OMGH: Online Manifold-Guided Hashing for Flexible Cross-Modal Retrieval

Xin Liu, *Senior Member, IEEE*, Jinhan Yi, Yiu-ming Cheung, *Fellow, IEEE*, Xing Xu, *Member, IEEE*, and Zhen Cui, *Member, IEEE*

***Abstract*—Cross-modal hashing has recently gained an increasing attention for its efficiency and fast retrieval speed in indexing the multimedia data across different modalities. Nevertheless, the multimedia data points often emerge in a streaming manner, and existing online methods often lack of learning capacity to handle both labeled and unlabeled data. To alleviate these concerns, this paper proposes an Online Manifold-Guided Hashing (OMGH) framework, which can incrementally learn the compact hash code of streaming data while adaptively optimizing the hash function in a streaming manner. To be specific, OMGH first exploits a matrix tri-factorization framework to learn the discriminative hash codes for streaming multi-modal data. Then, an online anchor-based manifold structure is designed to sparsely represent the old data and adaptively guide the hash code learning process, which can well reduce the complexity in preserving the semantic correlation between the old data and streaming data. Meanwhile, such anchor-based manifold embedding is adaptive to the unsupervised and supervised learning strategies in a flexible way. Besides, an online discrete optimization method is efficiently addressed to incrementally update the hash functions and optimize the hash codes on streaming data points. As a result, the derived hash codes are more semantically meaningful for various online cross-modal retrieval tasks. Extensive experiments verify the advantages of the proposed OMGH model, by achieving and improving the state-of-the-art cross-modal retrieval performances on three benchmark datasets.**

***Index Terms*—Cross-modal hashing, streaming data, Anchor-based manifold structure, online discrete optimization.**

Manuscript received 30 July 2021; revised 2 January 2022; accepted 7 April 2022. Date of publication 12 April 2022; date of current version 8 September 2023. This work was supported in part by the Open Project of Zhejiang Lab under Grant 2021KH0AB01, in part by the National Science Foundation of China under Grants 61673185, 61672444, and 61976049, in part by NSFC/RGC Joint Research Scheme under Grant N_HKBU214/21, in part by RGC General Research Fund under Grant RGC/HKBU/12201321, in part by Hong Kong Baptist University under Grants RC-FNRA-IG/18-19/SCI/03 and RC-IRCMs/18-19/SCI/01, in part by the Innovation and Technology Fund of Innovation and Technology Commission of the Government of the Hong Kong SAR under Grant ITS/339/18, in part by the National Science Foundation of Fujian Province under Grant 2020J01084, and in part by the Natural Science Foundation of Shandong Province under Grant ZR2020LZH008. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. X. Li. *(Corresponding author: Xin Liu.)*

Xin Liu is with the Department of Computer Science, Huaqiao University, Xiamen, Fujian 361021, China, and also with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China (e-mail: xliu@hqu.edu.cn).

Jinhan Yi is with the Xiamen Key Laboratory of Computer Vision and Pattern Recognition and Fujian Key Laboratory of Big Data Intelligence and Security, Huaqiao University, Xiamen, Fujian 361021, China (e-mail: jhyi@stu.hqu.edu.cn).

Yiu-ming Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, Hong Kong (e-mail: ymc@comp.hkbu.edu.hk).

Xing Xu is with the Center for Future Multimedia and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China (e-mail: xing.xu@uestc.edu.cn).

Zhen Cui is with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China (e-mail: zhen.cui@njust.edu.cn).

Digital Object Identifier 10.1109/TMM.2022.3166668

I. Introduction

With the tremendous explosion of multimedia data, recent years have heightened the need of cross-modal retrieval techniques for scalable similarity search in many real applications. More specifically, a user can utilize a query item of one modality to retrieve the semantically relevant items in another different modality, e.g., users can find images that best illustrate the topic of a textual query, or textual descriptions that best explain the content of a visual query [1]. In recent years, cross-modal hashing [2] is gaining significant popularity due to its extremely low storage cost and high retrieval efficiency, which aims to transform the high-dimensional real-valued examples into compact binary codes while preserving the semantical similarity in the original feature space.

In the past few years, various kinds of cross-modal hashing methods have been proposed with impressive performance [3]. Nevertheless, most existing cross-modal hashing methods intuitively adopt the offline learning mechanism, which, inevitably, requiring the whole training data to be available before training [2], [4]. In practice, multimedia data usually continuously arrive in a streaming fashion. For instance, the popular social media websites uninterruptedly upload massive amounts of data every day, which are highly dynamic and frequently updated. Under such circumstances, these offline methods have to accumulate all the data samples to retrain the hash functions and recompute the hash codes of the entire data points, which, inevitably, result in a great deal of computational burden. Besides, if the accumulated training dataset is very large, it is impractical to load all data into memory for hash function learning. Therefore, these offline methods are unadaptable and inefficient to these frequently updated multimedia database.

Advanced hashing technique is essential in processing frequently changed media data. To be specific, online cross-modal hashing aims to incrementally update hash functions from sequentially arriving multi-modal data, and simultaneously encode streaming data into compact binary codes. Since hash

functions are updated only by newly arriving data stream, online cross-modal hashing significantly reduces the computation cost and memory requirement. Inspired by these merits, online cross-modal hashing receives great attention very recently and some works have been developed [5]–[7], in either supervised fashion where the labels are provided or unsupervised manner where the labels are unavailable. Nevertheless, learning a good hash function that can well reflect the semantic relevance between the streaming data and existing data (i.e., old data) is the bottleneck problem in online cross-modal retrieval, and the existing methods still face three challenges: 1) The modality heterogeneity leads to a great challenge in directly measuring the semantic relevance among the streaming multi-modal data; 2) Most online methods mainly consider the manifold structure within the new data to preserve the semantic correlation across different modalities, which often weaken the semantic correlation between the newly arrived data and existing data. Consequently, the learning models derived from these methods may induce the 'semantic forgetting' problem and performance degradation; 3) Current online cross-modal hashing methods generally lack of the efficiency and flexibility to work with both labeled and unlabeled multi-modal data simultaneously.

To address the aforementioned challenges, this paper proposes an online manifold-guided hashing (OMGH) for flexible and efficient cross-modal retrieval, which can incrementally learn the hash code of new coming data while optimizing the hash function in a streaming manner. Within the proposed OMGH framework, an online anchor-based manifold structure is flexibly embedded to guide the hash code learning process, while significantly reducing the computational complexity in preserving the semantic correlations between the streaming data and old data. Meanwhile, the label information can be selectively embedded into the proposed online learning framework, which can well adapt to the variations of data stream by adaptively training in an unsupervised or supervised manner. The main contributions are summarized as follows:

- An efficient online manifold-guided hashing framework is newly proposed to benefit cross-modal retrieval for streaming multi-modal data, which can significantly reduce the computational load and memory storage.
- An online anchor-based manifold embedding is flexibly proposed to guide the hash code learning process, which can well preserve the semantic correlations between the newly coming data and existing data, while being adaptive to the unsupervised and supervised scenarios.
- An online discrete optimization method is efficiently addressed to incrementally update the hash functions and optimize the hash codes on streaming data points.
- Extensive experiments on public benchmarks highlight the advantages of OMGH under various retrieval tasks and show its outstanding performances.

The rest of this paper is organized as follows: Section II surveys the face-voice association works, and Section III the proposed learning framework in detail. The extensive experiments and comparisons are introduced in Section IV, and we draw a conclusion in Section V.

II. Related Work

Cross-modal hashing has recently received wide attention due to its effectiveness in improving query speed and reducing memory cost. It is noted that the recent multi-modal hashing [8] is designed for multimedia search when multi-modal features are all provided at the query stage, while cross-modal hashing aims to retrieve the most relevant objects represented by other modalities for a given query characterized by one modality. In the following, we mainly survey the cross-modal hashing works, which can be generally divided into offline learning and online learning branches.

A. Offline Cross-Modal Hashing

The offline cross-modal hashing assumes that all the training data points are available before the hash function training process. In recent years, various cross-modal hashing attempts have been proposed, mostly in either unsupervised fashion where the labels are unavailable, or supervised fashion where the labels are explicitly provided. Unsupervised cross-modal hashing intuitively learns the hash codes from the original feature space to Hamming space. Along this line, cross-view hashing (CVH) [9] first extends the spectral hashing method from single-view to cross-view case, and then learns the hash codes from the paired training data to preserve the similarity across different modalities. Similarly, inter-media hashing (IMH) [10] first utilizes the inter-view and intra-view consistency to obtain a common hamming space, and then selects the linear regression to generate the hash codes. Besides, collective matrix factorization hashing (CMFH) [11], [12] employs the joint matrix factorization to learn the semantically correlated hash codes for multi-modal data, while the latent semantic sparse hashing (LSSH) [13] utilizes sparse coding to extract latent semantic features and quantizes such latent semantic features into hash code. Recently, fusion similarity hashing (FSH) [14] learns the semantically correlated hash codes from the fused similarity to achieve cross-modal retrieval. Although these methods are able to capture the semantic correlations between the heterogeneous modalities, the hash codes learned in an unsupervised way are not discriminative enough and their retrieval performances need further improvements.

Supervised cross-modal hashing primarily leverages the semantic label supervision to guide the hash code learning, which can well mitigate the semantic gap between heterogeneous modalities. Along this way, semantic correlation maximization (SCM) [15] preserves the label similarity to learn the hash codes while supervised matrix factorization hashing (SMFH) [16] embeds the label supervision to perform collective matrix factorization hashing. In addition, semantic preserved hashing (SePH) [17] and generalized semantic preserving hashing (GSePH) [18] both construct an affinity matrix via label information to approximate hash codes. To reduce the quantitative losses, discrete cross-modal hashing (DCH) [19] generates hash codes in a bit by bit manner from the semantic labels. Besides, recent deep cross-modal hashing works [20], [21] attempt to combine the advanced feature representation learning with hash code learning in an integrated way. It is noted that

these methods train the hash functions on all the accumulated data and learn the hash codes of the entire data points, which inevitably involve high computational complexity and memory costs when training large scale data. In case where the media data points continuously arrive in a streaming fashion, these methods learned in offline way need to recalculate the hash functions on the whole database, which are computationally inefficient.

B. Online Cross-Modal Hashing

Online hashing algorithms incrementally learn the hash function by processing the streaming data in a sequential order, which can avoid the high computational burden and memory costs when processing large-scale datasets. In the literature, the pioneer online hashing efforts mainly focus on single modality [22]–[24], and these methods cannot be directly extended to cross-modal retrieval scenarios. Specifically, online multi-modal hashing [25] trains the hash model in a batch-based mode and supports online hashing on query stage. Nevertheless, multi-modal hashing is designed for multimedia search when multi-modal features are all provided at the query stage, which cannot support the retrieval across different modalities. In contrast to this, cross-modal hashing aims to retrieve the most relevant objects represented by other different modalities. Therefore, online multi-modal hashing cannot be directly utilized for online cross-modal retrieval scenarios.

