also been proposed. Among them, cross-modal hashing method has wider application scenarios because of its retrieval efficiency and storage saving. One of the most impressive hashing methods is Inter-Media Hashing (IMH) [8] which adopts spectral hashing and proposes to simultaneously preserve the inter-modal and intra-modal consistency of the multi-modal data for cross-modal retrieval. Besides, Binary Set Embedding (BSE) [9] learns binary code directly from the local feature set and generates more semantically robust coordinated representation for multi-modal data. Z. Ye et al. [10] propose the multi-scale feature guided sequential hashing learning method as well as the multi-scale correlation mining strategy to exploit the correlations among the multi-scale features of multi-modal data and learn finer cross-modal hash code. Multi-pathway Generative Adversarial Hashing (MGAH) [11] utilizes the generative adversarial network to exploit the underlying manifold structure of cross-modal data so that the meaningful nearest neighbors of different modalities can be effectively captured, and more robust common representation can be generated. What's more, considering the inaccurate similarity problem caused by the insufficient description of the cross-modal feature to the complex data relationship, deep graph-neighbor coherence preserving network (DGCPN) [12] adopts the graph-neighbor coherence (GC) to explore the data intrinsic in a graph and effectively addresses the above problem.

Joint representation assumes that all modalities of one instance have the same semantic and accordingly combines uni-modal representation into a unified representation. The most representative application of joint representation is matrix factorization method which finds consistent representation for different modalities by matrix factorization. As a typical matrix factorization method, Latent Semantic Sparse Hashing (LSSH) [13] firstly jointly learns the latent features from images and texts with sparse coding, and then achieves the unified representation using matrix factorization. In addition, to more effectively construct strong connection between different modalities, Collective Matrix Factorization Hashing (CMFH) [14] learns consistent hash codes for multi-modal data in the shared latent semantic space by collective matrix factorization. Multi-modal graph regularized Smooth matrix Factorization Hashing (MSFH) [15] reconstructs the intra- and inter-modal similarity graph by symmetric nonnegative matrix factorization and accordingly preserves them into the common space. Similar with coordinated representation, joint representation can also be generated by similarity preservation methods. Deep Semantic-Alignment cross-modal Hashing (DSAH) [16] adopts a semantic alignment loss to fully exploit the co-occurrence information in different modalities and designs a feature reconstruction procedure to bridge the modality gap. Inspired by the idea of re-weighted discriminatively embedded K-means, Cluster-wise Unsupervised Hashing (CUH) [17] designs a multi-view clustering method which can simultaneously find the common cluster information and learn the unified compact hash code of multi-modal instance. Considering the generative nature of probabilistic graphical models, common representation can be also generated through building probabilistic graphical models. In [19], the authors adopt two separated deep Boltzmann machine (DBM) [20] to model the distribution over the features of different modalities, and the two models are combined by an additional layer on the top of them as the joint representation layer, which can learn the representation by computing joint distribution. Considering the high computational cost of probabilistic graphical model, Deep Binary Reconstruction (DBRC) [21] constructs two branches of auto-encoder linked with only one layer DBM, and designs a scalable tanh activation framework to avoid the tiny gradient problem when train their model end-to-end. Distinguished from probabilistic graphical models, some methods also construct relation graph to model the relationship between different instance. Fusion Similarity Hashing (FSH) [22] applies Nearest Neighbor Similarity to construct the Fusion Anchor Graph from text and image modalities to help learning binary codes. Deep Cross-modality Spectral Hashing (DCSH) [23] designs an anchor-to-anchor mapping to build a

more informative Laplacian graph model which helps generate more robust unified hash codes.

Although most of the above-mentioned methods can perform effective cross-modal retrieval, exploiting too much similarity relation when learning common representation may weaken the modality-specific semantics and introduce unexpected noise, which affects their performance.

3. Methodology

In this section, we delve into the specific implementation of CMSSR which is shown in Fig. 2. Firstly, the process of learning statistic representation of uni-modal data will be introduced. Then, we accordingly give the definition of uni-modal statistical manifold and common statistical manifold. Lastly, the representation completion strategy is proposed to learn common representation for the uni-modal query data.

3.1. Learning statistic representation of uni-modal data

As stated in the introduction, corresponding text and image describe the same scene from different views and have different manifestations. Specifically, in the computer, text data is the combination of words, which expresses the semantic information of the scene, while image data is composed of pixel values,which records the visual information of scene. Therefore, we can generally extract high-level semantic feature from text, but only low-level visual feature can be extracted from image. And thus, there exists a semantic gap between text and image data. To narrow the semantic gap between them, in this part, we adopt statistic methods to further abstract the low-level image feature and blur the precise semantics of the text. In this way, image and text can be jointly embedded from different semantic level into a unified statistic level.

As an impressive statistic method, Gaussian Mixture Model can effectively model the distribution of data by the weighted mixture of a number of Gaussian density functions. Due to the powerful modeling capacity and interpretability, it has been successfully applied in a broad spectrum of applications [32]. Therefore, in this paper, we adopt GMM to learn statistic representation for text and image data.

3.1.1. Text representation based on gaussian mixture model

Gaussian Mixture Model (GMM) is a probability model that assumes all the data points are generated from a mixture of a finite number of Gaussian density functions with unknown parameters. And each Gaussian density function can also be seen as a basic function for representing data. But it is difficult to directly represent text data as GMM because of its complexity. Considering text data is the combination of words, this process can be started from the perspective of word. Suppose all the words $W = \{x_i\}_{i=1}^{N_W}$ are lying in the same word space, the distribution of these words can be modeled based on GMM:

$$p(x|\boldsymbol{\pi}^T) = \sum_{i=1}^{K_T} \pi_i^T g_i^T \tag{1}$$

where $g_i^T = N(x|\mu_i^T, \sigma_i^T)$ gives the i^{th} Gaussian density function with mean μ_i^T and co-variance σ_i^T, and π_i^T gives the weight of the i^{th} Gaussian density function. In (1), the distribution of words is represented by K_T Gaussian density functions which make up a function set denoted as $\mathbb{G}^T = \left\{g_1^T, g_2^T, \ldots, g_{K_T}^T\right\}$. Since each text data can be seen as a subspace of word space [33], given text data X, we use the same set of Gaussian density function $\mathbb{G}^T$ to represent it as:

$$p(x|X, \boldsymbol{\xi}) = \sum_{i=1}^{K_T} \xi_i g_i^T \tag{2}$$

where ξ_i gives the weight of the i^{th} Gaussian density function for representing X. Since the semantic of each text is determined by the combination of words, the weight of g_i^T should be the sum of weights corresponding to each word in the text data. What's more, to ensure the condition of $\sum_{i=1}^{K_T} \xi_i = 1$, we calculate the weight coefficient of i^{th} Gaussian density function for text data as:

$$\xi_i = \frac{\sum_{x \in X} \pi_i^T g_i^T}{\sum_{j=1}^{K_T} \sum_{x \in X} \pi_j^T g_j^T}, i = 1, 2, \ldots, K_T \tag{3}$$

To obtain the statistic representation of text data, the statistic variable x should be a computational form of word. Generally, x can be generated from the one-hot encoding or other word embedding methods, such as word2vec [34] and glove [35]. Once the distribution of all the words in the word space is modeled as (1), each Gaussian density function in $\mathbb{G}^T$ can be accordingly seen as an abstract concept. And text data represented as (2) can be seen as a weighted mixture of abstract concepts. Thus, the precise semantic expressed by text is blurred after being represented through GMM.

3.1.2. Image representation based on Gaussian Mixture Model

Although visual feature extracted from image data only contains corner and texture information, we can similarly represent each image based on GMM. Given a set of image data, we firstly obtain multiple local features from this image set. After that, through implementing k-means or other cluster methods on the local feature set, we can obtain k different cluster centers as k visual words. Suppose we have obtained a set of visual words $V = \{y_i\}_{i=1}^{N_V}$, we can model the distribution of these visual words with GMM:

$$p(y|\boldsymbol{\pi}^I) = \sum_{i=1}^{K_I} \pi_i^I g_i^I \tag{4}$$

where $g_i^I = N(x|\mu_i^I, \sigma_i^I)$ gives the i^{th} Gaussian density function of image modal with mean μ_i^I and co-variance σ_i^I, and π_i^I gives the weight of the i^{th} Gaussian density function. In this way, the distribution of these visual words is modeled by K_I Gaussian density functions which make up a function set denoted as $\mathbb{G}^I = \left\{g_1^I, g_2^I, \ldots, g_{K_I}^I\right\}$. Similarly, given image data Y, it can be accordingly represented through GMM:

$$p(y|Y, \boldsymbol{\eta}) = \sum_{i=1}^{K_I} \eta_i g_i^I \tag{5}$$

where η_i gives the weighted coefficient of the i^{th} Gaussian density function. And it is similar with ξ_i that η_i can be calculated as:

$$\eta_i = \frac{\sum_{y \in Y} \pi_i^I g_i^I}{\sum_{j=1}^{K_I} \sum_{y \in Y} \pi_j^I g_j^I}, i = 1, 2, \ldots, K_I \tag{6}$$

In this way, each image data can be also represented by GMM with the Gaussian density function set $\mathbb{G}^I$. Since each Gaussian density function is regarded as an abstract concept, this is a process of further abstraction for low-level visual feature.

3.1.3. Estimating basic function for modeling uni-modal data

In the previous part, we represent uni-modal data by GMM and embed them from different semantic level into a unified statistic level, which narrows the semantic gap between them. However, we need to further seek the appropriate $\mathbb{G}^T$ and $\mathbb{G}^I$, so that text

and image data can be sufficiently represented by GMM. And this process can be done by using the EM algorithm [36] on the corresponding data set to estimate the parameters of GMM. The detail of this estimation process is shown in **Algorithm 1**.

Through the EM algorithm given as **Algorithm 1**, we obtained $\mathbb{G}^T$ and $\mathbb{G}^I$ for sufficiently representing text and image data, respectively. What's more, through representing uni-modal data as GMM, the important information can be highlighted while the value of irrelevant information or noise will be exponentially shrunk by the Gaussian density function.

3.2. Uni-modal statistical manifolds and common statistical manifolds

In the previous part, we learned statistic representation for image and text data based on GMM. Since each GMM is a probability model which can be regarded as a point lying in statistical manifold. To better study the following cross-modal problem between image and text model, we accordingly give a definition of uni-modal statistical manifold based on the statistical representation in image and text domain. After that, the distance metric in statistical manifold will also be discussed so that the similarity between different data lying in statistical manifold can be measured.

