Pattern Recognition 145 (2024) 109972

Contents lists available at ScienceDirect

Pattern Recognition

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

Robust online hashing with label semantic enhancement for cross-modal retrieval

Li Li [a], Zhenqiu Shu [a,*], Zhengtao Yu [a], Xiao-Jun Wu [b]

[a] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, China
[b] Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, China

ARTICLE INFO

Keywords:
Robust
Noise
Low-rank
Sparse
Multi-label semantic correlations
Similarity
Online hashing
Cross-modal retrieval

ABSTRACT

Online hashing technology has attracted extensive attention owing to its effectiveness and efficiency in processing large-scale streaming data. However, there are still some limitations: (1) In practical applications, the observed labels of multimedia data are obtained through manual annotation, which may inevitably introduce some noises into labels. This may lead to retrieval performance degradation when the noisy labels are directly applied to retrieval tasks. (2) The potential semantic correlation of multi-labels cannot be fully explored. To overcome these limitations, in this paper, we propose robust online hashing with label semantic enhancement (ROHLSE). Specifically, ROHLSE seeks to recover the clean labels from the provided noisy labels by imposing low-rank and sparse constraints. Meanwhile, it employs the representation of samples in the feature space to predict the labels via the dependency between sample instances and labels. To efficiently handle streaming data, ROHLSE preserves the similarity between new data, and establishes the semantic relationships between new and old data through chunk similarity, simultaneously. Furthermore, ROHLSE can fully utilize the semantic correlations between multiple labels of each instance. Extensive experiments are conducted on three benchmark datasets to demonstrate the superiority of the proposed ROHLSE approach.

1. Introduction

With the explosive growth of multimedia data, it is a widespread concern issue how to effectively process heterogeneous data from large-scale datasets [1,2]. At present, many studies have been devoted to cross-modal retrieval [3–5] applications. The key challenge in cross-modal retrieval is to bridge the heterogeneity gap between different modalities. A common strategy is to map multi-modalities into a common semantic space to learn their relations. To improve the retrieval efficiency of multi-modal data, hashing methods were proposed to encode them into compact binary hash codes, and then measure the similarity between multi-modal data in the Hamming space. Recently, hashing technology has been widely applied to cross-modal retrieval due to its high efficiency and low storage costs. Currently, cross-modal hashing methods (CMH) can achieve satisfactory performances in many real retrieval problems. However, multimedia data usually appears in the form of streaming data in many real-world scenarios. Once new multimedia data arrives, most existing hashing methods must retrain the model using all accumulated data because they adopt the batch-based strategy. Therefore, it results in expensive computational consumption for processing streaming data.

In recent years, online hashing methods have achieved more significant efficiency than offline hashing methods for cross-modal retrieval tasks on large-scale streaming data. Specifically, online cross-modal hashing (OCMH) methods aim to gradually update the hash functions of the constantly arriving multi-modal data and encode the streaming data into compact binary codes, simultaneously. In addition, online hashing methods maintain the validity of the hash codes of old data while updating the hash functions and hash codes of new streaming data. Therefore, online hashing methods usually achieve higher retrieval efficiency than offline hashing methods for large-scale streaming data. At present, OCMH methods are roughly divided into supervised methods [6–8] and unsupervised methods [9,10]. Many studies have shown that the retrieval performances of supervised methods are better than unsupervised methods. However, in practical applications, the labels of multi-modal data are obtained by automatic or manual annotation, and the supervision information may be weak due to missing or incorrect class assignments. Hence, the observed labels are directly used as supervisory information, which may lead to unsatisfactory retrieval performance for online cross-modal retrieval tasks. In addition, existing OCMH methods ignore the latent semantic correlations between multi-labels when utilizing the supervision information. For example, an

* Corresponding author.
E-mail address: shuzhenqiu@163.com (Z. Shu).

https://doi.org/10.1016/j.patcog.2023.109972
Received 4 February 2023; Received in revised form 7 September 2023; Accepted 12 September 2023
Available online 16 September 2023

image sample may contain multiple categories, such as blue sky, white cloud, bird, fish, river, etc. Obviously, there is a potential semantic correlation between these labels. Generally, when the label of blue sky is assigned to an image, the label of white cloud is also assigned to this image. This is because the relationship between the labels of both blue sky and white cloud is usually very close. Other labels, such as fish and river, blue sky, and bird also have potential correlations in many cases. However, the correlations between these labels are often ignored in cross-modal retrieval. Therefore, the performance may be significantly improved by fully utilizing this potential correlation of the labels.

Currently, existing CMH retrieval has the following limitations: (1) When new data arrives, offline cross-modal hashing must use all accumulated data to retrain the model, resulting in expensive computational costs for processing streaming data. (2) Existing OCMH methods cannot effectively deal with the noisy labels among multi-modalities, which limits their application in real scenarios. (3) The semantic correlation between multiple labels cannot be fully explored in most OCMH methods, which results in insufficient improvement in their retrieval performance. To address the aforementioned limitations, we propose a new OCMH approach in this paper, termed robust online hashing with label semantic enhancement (ROHLSE). Fig. 1 illustrates the overall framework of our ROHLSE approach. To mitigate the impact of noisy labels (e.g. missing labels, wrong labels), ROHLSE decomposes the observed label matrix into the low-rank clean label matrix and the sparse noise matrix. Meanwhile, the multi-modal features are used to predict the recovered label matrix. To achieve online learning for streaming data, our ROHLSE approach not only preserves the similarity of new data but also establishes the relationship between new and old data by constructing the chunk similarity. Furthermore, the proposed ROHLSE approach fully exploits the semantic correlations between labels to improve retrieval performance. Experiments are conducted on several benchmark datasets to verify the effectiveness of our ROHLSE approach.

The main contributions of this work are given as follows:

- We propose a novel robust OCMH approach for real retrieval problems. It can learn the hash codes of new data and update the hash functions via online learning way, effectively handling the streaming media data. Therefore, ROHLSE achieves more retrieval efficiency than traditional offline hashing technology.
- To the best of our knowledge, this is the first work to retrieve multi-modal data with noisy labels among online hashing approaches. Specifically, ROHLSE recovers the clean labels from the observed labels by imposing the low-rank and sparse constraints, and then adaptively learns the latent semantic correlations between the clean labels to guide the generation of hash codes.
- In ROHLSE, the learning of hash codes is guided by pairwise similarity based on new data. In addition, chunk similarity is constructed based on the semantic relationship between new data and the results of the previous round and it can avoid semantic forgetting. An efficient optimization algorithm is developed to solve the proposed model. Experimental results on several benchmark datasets have demonstrated the advantage of our ROHLSE approach.

The remainder of this paper is organized as follows: Section 2 introduces the related works. Section 3 details our ROHLSE method. The experimental results are presented and discussed in Section 4. The conclusions are given in Section 5.