Online cross-modal hashing is designed for supporting the efficient search across streaming multi-modal database, and only a few online cross-modal hashing methods are proposed to process the streaming multi-modal data. Specifically, online cross-modal hashing (OCMH) [5] first decomposes the hash code matrix into a shared latent code matrix and a transition matrix, and further utilizes the dynamic transfer matrix to incrementally update the hash codes of new data. Later, online collective matrix factorization hashing (OCMFH) [6] learns the hash codes for streaming data by collective matrix factorization in an online optimization scheme. It is noted that these two methods are unsupervised learning methods, and their derived hash codes are not discriminative enough for high retrieval performance. Remarkably, the semantic labels have demonstrated to be very useful on enhancing the discriminative capability and thus significantly improve the retrieval performance. Accordingly, online latent semantic hashing (OLSH) [7] maps the discrete labels into a continuous latent semantic space and utilizes the newly coming data points to retrain the hash functions. With the supervision of semantic labels, this supervised cross-modal hashing method has achieve impressive performance. Later, online adaptive supervised hashing (OASH) [26] regresses the class labels to binary hash codes and learn the hash functions in an online optimization scheme, while label embedding online hashing (LEMON) [27] builds a label embedding framework to produce the discriminative binary code and reduce computational complexity. Although these supervised cross-modal hashing methods have delivered very promising performance, they still suffer from the 'semantic forgetting' problem. To be specific, the semantic correlations between the streaming data and old data are not well preserved such that the relevant online retrieval performances are not competitive. Meanwhile, these supervised online methods still lack of learning capability to preserve the semantic correlations between the streaming data and old data, and these framework cannot simultaneously process the unlabelled newly coming data. Therefore, it is still desirable to study a flexible online cross-modal hashing technique that is capable of processing different kinds of multi-modal data in a variety of scenarios.

III. Online Manifold-Guided Hashing

This section describes the proposed OMGH framework in detail. Without loss of generality, as shown in Fig. 1, this section mainly focuses on online cross-modal hashing with only two modalities (i.e., image and text), and the proposed framework can be easily extended to more modalities.

A. Proposed OMGH Methodology

1) Cross-Modal Semantic Analysis: The objective of cross-modal retrieval is to obtain semantically relevant data samples in one modality for a query in another different modality. Suppose that the training database consists of image data $\mathbf{X}_1 \in \mathbb{R}^{d_1 \times N}$ and text data $\mathbf{X}_2 \in \mathbb{R}^{d_2 \times N}$, d_1, d_2 are the feature length and N is the training number, an intuitive way is to project these heterogeneous data into a common a common latent subspace, formally $\mathbf{X}_1 \rightarrow \mathbf{U}_1\mathbf{V}$, $\mathbf{X}_2 \rightarrow \mathbf{U}_2\mathbf{V}$, $\mathbf{U}_1 \in \mathbb{R}^{d_1 \times k}$, $\mathbf{U}_2 \in \mathbb{R}^{d_2 \times k}$. Then, the hash codes can be generated by quantizing $\mathbf{V} \in \mathbb{R}^{k \times N}$ from real values to $\{-1, 1\}$ by $\mathtt{sign}(\mathbf{V})$. Although the correlations between different modalities can be well connected via the common latent semantic representation, such rigid assumption may not discriminatively characterize the heterogeneous data of different modalities due to their different physical meaning, dimensionality and statistical properties. To alleviate this concern, we relax this assumption and assume that each modality in an instance generates similar, not exactly identical latent semantic subspace, $\mathbf{V}_1 \in \mathbb{R}^{k_1 \times N}$ for image and $\mathbf{V}_2 \in \mathbb{R}^{k_2 \times N}$ for text (in general $k_1 \neq k_2$), featuring on discriminative modality-specific representations.

For cross-modal hashing, the semantic representations of image and text modalities need to be further projected into a common Hamming space [28]. Thus, we assume that the projected representations $\mathbf{M}_1^{\mathrm{T}}\mathbf{V}_1$ and $\mathbf{M}_2^{\mathrm{T}}\mathbf{V}_2$ of two modalities can produce the same instance in the common Hamming space, i.e., $\mathbf{M}_1^{\mathrm{T}}\mathbf{V}_1 \rightarrow \mathbf{B}$ and $\mathbf{M}_2^{\mathrm{T}}\mathbf{V}_2 \rightarrow \mathbf{B}$, $\mathbf{M}_1 \in \mathbb{R}^{k_1 \times r}$, $\mathbf{M}_2 \in \mathbb{R}^{k_2 \times r}$, with objective functions written as $\min_{\mathbf{M}_1} \|\mathbf{B} - \mathbf{M}_1^{\mathrm{T}}\mathbf{V}_1\|_F^2$ and $\min_{\mathbf{M}_2} \|\mathbf{B} - \mathbf{M}_2^{\mathrm{T}}\mathbf{V}_2\|_F^2$, where $\|\cdot\|_F^2$ is the Frobenius norm of a matrix and r is the hash code length. It is noted that these two functions are the ordinary least square regression problems, which regress $\mathbf{V}_1$ or $\mathbf{V}_2$ to $\mathbf{B}$. As indicated in work [29], it is equivalent to change the regressing target with the least square regression formulation. Inspired by this finding, we propose to regress $\mathbf{B}$ to $\mathbf{V}_1$ or $\mathbf{V}_2$, i.e., $\min_{\mathbf{M}_1} \|\mathbf{V}_1 - \mathbf{M}_1\mathbf{B}\|_F^2$ and $\min_{\mathbf{M}_2} \|\mathbf{V}_2 - \mathbf{M}_2\mathbf{B}\|_F^2$, simplified as $\mathbf{M}_1\mathbf{B} \rightarrow \mathbf{V}_1$, $\mathbf{M}_2\mathbf{B} \rightarrow \mathbf{V}_2$, and the following matrix tri-factorization framework can be

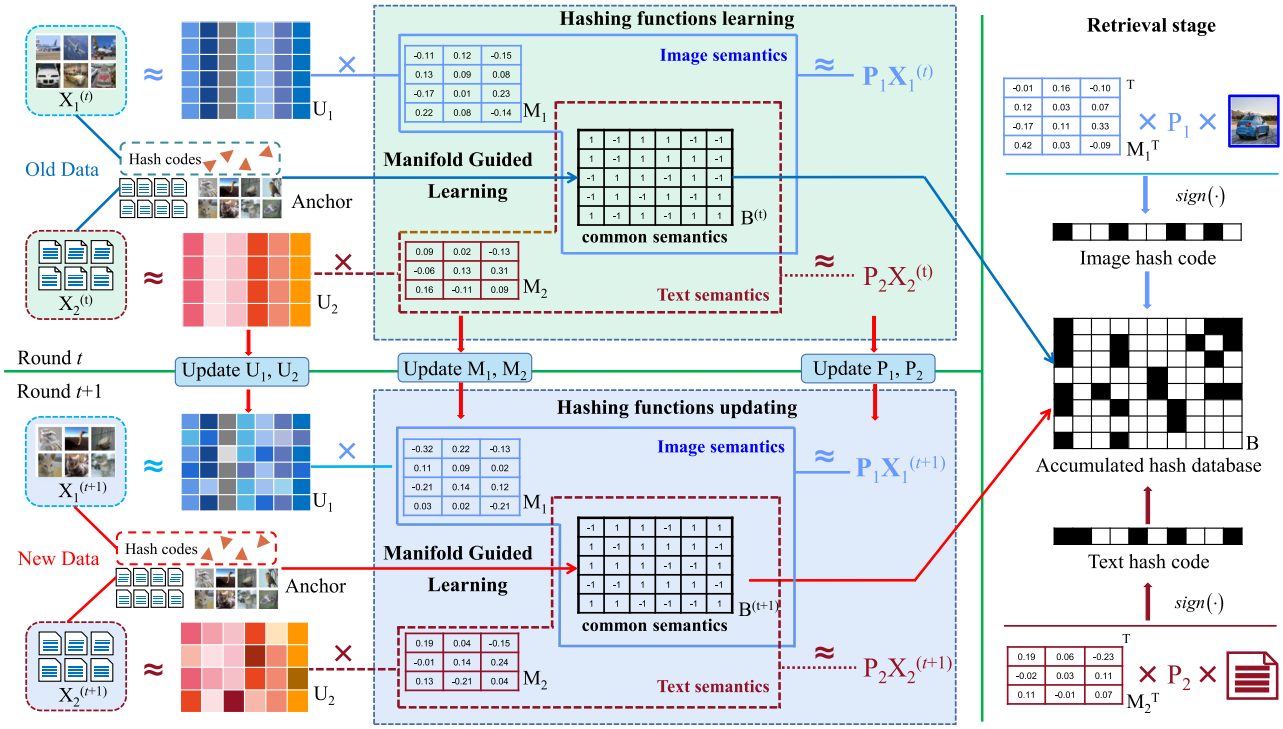


Fig. 1. The schematic pipeline of the proposed OMGH framework.

obtained:

$$\begin{aligned} \min \quad & \alpha||\mathbf{X}_1 - \mathbf{U}_1\mathbf{M}_1\mathbf{B}||_F^2 + (1-\alpha)||\mathbf{X}_2 - \mathbf{U}_2\mathbf{M}_2\mathbf{B}||_F^2 \\ & + \gamma R(\mathbf{U_1}, \mathbf{U_2}, \mathbf{M}_1, \mathbf{M}_2) \\ \text{s.t.} \quad & \mathbf{B} \in \{-1, 1\}^{r\times N}, \end{aligned} \tag{1}$$

where α is a trade-off parameter, $R(\cdot) = \|\cdot\|_F^2$ is the regularization term utilized to avoid overfitting, and γ is a penalty parameter. For cross-modal hashing, it is necessary to learn the modality-specific hash functions for out-of-sample instance. Similar to work [11], we assume that the original image and text features can be mapped into the latent semantic space by two projections, respectively, formulated as $\mathbf{P}_1\mathbf{X}_1 \to \mathbf{V}_1$ and $\mathbf{P}_2\mathbf{X}_2 \to \mathbf{V}_2$. The overall objective function, consisting of the collective matrix tri-factorization term, regression term and the projection term, is given as follows:

$$\begin{aligned} \min \quad & \alpha||\mathbf{X}_1 - \mathbf{U}_1\mathbf{M}_1\mathbf{B}||_F^2 + (1-\alpha)||\mathbf{X}_2 - \mathbf{U}_2\mathbf{M}_2\mathbf{B}||_F^2 \\ & + \mu(||\mathbf{P}_1\mathbf{X}_1 - \mathbf{M}_1\mathbf{B}||_F^2 + ||\mathbf{P}_2\mathbf{X}_2 - \mathbf{M}_2\mathbf{B}||_F^2) \\ & + \gamma R(\mathbf{U_1}, \mathbf{U_2}, \mathbf{M}_1, \mathbf{M}_2, \mathbf{P_1}, \mathbf{P_2}) \\ \text{s.t.} \quad & \mathbf{B} \in \{-1, 1\}^{r\times N}, \end{aligned} \tag{2}$$

where μ is a trade-off parameter.

2) Online Notation and Problem Formulation: Suppose that the training database consists of multiple streaming image-text data pairs. At each round t, a new data chunk $[\mathbf{X}_1^{(t)}, \mathbf{X}_2^{(t)}]$ of size N_t is added into the database, where $\mathbf{X}_1^{(t)} \in \mathbb{R}^{d_1\times N_t}$ and $\mathbf{X}_2^{(t)} \in \mathbb{R}^{d_2\times N_t}$, respectively, denote the feature matrices of newly coming image and text data. Let $\mathbf{Y}^{(t)} \in \{0,1\}^{c\times N_t}$ denote the semantic label of new data chunk and c be the number of all categories, the accumulated training dataset, consisting of old data and new data, can be formulated as $[\tilde{\mathbf{X}}_m^{(t-1)}, \mathbf{X}_m^{(t)}]$, where $\tilde{\mathbf{X}}_m^{(t-1)} \in \mathbb{R}^{d_m\times(N-N_t)}$ represents the accumulated old data before round t, and m is the index of different modalities, $m = 1, 2$. Similarly, $\tilde{\mathbf{Y}}^{(t-1)} \in \mathbb{R}^{c\times(N-N_t)}$ represents the label matrix of the accumulated old data. Accordingly, the hash code of accumulated data is written as $[\tilde{\mathbf{B}}^{(t-1)}, \mathbf{B}^{(t)}]$, where $\tilde{\mathbf{B}}^{(t-1)} \in \{-1, 1\}^{r\times(N-N_t)}$ is the hash code matrix of the accumulated old data, and $\mathbf{B}^{(t)} \in \{-1, 1\}^{r\times N_t}$ represents the hash code matrix of the new data pair arriving at round t.