Algorithm 1 Parameter Estimation for $\mathbb{G}^T$ and $\mathbb{G}^I$

Input: Word embedding set $W = \{x_i\}_{i=1}^{N_W}$ and visual words $V = \{y_i\}_{i=1}^{N_V}$ of the training data.

Output: Parameters $\pi_i^T, \mu_i^T, \sigma_i^T, i = 1, 2, \ldots, K_T$ of Gaussian density function set $\mathbb{G}^T$, and $\pi_i^I, \mu_i^I, \sigma_i^I, i = 1, 2, \ldots, K_I$ of $\mathbb{G}^I$.

1: **Initialization**: Implement k-means on W and V. And initialize the parameters $\pi_i^T, \mu_i^T, \sigma_i^T, i = 1, 2, \ldots, K_T$ and $\pi_i^I, \mu_i^I, \sigma_i^I, i = 1, 2, \ldots, K_I$ according to the cluster information.

2: **repeat**

3: **E-Step**: Utilize the current parameter values to evaluate the responsibilities γ_{ji}^T that the i^{th} Gaussian component of text domain takes for representing the j^{th} word vector x_j, and γ_{ji}^I for image domain:

$$\gamma_{ji}^T = \frac{\pi_i^T N(x_j|\mu_i^T, \sigma_i^T)}{\sum_{i=1}^{K_T} \pi_i^T N(x_j|\mu_i^T, \sigma_i^T)}, \quad \gamma_{ji}^I = \frac{\pi_i^I N(x_j|\mu_i^I, \sigma_i^I)}{\sum_{i=1}^{K_I} \pi_i^I N(x_j|\mu_i^I, \sigma_i^I)}$$

4: **M-Step**: Re-estimate the parameters of text domain and image domain using the current responsibilities γ_{ji}^T and γ_{ji}^I:

$$\pi_i^T = \frac{\sum_{j=1}^{N_W} \gamma_{ji}^T}{N_W}, \ \mu_i^T = \frac{\sum_{j=1}^{N_W} \gamma_{ji}^T x_j}{\sum_{j=1}^{N_W} \gamma_{ji}^T}, \ \sigma_i^T = \frac{\sum_{j=1}^{N_W} \gamma_{ji}^T (x_j - \mu_i^T)(x_j - \mu_i^T)^\top}{\sum_{j=1}^{N_W} \gamma_{ji}^T};$$

$$\pi_i^I = \frac{\sum_{j=1}^{M} \gamma_{ji}^I}{N_V}, \ \mu_i^I = \frac{\sum_{j=1}^{N_V} \gamma_{ji}^I y_j}{\sum_{j=1}^{N_V} \gamma_{ji}^I}, \ \sigma_i^I = \frac{\sum_{j=1}^{N_V} \gamma_{ji}^I (y_j - \mu_i^I)(y_j - \mu_i^I)^\top}{\sum_{j=1}^{N_V} \gamma_{ji}^I}$$

5: **until** the parameters of $\mathbb{G}^I$ and $\mathbb{G}^T$ become stable.

3.2.1. Definition of uni-modal statistical manifolds and common statistical manifolds

As introduced in [37], GMM belongs to the mixture family. And the statistical manifold of the mixture family can be defined as **Definition 1**.

Definition 1. Statistical Manifold of Mixture Family. Given n linearly independent probability density functions $g_1, g_2, \ldots, g_n$, they can compose a family of probability distribution represented as $p(u|\boldsymbol{\theta}) = \sum_{i=1}^n \theta_i g_i$ where $\sum_{i=1}^n \theta_i = 1, \theta_i > 0$. This is a statistical model called a mixture family. And all the probability distributions in this mixture family are lying in a same statistical manifold $\mathscr{M}$ which is spanned by $g_1, g_2, \ldots, g_n$, and its coordinate system can be accordingly denoted as $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$. And there exists a one-to-one mapping between $\boldsymbol{\theta}$ and $p(u|\boldsymbol{\theta})$.

According to 3.1.1, text data can be represented as GMM by a set of linearly independent Gaussian density function $\mathbb{G}^T = \left\{g_1^T, g_2^T, \ldots, g_{K_T}^T\right\}$. Therefore, according to **Definition 1**, after representing text data as $p(x|X, \boldsymbol{\xi}) = \sum_{i=1}^{K_T} \xi_i g_i^T$, text data can be embedded into text statistical manifold $\mathscr{M}^T$ spanned by $\mathbb{G}^T$. And the coordinate system of can be denoted as $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_{K_T})$. Also, there exists a one-to-one mapping between ξ and $p(x|X, \boldsymbol{\xi})$. In this way, different text data corresponds to data point with different coordinate in $\mathscr{M}^T$. And we denote text data $p(x|X, \boldsymbol{\xi})$ as $\boldsymbol{\xi}$ for simplification.

Similarly, according to 3.1.2, image data can also be represented as GMM by a set of linearly independent Gaussian density function $\mathbb{G}^I = \left\{g_1^I, g_2^I, \ldots, g_{K_I}^I\right\}$. Therefore, image data represented as $p(y|Y, \boldsymbol{\eta}) = \sum_{i=1}^{K_I} \eta_i g_i^I$ are lying in image statistical manifold $\mathscr{M}^I$ spanned by $\mathbb{G}^I$. And the coordinate system of $\mathscr{M}^I$ can be denoted as $\boldsymbol{\eta} = \left\{\eta_1, \eta_2, \ldots, \eta_{K_I}\right\}$. Also, there exists a one-to-one mapping between $\boldsymbol{\eta}$ and $p(y|Y, \boldsymbol{\eta})$. In this way, different image data corresponds to data point with different coordinate in $\mathscr{M}^I$. And we similarly denote image data $p(y|Y, \boldsymbol{\eta})$ as $\boldsymbol{\eta}$ for simplification.

For a specific scene, its corresponding text data X and image data Y have been represented as $\boldsymbol{\xi} = p(x|X, \boldsymbol{\xi}) = \sum_{i=1}^{K_T} \xi_i g_i^T$ and $\boldsymbol{\eta} = p(x|X, \boldsymbol{\eta}) = \sum_{i=1}^{K_I} \eta_i g_i^I$ in uni-modal statistical manifolds. Since text and image data describe the scene from two different views, we can fuse $\mathscr{M}^T$ and $\mathscr{M}^I$ to reconstruct common statistical manifold for the scene. Considering $\mathbb{G}^T$ and $\mathbb{G}^I$ are linearly independent, we reconstruct common statistical manifold by concatenating $\mathbb{G}^T$ and $\mathbb{G}^I$, and accordingly give the definition of common statistical manifold as **Definition 2**.

Definition 2. Common Statistical Manifold. Given basic function set $\mathbb{G}^T = \left\{g_1^T, g_2^T, \ldots, g_{K_T}^T\right\}$ of text modality and $\mathbb{G}^I = \left\{g_1^I, g_2^I, \ldots, g_{K_I}^I\right\}$ of image modality, common statistical manifold $\mathscr{M}^C$ is spanned by $\mathbb{G}^C = \{\mathbb{G}^T, \mathbb{G}^I\} = \{g_1^C, g_2^C, \ldots, g_K^C\}$ where $K = K_T + K_I$. And the corresponding coordinate system of $\mathscr{M}^C$ can be denoted as $\boldsymbol{\psi} = (\boldsymbol{\xi}, \boldsymbol{\eta}) = (\psi_1, \psi_2, \ldots, \psi_K)$. And text data can be specified in $\mathscr{M}^C$ when $\boldsymbol{\eta} = \mathbf{0}$ while image data can be also specified when $\boldsymbol{\xi} = \mathbf{0}$.

Based on **Definition 2**, each scene corresponding to a point in $\mathscr{M}^C$ should be accordingly represented as:

$$p(z|\boldsymbol{\psi}) = \sum_{i=1}^{K} \psi_i g_i^C \tag{7}$$

Since $\mathbb{G}^T$ and $\mathbb{G}^I$ contain important information of text and image modality respectively, each data point in $\mathscr{M}^C$ can sufficiently represent information of the corresponding scene.

3.2.2. distance metric in statistical manifold of mixture family

Previously, we gave the definition of both uni-modal statistical manifolds and common statistical manifold. To further perform cross-modal learning and retrieval in the statistical manifold, an effective distance metric is needed to measure the distance between different instance lying in the same statistical manifold. Therefore, in this section, we take $\mathscr{M}$ (defined in **Definition 1**) as an example to discuss the distance metric in statistical manifold

of mixture family. Given any two data points $\boldsymbol{p} = p(w|\theta^p)$ and $\boldsymbol{q} = p(w|\theta^q)$ lying in $\mathscr{M}$, their dissimilarity is usually measured by calculating their f-divergence:

$$D_f[\boldsymbol{p}:\boldsymbol{q}] = \sum \theta_i^p f\left(\frac{\theta_i^q}{\theta_i^p}\right) \tag{8}$$

where $f(.)$ is a convex function satisfying $f(1) = 0$. And when f is given as different certain functions, f-divergence can be accordingly derived as some commonly-used divergence, like KL-divergence and JS-divergence [37]. However, the calculation of these divergences requires a lot of sampling. According to [38], only after doing more than 500 sampling can KL-divergence be approximated to the true value of geodesic distance. It is too time-consuming if doing sampling for every similarity measurement. To conquer this problem, we consider each data in the statistical manifold as a function and directly calculate the functional distance between two data points. And the 2-norm functional distance between $\boldsymbol{p}$ and $\boldsymbol{q}$ is written as:

$$\begin{aligned} D_{finc}(p,q) &= \sqrt{\int \|\theta^p - \theta^q\|^2 du} = \sqrt{\int \sum_{i=1}^{n}\sum_{j=1}^{n} d_i d_j g_i g_j du} \\ &= \sqrt{(\theta^p - \theta^q)^\top \boldsymbol{\phi} (\theta^p - \theta^q)} \end{aligned} \tag{9}$$

where $d_i = \theta_i^p - \theta_i^q, g_i = N(u|\mu_i, \sigma_i)$ and $g_j = N\left(u|\mu_j, \sigma_j\right)$ represent the Gaussian density function with different parameters, and $\boldsymbol{\phi}$ is a $n \times n$ matrix with its element $r_{ij} \sim N\left(\mu_i|\mu_j, \sigma_i + \sigma_j\right)$. This matrix reflects the correlation between different Gaussian density function of the GMM. From [33], we know that the functional distance can be transformed as the geodesic distance in the statistical manifold if we replace matrix $\boldsymbol{\phi}$ with the Fisher information matrix. Considering the closed-form of Fisher information matrix is hard to directly calculate, we often use the Kronecker delta function as a replacement [39], i.e. $\boldsymbol{\phi} = \boldsymbol{I}$:

$$\phi_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Then, the approximated geodesic distance between $\boldsymbol{p}$ and $\boldsymbol{q}$ is written as:

$$D(\boldsymbol{p},\boldsymbol{q}) = \sqrt{(\theta^p - \theta^q)^\top \boldsymbol{\phi} (\theta^p - \theta^q)} = \|\theta^p - \theta^q\|_2 \tag{11}$$

It can be noticed that (11) is also the distance between $\boldsymbol{p}$ and $\boldsymbol{q}$ in the parametric manifold. And its effectiveness is also proven in [33,39,40], when using the 2-norm distance in parametric manifold to approximate the geodesic distance in the statistical manifold. Therefore, through (11), the similarity between different data in statistical manifold can be effectively and efficiently measured. It is worth pointing out that since geodesic distance takes the manifold structure into consideration, it can bring more accurate similarity measurement for multi-modal data.