2. Related work

2.1. Offline hashing

According to whether supervision information among multi-modalities is fully utilized, offline hashing technology can be divided into unsupervised [11–13], weakly supervised [14–16], and supervised [17–19] methods. Specifically, unsupervised offline hashing methods learn hash functions by transforming data features of different modalities into Hamming space. Linear cross-modal hashing [11] was proposed to explore the correlations of heterogeneous data using binary codes. However, it requires expensive computational costs to construct the similarity graph. Ding et al. [13] learned the common representation by collective matrix factorization and then generated its uniform hash codes. Zhou et al. [12] adopted sparse coding to extract the latent semantic features, and then used it to generate the hash codes.

Since supervised hashing methods fully consider the supervised information hidden in multi-modal data, their retrieval performance is usually superior to unsupervised hashing methods. Supervised matrix factorization hashing (SMFH) [20] learns hash codes by using matrix factorization framework, and adopts graph regularizer to maintain the similarity between multi-modal original features. Semantic Preserving Hashing (SePH) [21] utilizes KL-divergence to learn hash codes by transforming semantic information into probability distributions. Since SMFH and SePH need to construct a squared similarity matrix, it consumes expensive computational costs and large store space. Therefore, SMFH and SePH cannot deal with large-scale datasets. In addition, SMFH and SePH easily cause large quantization errors due to using the relaxation scheme to generate hash codes. Hence, various discrete hashing methods have been proposed in the past few years. Discrete Cross-modal Hashing (DCH) [22] obtains hash codes by projecting multi-modalities into a unified latent semantic space. Meanwhile, it learns a classifier to predict the class information by the label matrix. Although the quantization error is greatly reduced in DCH, the bit-by-bit optimization strategy for longer hash codes still requires expensive computational costs. Fang et al. [23] decomposed the nearest neighbor similarity graph of each modality and directly extracted the discrete common semantic representation of each modality. Therefore, it can reduce the quantization loss caused by binary relaxation. Currently, supervised offline hashing methods have achieved encouraging retrieval accuracy in many real problems.

However, the labels of multi-modal data inevitably mix with some wrong or missing labels in real scenarios. Therefore, the multi-modal data contain some weakly supervised information due to noise or outliers. It is directly used to guide the hash code learning in the cross-modal retrieval task, which may lead to unsatisfactory retrieval performance. To solve this problem, some weakly supervised offline hashing methods were proposed to retrieve these multi-modal data with noisy labels. Cui et al. [14] employed the loss function based on $l_{2,1}$-norm to learn hash codes, thus reducing the sensitivity of the model to noisy labels. Zhang et al. [16] imposed the low-rank and sparse constraints to mitigate the challenge caused by noisy labels, thereby improving the robustness of the algorithm.

Recently, offline hashing approaches have been successfully applied to cross-modal retrieval. However, most offline hashing methods are batch-based to learn the hash codes, thus requiring all data to retrain the model. For the continuous arrival of streaming data, they need to consume unacceptable computational costs and storage space to update the model. Hence, offline hashing technology is unsuitable for processing streaming data.

2.2. Online hashing

Existing online hashing methods can be roughly divided into online single-modal hashing methods and OCMH methods. Online single-modal hashing methods [24–26] mainly deal with the retrieval problem of single-modal streaming data. Online Hashing (OH) [25] updates the hashing model by adopting a prediction loss function, thus achieving zero prediction loss. Ding et al. [27] exploited the correlations between new and old data labels to learn the real-valued pseudo tag matrix, and used it to establish the similarity matrix between new and old data, thus effectively maintaining their semantic correlations. Lin et al. [26] used

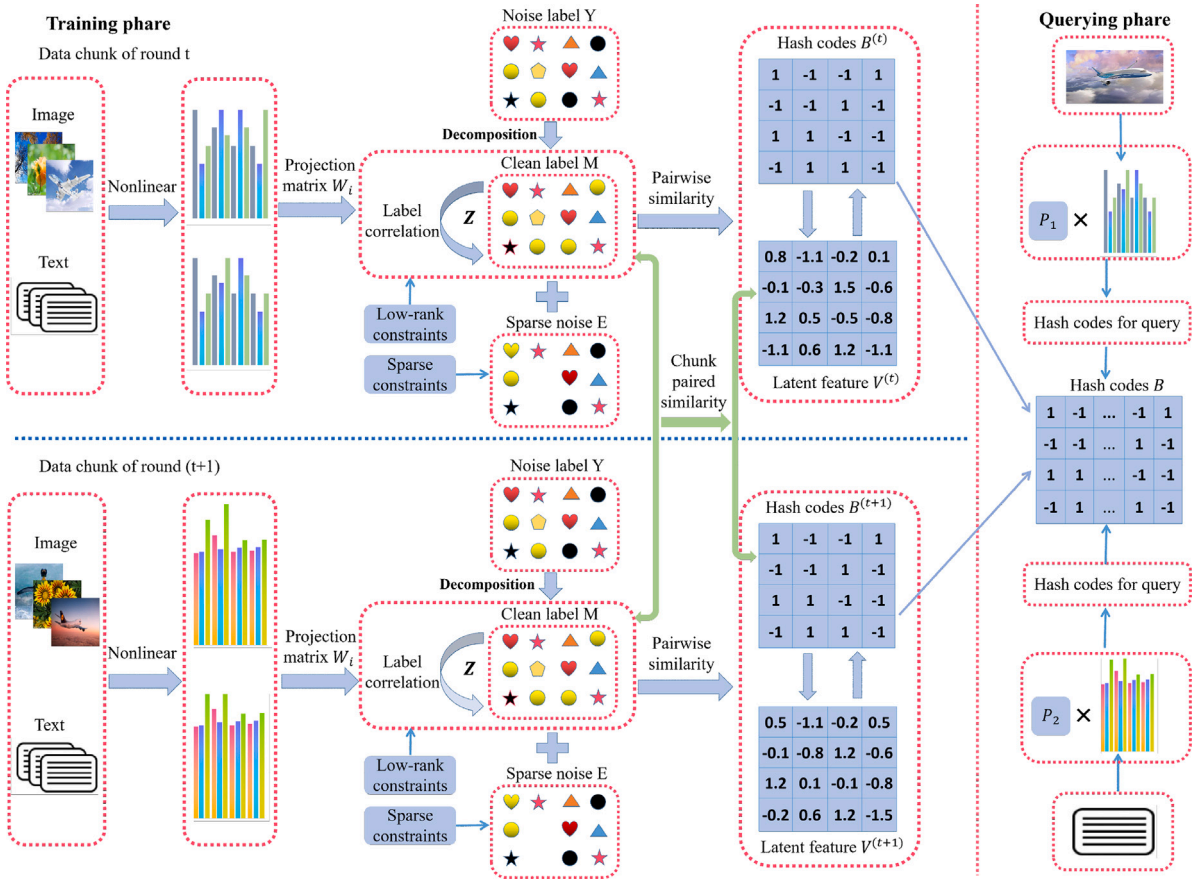


Fig. 1. The overall framework of our proposed ROHLSE approach. This framework mainly consists of the training phase and the query phase. In the training phase, it attempts to recover clean labels from the noisy labels of constantly arriving data chunks, and then learns the hash codes. In the query stage, the hash codes of the query samples are learned by the modality-specific projection, and then we use them to retrieve similar hash codes from the training set.

each column of the Hadamard matrix as the target codes of each category to guide the learning of hash codes. Although the aforementioned online single-modal hashing methods have achieved significant results, they cannot effectively perform cross-modal retrieval tasks.