3) Online Anchor-Based Manifold Embedding: The data across different modalities are inherently heterogeneous due to different physical representations. Therefore, it is beneficial for a retrieval model to preserve manifold structure that embedded in multimedia data, while at the same time preserving the neighborhood relationship to ensure semantic consistency. In recent years, Memory Block Prototype [30], Cross-Modal Prototypes [31] and Prototype-based Adaptive Network [32] are proposed to preserve the manifold structure across different modalities. Nevertheless, these methods cannot be directly applied to cross-modal hashing process. For hash code learning, IMH [10] and FSH [14] first compute an affinity matrix $\mathbf{S}$ between different samples, and then utilize $\mathbf{S}$ to guide the hash code learning. Nevertheless, the affinity matrix $\mathbf{S}$ derived from IMH and FSH is designed for the whole training dataset, and such mechanism is only adaptive to the offline cross-modal hashing works. Differently, online learning only attempts a limited number of samples at each round, and utilizes the newly arriving data

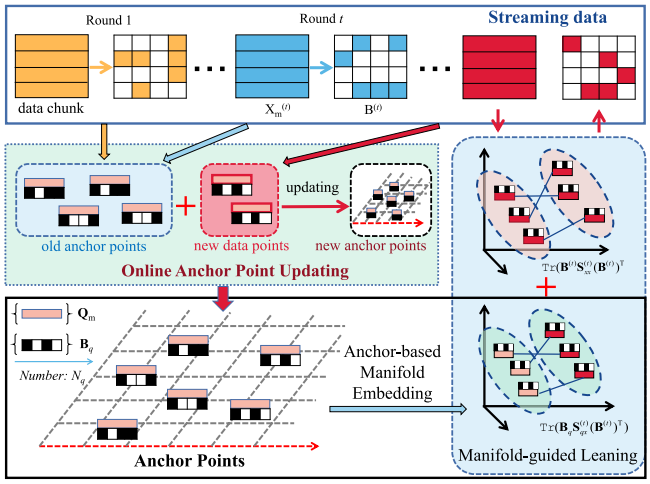


Fig. 2. The graphical illustration of the proposed anchor-based manifold embedding mechanism.

for adaptive learning. In particular, FOMH [25] first calculates the similarity of the new data $\mathbf{X}^{(t)}$ via the label information, and then embeds it into the hash code learning by the regularization $(\mathbf{B}^{(t)})^{\mathrm{T}}\mathbf{B}^{(t)} \rightarrow r\mathbf{S}^{(t)}$, featuring on preserving the semantic correlation of new samples. However, this method often weakens the semantic correlation between the new data and old samples, which may induce the 'semantic forgetting' problem and degrade its retrieval performance for the new coming data. To tackle this problem, we propose an online anchor-based manifold embedding to adaptively preserve the manifold structure that embedded in multimedia data, while enhancing the ability to preserve the semantic correlation between the new data and old samples.

As shown in Fig. 2, the data chunk $[\mathbf{Q}_1, \mathbf{Q}_2]$ of size N_q and its corresponding hash codes $\mathbf{B}_q$ are selected as the anchor points, which are heuristically chosen from the old samples in the memory. Note that, these anchors are the sparse representation of old data, which could reduce the computational complexity in measuring the semantic correlation between the old data and new data. To preserve the semantic consistency of data points across different modalities, the manifold regularization is often utilized to maintain the neighboring relationships during the hash code learning process. Given a normalized affinity matrix $\mathbf{S}^{(t)}$, the manifold regularization $\frac{1}{2}\sum_{i=1}^{N_t}\sum_{j=1}^{N_t}\|b_i^{(t)} - b_j^{(t)}\|_F^2\mathbf{S}_{ij}^{(t)} = -\mathtt{Tr}(\mathbf{B}^{(t)}\mathbf{S}^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}})$ is often utilized to preserve the neighboring relationships and semantic consistency among the new data samples [33].

Remarkably, the proposed anchor-based manifold embedding aims not only to preserve the neighboring relationships among the data points, but also to establish an internal semantic relationship between the hash codes of new data and old data. To be specific, when learning the hash code $\mathbf{B}^{(t)}$ of new data at each round t, the proposed framework not only preserves the correlation of new data embedded in itself (i.e., $\mathbf{S}_{xx}^{(t)} \rightarrow \mathcal{S}(\mathbf{X}_m^{(t)}, \mathbf{X}_m^{(t)}), m = 1, 2$), but also considers the semantic correlation between the anchor data and new data (i.e., $\mathbf{S}_{qx}^{(t)} \rightarrow \mathcal{S}(\mathbf{Q}_m, \mathbf{X}_m^{(t)})$). As shown in Fig. 2, these two manifold embeddings enable the online learning model to preserve the intra-modal and inter-modal manifold structure during the online updating process. By normalizing these two affinity matrices, the following objective function is obtained:

$$\begin{aligned} \min \quad & -\left(\mathtt{Tr}(\mathbf{B}^{(t)}\mathbf{S}_{xx}^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}}) + \mathtt{Tr}(\mathbf{B}_q\mathbf{S}_{qx}^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}})\right) \\ \text{s.t.} \quad & \mathbf{B}^{(t)} \in \{-1, 1\}^{r\times N_t}. \end{aligned} \tag{3}$$

For practical application, the designed framework also needs to provide a flexible online cross-modal hashing work, which can handle both supervised learning scenario and unsupervised learning scenario. To this end, we construct the formulation of $\mathbf{S}_{xx}^{(t)}$ and $\mathbf{S}_{qx}^{(t)}$ in either supervised case or unsupervised cases. In the following, we introduce the formulation $\mathbf{S}_{qx}^{(t)}$ in detail, and the $\mathbf{S}_{xx}^{(t)}$ can be obtained in the similar way.

Supervised scenario (OMGH-su): If the label information of data is provided, $\mathbf{S}_{qx}$ can be obtained directly from the cosine similarity of semantic label information. To enlarge the margin between the similar pair and dissimilar pair, we further regularize the similarity of similar examples as the positive direction and dissimilar examples as the negative direction, specifically formulated as follows:

$$\mathbf{S}_{qx}^{(t)}(i,j) = \begin{cases} \frac{l_q^{(i)}\cdot l_x^{(j)}}{\|l_q^{(i)}\|\,\|l_x^{(j)}\|} & \text{if } l_q^{(i)}\cdot l_x^{(j)} \neq 0 \\ -1 & \text{otherwise} \end{cases}, \tag{4}$$

where $l_q^{(i)}, l_x^{(j)} \in \{0,1\}^{c\times 1}$ are respectively the label information of the i-th sample in $\mathbf{Q}_m$ and j-th sample in $\mathbf{X}_m$, and $l_q^{(i)}\cdot l_x^{(j)} \neq 0$ means that the i-th sample of $\mathbf{Q}_m$ and j-th sample of $\mathbf{X}_m$ share at least one same category value. Meanwhile, the supervised solution of $\mathbf{S}_{xx}^{(t)}$ can be obtained in the similar way.

Unsupervised scenario (OMGH-un): If the label is unavailable, the manifold structure of one instance can be modeled by a nearest neighbor graph in the instance space. Similarly, we regularize the similarity of similar examples as the positive direction and dissimilar examples as the negative direction, whereby the margin between the similar and dissimilar pairs can be enlarged. For image and text samples, the local similarity is utilized to model the intra-modal similarity:

$$\mathbf{S}_{qx}^{1(t)}(i,j) = \begin{cases} 1 & \text{if } \mathbf{X}_{1,i}^{(t)} \in \mathbf{N}_k(\mathbf{Q}_{1,j}^{(t)}) \text{ or } \mathbf{Q}_{1,j}^{(t)} \in \mathbf{N}_k(\mathbf{X}_{1,i}^{(t)}) \\ -1 & \text{otherwise} \end{cases}, \tag{5}$$

$$\mathbf{S}_{qx}^{2(t)}(i,j) = \begin{cases} 1 & \text{if } \mathbf{X}_{2,i}^{(t)} \in \mathbf{N}_k(\mathbf{Q}_{2,j}^{(t)}) \text{ or } \mathbf{Q}_{2,j}^{(t)} \in \mathbf{N}_k(\mathbf{X}_{2,i}^{(t)}) \\ -1 & \text{otherwise} \end{cases}, \tag{6}$$

where $\mathbf{N}_k(\cdot)$ is the top-k nearest neighbor set. Consequently, we fuse these two similarities as $\mathbf{S}_{qx}^{(t)} = \alpha\mathbf{S}_{qx}^{1(t)} + (1-\alpha)\mathbf{S}_{qx}^{2(t)}$ to jointly exploit the semantic correlation between anchor data and new data, where α is a trade-off parameter as illustrated in (2). Meanwhile, the unsupervised solution of $\mathbf{S}_{xx}^{(t)}$ can be obtained in the similar way.

4) Overall Objective Function: The process of learning the discriminative hash representations and modality-specific hash functions can be conducted by jointly optimize the objective function illustrated in cross-modal semantic analysis and anchor-based manifold embedding. By integrating the online learning and anchor-based manifold embedding, the final objective function is formulated as:

$$\min_{\mathbf{U}_1,\mathbf{U}_2,\mathbf{M}_1,\mathbf{M}_2,\mathbf{P}_1,\mathbf{P}_2,\mathbf{B}^{(t)}} \mathcal{G}(\mathbf{U_1},\mathbf{U_2},\mathbf{M}_1,\mathbf{M}_2,\mathbf{P_1},\mathbf{P_2},\mathbf{B}^{(t)}) \tag{7}$$

where

$$\begin{aligned}\mathcal{G} = {} & \alpha||\mathbf{X}_1^{(t)}-\mathbf{U}_1\mathbf{M}_1\mathbf{B}^{(t)}||_F^2+(1-\alpha)||\mathbf{X}_2^{(t)}-\mathbf{U}_2\mathbf{M}_2\mathbf{B}^{(t)}||_F^2\\ & -\lambda\left(\mathtt{Tr}(\mathbf{B}^{(t)}\mathbf{S}_{xx}^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}})+\mathtt{Tr}(\mathbf{B}_q\mathbf{S}_{qx}^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}})\right)\\ & +\mu(||\mathbf{P}_1\mathbf{X}_1^{(t)}-\mathbf{M}_1\mathbf{B}^{(t)}||_F^2+||\mathbf{P}_2\mathbf{X}_2^{(t)}-\mathbf{M}_2\mathbf{B}^{(t)}||_F^2)\\ & +\gamma R(\mathbf{U_1},\mathbf{U_2},\mathbf{M}_1,\mathbf{M}_2,\mathbf{P_1},\mathbf{P_2})\\ \text{s.t. } & \mathbf{B}^{(t)}\in\{-1,1\}^{r\times N_t}.\end{aligned} \tag{8}$$