3.3. Representation completion in statistical manifold for cross-modal retrieval

Cross-modal retrieval aims to retrieve and return data of another modal which describes similar scene with the query data. So we expect to project query data into common statistical manifold $\mathscr{M}^C$ and perform cross-modal retrieval. As shown in Fig. 2, given any image (or text) data as query, we can model them as GMM with $\mathbb{G}^I$ (or $\mathbb{G}^T$) and embed into uni-modal statistical manifold $\mathscr{M}^I$ (or $\mathscr{M}^T$). However, according to **Definition 2**, $\mathscr{M}^I$ and $\mathscr{M}^T$ are the sub-manifolds of $\mathscr{M}^C$, and text and image data are specified as $(\xi_1, \ldots, \xi_{K_T}, \mathbf{0})$ and $\left(\mathbf{0}, \eta_1, \ldots, \eta_{K_I}\right)$ in $\mathscr{M}^C$, respectively. Obviously, the representation of query data in the uni-modal statistical manifold is incomplete and can not sufficiently represent the information of their corresponding scene. In order to sufficiently represent the scene and accordingly perform more accurate cross-modal retrieval, it is necessary to estimate the missing representation of text (image) data in $\mathscr{M}^I(\mathscr{M}^T)$ and accordingly complete their missing information for describing its scene. And this process can be done by projecting the data into statistical manifold of another modal and representing them with another set of basic function. So we need to seek two projections $P^{T\to I}$ and $P^{I\to T}$ for text and image data, respectively.

Given multiple pairs of image and text, we simply learn this two projections based on logistic regression. For a pair of text and image $(\xi, \boldsymbol{\eta})$, let $\hat{\boldsymbol{\eta}}$ and $\hat{\xi}$ be the projected point of ξ and $\boldsymbol{\eta}$, respectively. Text ξ can be projected into image statistical manifold using projection $P^{T\to I}$:

$$\hat{\boldsymbol{\eta}} = P^{T\to I}(\xi) = \frac{1}{S_T} e^{\boldsymbol{\omega}^\top \xi} \tag{12}$$

where $\boldsymbol{\omega}$ is a $K_T \times K_I$ projection matrix, and $S_T = \sum_{i=1}^{K_I} e^{\boldsymbol{\omega}_i^\top \xi}$ to ensure the condition $\sum_{i=1}^{K_I} \hat{\eta}_i = 1$. And similarly, image $\boldsymbol{\eta}$ can be projected into text statistical manifold using projection $P^{I\to T}$:

$$\hat{\xi} = P^{I\to T}(\boldsymbol{\eta}) = \frac{1}{S_I} e^{\boldsymbol{\varphi}^\top \eta} \tag{13}$$

where $\boldsymbol{\varphi}$ is a $K_I \times K_T$ projection matrix, and $S_I = \sum_{i=1}^{K_T} e^{\boldsymbol{\varphi}_i^\top \eta}$ to ensure the condition $\sum_{i=1}^{K_T} \hat{\xi}_i = 1$. Naturally, we expect $\hat{\boldsymbol{\eta}}$ and $\hat{\xi}$ to have similar representation with $\boldsymbol{\eta}$ and ξ. Also, $\boldsymbol{\omega}$ and $\boldsymbol{\varphi}$ should be sparse to prevent the overfitting problem and make the projection more robust. To sum up, we have the following objective functions:

$$\text{loss}^{T\to I} = \min_{\boldsymbol{\omega}} \|\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}\|_2 + \alpha \|\boldsymbol{\omega}\|_1 \tag{14}$$

and

$$\text{loss}^{I\to T} = \min_{\boldsymbol{\varphi}} \|\xi - \hat{\xi}\|_2 + \beta \|\boldsymbol{\varphi}\|_1 \tag{15}$$

where α and β are coefficients to control the sparse level of projection matrix. Obviously, this is two Lasso optimization problems which can be optimized by Coordinate Descent or Least Angle Regression [42]. Since this two optimization schemes may not achieve the global optimum, we simply implement two multi-layer neural network models to optimize this two problems and effectively learn $P^{T\to I}$ and $P^{I\to T}$, respectively.And the detail of learning $\boldsymbol{\omega}$ and $\boldsymbol{\varphi}$ for $P^{T\to I}$ and $P^{I\to T}$ is presented in the **Algorithm 2**.

After the appropriate $\boldsymbol{\omega}$ and $\boldsymbol{\varphi}$ is generated from **Algorithm 2**, text and image data can be projected into statistical manifold of another modality by $P^{T\to I}$ and $P^{I\to T}$. And their missing representation of another modality for describing the scene can be accordingly completed. Given text data X, it is firstly represented as $p(x|X,\xi)=\sum_{i=1}^{K_T}\xi_i g_i^T$ and specified as $\boldsymbol{\xi}=[\xi_1,\xi_2,\ldots,\xi_{K_T}]$ in $\mathscr{M}^T$. Then, we can project $\boldsymbol{\xi}$ from $\mathscr{M}^T$ into $\mathscr{M}^I$ and get its corresponding representation $\hat{\boldsymbol{\eta}}=[\hat{\eta}_1,\hat{\eta}_2,\ldots,\hat{\eta}_{K_I}]$ in $\mathscr{M}^I$ using (12). After that, its representation in the common statistical manifold can be generated for cross-modal retrieval through concatenating this two representation as $\boldsymbol{\psi}^t=(\xi_1,\ldots,\xi_{k_T},\hat{\eta}_1,\ldots,\hat{\eta}_{K_I})$. For its corresponding image data $p(y|Y,\boldsymbol{\eta})=\sum_{i=1}^{K_I}\eta_i g_i^I$, we can similarly generate its representation in common statistical manifold as $\boldsymbol{\psi}^i=\left(\hat{\xi}_1,\ldots,\hat{\xi}_{k_T},\eta_1,\ldots,\eta_{K_I}\right)$.

Algorithm 2 Learning Projection Matrix $\boldsymbol{\omega}$ and $\boldsymbol{\varphi}$ for Representation Completion of Uni-modal Data

Input: Training data $\{X_i,Y_i\}_{i=1}^{N_D}$ and their corresponding word embedding set $\{x_i\}_{i=1}^{N_W}$, visual words $\{y_i\}_{i=1}^{N_V}$. Parameters K_1, K_2, α, β and epoch E_{RC}.
Output: Projection matrix $\boldsymbol{\omega}$ and $\boldsymbol{\varphi}$.
1: Estimate the parameter of $\mathbb{G}^T$ and $\mathbb{G}^I$ by inputting $\{x_i\}_{i=1}^{N_W}$ and $\{y_i\}_{i=1}^{N_V}$ into **Algorithm 1**.
2: Generate statistic representation for text data by (2), and image data by (5). And training data are represented as $\{\boldsymbol{\xi}_i,\boldsymbol{\eta}_i\}_{i=1}^{N_D}$.
3: Randomly initialize the projection matrix $\boldsymbol{\omega}^1$ and $\boldsymbol{\varphi}^1$.
4: **for** e = 1 to E_{RC} **do**
5: Compute (14) and (15) with $\boldsymbol{\omega}^e$, $\boldsymbol{\varphi}^e$ and $\{\boldsymbol{\xi}_i,\boldsymbol{\eta}_i\}_{i=1}^{N_D}$.
6: Compute the gradient of (14) and (15) with respect to $\boldsymbol{\omega}$ and $\boldsymbol{\varphi}$ with Adam optimizer [41], respectively.
7: Update the projection matrix with the gradient and obtain $\boldsymbol{\omega}^{e+1}$, $\boldsymbol{\varphi}^{e+1}$.
8: **end for**
9: **return** $\boldsymbol{\omega}^{e+1}$ and $\boldsymbol{\varphi}^{e+1}$.

3.4. Generating binary code of sufficient scene representation

After generating SSR for multi-modal data, they can be embedded into common statistical manifold in which their similarity can be accurately measured based on (11). For more efficient retrieval and storage saving over large scale multi-modal data, it is necessary to further generate compact binary code for sufficient scene representation of multi-modal data. Considering the balance between quantization loss and model complexity, this paper utilizes ITQ to learn similarity preserving binary code for cross-modal retrieval. Given SSR of all the training data, ITQ learns a matrix to rotate the data so as to minimize the quantization loss and then generate compact binary code for multi-modal data through quantifying the rotated data by sign function sgn(·). And this corresponding process is elaborated in the **Algorithm 3**.

After obtaining the rotation matrix R through **Algorithm 3**, given SSR of any query data ψ, it's low-dimensional representation $h\in\Re^{1\times L}$ can be generated through the pre-trained PCA. Then, it's corresponding binary code can be computed as:

$$b=\mathrm{sgn}(hR) \tag{16}$$

Algorithm 3 Generating Binary Code for Sufficient Scene Representation with ITQ

Input: Corresponding SSR of all the training data $\{\psi_i\}_{i=1}^{N_S}$, the length of binary code L and the number of iteration E_{ITQ}.
Output: Rotation matrix R, binary code of all the training data $B=\{b_i\}_{i=1}^{N_S}$.
1: Implement Principal Component Analysis (PCA) to project all the input data $\{\psi_i\}_{i=1}^{N_S}$ into L-dimension space and denote the generated low-dimensional representation as matrix $H\in\Re^{N_S\times L}$.
2: Initialize $R\in\Re^{L\times L}$ as a random orthogonal matrix.
3: Quantify H with the sign operation: $B=\mathrm{sgn}(H)\in\{0,1\}^{N_S\times L}$.
4: Compute the quantization loss using squared F-norm $\|\cdot\|_F^2$: $Q(B,H)=\|B-HR\|_F^2$
5: Utilize ITQ to find a local minimum of the $Q(B,H)$. In each iteration, each data point is first assigned to the nearest vertex of the binary hyper-cube, and then R is updated to minimize the quantization loss given this assignment. According to [26], the detail of this two alternating steps are as follows:
6: **for** e = 1 to E_{ITQ} **do**
7: Let $\widetilde{H}=HR$, fix R and update B as: $B_{ij}=\begin{cases}1, & \widetilde{H}_{ij}\geqslant 0\\ 0, & \widetilde{H}_{ij}<0\end{cases}$
8: Fix B and update R: Compute the singular value decomposition of matrix B^TH as $S\Omega\widehat{S}^\top$ and update R as $\widehat{S}S^\top$.
9: **end for**
10: **return** R and B.