In the past few years, several online hashing methods [28,29] have been put forward for cross-modal retrieval tasks. According to whether the supervised information is utilized, existing OCMH methods can be roughly divided into unsupervised and supervised methods. Unsupervised OCMH methods attempt to utilize the data correlation between different modal instances, which cannot usually achieve remarkable performances in real cross-modal retrieval. Supervised OCMH methods make full use of the semantic labels of multi-modalities, thus they can usually achieve better performances compared to unsupervised methods due to the use of labels or semantic similarity. OCMH [9] decomposes the feature matrix into a shared latent semantic matrix and transfer matrix that supports online learning, and efficiently updates the hash codes of new data via shared latent semantic matrix and dynamic transfer matrix. OCMFH [10] learns the hash functions and the hash codes of streaming data online through collective matrix factorization. Zhan et al. [8] proposed to directly preserve the similarity between new arrival data and existing old data in Hamming space. Furthermore, the fine-grained semantic information is fully exploited by label embedding. However, it adopts a bit-by-bit discrete online optimization scheme to learn the hash codes, thus needing a large number of iterations to solve the model. Online manifold-guided hashing [29] constructs online anchors based on manifold structure to guide the learning of hash codes. Therefore, it effectively preserves the semantic correlations between old data and new data. Wang et al. [30] built an online learning framework by preserving label similarity and label embedding. Furthermore, the relationship between old and new data is maintained by constructing chunk similarity. Shu et al. [31] employed label embedding to learn the shared and modal-specific latent semantic representation and then constructed the relationship between old data and new arrival data. Therefore, it can generate more discriminative hash codes for cross-modal retrieval. Compared with offline hashing, online hashing is more suitable for large-scale multi-modal streaming data retrieval.

3. Proposed approach

In this section, we introduce our ROHLSE approach in detail. For convenience, we only describe two representative modalities (i.e., image and text) in our proposed method.

3.1. Problem definition

Assume that the training dataset consisting of text and image modalities comes in the form of streaming data. In each round t, data chunk $X^{(t)} = [X_1^{(t)}, X_2^{(t)}]$ are added to the training dataset, and the associated observed label matrix is $L^{(t)} \in \{0,1\}^{c\times n_t}$. $X_1^{(t)} = \left[x_1^{1(t)}, x_1^{2(t)}, \ldots, x_1^{n_t(t)}\right] \in \mathbb{R}^{d_1\times n_t}$ and $X_2^{(t)} = \left[x_2^{1(t)}, x_2^{2(t)}, \ldots, x_2^{n_t(t)}\right] \in \mathbb{R}^{d_2\times n_t}$ respectively represents the image feature matrix and the text feature matrix, where d_1 and d_2 denotes the dimensionality of image and text matrix, respectively.

n_t and c are the new data chunk size and the number of categories, respectively. Furthermore, the old data chunks before tth round can be expressed as $\widetilde{X}^{(t-1)} = [\widetilde{X}_1^{(t-1)}, \widetilde{X}_2^{(t-1)}]$ containing $\widetilde{n}_{t-1}$ sample pairs, and the associated observed labels are $\widetilde{L}^{(t-1)} \in \{0,1\}^{c\times\widetilde{n}_{t-1}}$. Therefore, $X = [\widetilde{X}^{(t-1)}, X^{(t)}]$ is the current total dataset with $n = n_t + \widetilde{n}_{t-1}$ pairs samples, and its associated observed label matrix is denoted as $L = \left[\widetilde{L}^{(t-1)}, L^{(t)}\right] \in \{0,1\}^{c\times n}$.

3.2. Proposed model

3.2.1. Label recovery

In practical applications, the observed labels in multi-modalities are obtained by automatic or manual annotation, which inevitably leads to missing and wrong labels. Moreover, the observed labels are used to guide the learning of hash codes, thereby achieving unsatisfactory retrieval performances in real tasks. Candès et al. [32] proposed robust principal component analysis (RPCA), which can recover underlying clean data with the low-rank structure from the corrupted observations by solving a simple convex optimization problem. Therefore, it aims to separate the observed data into the low-rank part and the sparse part by imposing the low-rank constraint and the sparse constraint, respectively. Our proposed method can decompose the observed label matrix into a low-rank clean label matrix and a sparse noise matrix as follows:

$$\min_{M,E} \operatorname{rank}(M) + \beta\|E\|_1 \qquad s.t. L = M + E, \tag{1}$$

where $M \in \mathbb{R}^{c\times n}$ denotes the clean label matrix recovered from the observed labels L, and $E \in \mathbb{R}^{c\times n}$ denotes the sparse noise matrix. $\operatorname{rank}(M)$ represents the lowest rank representation of matrix M, and $\|E\|_1$ denotes the l_1-norm of noise matrix E. β is a balance parameter.

Since the optimization of $\operatorname{rank}(M)$ is a non-convex problem, it is difficult to directly obtain the solution of Eq. (1). Therefore, $\operatorname{rank}(M)$ is usually replaced by nuclear norm $\|M\|_*$ in the practical optimization procedure, we rewrite Eq. (1) as the following problem:

$$\min_{M,E} \|M\|_* + \beta\|E\|_1 \qquad s.t. L = M + E. \tag{2}$$

To recover the clean label matrix M more accurately, we fully exploit the dependency between sample instances and labels to predict the clean labels via the representation of samples in Hilbert space. Therefore, it can be described as follows:

$$\min_{W_i,M} \sum_{i=1}^{2} \left(\mu \left\| W_i \phi\left(X_i\right) - M \right\|_F^2 + \lambda \left\| W_i \right\|_F^2 \right), \tag{3}$$

where $W_i \in \mathbb{R}^{c\times m}$ is the projection matrix. μ and λ denote the trade-off parameters. $\phi(\cdot)$ represents the RBF kernel function, and m is the number of anchor points. $\phi\left(x_i\right) \in \mathbb{R}^{m\times n}$ is the kernel feature matrix of the image or text modality, which is used to explore the nonlinear structure of training data. Here, $\phi\left(x_i\right)$ is defined as follows:

$$\phi\left(x_i\right) = \left[\exp\left(-\frac{\left\| x_i - \alpha_1^{(i)} \right\|^2}{2\sigma_{(i)}^2} \right), \ldots, \exp\left(-\frac{\left\| x_i - \alpha_m^{(i)} \right\|^2}{2\sigma_{(i)}^2} \right) \right], \tag{4}$$

where $\left\{\alpha_j^{(i)}\right\}_{j=1}^{m}$ represents m anchor points, and $\sigma_{(i)} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} \left\| x_i - \alpha_j^{(i)} \right\|$ is the Gaussian kernel parameter.