B. Online Discrete Optimization

The main objective of the proposed OMGH framework is to sequentially process the arriving data chunks, while incrementally updating the hash function and producing high-quality hash codes. That is, if the new data chunk $[\mathbf{X}_1^{(t)},\mathbf{X}_2^{(t)}]$ is accumlated at round t, the online representation of the objective function (8) can be expressed as:

$$\begin{aligned}\min\ \mathcal{G}^{(t)} = {} & \min\mathcal{G}^{(t-1)}+\alpha||\mathbf{X}_1^{(t)}-\mathbf{U}_1\mathbf{M}_1\mathbf{B}^{(t)}||_F^2\\ & +(1-\alpha)||\mathbf{X}_2^{(t)}-\mathbf{U}_2\mathbf{M}_2\mathbf{B}^{(t)}||_F^2\\ & -\lambda\left(\mathtt{Tr}(\mathbf{B}^{(t)}\mathbf{S}_{xx}^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}})+\mathtt{Tr}(\mathbf{B}_q\mathbf{S}_{qx}^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}})\right)\\ & +\mu(||\mathbf{P}_1\mathbf{X}_1^{(t)}-\mathbf{M}_1\mathbf{B}^{(t)}||_F^2+||\mathbf{P}_2\mathbf{X}_2^{(t)}\\ & -\mathbf{M}_2\mathbf{B}^{(t)}||_F^2)+\gamma R(\mathbf{U_1},\mathbf{U_2},\mathbf{M}_1,\mathbf{M}_2,\mathbf{P_1},\mathbf{P_2})\\ \text{s.t. } & \mathbf{B}^{(t)}\in\{-1,1\}^{r\times N_t}.\end{aligned} \tag{9}$$

The optimization problem in (9) is non-convex and intractable due to its discrete constraint, and it is very difficult to learn the discrete hash code directly. Fortunately, the objective function is convex to any one variable while fixing the others, and such optimization problem can be handled by using an alternating optimization, i.e., updating one variable while fixing the others until convergence. The detailed online discrete optimization procedure is elaborated as follows:

Update $\mathbf{U}_1$, $\mathbf{U}_2$: Learn $\mathbf{U}_1$ and $\mathbf{U}_2$ by fixing other variables. Since the solution of $\mathbf{U}_2$ is exactly similar to the solution of $\mathbf{U}_1$, we first clarify the detailed steps of updating the $\mathbf{U}_1$ and then give the solution of $\mathbf{U}_2$ directly. Since the projection $\mathbf{U}_1$ is relevant to all accumulated data, (9) can be simplified as:

$$\begin{aligned}\min_{\mathbf{U}_1}\ & \alpha||[\tilde{\mathbf{X}}_1^{(t-1)},\mathbf{X}_1^{(t)}]-\mathbf{U}_1\mathbf{M}_1[\tilde{\mathbf{B}}^{(t-1)},\mathbf{B}^{(t)}]||_F^2+\gamma||\mathbf{U}_1||_F^2\\ \text{s.t. } & \mathbf{B}^{(t)}\in\{-1,1\}^{r\times N_t}.\end{aligned} \tag{10}$$

By setting the derivative of (10) w.r.t $\mathbf{U}_1$ to 0, the analytic solution can be obtained:

$$\mathbf{U}_1=\mathbf{E}_1^{(t)}\mathbf{M}_1^{\mathrm{T}}\left(\mathbf{M}_1\mathbf{H}^{(t)}\mathbf{M}_1^{\mathrm{T}}+\frac{\gamma}{\alpha}\mathbf{I}\right)^{-1}, \tag{11}$$

$$\mathbf{E}_1^{(t)}=[\tilde{\mathbf{X}}_1^{(t-1)},\mathbf{X}_1^{(t)}]\begin{bmatrix}(\tilde{\mathbf{B}}^{(t-1)})^{\mathrm{T}}\\(\mathbf{B}^{(t)})^{\mathrm{T}}\end{bmatrix}=\mathbf{E}_1^{(t-1)}+\mathbf{X}_1^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}}, \tag{12}$$

$$\mathbf{H}^{(t)}=[\tilde{\mathbf{B}}^{(t-1)},\mathbf{B}^{(t)}]\begin{bmatrix}(\tilde{\mathbf{B}}^{(t-1)})^{\mathrm{T}}\\(\mathbf{B}^{(t)})^{\mathrm{T}}\end{bmatrix}=\mathbf{H}^{(t-1)}+\mathbf{B}^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}}, \tag{13}$$

where $\mathbf{E}_1^{(t-1)}=\tilde{\mathbf{X}}_1^{(t-1)}(\tilde{\mathbf{B}}^{(t-1)})^{\mathrm{T}}$, $\mathbf{H}^{(t-1)}=\tilde{\mathbf{B}}_1^{(t-1)}(\tilde{\mathbf{B}}^{(t-1)})^{\mathrm{T}}$, respectively, correspond to the results related to the accumulated old image data and their values can be directly obtained in previous learning round. Therefore, this updating step only needs to calculate $\mathbf{X}_1^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}}$ and $\mathbf{B}^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}}$.

Similar to $\mathbf{U}_1$, the solution expression for $\mathbf{U}_2$ is:

$$\mathbf{U}_2=\mathbf{E}_2^{(t)}\mathbf{M}_2^{\mathrm{T}}\left(\mathbf{M}_2\mathbf{H}^{(t)}\mathbf{M}_2^{\mathrm{T}}+\frac{\gamma}{\alpha}\mathbf{I}\right)^{-1}, \tag{14}$$

where $\mathbf{E}_2^{(t)}=\tilde{\mathbf{X}}_2^{(t-1)}(\tilde{\mathbf{B}}^{(t-1)})^{\mathrm{T}}+\mathbf{X}_2^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}}$, and the item $\tilde{\mathbf{X}}_2^{(t-1)}(\tilde{\mathbf{B}}^{(t-1)})^{\mathrm{T}}$ corresponds to the result related to the accumulated old text data. Similarly, its value can be directly obtained in previous learning round.

Update $\mathbf{M}_1$, $\mathbf{M}_2$: Learn $\mathbf{M}_1$ and $\mathbf{M}_2$ by fixing other irrelevant variables. Similarly, we first clarify the detailed steps of updating the $\mathbf{M}_1$ and then give the solution of $\mathbf{M}_2$ directly. Accordingly, (9) can be simplified as:

$$\begin{aligned}\min_{\mathbf{M}_1}\ & \alpha||[\tilde{\mathbf{X}}_1^{(t-1)},\mathbf{X}_1^{(t)}]-\mathbf{U}_1\mathbf{M}_1[\tilde{\mathbf{B}}^{(t-1)},\mathbf{B}^{(t)}]||_F^2\\ & +\mu||\mathbf{P}_1[\tilde{\mathbf{X}}_1^{(t-1)},\mathbf{X}_1^{(t)}]-\mathbf{M}_1[\tilde{\mathbf{B}}^{(t-1)},\mathbf{B}^{(t)}]||_F^2\\ \text{s.t. } & \mathbf{B}^{(t)}\in\{-1,1\}^{r\times N_t}.\end{aligned} \tag{15}$$

By setting the partial derivative of (15) w.r.t $\mathbf{M}_1$ to 0, the analytic solution can be obtained:

$$\mathbf{M}_1=(\alpha\mathbf{U}_1^{\mathrm{T}}\mathbf{U}_1+\mu\mathbf{I})^{-1}(\alpha\mathbf{U}_1^{\mathrm{T}}\mathbf{E}_1^{(t)}+\mu\mathbf{P}_1\mathbf{E}_1^{(t)})(\mathbf{H}^{(t)})^{-1} \tag{16}$$

Similarly, the solution expression for $\mathbf{M}_2$ is formulated as:

$$\mathbf{M}_2=(\alpha\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_2+\mu\mathbf{I})^{-1}(\alpha\mathbf{U}_2^{\mathrm{T}}\mathbf{E}_2^{(t)}+\mu\mathbf{P}_2\mathbf{E}_2^{(t)})(\mathbf{H}^{(t)})^{-1}. \tag{17}$$

Update $\mathbf{P}_1$, $\mathbf{P}_2$: Learn $\mathbf{P}_1$ and $\mathbf{P}_2$ by fixing other irrelevant variables. Also, we first elaborate the detailed solution of $\mathbf{P}_1$ and then show the solution of $\mathbf{P}_2$ directly. Similarly, (9) can be simplified as:

$$\begin{aligned}\min_{\mathbf{P}_1}\ & \mu||\mathbf{P}_1[\tilde{\mathbf{X}}_1^{(t-1)},\mathbf{X}_1^{(t)}]-\mathbf{M}_1[\tilde{\mathbf{B}}^{(t-1)},\mathbf{B}^{(t)}]||_F^2+\gamma||\mathbf{P}_1||_F^2\\ \text{s.t. } & \mathbf{B}^{(t)}\in\{-1,1\}^{r\times N_t}.\end{aligned} \tag{18}$$

By setting the partial derivative of (18) w.r.t $\mathbf{P}_1$ to 0, the analytic solution can be obtained:

$$\mathbf{P}_1=\mathbf{M}_1(\mathbf{E}_1^{(t)})^{\mathrm{T}}\left(\mathbf{E}_3^{(t)}+\frac{\gamma}{\mu}\mathbf{I}\right)^{-1}, \tag{19}$$

where $\mathbf{E}_3^{(t)} = \tilde{\mathbf{X}}_1^{(t-1)}(\tilde{\mathbf{X}}_1^{(t-1)})^{\mathrm{T}} + \mathbf{X}_1^{(t)}(\mathbf{X}_1^{(t)})^{\mathrm{T}}$, and the item $\tilde{\mathbf{X}}_1^{(t-1)}(\tilde{\mathbf{X}}_1^{(t-1)})^{\mathrm{T}}$ corresponds to the result related to the accumulated old image data and its value can be directly obtained in previous learning round.

Similar to $\mathbf{P}_1$, the solution expression for $\mathbf{P}_2$ is:

$$\mathbf{P}_2 = \mathbf{M}_2(\mathbf{E}_2^{(t)})^{\mathrm{T}}\left(\mathbf{E}_4^{(t)} + \frac{\gamma}{\mu}\mathbf{I}\right)^{-1}, \tag{20}$$

where $\mathbf{E}_4^{(t)} = \tilde{\mathbf{X}}_2^{(t-1)}(\tilde{\mathbf{X}}_2^{(t-1)})^{\mathrm{T}} + \mathbf{X}_2^{(t)}(\mathbf{X}_2^{(t)})^{\mathrm{T}}$, and the item $\tilde{\mathbf{X}}_2^{(t-1)}(\tilde{\mathbf{X}}_2^{(t-1)})^{\mathrm{T}}$ corresponds to the result related to the accumulated old text data, and its value can be directly obtained in previous learning round.