4. Experiment

In this section, plenty of experiments are conducted to convince the speciality of CMSSR. Firstly, some experimental settings and basic situation are introduced. Then, CMSSR is compare with several state-of-the-art methods on three standard benchmark datasets. Further, ablation experiments are conducted in allusion to the key module of CMSSR. Finally, we discuss the effects of the Number of Gaussian Density Function to CMSSR.

4.1. Datasets and evaluation metric

4.1.1. Datasets

To validate the effectiveness of CMSSR, a series of experiments will be conducted on the following three standard benchmark datasets: Wiki is the most widely-used dataset for evaluating the performance cross-modal retrieval method. It is collected from "Wikipedia featured articles" and consists of 2,866 image-text pairs grouped into 10 semantic categories. Following [43], 693 image-text pairs will be randomly chosen as the query set, and the remaining as the training set and the database.

Mir-Flickr25k [44], collected from the Flickr website, which contains 25,000 images annotated from the provided 24 labels. And each image is described by more than one textual tag. We remove stop words and non-English words, and only keep those tags that appear more than 10 times. After that, those images without textual tags are filtered, and we finally construct 21621 image-text pairs. Among them, 2000 pairs are randomly selected as the query set while the remaining is used as the training set and the database.

IAPRTC-12 [45] consists of 20000 images collecting from a wide variety of domain, such as landscapes, portraits, indoor and sports scenes. Each image is associated with a textual description in up to three different languages (English, German and Spanish). But we only use the English version. For evaluation, we follow the common practice in [3,23] and select the subset of the top 22 frequent labels from the 275 concepts obtained from the segmentation task as the label set. Then, those images without label are filtered while the remaining 18685 image-text pairs are kept. Similar to the MirFlickr25k dataset, 2,000 image-text pairs are selected as the query set and the remaining pairs are the training set and the database.

4.1.2. Evaluation metrics

Following the setting of most of the impressive unsupervised cross-modal hashing methods [24–27], the numeric evaluation of the cross-modal retrieval performance is shown by the mean average precision (mAP) score which has good discriminated power and stability. Given a query data and its N_R retrieved instances, the average precision (AP) for this query is calculated as:

$$AP = \frac{1}{N_L}\sum_{n=1}^{N_R} P(n)\varepsilon(n) \tag{17}$$

where N_L is the number of related instances in the retrieved set, $P(n)$ denotes the precision of the top n retrieved instances, and $\varepsilon(n) = 1$ if the retrieved item at rank n is relevant, $\varepsilon(n) = 0$ otherwise. Then, the mAP score can be calculated by averaging the AP value of all the queries. Generally, the better the retrieval performance of a cross-modal retrieval method, the higher of its corresponding mAP score. The same with [9,12,15,23], N_R is set to 50. To present the evaluation result from the global scope, we further plot the precision-recall (PR) curves for different retrieval tasks (image query text, text query image, image query image and text query text) in different datasets to show the precision in different recall levels.

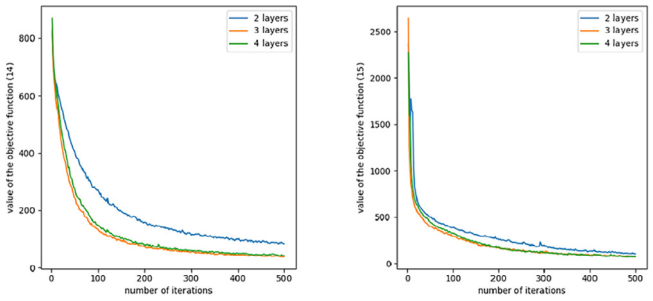
In addition to the above two metrics for measuring the retrieval performance, this paper utilizes the Normalization Fisher's Discriminant Ratio (NFDR) [46] to reflect the chaotic level of the common space. Actually, NFDR mainly measures the class overlap level of data. The higher of the class overlap level, the more difficult of the classification. Suppose there are totally N_C classes in a dataset, the NFDR of the common space generated by any cross-modal retrieval methods can be calculated as:

$$\boldsymbol{F} = \frac{1}{N_C^2}\sum_{i=1}^{N_C}\sum_{j=1}^{N_C} F_{i,j} \tag{18}$$

where $F_{i,j}$ is the overlap level between data from the i^{th} and the j^{th} class. And it can be calculated as:

$$F_{i,j} = \frac{\left\|\mu_i - \mu_j\right\|_2}{\|\Sigma_i\|_2 + \|\Sigma_j\|_2 + \left\|\mu_i - \mu_j\right\|_2} \tag{19}$$

where $\|\cdot\|_2$ is the L2-norm of the matrix or vector, and μ_i, Σ_i and μ_j, Σ_j are the mean vectors and co-variance matrices of data from the i^{th} and the j^{th} class, respectively. From Eq. (18) and (19), we know that $F_{i,j}$ will be larger if the distance between the mean vector of two classes is farther and data point in the same class are more compact. On the contrary, if the distance between data of two class is too close and data in the same class distribute sparsely, the value of $F_{i,j}$ will get smaller. Therefore, $F_{i,j}$ can intuitively measure the overlap degree of data from two different classes. As the accumulation of $F_{i,j}$, $\boldsymbol{F}$ can be used in measuring the chaotic level of a common space. And lower $\boldsymbol{F}$ value reflects higher chaotic level of a common space. In the comparison experiment, we will calculate the corresponding $\boldsymbol{F}$ value of image data, text data and all the multi-modal data in the common space to help analyse the chaotic distribution problem.



(a) Image-to-text projection (b) Text-to-image projection

Fig. 3. The convergence curves corresponding to different numbers of network layer in Wiki dataset.

4.2. Compared methods and implementation details

To better show the characteristic of CMSSR, we compare CMSSR against several state-of-the-art unsupervised cross-modal retrieval methods: IMH [8], BSE [9], FSH [22], LSSH [13], CMFH [14], MSFH [15], DBRC [21], DCSH [23], DSAH [16], DGCPN [12]. For fair comparison, all of these methods are implemented on the same set of 128-d SIFT features (d for dimension) and 200-d word vectors in the image and text domain, respectively. Considering some methods, such as IMH, LSSH, CMFH, FSH, MSFH and CUH, only accept global feature as their input, we further generate integrated representation as their input based on the above local feature set. Specifically, 500-d bag of feature (BoF) representation based on the 128-d SIFT feature is generated for image data while 128-d BoF representation based on 200-d word vectors is generated for text data. Besides, for the methods based on deep feature, such as DCSH, DSAH and DGCPN, we replace their image feature extractor (such as AlexNet [47]) and text feature with the above integrated feature of image and text modality, respectively. As for the rest of methods, we directly input the local feature (i.e. 128-d SIFT features and 200-d word vectors). Since the input features are different from which in their original publications, we carefully tune their parameters and retain the best result for comparison. When comparing CMSSR with those compared methods, we empirically set the number of Gaussian density function K_1 for representing text data to 32 and K_1 for image modality to 128. Besides, to decide how many fully connected neural networks should be used in the image-to-text projection and text-to-image projection, the corresponding convergence curves with respect to different numbers of layers of different projections in the Wiki dataset are also plotted. As is shown in the Fig. 3, CMSSR can well converge to a relatively small value in both the image branch and text branch when we use 3 or 4-layers neural networks. Considering the scale of our model and the training efficiency, 3-layers neural network is used in both the image-to-text projection and text-to-image projection. Since the dimension of uni-modal statistical manifold is different, the values of coefficient α and β are set as 10^{-2} and 10^{-1}, respectively. Furthermore, the learning rate of Adam optimizer of the image-to-text projection and the text-to-image projection are set as 10^{-3} and 10^{-4}, respectively.

4.3. Comparisons with state-of-the-art methods

In this part, CMSSR is compared with the above-mentioned eight unsupervised cross-modal hashing methods. And we will discuss the experimental results on varying datasets.

Table 1
The comparison results of different cross-modal hashing methods with respect to mAP values on Wiki dataset.

Methods	Image-to-text retrieval						Text-to-image retrieval					
	16 bits	32 bits	48 bits	64 bits	80 bits	96 bits	16 bits	32 bits	48 bits	64 bits	80 bits	96 bits
IMH	0.2166	0.2316	0.2330	0.2329	0.2252	0.2203	0.3029	0.3622	0.3944	0.4177	0.4465	0.4584
BSE	0.2355	0.2388	0.2418	0.2456	0.2496	0.2542	0.5131	0.5317	0.5532	0.5798	0.5871	0.6094
LSSH	0.2367	0.2450	0.2527	0.2371	0.2484	0.2563	0.5247	0.5406	0.5597	0.5615	0.5683	0.5779
CMFH	0.2276	0.2184	0.2545	0.2418	0.2426	0.2430	0.4867	0.4964	0.5120	0.5232	0.5391	0.5469
MSFH	0.2258	0.2337	0.2372	0.2427	0.2442	0.2467	0.4963	0.5112	0.5184	0.5291	0.5571	0.5821
FSH	0.2371	0.2421	0.2442	0.2470	0.2487	0.2505	0.4696	0.4712	0.4896	0.5126	0.5434	0.5557
DBRC	0.2343	0.2388	0.2417	0.2432	0.2448	0.2455	0.5168	0.5088	0.5219	0.5401	0.5442	0.5617
DCSH	0.2421	0.2432	0.2447	0.2454	0.2466	0.2489	0.5227	0.5473	0.5514	0.5602	0.5615	0.5884
DSAH	0.2371	0.2398	0.2433	0.2476	0.2506	0.2550	0.5314	0.5456	0.5512	0.5654	0.5791	0.5918
DGCPN	0.2453	0.2488	0.2512	0.2540	0.2578	0.2613	0.5368	0.5501	0.5692	0.5761	0.5805	0.6024
CMSSR	**0.2533**	**0.2543**	**0.2552**	**0.2571**	**0.2621**	**0.2657**	**0.5718**	**0.6319**	**0.6298**	**0.6488**	**0.6562**	**0.6781**

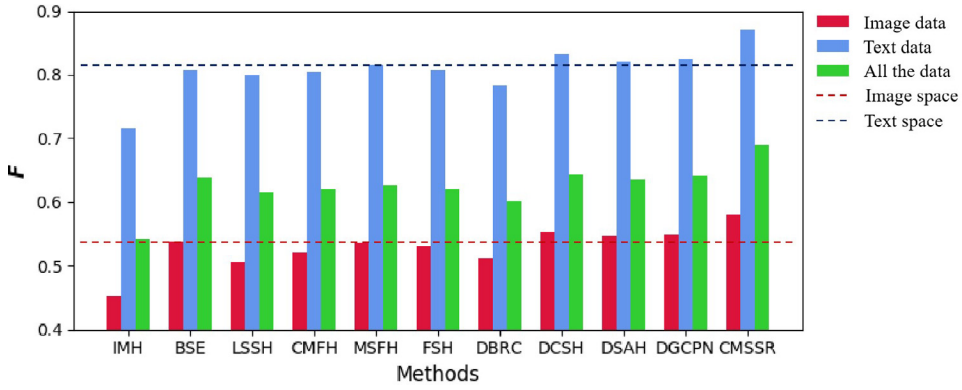


Fig. 4. The ***F*** value of common spaces generated by different methods on Wiki dataset.