3.2.2. Multi-label semantic correlations

Generally, each instance may be assigned multiple labels in a multi-modal dataset, and there is a latent semantic correlation between these labels. Therefore, the retrieval performances can be greatly improved by utilizing this additional semantic correlation of labels [33]. Here, the label correlation of each instance can be expressed as follows:

$$\begin{aligned} \bar{M}_{ij} &= Z_{i1} \times M_{1j} + Z_{i2} \times M_{2j} + \cdots + Z_{ic} \times M_{cj} \\ &= \sum_{m=1}^{c} Z_{im} \times M_{mj}, \end{aligned} \tag{5}$$

where $Z \in \mathbb{R}^{c\times c}$ represents the latent semantic correlation matrix between labels, and $\bar{M}_{ij}$ denotes the correlation score of the jth sample belonging to the ith class. To fully utilize the correlation between semantic labels, Eq. (3) can be rewritten as follows:

$$\begin{aligned} \min_{W_i M} \sum_{i=1}^{2} \left(\mu \left\| W_i \phi\left(X_i\right) - ZM \right\|_F^2 + \lambda \left\| W_i \right\|_F^2 \right) \\ + \gamma \|M - ZM\|_F^2, \end{aligned} \tag{6}$$

where $ZM \in \mathbb{R}^{c\times n}$ aims to take full advantage of the latent semantic correlation of labels, and the term $\|M - ZM\|_F^2$ uses ZM to approximate the clean matrix M. γ is the regularization parameter. Therefore, Eq. (2) can be rewritten as follows:

$$\min_{ZM,E} \|ZM\|_* + \beta\|E\|_1 \qquad s.t. L = ZM + E. \tag{7}$$

3.2.3. Hash codes learning

The CMH method seeks to find a shared latent semantic representation of multi-modalities, and then generate unified hash codes by binarizing the shared latent semantic representation. However, the binary hash codes may lead to large quantization errors and semantic loss in this process. To reduce the loss of semantic information, we adopt cosine distance to measure the semantic similarity, i.e. $s_{ij} = \cos\left(G_i, G_j\right)$, where G is considered as the normalized clean label matrix ZM and G_i is the label vector of G. The pairwise similarity between two modalities can be represented as $G^T G$. To embed the pairwise similarity into hash codes effectively, we try to minimize the following problem:

$$\begin{aligned} &\min_{B} \left\| B^T B - rG^T G \right\|_F^2, \\ &s.t. B \in \{-1,1\}^{r\times n}, BB^T = nI_r, B1_n = 0_r. \end{aligned} \tag{8}$$

However, it is difficult to optimize the minimization problem (8), thereby limiting the ability of the algorithm to model the pairwise similarity of multi-modalities. To facilitate the optimization of the model, various hashing methods attempt to relax the binary constraint, which leads to large quantization errors. To solve this problem, the asymmetric strategy [34] was proposed by introducing a continuous variable matrix $V \in \mathbb{R}^{r\times n}$ to replace a matrix B of Eq. (8). Thus, Eq. (8) can also be expressed the following minimization problem:

$$\begin{aligned} &\min_{V,B} \left\| V^T B - rG^T G \right\|_F^2 + \delta\|V - B\|_F^2, \\ &s.t. B \in \{-1,1\}^{r\times n}, VV^T = nI_r, V1_n = 0_r, \end{aligned} \tag{9}$$

where δ is a balance parameter.

3.2.4. Online learning

Our proposed ROHLSE method processes streaming data through online learning, hence the whole training dataset for online learning includes new arrival data and old data. By integrating Eqs. (6), (7), and (9), the overall objective function of our ROHLSE approach is given as follows:

$$\begin{aligned} \mathcal{L} &= \widetilde{\mathcal{L}} \\ &+ \sum_{i=1}^{2} \left(\mu \left\| W_i^{(t)} \phi(X_i^{(t)}) - Z^{(t)} M^{(t)} \right\|_F^2 + \lambda \left\| W_i^{(t)} \right\|_F^2 \right) \\ &+ \alpha \left(\left\| Z^{(t)} M^{(t)} \right\|_* + \beta \left\| E^{(t)} \right\|_1 \right) + \gamma \left\| M^{(t)} - Z^{(t)} M^{(t)} \right\|_F^2 \\ &+ \delta \left\| V^{(t)} - B^{(t)} \right\|_F^2 + \left\| V^{(t)T} B^{(t)} - rG^{(t)T} G^{(t)} \right\|_F^2, \\ &\quad s.t. L^{(t)} = Z^{(t)} M^{(t)} + E^{(t)}, B^{(t)} \in \{-1,1\}^{r\times n_t}, \\ &\quad V^{(t)} V^{(t)T} = n_t I_r, V 1_{n_t} = 0. \end{aligned} \tag{10}$$

where $\widetilde{\mathcal{L}}$ is formulated as follows:

$$\begin{aligned}\widetilde{\mathcal{L}} &= \mu\sum_{i=1}^{2}\left\|W_i^{(t)}\phi(\widetilde{X}_i^{(t-1)}) - \widetilde{Z}^{(t-1)}\widetilde{M}^{(t-1)}\right\|_F^2 \\ &+ \alpha\left(\left\|\widetilde{Z}^{(t-1)}\widetilde{M}^{(t-1)}\right\|_* + \beta\left\|\widetilde{E}^{(t-1)}\right\|_1\right) \\ &+ \gamma\left\|\widetilde{M}^{(t-1)} - \widetilde{Z}^{(t-1)}\widetilde{M}^{(t-1)}\right\|_F^2 + \delta\left\|\widetilde{V}^{(t-1)} - \widetilde{B}^{(t-1)}\right\|_F^2 \\ &+ \left\|\widetilde{V}^{(t-1)T}B^{(t)} - r\widetilde{G}^{(t-1)T}G^{(t)}\right\|_F^2\end{aligned} \tag{11}$$

The last item in Eq. (11) is to establish the semantic relationship between new and old data by constructing the chunk similarity.