Update $\mathbf{B}^{(t)}$: Learn $\mathbf{B}^{(t)}$ by fixing other irrelevant variables, (9) can be rewritten as follows:

$$\begin{aligned}\min_{\mathbf{B}^{(t)}}\ & \alpha||\mathbf{X}_1^{(t)} - \mathbf{U}_1\mathbf{M}_1\mathbf{B}^{(t)}||_F^2 + (1-\alpha)||\mathbf{X}_2^{(t)} - \mathbf{U}_2\mathbf{M}_2\mathbf{B}^{(t)}||_F^2 \\ & - \lambda\left(\mathtt{Tr}(\mathbf{B}^{(t)}\mathbf{S}_{xx}^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}}) + \mathtt{Tr}(\mathbf{B}_q\mathbf{S}_{qx}^{(t)}(\mathbf{B}^{(t)})^{\mathrm{T}})\right) \\ & + \mu(||\mathbf{P}_1\mathbf{X}_1^{(t)} - \mathbf{M}_1\mathbf{B}^{(t)}||_F^2 + ||\mathbf{P}_2\mathbf{X}_2^{(t)} - \mathbf{M}_2\mathbf{B}^{(t)}||_F^2) \\ \text{s.t. } & \mathbf{B}^{(\mathbf{t})} \in \{-1, 1\}^{r\times N_t}.\end{aligned} \tag{21}$$

To solve such minimization problem, the discrete optimization algorithm is selected. Specifically, the constant terms in (21) is removed, and its formulation can be simplified as:

$$\begin{aligned}\min_{\mathbf{B}^{(t)}}\ & -\mathtt{Tr}((\mathbf{B}^{(t)})^{\mathrm{T}}(\alpha\mathbf{M}_1^{\mathrm{T}}\mathbf{U}_1^{\mathrm{T}}\mathbf{X}_1^{(t)} + (1-\alpha)\mathbf{M}_2^{\mathrm{T}}\mathbf{U}_2^{\mathrm{T}}\mathbf{X}_2^{(t)} \\ & + \mu\mathbf{M}_1^{\mathrm{T}}\mathbf{P}_1\mathbf{X}_1^{(t)} + \mu\mathbf{M}_2^{\mathrm{T}}\mathbf{P}_2\mathbf{X}_2^{(t)} + \lambda\mathbf{B}^{(t)}\mathbf{S}_{xx} + \lambda\mathbf{B}_q\mathbf{S}_{qx})) \\ \text{s.t. } & \mathbf{B}^{(t)} \in \{-1, 1\}^{r\times N_t}.\end{aligned} \tag{22}$$

Consequently, an efficient close-form solution of $\mathbf{B}^{(t)}$ can be approximated by:

$$\begin{aligned}\mathbf{B}^{(t)} = \mathtt{sign}(&\alpha\mathbf{M}_1^{\mathrm{T}}\mathbf{U}_1^{\mathrm{T}}\mathbf{X}_1^{(t)} + (1-\alpha)\mathbf{M}_2^{\mathrm{T}}\mathbf{U}_2^{\mathrm{T}}\mathbf{X}_2^{(t)} \\ & + \mu\mathbf{M}_1^{\mathrm{T}}\mathbf{P}_1\mathbf{X}_1^{(t)} + \mu\mathbf{M}_2^{\mathrm{T}}\mathbf{P}_2\mathbf{X}_2^{(t)} + \lambda\mathbf{B}^{(t)}\mathbf{S}_{xx} + \lambda\mathbf{B}_q\mathbf{S}_{qx}).\end{aligned} \tag{23}$$

C. Online Anchor Point Updating

For online cross-modal retrieval, multimedia data points often continuously arrive in a streaming fashion, and the anchor-based manifold embedding should be adaptive to such fashion. At round $t-1$, suppose we have the anchor data chunk $[\mathbf{Q}_1, \mathbf{Q}_2]$ of size N_q and its corresponding hash code matrix $\mathbf{B}_q$, a new data chunk $[\mathbf{X}_1^{(t)}, \mathbf{X}_2^{(t)}]$ is arrived at round t and its corresponding hash code matrix $\mathbf{B}^{(t)}$ is calculated by the proposed framework. To be specific, we take the image data for reference and adaptively update the anchor points as follows:

$$\mathbf{Q}_1 = \left\{\mathcal{R}and\left(\mathbf{Q}_1, \frac{N_q\cdot N_q}{N_q+N_t}\right),\ \mathcal{R}and\left(\mathbf{X}_1^{(t)}, \frac{N_q\cdot N_t}{N_q+N_t}\right)\right\},$$

$$\mathbf{Q}_2 = \left\{\mathcal{C}orr\left(\mathbf{Q}_2, \frac{N_q\cdot N_q}{N_q+N_t}\right),\ \mathcal{C}orr\left(\mathbf{X}_2^{(t)}, \frac{N_q\cdot N_t}{N_q+N_t}\right)\right\},$$

Algorithm 1: Online Discrete Optimization for OMGH

Input: new data chunk $[\mathbf{X}_1^{(t)}, \mathbf{X}_2^{(t)}]$ at round t;
Output: $\mathbf{B}^{(t)}$, $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{P}_1$, $\mathbf{P}_2$;

```
Obtain U_1, U_2, M_1, M_2, P_1, P_2 in round r − 1;
Compute S_xx and S_qx via Section III-A3;
Initialize B^(t) with random values;
repeat
  Update U_1 via (11), and U_2 via (14);
  Update M_1 via (16), and M_2 via (17);
  Update P_1 via (19), and P_2 via (20);
  Compute B^(t) via (23);
until (convergency or reaching maximum iterations)
Update anchor data points via (24);
Put B^(t) into the hash table B = [B, B^(t)]:
return U_1, U_2, M_1, M_2, P_1, P_2 and B^(t);
```

$$\mathbf{B}_q = \left\{\mathcal{C}orr\left(\mathbf{B}_q, \frac{N_q\cdot N_q}{N_q+N_t}\right),\ \mathcal{C}orr\left(\mathbf{B}^{(t)}, \frac{N_q\cdot N_t}{N_q+N_t}\right)\right\}, \tag{24}$$

where function $\mathcal{R}and(\mathcal{X}, N_{num})$ represents a random selection of N_{num} examples from data $\mathcal{X}$, and $\mathcal{C}orr(\mathcal{X}, N_{num})$ denotes the corresponding selection related to the index of random selection. Accordingly, the hash code matrix $\mathbf{B}_q$ of the anchor points can be adaptively updated to the hash codes corresponding to the updated anchor data points $[\mathbf{Q}_1, \mathbf{Q}_2]$. Since the number of anchor points is fixed to be N_q, no additional storage space is needed.

D. Complexity Analysis

The optimization process of the proposed OMGH framework is shown in Algorithm 1. As the proposed OMGH framework incrementally learns hash functions and hash codes in a streaming manner, we analyze its computational complexity at each learning round. Although the matrix variables $\mathbf{U}_1$, $\mathbf{U}_2$ $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{P}_1$ and $\mathbf{P}_1$ are updated at each learning round, the auxiliary matrix variables $\mathbf{E}_1^{(t)}$, $\mathbf{E}_2^{(t)}$, $\mathbf{E}_3^{(t)}$ and $\mathbf{E}_4^{(t)}$ retain the results related to the accumulated old data and the time complexity of updating these variables is only related to the new coming data chunk, i.e., $O(N_t)$. Since the number N_q of anchor data set is fixed and relatively small, the time complexities of calculating these variables are $O(N_t)$. Experimentally, the proposed framework requires very fewer iterations and it is appropriate to set the iteration number at 2 in the implementation. Therefore, the overall complexity of each learning round is linear to the new data size, which is very practical for online cross-modal hashing tasks.

E. Hash Codes for Out-of-Sample Extension

For any image or text data that is not enrolled in the training set, we can obtain its semantic representation by the modality-specific projections. To be specific, for any unseen instances $\mathbf{x}_1$ and $\mathbf{x}_2$, we can obtain their corresponding hash codes $\mathbf{b}_1$ and $\mathbf{b}_2$ as follows:

$$\mathbf{b}_1 = \mathtt{sign}(\mathbf{M}_1^{\mathrm{T}}\mathbf{P}_1\mathbf{x}_1),\quad \mathbf{b}_2 = \mathtt{sign}(\mathbf{M}_2^{\mathrm{T}}\mathbf{P}_2\mathbf{x}_2) \tag{25}$$

IV. Experiments

In this section, we conduct a series of quantitative experiments on public benchmarks and verify the effectiveness of the proposed OMGH approach on various cross-modal retrieval tasks, including retrieving text with given image (I $\rightarrow$ T) and retrieving image with given text (T $\rightarrow$ I).

A. Datasets and Evaluation Protocol

In the experiments, three popular image-text datasets, i.e., MIRFlickr, PASCAL-VOC and NUS-WIDE are selected for evaluation, and their brief descriptions are stated as follows:

MIRFlickr [34] dataset consists of 25,000 image-text pairs annotated from 24 semantic labels and taken from the Flickr website. Each image is represented by 150-dimensional edge histogram vector, while each text is characterized by 500-dimensional PCA feature vector. Similar to the work [35], we remove the instances without labels or textual tags appearing less than 20 times, resulting the 16,738 instances in total. Accordingly, we randomly select 836 image-text pairs to serve as the query set, while leaving the remaining pairs as the training set. For online learning, the training set is split into 16 data chunks, each of the first 15 data chunks contains 1000 pairs, and the last chunk contains 902 pairs.

PASCAL-VOC[36] dataset includes 9963 images of 20 categories and each image is annotated with 399 semantic tags. This data set is divided into train, val, and test subsets, and we conduct experiments on trainval and test splits. By dropping those pairs without text annotation, the trainval and test splits, which, respectively, contain 5,000 training pairs and 4,919 test pairs, are selected for evaluation. Each image is represented as a 4096-dimensional CNN feature vector derived from the last fully connected layer of VGG19 model [37], while each text is characterized by a 399-dimensional bag-of-words vector. For online learning, the training set is split to 10 data chunks, and each of the chunk contains 500 pairs.

NUS-WIDE [38] dataset contains 269,648 image-text pairs with 81 semantic concepts. Each image is characterized by a 500-dimensional SIFT feature vector, while the text is described by a 1,000-dimensional bag-of-words vector. Since a large part of semantic concepts contain little samples, we select pictures from top 10 most frequent concepts, and finally obtain 186,577 examples. Accordingly, we randomly select 100,000 labeled image-text pairs for evaluation, with 5% percent pairs as the query set and the remaining parts as the training set. For online learning, the training set is split to 10 data chunks, each of the first 9 chunks contains 10000 pairs, and the last chunk contains 5000 pairs.

The popular mean Average Precision (mAP), precision-recall curve and topK-precision [39], [40] are utilized to evaluate the cross-modal retrieval performance. In particular, mAP@100 is selected to evaluate the retrieval effectiveness because the similar data samples are expected to be indexed in the top retrieval list. In general, a higher topK-precision curve or precision-recall curve indicates better performance.

B. Baseline and Experimental Settings

As surveyed in Section II, there exist limited online cross-modal hashing works, and five well known online cross-modal hashing methods, i.e., OCMH [5], OCMFH [6], OLSH [7], OASH [26], LEMON [27], are selected for evaluation. Meanwhile, we also train the offline model with all training dataset in only one round, and select eight offline methods (i.e., IMH [10], CMFH [11], FSH [14], SCM [15], SMFH [16], SePH [17], GSePH [18] and DCH [19]) for meaningful comparisons. Remarkably, OCMH, OCMFH, IMH, CMFH and FSH methods are unsupervised learning approaches, while OLSH, SCM, SMFH, SePH, GSePH, and DCH m are supervised learning methodologies. For selected baselines, we utilize the source codes kindly provided by the respective authors, and the parameters are initialized as the authors have given in their original papers. Besides, we refer to the batch based training scheme [6] and also enable some representative offline methods to work with the streaming data chunks. That is, the hash function obtained in the previous round is utilized as the initializer for the next training round, specially abbreviated as IMH-b [10], CMFH-b [11], FSH-b [14], GSePH-b [18] and DCH-b [19]. It is noted that some recent unsupervised cross-modal retrieval work [41] cannot incrementally update the hash functions from sequentially arriving multi-modal data, and it is inappropriate to select these approaches as the baselines. Within the proposed OMGH framework, we fix $\alpha = 0.5, \lambda = 100, \mu = 1, \gamma = 10^{-3}$ and $N_q = 500$ in the experiments. Meanwhile, the dimensions of image and text semantic representation are set at $k_1 = 100$ and $k_2 = 50$. For unsupervised learning, top-10 nearest neighbors are selected to construct the manifold structure.