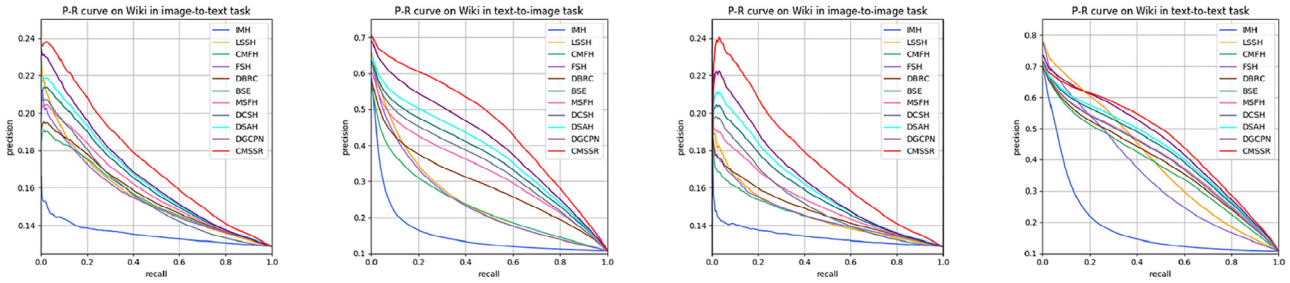


Fig. 5. Precision-Recall curves on Wiki dataset by varying retrieval task with the hash code length of 96 bits.

4.3.1. Experimental results on Wiki dataset

Wiki dataset is one of the most challenging cross-modal dataset. All of the textual descriptions in Wiki are long text with averaging 544 words per instance, which means text data of Wiki contains much more information than image data. And this tremendous imbalance enlarges the semantic gap between text and image data. Table 1 shows the performance of hash with different length generated by different methods on the Wiki dataset. In addition, we calculate the ***F*** value of 96-bits common representation generated by varying methods and record them in the histogram shown in Fig. 4. And the corresponding P-R curves of the 96 bits hash code of all the experimental methods are plotted in Fig. 5. From [25], we know that maximizing the intra- and inter-modal similarity are mutually constrained, preserving inter-modal similarity will destroy the intra-modal similarity relation, and vice versa. In Fig. 4, we additionally plot the dash line in red and blue to represent the ***F*** value of the original image and text space. It can be observed that most of the experimental methods (such as IMH, BSE, LSSH, CMFH, FSH and DBRC) get lower ***F*** value of image and text data in the common space than which in the original feature space, that means both image and text data get more chaotic distribution after being projected into common space. This tells that these methods can not well preserve the intra-modal similarity. Meanwhile, it can be found that they also get low ***F*** value of all the multi-modal data in common space, which indicates that they can not well preserve the inter-modal similarity either. Although the rest of compared methods (i.e. MSFH, DCSH, DSAH, DGCPN) get higher ***F*** value in common space than which in the modal-specific feature space, their increases are imperceptible. Unlike these methods, CMSSR generates SSR through the representation completion strategy without the consideration of too much similarity relation. Therefore, it can be noticed that the ***F*** value of image and text data in the common space of CMSSR considerably higher than which of the original fea-

Table 2
The comparison results of different cross-modal hashing methods with respect to mAP values on Mir-Flickr25k dataset.

Methods	Image-to-text retrieval						Text-to-image retrieval					
	16 bits	32 bits	48 bits	64 bits	80 bits	96 bits	16 bits	32 bits	48 bits	64 bits	80 bits	96 bits
IMH	0.5777	0.5697	0.5687	0.5643	0.5628	0.5621	0.5857	0.5784	0.5751	0.5723	0.5712	0.5715
BSE	0.6101	0.6187	0.6198	0.6232	0.6254	0.6365	0.6387	0.6497	0.6517	0.6689	0.6761	0.6834
LSSH	0.5929	0.5962	0.5953	0.6025	0.5959	0.6011	0.6175	0.6245	0.6267	0.6277	0.6324	0.6277
CMFH	0.5931	0.6031	0.6044	0.6114	0.6055	0.6117	0.6007	0.6161	0.6179	0.6187	0.6188	0.6295
MSFH	0.6066	0.6093	0.6114	0.6185	0.6121	0.6255	0.6337	0.6461	0.6482	0.6587	0.6592	0.6612
FSH	0.6005	0.6085	0.6179	0.6144	0.6221	0.6295	0.6248	0.6349	0.6468	0.6416	0.6573	0.6520
DBRC	0.5932	0.5972	0.6085	0.6144	0.6058	0.6179	0.6038	0.6176	0.6135	0.6168	0.6287	0.6339
DCSH	0.6030	0.6095	0.6177	0.6201	0.6229	0.6322	0.6210	0.6362	0.6433	0.6587	0.6661	0.6751
DSAH	0.6269	0.6285	0.6316	0.6342	0.6359	0.6371	0.6572	0.6631	0.6704	0.6743	0.6781	0.6830
DGCPN	0.6297	0.6305	0.6331	0.6359	0.6388	0.6405	0.6593	**0.6659**	**0.6773**	0.6824	0.6887	0.6953
CMSSR	**0.6343**	**0.6399**	**0.6396**	**0.6439**	**0.6487**	**0.6511**	**0.6624**	0.6647	0.6751	**0.6868**	**0.6965**	**0.7112**

ture space, which means the uni-modal data become more discriminative after being projected into common space by CMSSR. And this also benefits the distribution of all the multi-modal data in the common space. Thus, from Table 1, we can see that CMSSR outperforms the second-best competitor, DGCPN, by 0.44% on 96 bits for image-to-text retrieval, and by 7.57% on 96 bits for text-to-image retrieval. Moreover, CMSSR also shows its superior performance compared with other methods on cross-modal retrieval task of all the bits. In addition to the above phenomenon, there are a few points in Fig. 5 worth noticing. From Fig. 5, we can see that the area below the PR curves of CMSSR are larger than which of the other compared methods in the cross-modal retrieval tasks. Especially, the performance of CMSSR is dramatically better than the other methods on text-to-image retrieval, shown as Fig. 5 (b). This situation may come from the following reasons. Gaussian density functions of GMM can exponentially shrink the influence of that useless information and highlight that useful information. As stated before, all the text data in Wiki are long text containing much irrelevant information and noise. Representing text data with GMM effectively eliminates this irrelevant information and brings more robust representation. Besides, long text is beneficial to the statistical method for representing data, since long text contains more abundant statistical sample. Thus, CMSSR shows its superior performance on text-to-image retrieval task. On the other hand, CMSSR does not exert strong similarity constrain for preserving intra-modal similarity. Therefore, from Fig. 5 (d), we can find that the PR curve of CMSSR is under the curves of three other compared methods in the text-to-text retrieval when the recall is low. But the downward trend of the precision of CMSSR is smaller than the other methods with the increasing of recall. We also notice that the joint representation generated by CMSSR has considerable improvement on image-to-image retrieval compared with the other methods, which is shown in Fig. 5 (c). This owes to the process of representation completion which completes the complementary information from text domain for image data and brings more accurate representation. And thanks to this accurate repre-

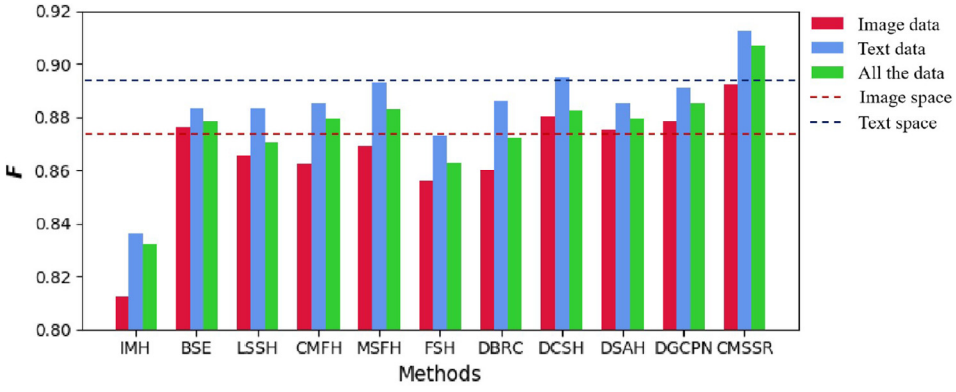
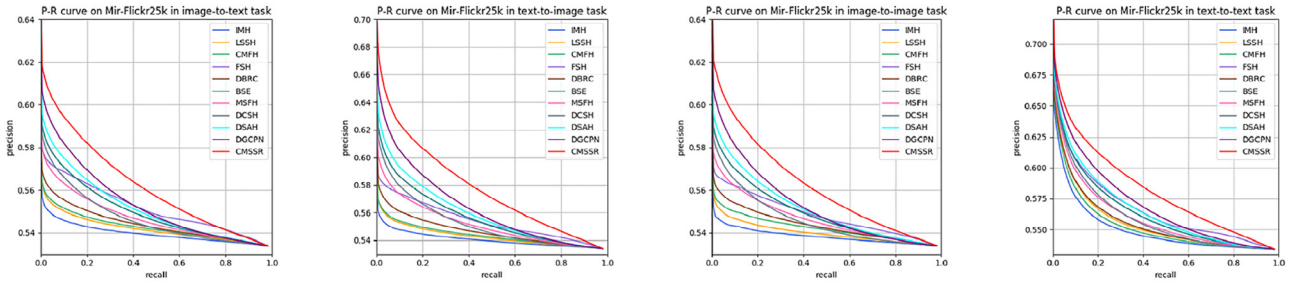


Fig. 6. The ***F*** value of common spaces generated by different methods on Mir-Flickr25k dataset.