3.3. Online optimization

In this subsection, we develop an iterative updating scheme to solve the problem (10). For convenience, an auxiliary variable $N^{(t)} \in \mathbb{R}^{c\times n_{(t)}}$ is introduced to optimize the proposed method. Therefore, the problem (10) can be rewritten as follows:

$$\begin{aligned}\mathcal{L} &= \widetilde{\mathcal{L}} \\ &+ \sum_{i=1}^{2}\left(\mu\left\|W_i^{(t)}\phi(X_i^{(t)}) - Z^{(t)}M^{(t)}\right\|_F^2 + \lambda\left\|W_i^{(t)}\right\|_F^2\right) \\ &+ \alpha\left(\left\|N^{(t)}\right\|_* + \beta\left\|E^{(t)}\right\|_1\right) + \gamma\left\|M^{(t)} - Z^{(t)}M^{(t)}\right\|_F^2 \\ &+ \delta\left\|V^{(t)} - B^{(t)}\right\|_F^2 + \left\|V^{(t)T}B^{(t)} - rG^{(t)T}G^{(t)}\right\|_F^2, \\ s.t.\quad & L^{(t)} = Z^{(t)}M^{(t)} + E^{(t)}, Z^{(t)}M^{(t)} = N^{(t)}, \\ & B^{(t)} \in \{-1,1\}^{r\times n_t}, V^{(t)}V^{(t)T} = n_t I_r, V1_{n_t} = 0.\end{aligned} \tag{12}$$

Eq. (12) can be transformed into an augmented Lagrangian function. Thus, we further rewrite problem (12) as follows:

$$\begin{aligned}\mathcal{L} &= \widetilde{\mathcal{L}} \\ &+ \sum_{i=1}^{2}\left(\mu\left\|W_i^{(t)}\phi(X_i^{(t)}) - Z^{(t)}M^{(t)}\right\|_F^2 + \lambda\left\|W_i^{(t)}\right\|_F^2\right) \\ &+ \alpha\left(\left\|N^{(t)}\right\|_* + \beta\left\|E^{(t)}\right\|_1\right) + \gamma\left\|M^{(t)} - Z^{(t)}M^{(t)}\right\|_F^2 \\ &+ \delta\left\|V^{(t)} - B^{(t)}\right\|_F^2 + \left\|V^{(t)T}B^{(t)} - rG^{(t)T}G^{(t)}\right\|_F^2 \\ &+ \frac{\rho}{2}\left\|L^{(t)} - Z^{(t)}M^{(t)} - E^{(t)} + \frac{F_1}{\rho}\right\|_F^2 \\ &+ \frac{\rho}{2}\left\|Z^{(t)}M^{(t)} - N^{(t)} + \frac{F_2}{\rho}\right\|_F^2 \\ s.t.\quad & B^{(t)} \in \{-1,1\}^{r\times n_t}, V^{(t)}V^{(t)T} = n_t I_r, V^{(t)}1_{n_t} = 0_r,\end{aligned} \tag{13}$$

where $F_1 \in \mathbb{R}^{c\times n_t}$ and $F_2 \in \mathbb{R}^{c\times n_t}$ are Lagrangian multipliers, and ρ is the regularization parameter.

It is clear to see that the objective function (13) is non-convex for all variables. Fortunately, it is convex to one variable while fixing the others. The optimization procedure of each variable is given as follows:

(1) Update $W_i^{(t)}(i = 1, 2)$ by fixing other variables. The model (13) can be rewritten as follows:

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{2}\left(\mu\left\|W_i^{(t)}\phi(X_i^{(t)}) - Z^{(t)}M^{(t)}\right\|_F^2 + \lambda\left\|W_i^{(t)}\right\|_F^2\right) \\ &+ \mu\sum_{i=1}^{2}\left\|W_i^{(t)}\phi(\widetilde{X}_i^{(t-1)}) - \widetilde{Z}^{(t-1)}\widetilde{M}^{(t-1)}\right\|_F^2.\end{aligned} \tag{14}$$

By setting $\frac{\partial\mathcal{L}}{\partial W_i^{(t)}} = 0$, we have

$$W_i^{(t)} = \mu C_i^{(t)}\left(\mu D_i^{(t)} + \lambda I\right)^{-1}, \tag{15}$$

where $C_i^{(t)}$ and $D_i^{(t)}$ can be respectively represented as follows:

$$\begin{aligned}C_i^{(t)} &= Z^{(t)}M^{(t)}\phi(X_i^{(t)})^T + \widetilde{Z}^{(t-1)}\widetilde{M}^{(t-1)}\phi(\widetilde{X}_i^{(t-1)})^T \\ &= Z^{(t)}M^{(t)}\phi(X_i^{(t)})^T + \widetilde{C}_i^{(t-1)},\end{aligned} \tag{16}$$

$$\begin{aligned}D_i^{(t)} &= \phi(X_i^{(t)})\phi(X_i^{(t)})^T + \phi(\widetilde{X}_i^{(t-1)})\phi(\widetilde{X}_i^{(t-1)})^T \\ &= \phi(X_i^{(t)})\phi(X_i^{(t)})^T + \widetilde{D}_i^{(t-1)},\end{aligned} \tag{17}$$

(2) Update $Z^{(t)}$ and $M^{(t)}$ by fixing other variables. Eq. (13) can be rewritten as follows:

$$\begin{aligned}\mathcal{L} &= \mu\sum_{i=1}^{2}\left\|W_i^{(t)}\phi(X_i^{(t)}) - Z^{(t)}M^{(t)}\right\|_F^2 \\ &+ \gamma\left\|M^{(t)} - Z^{(t)}M^{(t)}\right\|_F^2 \\ &+ \frac{\rho}{2}\left\|L^{(t)} - Z^{(t)}M^{(t)} - E^{(t)} + \frac{F_1}{\rho}\right\|_F^2 \\ &+ \frac{\rho}{2}\left\|Z^{(t)}M^{(t)} - N^{(t)} + \frac{F_2}{\rho}\right\|_F^2.\end{aligned} \tag{18}$$

By setting $\frac{\partial\mathcal{L}}{\partial Z^{(t)}} = 0$ and $\frac{\partial\mathcal{L}}{\partial M^{(t)}} = 0$, respectively, the updating rules of $Z^{(t)}$ and $M^{(t)}$ are derived as follows:

$$\begin{aligned}Z^{(t)} = &\left(\left(\mu\sum_{i=1}^{2}W_i^{(t)}\phi(X_i^{(t)}) + \frac{\rho}{2}(L^{(t)} - E^{(t)} + \frac{F_1}{\rho})\right.\right. \\ &\left.\left. + \frac{\rho}{2}(N^{(t)} - \frac{F_2}{\rho}) + \gamma M^{(t)}\right)M^{(t)T}\right)\cdot A^{-1},\end{aligned} \tag{19}$$

$$\begin{aligned}M^{(t)} = &\left((2\mu+\gamma+\rho)Z^{(t)T}Z^{(t)} + \gamma(I - Z^{(t)})\right)^{-1}\cdot \\ &\left(Z^{(t)T}\left(\mu\sum_{i=1}^{2}W_i^{(t)}\phi(X_i^{(t)}) + \frac{\rho}{2}(N^{(t)} - \frac{F_2}{\rho}) + \frac{\rho}{2}(L^{(t)} - E^{(t)} + \frac{F_1}{\rho})\right)\right),\end{aligned} \tag{20}$$

where $A = (2\mu + \gamma + \rho)M^{(t)}M^{(t)T}$.