C. Results of Retrieval Accuracy

The quantitative comparisons with state-of-the arts on three datasets are summarized in Table I, where the upper half parts of each retrieval task categorize the unsupervised methods and the lower half parts aggregate the supervised methods. It can be found that the proposed OMGH approach have delivered very promising cross-modal retrieval performances in different learning strategies and outperforms most baselines on different datasets. Comparing with traditional offline learning methods, the online method is prone to loss of retrieval accuracy due to the limited training numbers at each training round. Fortunately, the mAP@100 scores obtained by the proposed OMGH-un and OMGH-su method are still competitive to the results obtained by the competing offline methods that require the whole training data to train the hash functions, e.g., SMFH, FSH, GSePH. Specifically, DCH has reported the better T $\rightarrow$ I performance on 16 and 32 bits. Note that, DCH is also an offline learning algorithm, which require all the accumulated data to train the hash functions. Meanwhile, these offline methods significantly degrade their performance when the hash functions are trained in a streaming manner, i.e., DCH-b. For instance, the competitive DCH-b approach has reported the lower accuracy than the result obtained by traditional DCH. That is, these offline methods are unadaptable to the multimedia data points that continuously

TABLE I
THE MAP@100 SCORES TESTED ON MIRFLICKR, PASCAL-VOC AND NUS-WIDE DATASETS

Task	Method	MIRFlick				PASCAL-VOC				NUS-WIDE			
		16	32	64	128	16	32	64	128	16	32	64	128
I→T	IMH-b [11]	0.5792	0.5803	0.5742	0.5928	0.3479	0.3539	0.3524	0.3631	0.3920	0.4031	0.3956	0.4124
	CMFH-b [12]	0.5905	0.5917	0.6023	0.5990	0.3526	0.3748	0.3849	0.3904	0.4099	0.4345	0.4441	0.4567
	FSH-b [15]	0.5793	0.5829	0.5750	0.5879	0.3649	0.3856	0.3920	0.4120	0.4102	0.4266	0.4523	0.4658
	OCMH [6]	0.5746	0.5749	0.5602	0.5889	0.3594	0.3737	0.3839	0.3913	0.4784	0.4733	0.4178	0.4334
	OCMFH [7]	0.6386	0.6382	0.6287	0.6291	0.3791	0.3909	0.4032	0.4158	0.4088	0.4353	0.4496	0.4426
	OMGH-un	**0.6401**	**0.6451**	**0.6411**	**0.6393**	**0.4285**	**0.4391**	**0.4419**	**0.4323**	**0.4787**	**0.4735**	**0.4702**	**0.4728**
	SCM [16]	0.6943	0.6953	0.6961	0.7099	0.4667	0.4648	0.4864	0.4905	0.5780	0.6126	0.6029	0.6332
	SMFH [17]	0.6913	0.6831	0.6830	0.6867	0.4165	0.4206	0.4254	0.3459	0.3831	0.3973	0.4070	0.4033
	SePH [18]	0.7406	0.7545	0.7606	0.7656	0.5568	0.5678	0.5757	0.5888	0.5595	0.5729	0.5838	0.5853
	GSePH [19]	0.7374	0.7388	0.7551	0.7563	0.5720	0.5869	0.6026	0.6093	0.5583	0.5724	0.5819	0.5890
	DCH [20]	0.7476	0.7546	0.7825	0.7632	0.4503	0.4567	0.4695	0.4558	0.6128	0.6088	0.6091	0.6453
	GSePH-b [19]	0.6741	0.6604	0.6847	0.6986	0.4538	0.4694	0.4837	0.5112	0.5206	0.5293	0.5346	0.5445
	DCH-b [20]	0.5353	0.5640	0.5874	0.5983	0.3796	0.3827	0.3945	0.4036	0.3732	0.4863	0.5415	0.5004
	OLSH [8]	0.7973	0.8136	0.7882	0.7981	0.4461	0.4510	0.4614	0.4703	0.6799	0.6897	0.6857	0.7011
	OASH [27]	0.7847	0.7892	0.7938	0.7989	0.5348	0.5231	0.5437	0.5541	0.6883	0.6922	0.7048	0.7182
	LEMON [28]	0.8043	**0.8149**	0.7936	0.8012	0.5763	0.5832	0.5839	0.5757	0.6574	0.6648	0.6802	0.6746
	OMGH-su	**0.8071**	0.7976	**0.7959**	**0.8080**	**0.5837**	**0.5986**	**0.6003**	**0.5937**	**0.7182**	**0.7219**	**0.7210**	**0.7292**
T→I	IMH-b [11]	0.5938	0.6044	0.6096	0.6129	0.6034	0.6213	0.6329	0.6547	0.4037	0.4126	0.4112	0.4159
	CMFH-b [12]	0.5797	0.5850	0.5939	0.6010	0.6893	0.6948	0.6859	0.6722	0.4100	0.4332	0.4481	0.4526
	FSH-b [15]	0.5733	0.5881	0.5863	0.6019	0.6354	0.6574	0.6593	0.6438	0.4573	0.4465	0.4683	0.4712
	OCMH [6]	0.5796	0.5792	0.6001	0.6021	0.6723	0.6864	0.6999	0.7115	0.4920	0.5144	0.4331	0.4704
	OCMFH [7]	0.7025	0.7247	**0.7559**	**0.7745**	0.6903	0.7050	0.7109	0.7244	0.5141	0.5482	0.5568	0.5712
	OMGH-un	**0.7301**	**0.7327**	0.7412	0.7587	**0.7517**	**0.7525**	**0.7581**	**0.7630**	**0.5428**	**0.5666**	**0.5772**	**0.5739**
	SCM [16]	0.7038	0.7042	0.7120	0.7263	0.6258	0.5149	0.6598	0.6680	0.5312	0.6251	0.6302	0.6401
	SMFH [17]	0.6576	0.6711	0.6755	0.6866	0.5076	0.5700	0.5702	0.3468	0.3780	0.3876	0.3853	0.4142
	SePH [18]	0.8467	0.8574	0.8661	0.8743	0.8139	0.8375	0.8427	0.8501	0.7159	0.7431	0.7648	0.7616
	GSePH [19]	0.8598	0.8649	0.8745	0.8823	0.8328	0.8519	0.8637	0.8594	0.7335	0.7467	0.7613	0.7703
	DCH [20]	**0.9025**	**0.9117**	0.9078	0.9070	0.8023	0.8149	0.8273	0.8355	0.8090	0.8172	0.8101	0.8239
	GSePH-b [19]	0.7417	0.7575	0.7843	0.7859	0.6993	0.7163	0.7238	0.7349	0.6492	0.6826	0.7053	0.7092
	DCH-b [20]	0.6200	0.6150	0.6142	0.6065	0.7532	0.7775	0.7838	0.7746	0.6982	0.8015	0.7951	0.7270
	OLSH [8]	0.7890	0.8587	0.8693	0.8521	0.7812	0.7936	0.8003	0.8176	0.8282	0.8312	0.8556	0.8465
	OASH [27]	0.8039	0.8217	0.8325	0.8357	0.7922	0.8150	0.8275	0.8337	0.8039	0.8120	0.8177	0.8239
	LEMON [28]	0.8737	0.8884	0.8993	0.8921	0.8296	0.8424	0.8521	0.8657	0.8442	0.8536	0.8594	0.8676
	OMGH-su	0.8834	0.9012	**0.9123**	**0.9102**	**0.8467**	**0.8596**	**0.8670**	**0.8751**	**0.8692**	**0.8714**	**0.8793**	**0.8806**

arrive in a streaming fashion. In contrast to this, the proposed OMGH approach only select a limited amount of data for training at each learning round, and have delivered the comparable or in most cases the better performance than that generated by DCH. Importantly, the proposed OMGH approach is designed for processing of streaming multi-modal data, while being adaptive to the unsupervised and supervised scenarios.

Comparing with the competitive online learning methods, i.e., OCMH, OCMFH, OLSH, OASH and LEMON, the proposed OMGH-un and OMGH-su methods generally yield higher mAP@100 scores in most cases, respectively, evaluated on unsupervised and supervised learning mechanisms. On the one hand, the proposed OMGH-un approach has delivered comparable or in most cases the better performance than that generated by unsupervised online learning methods, e.g., OCMH [5] and OCMFH [6]. The main reason lies in that the proposed OMGH-un framework seamlessly preserve the manifold correlation between the anchor data and new data during the online learning process, whereby the learnt hash codes are semantically meaningful to correlate the old data and new arriving data. As a result, the proposed OMGH-un approach generally boosts the retrieval performances in different hash length settings, especially when tested on PASCAL-VOC and NUS-WIDE datasets. On the other hand, the proposed OMGH-su approach also yields competitive or even the better retrieval performances than that generated by the supervised online methods, i.e., OLSH [7], OASH [26] and LEMON [27]. For instance, the proposed OMGH-su method has delivered the best I $\rightarrow$ T or T $\rightarrow$ I retrieval performance on PASCAL-VOC and NUS-WIDE datasets. The possible reasons contribute these competing performances are threefold: 1) The hash codes derived from matrix tri-factorization framework retain more semantic information between the high-dimensional feature space and binary space; 2) The embedding of anchor-based manifold structure is able to efficiently guide the hash code learning process, which can well preserve the semantic correlations between the streaming data and old data; 3) The proposed online discrete optimization can well optimize the hash codes with less quantization errors. The experimental results demonstrates the flexibility and effectiveness of the learning framework on various cross-modal retrieval scenarios.

Further, the precision-recall curves and topK-precision curves tested on different datasets are shown in Figs. 3 and 4, respectively. On the one hand, it can be observed that the precision-recall curves show that the proposed OMGH-un and OMGH-su methods have delivered the better retrieval performances in most cases, respectively, than the results generated by the unsupervised and supervised baselines. On the other hand, topK-precision curves indicates the change of precision with respect

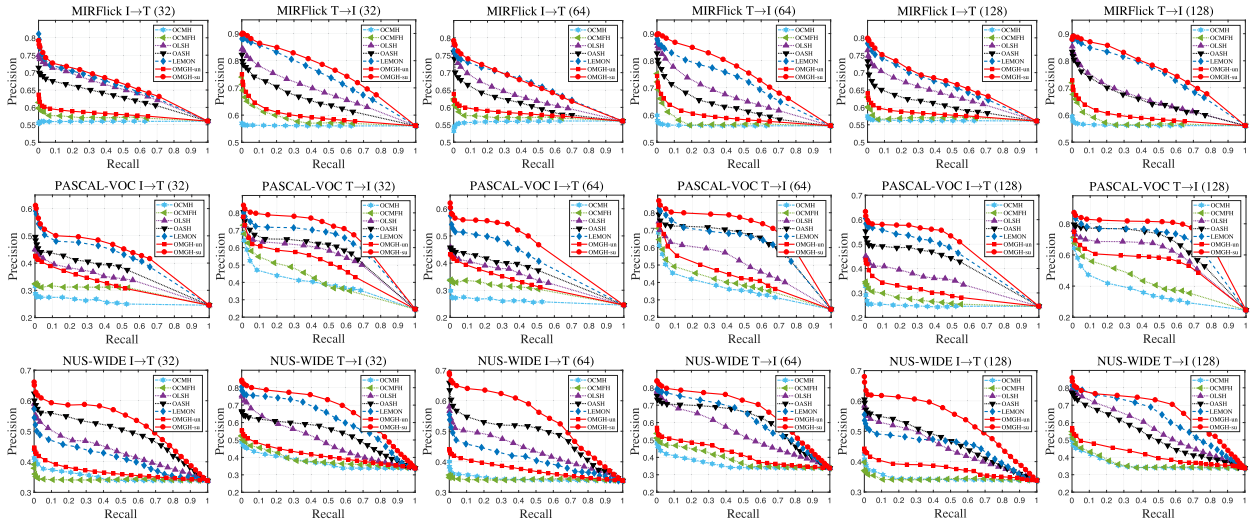


Fig. 3. The precision-recall curves evaluated on MIRFlick, PASCAL-VOC and NUS-WIDE datasets.