(a) Image-to-text (b) Text-to-image (c) Image-to-image (d) Text-to-text

Fig. 7. Precision-Recall curves on Mir-Flickr25k dataset by varying retrieval task with the hash code length of 96 bits.

Table 3
The comparison results of different cross-modal hashing methods with respect to mAP values on IAPRTC dataset.

Methods	Image-to-text retrieval						Text-to-image retrieval					
	16 bits	32 bits	48 bits	64 bits	80 bits	96 bits	16 bits	32 bits	48 bits	64 bits	80 bits	96 bits
CUH	0.3711	0.3723	0.3863	0.3916	0.4009	0.4067	0.4422	0.4154	0.4311	0.4223	0.4339	0.4338
IMH	0.4269	0.4249	0.4182	0.4132	0.4113	0.4103	0.4567	0.4471	0.4414	0.4403	0.4379	0.4351
BSE	0.4483	0.4592	0.4662	0.4688	0.4707	0.4761	0.5076	0.5119	0.5281	0.5300	0.5387	0.5507
LSSH	0.4593	0.4565	0.4617	0.4667	0.4670	0.4746	0.4744	0.4960	0.4997	0.5072	0.5169	0.5149
CMFH	0.4449	0.4567	0.4573	0.4565	0.4577	0.4597	0.4983	0.5101	0.5151	0.5159	0.5156	0.5172
MSFH	0.4311	0.4421	0.4488	0.4542	0.4605	0.4653	0.4955	0.4912	0.5023	0.5091	0.5167	0.5211
MSFH	0.4416	0.4428	0.4414	0.4528	0.4512	0.4517	0.4500	0.4685	0.4752	0.4758	0.4834	0.4835
DBRC	0.4424	0.4453	0.4494	0.4574	0.4582	0.4601	0.4961	0.5023	0.5147	0.5089	0.5106	0.5184
DCSH	0.4398	0.4460	0.4527	0.4602	0.4667	0.4711	0.5013	0.5066	0.5097	0.5122	0.5168	0.5203
DCSH	0.4412	0.4478	0.4552	0.4631	0.4704	0.4759	0.5109	0.5183	0.5227	0.5269	0.5314	0.5366
DCSH	0.4539	0.4583	0.4638	0.4703	0.4755	0.4822	0.5155	0.5201	0.5238	0.5286	0.5332	0.5397
CMSSR	**0.4770**	**0.4898**	**0.4948**	**0.4975**	**0.4989**	**0.4912**	**0.5336**	**0.5401**	**0.5640**	**0.5704**	**0.5751**	**0.5821**

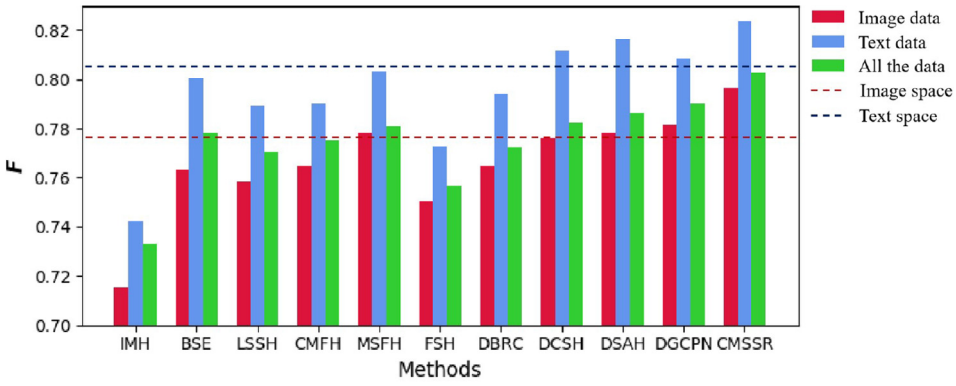


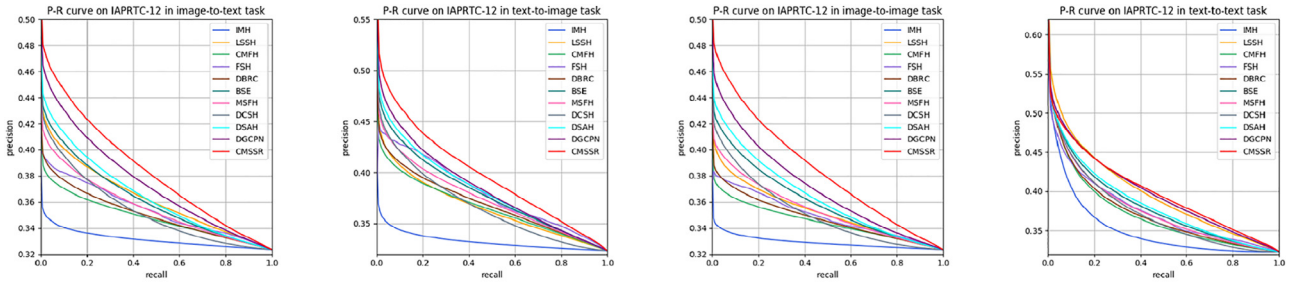
Fig. 8. The ***F*** value of common spaces generated by different methods on IAPRTC dataset.

sentation, CMSSR can effectively narrow the semantic gap and gain higher cross-modal retrieval results.

4.3.2. Experimental results on Mir-Flickr25k dataset

The mAP values of different methods with different code length in the Mir-Flickr25k dataset are shown in Table 2. Besides, the ***F*** value of 96-bits common representation generated by varying methods are recorded in the histogram shown in Fig. 6. And the corresponding P-R curves of the 96 bits hash code of all the experimental methods are plotted in Fig. 7. Unlike the Wiki dataset, all the text data in Mir-Flickr25k are composed of multiple different discrete textual tags. Text data can well describe the content in the corresponding image, while the relationship between different objects in the image cannot be reflected by the text.

Since the combination of discrete tags can well describe the image data, the semantic gap between image and text in Mir-Flickr25k is minor compared with Wiki. Therefore, as is shown in Table 2, the overall experimental results of Mir-Flickr25k are better than which in the Wiki. Also, the imbalance situation between results of image-to-text retrieval and text-to-image retrieval in the Wiki dataset does not exist in the Mir-Flickr25k dataset. In combination with the analysis in Wiki, those methods which focus too much on preserving similarity relationship still suffer from the chaotic distribution problem and get lower ***F*** value shown in Fig. 6. Compared with other experimental methods, CMSSR gets highest ***F*** value. This indicates the inter-class distribution of our generated common space is more discriminative, which does good to the retrieval task. And it can be found from the P-R curves shown in Fig. 7 that CMSSR possesses obvious superiority on four different retrieval tasks for 96-bits length hash code. Likewise, from the Table 2, although DGCPN slightly outperforms CMSSR by 0.12% and 0.22% on the 32-bits and 48-bits text-to-image retrieval task, CMSSR still keeps its leading position on the overall retrieval performance among all the compared methods.



(a) Image-to-text (b) Text-to-image (c) Image-to-image (d) Text-to-text

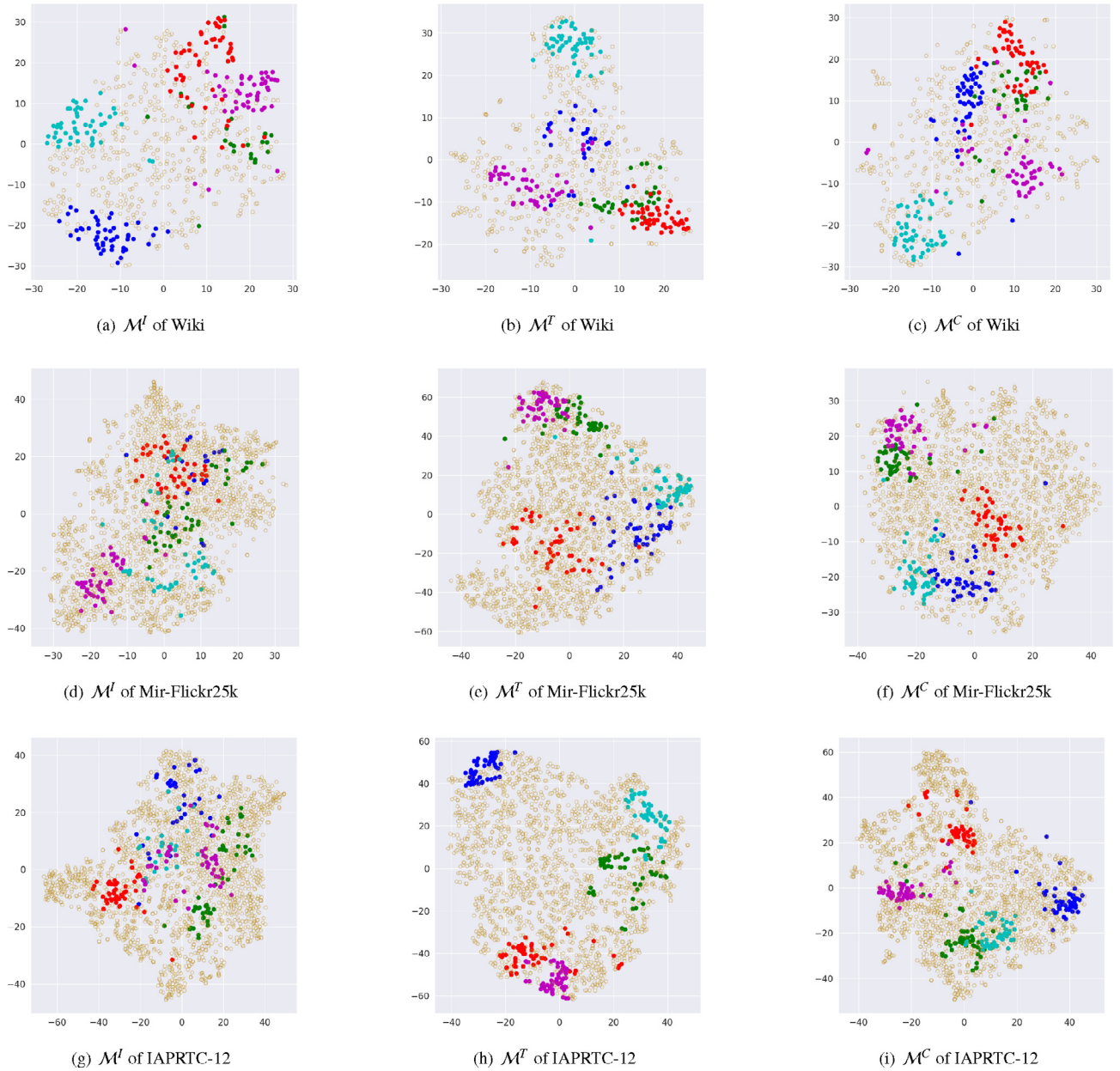
Fig. 9. Precision-Recall curves on IAPRTC dataset by varying retrieval task with the hash code length of 96 bits.