(3) Update $E^{(t)}$ by fixing other variables. The model (13) can be rewritten as follows:

$$\mathcal{L} = \alpha\beta\left\|E^{(t)}\right\|_1 + \frac{\rho}{2}\left\|L^{(t)} - Z^{(t)}M^{(t)} - E^{(t)} + \frac{F_1}{\rho}\right\|_F^2. \tag{21}$$

According to shrinkage operator [35], the closed solution of $E^{(t)}$ can be obtained as follows:

$$E^{(t)} = K_{\alpha\beta/\rho}\left(L^{(t)} - Z^{(t)}M^{(t)} + \frac{F_1}{\rho}\right), \tag{22}$$

where $K_{\alpha\beta/\rho}(A_1)$ is the shrinkage operator, and its definitions is given as follows:

$$K_{\alpha\beta/\rho}(A_1) = \mathrm{sgn}(A_1)\cdot\max(|A_1| - \alpha\beta/\rho, 0). \tag{23}$$

(4) Update $N^{(t)}$ by fixing other variables. Eq. (13) is expressed as the following problem:

$$N^{(t)} = \arg\min\alpha\|N^{(t)}\|_* + \frac{\rho}{2}\left\|Z^{(t)}M^{(t)} - N^{(t)} + \frac{F_2}{\rho}\right\|_F^2, \tag{24}$$

The updating rule of $N^{(t)}$ can be derived by using the singular value thresholding (SVT) [36] operator as follows:

$$N^{(t)} = \kappa_{\alpha/\rho}\left(Z^{(t)}M^{(t)} + \frac{F_2}{\rho}\right), \tag{25}$$

where $\kappa_\mu[J]$ is the singular value thresholding, and is defined as follows:

$$\kappa_\mu[J] = U\Sigma_{\mu+}H^T, \Sigma_{\mu+} = \mathrm{diag}\left(\{\delta - \mu\}_+\right), \tag{26}$$

where U and H are the left-singular and right-singular matrices of J with orthogonal columns, respectively. δ is the singular value of the matrix J and $\{\delta - \mu\}_+ = \max(0, \delta - \mu)$.

(5) The Lagrangian multiplier F_1, F_2 and the penalty factor ρ can be updated by

$$F_1 = F_1 + \rho\left(L^{(t)} - Z^{(t)}M^{(t)} - E^{(t)}\right), \tag{27}$$

$$F_2 = F_2 + \rho\left(Z^{(t)}M^{(t)} - N^{(t)}\right), \tag{28}$$

$$\rho = \min\left(\rho_{\max}, \varepsilon\rho\right), \tag{29}$$

where $\rho_{\max}$ is a pre-defined large number and ε is a constant greater than 1.

(6) Update $V^{(t)}$ by fixing other variables. The sub-problem w.r.t $V^{(t)}$ is expressed as follows:

$$\begin{aligned} &\mathcal{L} = \delta\left\|V^{(t)} - B^{(t)}\right\|_F^2 + \left\|V^{(t)T}B^{(t)} - rG^{(t)T}G^{(t)}\right\|_F^2 \\ &s.t. \quad V^{(t)}V^{(t)T} = n_t I_r, V^{(t)}1_{n_t} = 0_r, \end{aligned} \tag{30}$$

Eq. (30) can be converted into the following problem:

$$\begin{aligned} &\max_{V} \operatorname{Tr}\left(RV^{(t)T}\right), \\ &s.t. \quad V^{(t)}V^{(t)T} = n_t I_r, V^{(t)}1_{n_t} = 0_r, \end{aligned} \tag{31}$$

where $R = rB^{(t)}G^{(t)T}G^{(t)} + \delta B^{(t)}$. Eq. (31) can be solved by performing singular value decomposition on RU_1R^T as follows:

$$RU_1R^T = [Q, \widetilde{Q}]\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}[Q, \widetilde{Q}]^T, \tag{32}$$

where U_1 is defined as $U_1 = I_{n_t} - \frac{1}{n_t}1_{n_t}1_{n_t}^T$, and $\Sigma \in \mathbb{R}^{r_s \times r_s}$ and $Q \in \mathbb{R}^{r \times r_s}$ are diagonal matrices with positive eigenvalues and corresponding eigenvectors, respectively. r_s is the rank of RU_1R^T. By performing a Gram–Schmidt process on $\widetilde{Q}$, the orthogonal matrix $\overline{\mathbf{Q}} \in \mathbb{R}^{r \times (r - r_s)}$ can be obtained. Thus, we have $\Gamma = U_1R^TQ\Sigma^{-1/2}$ and then generate a random orthogonal matrix $\overline{\Gamma} \in \mathbb{R}^{n_t \times (r - r_s)}$. According to Ref. [37], the updating rule of $V^{(t)}$ is given as follows:

$$V^{(t)} = \sqrt{n_t}\begin{bmatrix} Q & \overline{Q} \end{bmatrix}\begin{bmatrix} \Gamma & \overline{\Gamma} \end{bmatrix}^T. \tag{33}$$

In Eq. (33), if $r = r_s$, $\overline{Q}$ and $\overline{\Gamma}$ are empty.

(7) Update $B^{(t)}$ by fixing other variables. Eq. (13) can be rewritten as follows:

$$\begin{aligned} &\mathcal{L} = \delta\left\|V^{(t)} - B^{(t)}\right\|_F^2 + \left\|V^{(t)T}B^{(t)} - rG^{(t)T}G^{(t)}\right\|_F^2 \\ &\quad + \left\|\widetilde{V}^{(t-1)T}B^{(t)} - r\widetilde{G}^{(t-1)T}G^{(t)}\right\|_F^2 \\ &s.t. \quad B^{(t)} \in \{-1, 1\}^{r \times n_t}. \end{aligned} \tag{34}$$

We can get the closed solution of $B^{(t)}$ as follows:

$$B^{(t)} = \operatorname{sgn}\left(\delta V^{(t)} + rC_3^{(t)}G^{(t)}\right), \tag{35}$$

$$\begin{aligned} C_3^{(t)} &= V^{(t)}G^{(t)T} + \widetilde{V}^{(t-1)}\widetilde{G}^{(t-1)T} \\ &= V^{(t)}G^{(t)T} + \widetilde{C}_3^{(t-1)}, \end{aligned} \tag{36}$$

where $sgn(\cdot)$ is a symbolic function.

The optimization algorithm of our ROHLSE approach is summarized in Algorithm 1.