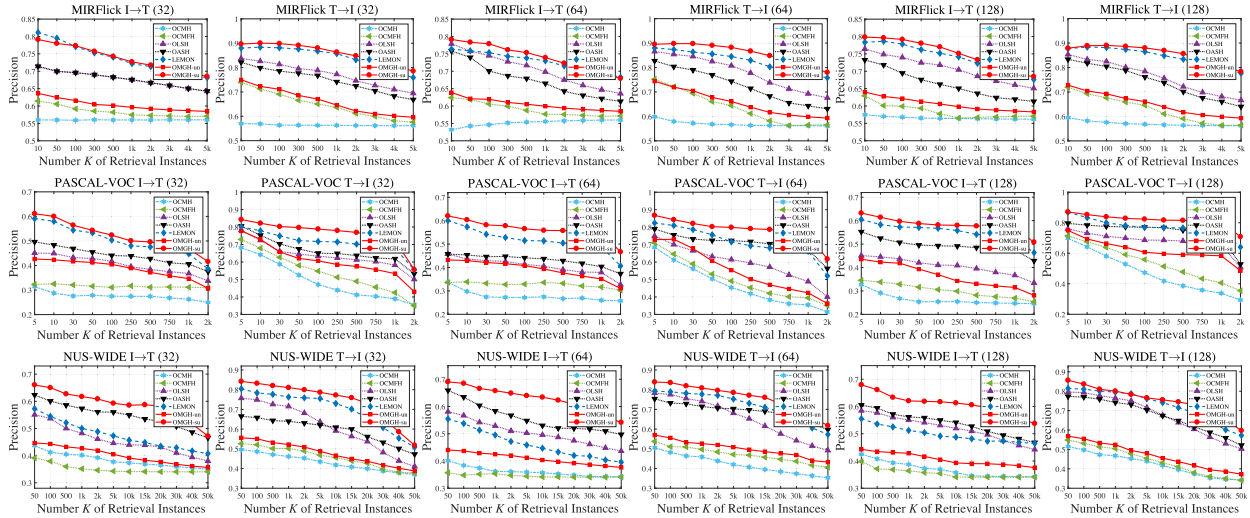


Fig. 4. The representative topK-precision curves tested on MIRFlick, PASCAL-VOC and NUS-WIDE datasets.

to the number of top-ranked K instances indexed by the searching algorithm. It can be observed that the proposed OMGH-un and OMGH-su methods generally have yielded the highest precision scores than the corresponding baselines with the number of retrieved instance (K) changes, both in unsupervised and supervised learning strategies. This indicates that the proposed OMGH approach is able to search much more similar samples at the beginning, which is of crucial importance to a practical retrieval application.

Besides, we further evaluate the online retrieval performance on different learning rounds, in which only the new data chunk is added to train the hash function and optimize the hash code of new arriving data. Fig. 5 shows the mAP@100 scores of online cross-modal hashing methods that evaluated at each learning round (hash length: 32 bits). It can be observed that the mAP@100 scores derived from the online methods increase with the growth of available training data points, and gradually achieve a stable value when the round number is large. This indicates that the online methods are adaptive to process the streaming multi-modal data. Remarkably, the proposed OMGH-un and OMGH-su methods often perform better than the corresponding baselines with the increasing of learning round, while exhibiting a more stable curve. For instance, the mAP@100 score does not increase significantly when the learning round is larger than 8 when tested on NUS-WIDE dataset. That is, the proposed OMGH method always converge faster to the better results with less training data, which is of crucial importance to the online retrieval system.

D. Result of Training Time

The computational complexity of the proposed OMGH framework mainly accumulates from the online training process, which only involves the newly arriving data for learning. Note

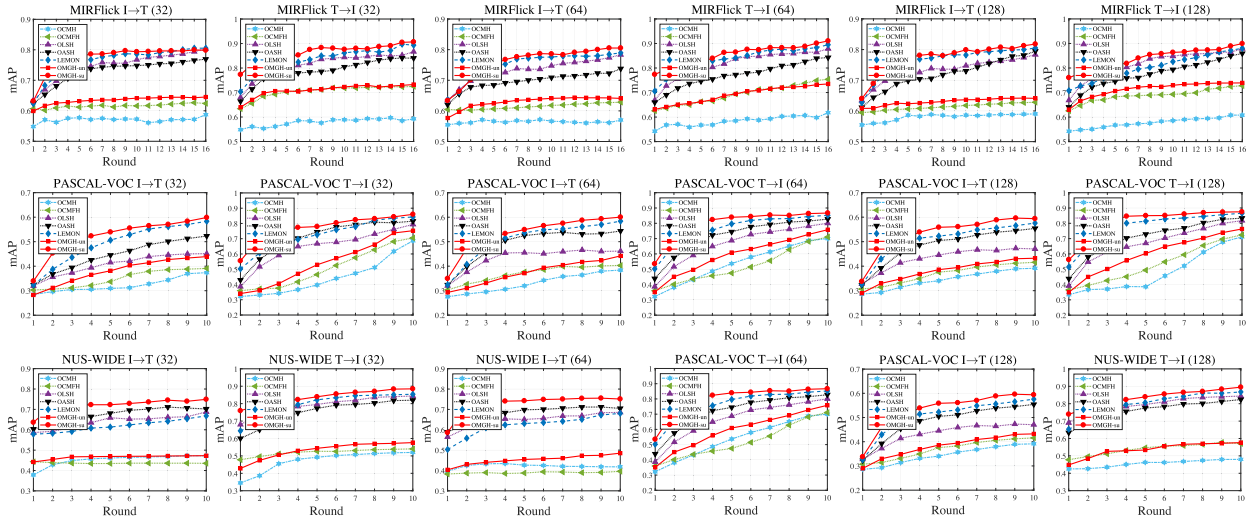


Fig. 5. The mAP scores tested on MIRFlick, PASCAL-VOC and NUS-WIDE datasets at each learning round.

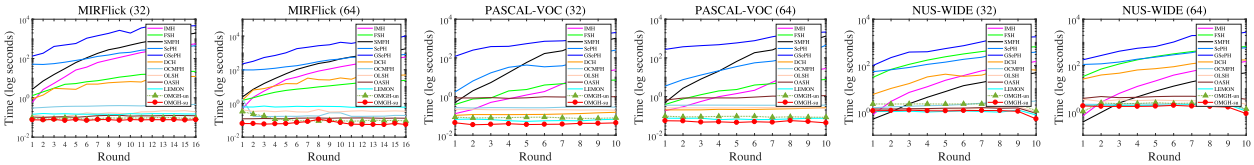


Fig. 6. The training times obtained by different approaches at each learning round.

TABLE II
THE TOTAL TRAINING TIME (IN SECOND) TESTED ON MIRFLICK DATASET

Learning	Training	Method	MIRFlick dataset		
			16	32	64
Unsupervised	Offline	IMH [11]	507.56	486.50	479.17
		FSH [15]	24.38	23.56	24.61
	Online	OCMFH [7]	3.59	4.49	5.74
		OMGH-un	**2.36**	**2.57**	**3.39**
Supervised	Offline	SMFH [17]	470.19	460.53	423.26
		SePH [18]	960.02	1125.51	1468.58
		GSePH [19]	10438.77	11464.38	31158.06
		DCH [20]	8.55	14.04	49.24
	Online	OLSH [8]	2.11	2.63	5.60
		OASH [27]	1.92	2.39	4.52
		LEMON [28]	6.12	6.27	8.84
		OMGH-su	**1.64**	**1.97**	**2.71**

that, the offline methods need to reload all the accumulated data for training. Fig. 6 records the training time of representative baselines on different learning rounds. Since the training times of offline methods are much larger than that obtained by the online methods, the figures draw the log value of seconds to represent the y-coordinate. It can be observed that the execution times obtained from the offline learning methods increase significantly with the increase of the training data sizes, because all the accumulated training data points are enrolled to train hash functions at each round. In contrast to this, the online learning methods perform sufficient fast over the offline methods, and generally show a relatively stable curve on different learning rounds.

Table II displays the total time evaluated on MIRFlick dataset. It can be found that the proposed OMGH-un and OMGH-su methods always perform faster than the corresponding baselines. The main reason lies that the anchor-based manifold structure is able to significantly reduce the size of affinity matrix, which can well reduce the computational load during the online updating process. In addition, the proposed online discrete optimization method has a close-form solution to the hash code learning, which often requires fewer iterations to optimize hash codes. That is, the proposed OMGH framework not only achieves the high cross-modal retrieval accuracy, but also holds a competitive training efficiency.

E. Result of Ablation Studies

The proposed OMGH framework assumes that each modality in an instance generates similar, not exactly identical latent semantic subspace, i.e., $\mathbf{M}_1\mathbf{B}$ for image and $\mathbf{M}_2\mathbf{B}$ for text. To evaluate its effectiveness, we further utilize the same semantic representation for both modalities and therefore let $\mathbf{V}_1 = \mathbf{V}_2$ to learn the hash codes (abbreviated as OMGH1). Fig. 7 reports the mAP@100 values of OMGH and OMGH1 tested on different datasets. It can be seen that the proposed OMGH method always performs better than OMGH1, both in unsupervised and supervised cases. This indicates that the relaxed assumption is able to discriminatively represent the heterogeneous modalities, and the hash codes derived from the matrix tri-factorization framework are more semantically meaningful for better performance.

Meanwhile, we vary the dimension values (k_1 and k_2) of semantic representation in different modalities and report the cross-modal retrieval results by different combinations. Fig. 8 reports the mAP@100 scores with different k_1 and k_2 values,

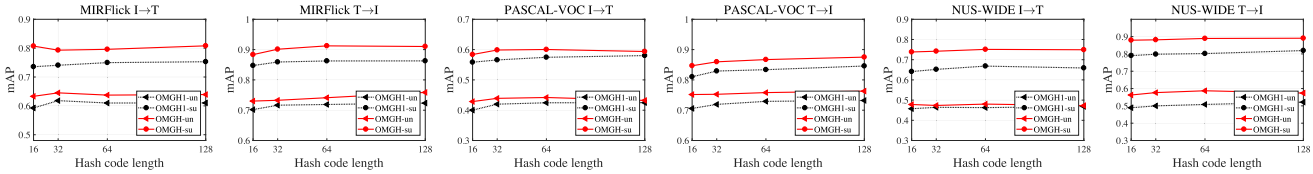
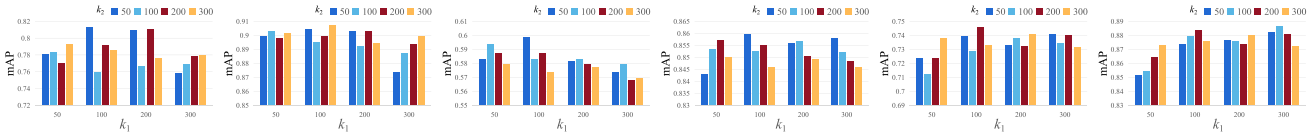


Fig. 7. The ablation results tested on three datasets with different hash code lengths.



(a) MIRFlick I→T (b) MIRFlick T→I (c) PASCAL-VOC I→T (d) PASCAL-VOC T→I (e) NUS-WIDE I→T (f) NUS-WIDE T→I

Fig. 8. The mAP scores obtained by different k_1 and k_2 on three datasets (32 bits).