Table 4
The mAP values of four different tasks in original feature space, uni-modal statistical manifolds and common statistical manifold in Mir-Flickr25k dataset.

	Cross-modal retrieval						Uni-modal retrieval					
	Image-to-text retrieval			Text-to-image retrieval			Image-to-image retrieval			Text-to-text retrieval		
	Wiki	Flickr	IAPRTC	Wiki	Flickr	IAPRTC	Wiki	Flickr	IAPRTC	Wiki	Flickr	IAPRTC
OFS	–	–	–	–	–	–	0.2047	0.6079	0.4136	0.6413	0.7313	0.5957
$\mathscr{M}^I$	0.2512	0.6318	0.4493	0.6057	0.6259	0.4519	0.2337	0.6345	0.4465	–	–	–
$\mathscr{M}^T$	0.2455	0.6191	0.4741	0.6714	0.7086	0.5893	–	–	–	0.6923	0.7335	0.6001
$\mathscr{M}^C$	**0.2721**	**0.6613**	**0.4984**	**0.6976**	**0.7207**	**0.6045**	**0.2661**	**0.6468**	**0.4913**	**0.7103**	**0.7422**	**0.6157**

4.3.3. Experimental results on IAPRTC dataset

In Table 3, we list the mAP values of different methods with different code length in the IAPRTC-12 dataset. What's more, the $\boldsymbol{F}$ value of 96-bits common representation generated by varying methods are recorded in the histogram shown in Fig. 8. And the corresponding P-R curves of the 96 bits hash code of all the experimental methods are plotted in Fig. 9. Different from Wiki and Mir-Flickr25k, text data in IAPRTC-12 are short text which can precisely describe the image data. But just as sentence in IAPRTC-12 is more precise semantic than the combination of multiple discrete tags,



(a) $\mathcal{M}^I$ of Wiki (b) $\mathcal{M}^T$ of Wiki (c) $\mathcal{M}^C$ of Wiki

(d) $\mathcal{M}^I$ of Mir-Flickr25k (e) $\mathcal{M}^T$ of Mir-Flickr25k (f) $\mathcal{M}^C$ of Mir-Flickr25k

(g) $\mathcal{M}^I$ of IAPRTC-12 (h) $\mathcal{M}^T$ of IAPRTC-12 (i) $\mathcal{M}^C$ of IAPRTC-12

Fig. 10. The t-SNE Visualization of the testing data of Mir-Flickr25k in uni-modal statistical manifolds and common statistical manifold.

the correlation between image and text data in IAPRTC-12 is more complex and difficult to be learned. So the performance of all methods in IAPRTC-12 dataset is inferior to that in Mir-Flickr25k. In spite of this, CMSSR still gets the highest $\boldsymbol{F}$ value of its common space, which can be observed from Fig. 8. And accordingly, CMSSR outclasses the other compared methods on both cross-modal retrieval and uni-modal retrieval, which is presented in Table 3 and Fig. 9.

4.4. Ablation analysis

To validate the effectiveness of the GMM representation and the proposed common statistical manifold respectively, we conduct different retrieval tasks in the following four different data spaces, i.e. original feature space (OFS), image statistical manifold $\mathscr{M}^I$, text statistical manifold $\mathscr{M}^T$, and common statistical manifold $\mathscr{M}^C$. Due to the heterogeneity between data of different modalities, only uni-modal retrieval tasks (i.e. image-to-image retrieval and text-to-text retrieval) are conducted in the original feature space. Using $P^{T\to I}$, text data can be projected into Image Statistical Manifold. So we only conduct cross-modal retrieval and image-to-image retrieval in Image Statistical manifold. Similarly, image data can be projected into text Statistical Manifold using $P^{I\to T}$.And only cross-modal retrieval and text-to-text retrieval are conducted in this manifold. As for common statistical manifold, we can implement both cross-modal retrieval tasks and uni-modal retrieval tasks since data from different modalities can be all projected into this manifold. These experiments are conducted under three different datasets, respectively. And the experimental results are calculated with the representation before quantization.

4.4.1. Benefits of GMM representation

Through implementing uni-modal retrieval in both original feature space and uni-modal statistical manifolds, we attempt to figure out the impact of the GMM representation. From Table 4, it can be observed that uni-modal statistical manifolds outperform original feature space on both image-to-image and text-to-text retrieval task in all the dataset. This indicates that the GMM representation is more robust and can more sufficiently represent each instance. As we stated in the Section 3.1, each component of the GMM can be regarded as an abstract concept. After representing as GMM, image data are further abstracted from low-level visual features to higher-level representation which is closer with the label domain. Therefore, the mAP value of $\mathscr{M}^I$ has about 2.66% to 3.29% improvement compared with OFS in the task of image-to-image retrieval. As for text data, this process blurs the precise semantic of text and mainly reduces the impact of noise with the Gaussian density function. It can be seen from Table 4 that the mAP value of $\mathscr{M}^T$ has improvement compared with which of the OFS in three datasets. Specifically, the mAP value of $\mathscr{M}^T$ has about 5.1% increase in the Wiki dataset. And the effect of GMM representation is more conspicuous in the Wiki dataset because the text data of Wiki are long text containing a mass of noise.

4.4.2. Effectiveness of common statistical manifold

We additionally investigate the effectiveness of our abstracted and constructed common statistical manifold. From Table 4, we find that the constructed common statistical manifold outperforms uni-modal statistical manifolds and original feature space on uni-modal retrieval in all the datasets. The mAP value of common statistical manifold of three datasets has an average of 2.98% and 1.41% improvement compared with which of the other data manifolds in image-to-image and text-to-text retrieval, respectively. Besides, if we concentrate on the cross-modal retrieval task, we can find that the mAP value of common statistical manifold of three datasets has an average of 3.21% and 6.54% improvement compared with which of the other data manifolds on image-to-text and text-to-image retrieval task, respectively. This demonstrates that the proposed representation completion can effectively complete the missing representation of another modal, which helps to narrow the semantic gap between multi-modal data. And this also proves the correctness of our modeling of SSR. To more intuitively show the effectiveness of our proposed common statistical manifold, we visualize the testing data (including both image and text data) of different datasets in three different data statistical manifolds using t-SNE, which is shown in Fig. 10.

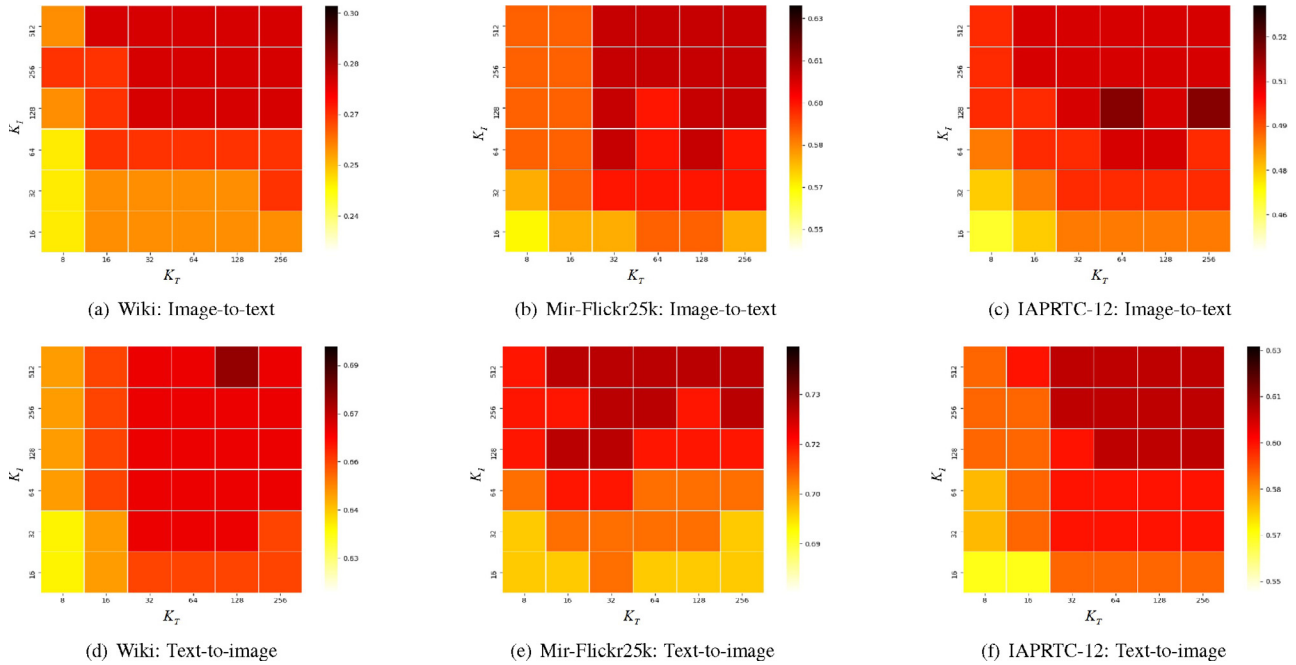


Fig. 11. Cross-modal retrieval performance of CMSSR with varying number of Gaussian density function K_T and K_I on three different datasets.

Observing the scatter diagram of image statistical manifold, it can be noticed that data in the same class are not dense enough and the data in different class are not scattered enough. Comparing the scatter diagram between image and text statistical manifold, we can also find that the data distribution in text statistical manifold is better than image statistical manifold. When coming to the common statistical manifold, we can see that our proposed common statistical manifold has better data distribution than image and text statistical manifold in three different datasets which illustrates the advantages of our generated SSR.