3.4. Hash functions learning

In the query stage, the ROHLSE method can convert high-dimensional multi-modal data into binary codes. Here, we learn the hash functions through the least-squares regression. To achieve online learning for streaming data, the hash functions is given as follows:

$$\begin{aligned} \min_{P_1^{(t)}, P_2^{(t)}} &\sum_{i=1}^{2}\left(\left\|B^{(t)} - P_i^{(t)}\phi(X_i^{(t)})\right\|_F^2 + \lambda\left\|P_i^{(t)}\right\|_F^2\right) \\ &+ \sum_{i=1}^{2}\left\|\widetilde{B}^{(t-1)} - P_i^{(t)}\phi(\widetilde{X}_i^{(t-1)})\right\|_F^2, \end{aligned} \tag{37}$$

where $P_i^{(t)} \in \mathbb{R}^{r \times m}$ is the projection matrix.

Let the derivatives of Eq. (37) w.r.t $P_i^{(t)}$ be equal to zero, and we can get the solution of $P_i^{(t)}$ as follows:

$$P_i^{(t)} = T_i^{(t)}\left(D_i^{(t)} + \gamma I\right)^{-1}, \tag{38}$$

Algorithm 1 ROHLSE

Training stage

Input: Data chunks $\left[X^{(1)}, X^{(2)}, \ldots, X^{(t)}\right]$ and the associated observed labels $\left[L^{(1)}, L^{(2)}, \ldots, L^{(t)}\right]$, parameters λ, α, β, μ, δ and γ, and the maximum iteration times τ.

Output: Hash codes B, projection matrix $P_1^{(t)}$ and $P_2^{(t)}$.

Procedure:

1. Calculate $\phi(X_i^{(t)})$.
2. **Initialize** $C_1^{(0)}$, $C_2^{(0)}$, $C_3^{(0)}$, $D_1^{(0)}$, $D_2^{(0)}$, $T_1^{(0)}$ and $T_2^{(0)}$ by zero matrices.
3. **for** $chunk = 1 \to t$

(1)**Initialize** $Z^{(t)}$, $M^{(t)}$, $V^{(t)}$, F_1, F_2, $E^{(t)}$ and $B^{(t)}$ matrices.

(2) Update $D_i^{(t)}(i = 1, 2)$ by Eq.(17).

Repeat

(3) Update $B^{(t)}$ by Eq.(35).

(4) Update $Z^{(t)}$ and $M^{(t)}$, by Eqs.(19) and (20).

(5) Update $N^{(t)}$ by Eq.(25).

(6) Update $W_i^{(t)}(i = 1, 2)$ by Eq.(15).

(7) Update $E^{(t)}$ by Eq.(22)

(8) Update $V^{(t)}$ by Eq.(33)

(9) Update F_1, F_2 and ρ by Eqs.(27), (28) and (29).

Until convergence or reaching the maximum iterations.

(10) Update $P_i^{(t)}(i = 1, 2)$ by Eq.(38).

end for

Retrieval Stage

Input: Feature matrix X_{1quey} or X_{2quey} of the query data, projection matrix $P_1^{(t)}$ and $P_2^{(t)}$.

Output: B_{1query} and B_{2query}.

1. Calculate $\phi(X_{1quey})$ and $\phi(X_{2quey})$.
2. For X_{1quey}: calculate hash codes by Eq.(40).
3. For X_{2quey}: calculate hash codes by Eq.(41).

where $T_i^{(t)}$ is expressed as follows:

$$\begin{aligned} T_i^{(t)} &= B^{(t)}\phi(X_i^{(t)})^T + \widetilde{B}^{(t-1)}\phi(\widetilde{X}_i^{(t-1)})^T \\ &= B^{(t)}\phi(X_i^{(t)})^T + \widetilde{T}_i^{(t-1)}, \end{aligned} \tag{39}$$

For the query samples x_{1query} and x_{2query} from different modalities, ROHLSE generates their corresponding hash codes by the following formulas:

$$B_{1query} = \operatorname{sgn}\left(P_1^{(t)}\phi\left(x_{1query}\right)\right), \tag{40}$$

$$B_{2query} = \operatorname{sgn}\left(P_2^{(t)}\phi\left(x_{2query}\right)\right), \tag{41}$$

where $\phi\left(x_{1\text{query}}\right)$ and $\phi\left(x_{2\text{query}}\right)$ are the nonlinear embedding of $x_{1\text{query}}$ and $x_{2\text{query}}$, respectively.

3.5. Complexity analysis

The proposed ROHLSE approach can learn hash functions and hash codes incrementally for streaming data. Therefore, we give the computational complexity of updating each variant in the tth round. Here, τ denotes the number of iterations. Specifically, the computational complexity of updating variables $B^{(t)}$ and $W_i^{(t)}$ is $O(crn\tau)$ and $O\left(\left(\left(c^2 + m^2 + cm\right)n + m^2c + m^3\right)\tau\right)$, respectively. In addition, the computational complexity of updating variables $Z^{(t)}$ and $M^{(t)}$ is $O\left(\left(\left(6c^2 + 2cm\right)n + c^3\right)\tau\right)$ and $O\left(\left(\left(5c^2 + 2cm\right)n + 2c^3\right)\tau\right)$, respectively. The computational complexity of updating variables $E^{(t)}$ and $V^{(t)}$ is $O\left(c^2n\tau\right)$ and $O(crn\tau)$, respectively. Since m, r and $c \ll n$, we set $a = \max\{m, r, c\}$. Therefore, the computational complexity of ROHLSE in the tth round is $O\left(a^2n\tau\right)$. It can be seen that the overall complexity of our ROHLSE approach in each round is linear with the training set size n.

Table 1
Training set division of each dataset.

Division details	MIRFlickr	NUS-WIDE	IAPR TC-12
Training set size	15,902	184,710	18,000
Number of chunks	8	19	9
Chunk size except the last	2,000	10,000	2,000
Last chunk size	1,902	4,710	2,000

4. Experiments

In this section, we conduct extensive experiments on three benchmark datasets to verify the effectiveness of the proposed ROHLSE approach by comparing it with other state-of-the-art OCMH methods.

4.1. Datasets

MIRFlickr [38] consists of 25,000 image-text pairs from 24 categories. Here, we choose 16,738 pairs as the experimental dataset. Each image and text sample is represented by a 150-dimensional GIST feature vector and a 500-dimensional bag-of-words vector, respectively. 836 image-text pairs are randomly selected as the retrieval dataset, and the remaining 15,902 image-text pairs are used as the training dataset.

NUS-WIDE [39] contains 269,648 instances collected from Flickr. Here, we choose the top 10 common classes and the corresponding 186,577 image-text pair instances. Specifically, 1867 instances are randomly selected as retrieval dataset and the rest as training dataset. The samples from image and text modalities are characterized by 500-dimensional and 1000-dimensional feature vectors, respectively.