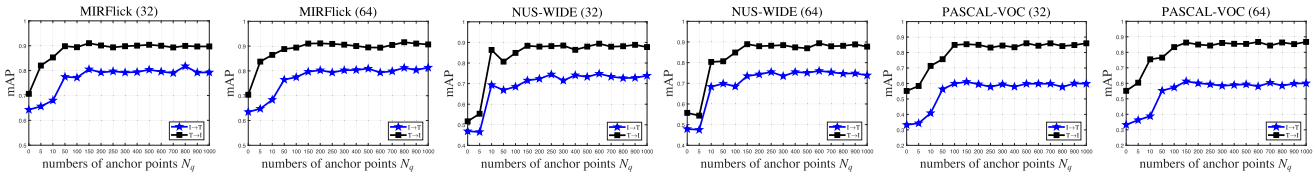


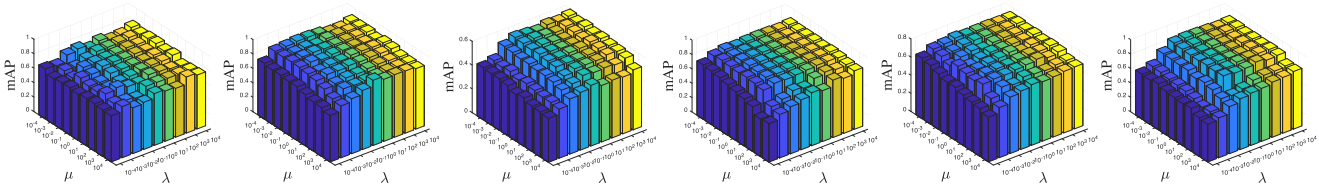
Fig. 9. The mAP scores obtained by different numbers of anchor points (N_q).

it can be found that the larger dimension does not always improve the cross-modal retrieval performance and the optimum retrieval results are not usually achieved by the equal dimensions. For instance, the best retrieval results tested on the MIRFlickr amd PASCAL-VOC dataset are generated by combination $k_1 = 100$, $k_2 = 50$, while the best retrieval results tested on the NUS-WIDE dataset are generated by combination $k_1 = 100$, $k_2 = 200$. Therefore, the proposed matrix tri-factorization framework is able to provide different semantic representations for heterogeneous modalities, which could discriminatively and flexibly characterize the heterogeneous modalities for high retrieval performance.

In addition, the proposed OMGH framework innovates an anchor-based manifold embedding to guide the hash code learning, while simultaneously preserving the semantic correlation between the streaming data and old data. Specifically, the anchors are the sparse representation of old data, and the anchor number N_q balances the importance between the semantic correlation mining and computational complexity. Further, we select OMGH-su to report the mAP@100 scores with different anchor number N_q, and representative results are shown in Fig. 9. It can be observed that a small number of N_q often degrades the retrieval performances, for reason that the limited anchor points cannot comprehensively reveal the semantic manifold information embedded the old data. Meanwhile, a large number N_q does not significantly boost the retrieval performance because some of the anchor points almost have no contribution to the semantic correlation due to their redundancy. Note that, a large number could bring more computational load during the online updating process. Fortunately, the proposed framework shows a relatively high mAP@100 score when N_q is greater than 500. This indicates that the proposed OMGH framework only require a limited number of anchor points to adaptively guide the hash code learning during the online training process. The experimental results have shown its outstanding performance.

F. Parameter Sensitivity Analysis

There are four main parameters involved in OMGH, i.e., α, λ, μ, and γ. Specifically, α balances the importance of each modality. Since our work aims to achieve cross-modal retrieval between image and text, it is natural to set $\alpha = 0.5$ for balancing two modalities. Similar to work [6], γ is the regularization parameter to prevent overfitting, it is generally set $\gamma = 10^{-3}$ in most cases. Specifically, λ controls the learning influence of manifold embedding module, while μ regularizes the influence of semantic projections. Accordingly, we further evaluate the proposed OMGH-su method with different λ and μ values, and the mAP@100 scores tested on different datasets are shown in Fig. 10. It can be observed that the proposed method has achieved very stable performance when λ is greater than 1, which validates the importance of the proposed anchor-based manifold embedding module. In addition, different settings of μ just induce a minor fluctuation on the retrieval performance, and its value can be selected within a wide range such as $[10^{-4}, 10]$. Therefore, μ is insensitive to the cross-modal retrieval performance.



(a) MIRFlick I→T (b) MIRFlick T→I (c) PASCAL-VOC I→T (d) PASCAL-VOC T→I (e) NUS-WIDE I→T (f) NUS-WIDE T→I

Fig. 10. The mAP scores tested with different λ and μ values on three datasets (32 bits).

V. Conclusion

In this paper, we present an efficient and flexible online manifold-guided hashing method to benefit cross-modal retrieval, which incrementally learns hash codes for the current arriving data and adaptively updates hash function in a streaming manner. Specifically, a matrix tri-factorization framework is efficiently developed to decompose the high-dimensional features into more effective modality-specific semantic representation and more discriminative hash codes. In addition, an anchor-based manifold structure is newly proposed to guide hash code learning process, which can well preserve the correlation between the streaming data and old data. Meanwhile, the proposed manifold embedding module is adaptive to unsupervised and supervised cross-modal retrieval scenarios. Further, the proposed discrete optimization algorithm could directly solve the binary optimization problem without relaxation, which can well reduce the quantization error for discriminative hash code learning. Extensive experiments have shown its outstanding performance.

Along the line of the present work, several open problems also deserve our further research. For example, the current online learning model mainly focus on dealing with the balanced multi-modal data collections, which may not be directly applied to deal with the imbalanced multi-modal data. Therefore, it is also imperative to pay attention on training different kinds of imbalanced data collections. Besides, the fine-grained correlation learning would also have an influence on the cross-modal retrieval results, and more robust correlation mining methods deserve further investigation. We shall leave these studies in our future works.

References

[1] 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.

[2] D. Hu, F. Nie, and X. Li, "Deep binary reconstruction for cross-modal hashing," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 973–985, Apr. 2019.

[3] J. D. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 769–790, Apr. 2018.

[4] X. Ma, T. Zhang, and C. Xu, "Multi-level correlation adversarial hashing for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3101–3114, Dec. 2020.

[5] L. Xie, J. L. Shen, and L. Zhu, "Online cross-modal hashing for web image retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 294–300.

[6] D. Wang, Q. Wang, Y. Q. An, X. B. Gao, and Y. M. Tian, "Online collective matrix factorization hashing for large-scale cross-media retrieval," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 1409–1418.

[7] T. Yao *et al.*, "Online latent semantic hashing for cross-media retrieval," *Pattern Recognit.*, vol. 89, pp. 1–11, 2019.

[8] L. Zhu *et al.*, "Efficient multi-modal hashing with online query adaption for multimedia retrieval," *ACM Trans. Inf. Syst.*, vol. 40, no. 2, 2021, Art. no. 41.

[9] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1360–1365.

[10] J. K. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 785–796.

[11] G. G. Ding, Y. C. Guo, and J. L. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2075–2082.

[12] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5427–5440, Nov. 2016.

[13] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 415–424.

[14] H. Liu, R. R. Ji, Y. J. Wu, F. Y. Huang, and B. C. Zhang, "Cross-modality binary code learning via fusion similarity hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7380–7388.

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

[16] J. Tang, K. Wang, and L. Shao, "Supervised matrix factorization hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3157–3166, Jul. 2016.

[17] Z. Lin, G. Ding, J. Han, and J. Wang, "Cross-view retrieval via probability-based semantics-preserving hashing," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4342–4355, Dec. 2017.

[18] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 102–112, Jan. 2019.

[19] X. Xu, F. M. Shen, Y. Yang, H. T. Shen, and X. L. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2494–2507, May 2017.

[20] C. Li *et al.*, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4242–4251.

[21] L. Wang, L. Zhu, E. Yu, J. D. Sun, and H. X. Zhang, "Fusion-supervised deep cross-modal hashing," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 37–42.

[22] F. Cakir, K. He, S. Adel Bargal, and S. Sclaroff, "Mihash: Online hashing with mutual information," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 437–445.

[23] Z. Y. Weng and Y. S. Zhu, "Online supervised sketching hashing for large-scale image retrieval," *IEEE Access*, vol. 7, pp. 88369–88379, 2019.

[24] X. X. Chen, H. Q. Yang, S. L. Zhao, M. R. Lyu, and I. King, "Making online sketching hashing even faster," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 3, pp. 1089–1101, Mar. 2021.

[25] X. Lu *et al.*, "Flexible online multi-modal hashing for large-scale multimedia retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 1129–1137.

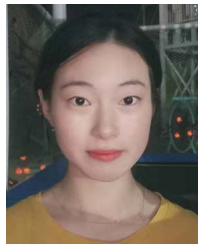
[26] R. Q. Su, D. Wang, Z. Huang, Y. Liu, and Y. Q. An, "Online adaptive supervised hashing for large-scale cross-modal retrieval," *IEEE Access*, vol. 8, pp. 206360–206370, 2020.

[27] Y. X. Wang, X. Luo, and X. S. Xu, "Label embedding online hashing for cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 871–879.

[28] X. Liu, X. Wang, and Y.-m. Cheung, "FDDH: Fast discriminative discrete hashing for large-scale cross-modal retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2021.3076684.

[29] J. Gui, T. L. Liu, Z. Sun, D. C. Tao, and T. N. Tan, "Fast supervised discrete hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 490–496, Feb. 2018.
[30] G. Song, D. Wang, and X. Tan, "Deep memory network for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1261–1275, May 2019.
[31] Y. Liu, Q. Chen, and S. Albanie, "Adaptive cross-modal prototypes for cross-domain visual-language retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14949–14959.
[32] Z. Zeng, S. Wang, N. Xu, and W. Mao, "PAN: Prototype-based adaptive network for robust cross-modal retrieval," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1125–1134.
[33] V. E. Liong, J. Lu, and Y.-P. Tan, "Cross-modal discrete hashing," *Pattern Recognit.*, vol. 79, pp. 114–129, 2018.
[34] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 39–43.
[35] Z. J. Lin, G. G. Ding, M. Q. Hu, and J. M. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3864–3872.
[36] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[37] Y. Wei *et al.*, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.
[38] T. Chua *et al.*, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, pp. 1–9.
[39] Y. X. Peng, X. Huang, and Y. Z. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.
[40] X. Liu, Z. Hu, H. Ling, and Y. Cheung, "MTFH: A matrix tri-factorization hashing framework for efficient cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 964–981, Mar. 2021.
[41] P. F. Zhang, J. Duan, Z. Huang, and H. Yin, "Joint-teaching: Learning to refine knowledge for resource-constrained unsupervised cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 1517–1525.

Xin Liu (Senior Member, IEEE) received the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, in 2013. From 2017 to 2018, he was a Visiting Scholar with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. He is currently a Full Professor with the Department of Computer Science and Technology, Huaqiao University, Xiamen, China, a Research Fellow with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing, China, and also a Research Fellow with Zhejiang Lab, Hangzhou, China. His research interests include multimedia data analysis, pattern recognition and machine learning.



Jinhan Yi is currently with the Department of Computer Science, Huaqiao University, Xiamen, China, and also a Research Fellow with the Xiamen Key Laboratory of Computer Vision and Pattern Recognition & Fujian Key Laboratory of Big Data Intelligence and Security, Xiamen, China. Her research interests include multimedia content analysis, pattern recognition, and deep learning.

Yiu-ming Cheung (Fellow, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong. He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His research interests include machine learning, pattern recognition, and visual computing. Prof. Cheung is the Founding Chairman of the Computational Intelligence Chapter of the IEEE Hong Kong Section. He is an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE, *Pattern Recognition*, *Knowledge and Information Systems*, and *Neurocomputing*. He is an IET Fellow, a RSA Fellow, and BCS Fellow.

Xing Xu (Member, IEEE) received the B.E. and M.E. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2012, respectively, and the Ph.D. degree from Kyushu University, Fukuoka, Japan, in 2015. He is currently with the School of Computer Science and Engineering, University of Electronic of Science and Technology of China, Chengdu, China. His research interests include information retrieval, pattern recognition, and computer vision.

Zhen Cui (Member, IEEE) received the B.S. degree from Shandong Normal University, Jinan, China, in 2004, the M.S. degree from Sun Yat-sen University, Guangzhou, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2014. From June 2012 to December 2012, he was a Research Assistant (half a year) with Nanyang Technological University, Singapore. From 2014 to 2015, he was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. He is currently a Professor with the Nanjing University of Science and Technology, Nanjing, China. His research interests include deep learning, computer vision, and pattern recognition.