4.5. Discussion of the number of Gaussian density function

To analyse the sensitivity of CMSSR with respect to the number of Gaussian density function of GMM, i.e. K_T and K_I, we conduct cross-modal retrieval with the combination of different K_T and K_I on three different datasets. And the experimental results are intuitively shown with heat map in Fig. 11. From Fig. 11, we have the following observations. (i) When a few of Gaussian density functions are used, GMM cannot fully represent each data. And thus, we can accordingly find that CMSSR performs worse when the value of K_T and K_I are small. (ii) The performance on each dataset can achieve a relatively high value when $K_T \geqslant 32$ and $K_I \geqslant 128$. Besides, it can be observed that even when K_T and K_I are set to higher values, the change of the performance of CMSSR is imperceptible, i.e. the rangeability are less than about 1.5% in different retrieval task among different dataset. This shows the stability of CMSSR with respect to the number of Gaussian density function.

5. Conclusion

This paper proposes to model the common representation of multi-modal data as the sufficient scene representation and perform cross-modal retrieval from the perspective of scene. Through integrating information from different modalities, we directly generate such a common representation for multi-modal data without the consideration of simultaneously preserving inter- and intra-modal similarity. Hence, CMSSR does not suffer from the chaotic distribution problem and has better performance than several similarity preservation methods. Also, such representation is more semantically robust and can promote to narrow the semantic gap since they contain more complete information from different modalities. Therefore, even when comparing with other state-of-the-art unsupervised cross-modal hashings, CMSSR has superior retrieval performance. In the ablation analysis, common statistical manifold has shown considerable improvement compared with uni-modal statistical manifold, which has also demonstrated the effectiveness of such an informatively complete representation. Besides, the significant performance of CMSSR in extensive experiments has confirmed the availability of our proposed representation completion strategy for completing missing information of another modality. Lastly, we also find that CMSSR is stable with respect to the number of Gaussian density function.

CRediT authorship contribution statement

Jieting Luo: Methodology, Software, Writing - original draft. **Yan Wo:** Conceptualization, Writing - review & editing, Supervision. **Bicheng Wu:** Data curation, Writing - review & editing. **Guoqiang Han:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is supported by National Natural Science Foundation of Guangdong [Grant No.2018A030313994, Grant No.2021A1515012020, Grant No.2017A030312008], and Guangzhou science and technology plan project [Grant No.202002030298].

References

[1] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: Proceedings of ACM International Conference on Multimedia, 2010, pp. 251–260.
[2] Y. Peng, X. Huang, J. Qi, Cross-media shared representation by hierarchical learning with multiple deep networks, in: Proceedings of International Joint Conference on Artificial Intelligence, 2016, pp. 3846–3853.
[3] V.E. Liong, J. Lu, Y.-P. Tan, J. Zhou, Cross-modal deep variational hashing, in: Proceedings of IEEE International Conference on Computer Vision, 2017, pp. 4077–4085.
[4] C. Yan, X. Bai, S. Wang, J. Zhou, E.R. Hancock, Cross-modal hashing with semantic deep embedding, Neurocomputing 337 (2019) 58–66.
[5] D. Mandal, K.N. Chaudhury, S. Biswas, Generalized semantic preserving hashing for cross-modal retrieval, IEEE Transactions on Image Processing 28 (1) (2019) 102–112.
[6] Q. Lin, W. Cao, Z.h. He, Z.q. He, Semantic deep cross-modal hashing, Neurocomputing, 396 (2020) 113-122..
[7] X. Gu, G. Dong, X. Zhang, L. Lan, Z. Luo, Semantic-consistent cross-modal hashing for large-scale image retrieval, Neurocomputing 433 (2020) 181–198.
[8] J. Song, Y. Yang, Y. Yang, et al., Inter-media hashing for large-scale retrieval from heterogeneous data sources, in: Proceedings of ACM SIGMOD International Conference on Management of Data, 2013, pp. 785–796.
[9] M. Yu, L. Liu, L. Shao, Binary set embedding for cross-modal retrieval, IEEE Transactions on Neural Networks and Learning Systems 28 (12) (2017) 1–12.
[10] Z. Ye, Y. Peng, Sequential cross-modal hashing learning via multi-scale correlation mining, ACM Transactions on Multimedia Computing Communications and Applications 15 (4) (2019) 1–20.
[11] J. Zhang, Y. Peng, Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval, IEEE Transactions on Multimedia 22 (1) (2020) 174–187.
[12] Yu J, Zhou H, Zhan Y, Comprehensive graph-conditional similarity preserving network for unsupervised cross-modal hashing. arXiv preprint arXiv: 2012.13538 (2021)..
[13] J. Zhou, G. Ding, Y. Guo, Latent semantic sparse hashing for cross-modal similarity search, in: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2014, pp. 415–424.
[14] G. Ding, Y. Guo, J. Zhou, Y. Gao, Large-scale cross-modality search via collective matrix factorization hashing, IEEE Transactions on Image Processing 25 (11) (2016) 5427–5440.
[15] Y. Fang, H. Zhang, Y. Ren, Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing, Knowledge Based Systems 171 (2019) 69–80.
[16] D. Yang, D. Wu, W. Zhang, H. Zhang, B. Li, Deep semantic-alignment hashing for unsupervised cross-modal retrieval, in: Proceedings of International Conference on Multimedia Retrieval, 2020, pp. 44–52.
[17] L. Wang, J. Yang, M. Zareapoor, Z. Zheng, Cluster-wise unsupervised hashing for cross-modal similarity search, Pattern Recognition 111 (2021) 107732.
[18] J. Xu, J. Han, F. Nie, X. Li, Re-weighted discriminatively embedded k-means for multi-view clustering, IEEE Transactions on Image Processing 26 (6) (2017) 3016–3027.
[19] N. Srivastava, R.R. Salakhutdinov, Multimodal learning with deep Boltzmann machines, in: Proceedings of 28th International Conference on Neural Information Processing Systems, 2012, pp. 2949–2980.
[20] R.R. Salakhutdinov, G.E. Hinton, Deep Boltzmann machines, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2009, pp. 448–455.
[21] D. Hu, F. Nie, X. Li, Deep binary reconstruction for cross-modal hashing, IEEE Transactions on Multimedia 21 (4) (2019) 973–985.
[22] H. Liu, R. Ji, Y. Wu, F. Huang, B. Zhang, Cross-modality binary code learning via fusion similarity hashing, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7380–7388.
[23] T. Hoang, T. Do, T.V. Nguyen, N. Cheung, Unsupervised deep cross-modality spectral hashing, IEEE Transactions on Image Processing 29 (2020) 8391–8406.

[24] Y. Peng, J. Chi, Unsupervised cross-media retrieval using domain adaptation with scene graph, IEEE Transactions on Circuits and Systems for Video Technology 30 (11) (2020) 4368–4379.
[25] J. Huang, C. Min, L. Jing, Unsupervised deep fusion cross-modal hashing, in: Proceedings of International Conference on Multimodal Interaction, 2019, pp. 358–366.
[26] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: Proceedings of International Conference on Machine Learning, 2013, pp. 1247–1255.
[27] N. Rasiwasia, D. Mahajan, V. Mahadevan, G. Aggarwal, Cluster canonical correlation analysis, in: Proceedings of International Conference on Artificial Intelligence and Statistics, 2014, pp. 823–831.
[28] Y. Gong, Q. Ke, M. Isard, S. Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, Proceedings of International Journal of Computer Vision 106 (2) (2014) 210–233.
[29] Y. Gong, S. Lazebnik, Iterative quantization: a procrustean approach to learning binary codes, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2011, pp. 814–824.
[30] T. Baltrusaitis, C. Ahuja, L. Morency, Multimodal machine learning: a survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2) (2017) 423–443.
[31] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 321–377.
[32] H.J. Kim, N. Adluru, M. Banerjee, B.C. Vemuri, V. Singh, Interpolation on the manifold of K component GMMs, in: Proceedings of IEEE International Conference on Computer Vision, 2015.
[33] B. Jiang, Z. Li, H. Chen, A.G. Cohn, Latent topic text representation learning on statistical manifolds, IEEE Transactions on Neural Networks and Learning Systems 29 (11) (2018) 5643–5654.
[34] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013)..
[35] Jeffrey Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.
[36] A. Dempster, N. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society 39 (1) (1977) 1–38.
[37] S. Amari, Information Geometry and Its Applications, Applied Mathematical Sciences, Springer, Japan, 2016..
[38] K.M. Carter, Dimensionality Reduction on Statistical Manifolds, ProQucst, American, 2009.
[39] L. Maaten, Learning discriminative fisher kernels, in: Proceedings of International Conference on Machine Learning, 2011, pp. 217–224.
[40] B. Zhang, Machine Learning on Statistical Manifold, HMC Senior Theses, American, 2017.
[41] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proceedings of International Conference for Learning Representations, 2014.
[42] Y. Zeng, T. Yang, P. Breheny, Hybrid safe-strong rules for efficient optimization in lasso-type problems, Computational Statistics and Data Analysis 153 (2021) 107063.
[43] J.C. Pereira, E. Coviello, G. Doyle, On the role of correlation and abstraction in cross-modal multimedia retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (3) (2014) 521–535.
[44] M.J. Huiskes, M.S. Lew, The Mir flickr retrieval evaluation, in: Proceedings of ACM International Conference on Multimedia Information Retrieval, 2008, pp. 39–43.
[45] M. Grubinger, P. Clough, H. Muller, T. Deselaers, The IAPR TC12 benchmark: a new evaluation resource for visual information systems, in: International Conference on Language Resources and Evaluation, 2006.
[46] Y. Xin, J. Chen, X. Jia, X. Wang, Evaluation of class overlap measures on imbalance data classification, Journal of Data Acquisition and Processing 33 (5) (2018) 936–944.
[47] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of Conference on Neural Information Processing Systems, 2012, pp. 1106–1114.

Jieting Luo received the B.S. degree in Computer Science and Technology from Guangdong University of Foreign Study, in 2018. He is currently pursuing an M.S. degree in Department of Computer Science and Engineering of South China University of Technology, China. His current research interests include machine learning and cross-modal retrieval.

Yan Wo received the M.S. degree in Computer Science from Lanzhou University, in 1999, and a ph.D. degree in Computer Science from South China University of Technology in 2004. She is now a Professor of Department of Computer Science and Engineering of South China University of Technology, China. Her current research interests are in the fields of image processing, information security, and pattern recognition.

Bicheng Wu received a B.S degree in Computer Science and Technology from South China Normal University, in 2020. He is currently pursuing an M.S. degree in Department of Computer Science and Engineering of South China University of Technology, China. His current research interests are in the fields of deep learning and multi-modal learning.

Guoqiang Han received the M.S. degree in Computer Science from Sun Yat-sen University, in 1985, and a Ph. D. degree in Computer Science from Sun Yat-sen University in 1988. He is now a Professor of Department of Computer Science and Engineering of South China University of Technology, China. His current research interests include multimedia, computational intelligence, machine learning and computer graphics.