IAPR TC-12 [40] includes 20,000 image-text pairs belonging to 255 categories. Each image and text is represented as a 512-dimensional GIST feature vector and a 2912-dimensional bag-of-words vector, respectively. We randomly sample 2000 image-text pairs as the retrieval dataset and the rest as the training dataset.

To simulate steaming data on the aforementioned benchmark datasets, we divide them into several data chunks. The division details of the training dataset are provided in Table 1.

4.2. Baselines and evaluation metrics

To verify the effectiveness of our ROHLSE approach, we compare it with several state-of-the-art OCMH methods. These methods include OASH [6], OLSH [7], DOCH [8], FOMH [28], LEMON [30] and OS-CMFH [31] . It is noteworthy that these methods belong to supervised learning methods, thus utilizing the supervision information among multi-modalities.

We adopt the mean Average Precision (mAP) to evaluate the retrieval performances of all methods, and its definition can be given as follows:

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{N}\sum_{r=1}^{R}P_q(r)\delta_q(r), \tag{42}$$

where Q and N are the query instance and the number of relevant instances in the retrieval set, respectively, and R is the total number of retrieved data points. $P_q(r)$ represents the retrieval accuracy of top-r. Here, $\delta_q(r) = 1$ if the qth query instance is related to the rth instance, otherwise $\delta_q(r) = 0$.

In the experiment, we first give the mAP results after the last data chunk arrives. Then, we give the mAP results of online hashing approaches after each new data chunk arrives.

4.3. Experimental settings

There are several parameters in our ROHLSE approach, such as α, β, μ, γ, δ, and λ. In the following experiments, we set them as $\{1, 10^{-1}, 10^{-1}, 1, 10^{-3}, 10^{-2}\}$ on the MIRFlickr dataset, $\{10^{-2}, 10^{2}, 10^{-1}, 1, 10^{-3}, 10^{-3}\}$ on the NUS-WIDE dataset and $\{10^{-2}, 10^{2}, 10^{-1}, 10^{-3}, 10^{-3}, 10^{-1}\}$ on the IAPR TC-12 dataset, respectively. Furthermore, the number of anchors m is set to 1000 on three datasets. In the experiments, we run each method five times and record the average performance as the result.

To evaluate the robustness of the ROHLSE approach, we set different noise ratios (NR) for the label matrix to simulate missing and wrong labels. When NR is set to 0.2, it indicates that 20% missing labels and 20% wrong labels are included in the label matrix, respectively. The settings of missing labels and wrong labels are given as follows: (1) A certain ratio of labels is randomly selected from the training dataset to simulate missing labels. If the total number of elements with the value of "1" in the label matrix is z, then the $z * \mathrm{NR}$ elements with the value of "1" in the label matrix are set to "0"; (2) Some labels are randomly chosen from training dataset to simulate the wrong label. Then $z * \mathrm{NR}$ elements with the value "0" in the label matrix are modified to "1". Therefore, the processed label matrix contains missing labels and wrong labels. Finally, we employ the processed training dataset to train all hashing models.

4.4. Experimental results

To evaluate the robustness of our ROHLSE approach under different noise conditions, we set the NR value to 0.2, 0.3, and 0.4, respectively. Tables 2–4 show the mAP values of all methods on the MIRFickr, NUS-WIDE, and IAPR TC-12 datasets with different hash code lengths, respectively. From the experiment results on three multi-modal datasets, we can get the following observations:

(1) Our ROHLSE approach achieves the best performance under different noise ratio settings in different retrieval tasks. It indicates that our ROHLSE approach has shown strong robustness for large-scale multi-modal data with noisy labels in cross-modal retrieval. In addition, it can be found that LEMON outperforms DOCH in different retrieval tasks on three datasets. One main reason may be that DOCH only selects some anchors of old data to establish chunk similarity with new data, while LEMON constructs the chunk similarity using new data and all old data. Thus, the learned hash codes of LEMON embed more semantic information than DOCH.

(2) It can be seen that the mAP values of most hashing approaches decrease as the noise ratio increases. It indicates that the noisy labels directly lead to retrieval performance degradation of hashing approaches. Furthermore, our ROHLSE approach always achieves the best performance under different noise ratios, which shows that our ROHLSE approach is more suitable for real tasks with noisy labels. In addition, the retrieval performances of most approaches are constantly improving with the increase of hash code length. This is because longer hash codes can embed more semantic information hidden in multi-modal data.

(3) We can know that LEMON, DOCH, and ROHLSE can achieve better performances than other comparison approaches on three datasets. This is because these three hashing approaches construct the pairwise similarity using label information, and then employ it to guide the learning of hash codes. Therefore, they can generate more discriminative hash codes than other competitors, and hence achieve better retrieval performances on different datasets.

(4) It is clear that the performances of most approaches in text retrieval image tasks are better than in image retrieval text tasks. One possible reason is that the samples from the image modality may lose relatively less information than from the text modality when they are mapped to hash codes. In other words, visual features contain

Table 2
The mAP values of each approach on the MIRFlickr dataset.

Task	Method	NR = 0.2				NR = 0.3				NR = 0.4			
		16	32	64	128	16	32	64	128	16	32	64	128
Image to Text	OLSH	0.5945	0.6015	0.6020	0.6098	0.5879	0.5968	0.5988	0.6032	0.5823	0.5951	0.6016	0.5996
	FOMH	0.5869	0.6105	0.6213	0.6268	0.5778	0.5972	0.6111	0.6184	0.5712	0.5881	0.5903	0.6052
	OASH	0.6365	0.6500	0.6631	0.6701	0.6294	0.6440	0.6563	0.6570	0.6071	0.6258	0.6330	0.6421
	DOCH	0.6732	0.6803	0.6829	0.6840	0.6526	0.6548	0.6620	0.6614	0.6315	0.6374	0.6422	0.6488
	LEMON	0.6833	0.6932	0.7001	0.7031	0.6640	0.6746	0.6824	0.6866	0.6416	0.6509	0.6593	0.6639
	OSCMFH	0.6538	0.6608	0.6607	0.6638	0.6440	0.6456	0.6627	0.6549	0.6298	0.6381	0.6385	0.6390
	ROHLSE	**0.7227**	**0.7308**	**0.7337**	**0.7366**	**0.6983**	**0.7088**	**0.7143**	**0.7168**	**0.6689**	**0.6789**	**0.6883**	**0.6887**
Text to Image	OLSH	0.5893	0.5973	0.6029	0.6110	0.5865	0.5965	0.6025	0.6101	0.5825	0.5845	0.5947	0.5998
	FOMH	0.5961	0.6323	0.6503	0.6553	0.5832	0.6102	0.6272	0.6412	0.5789	0.5978	0.6083	0.6187
	OASH	0.6834	0.7116	0.7303	0.7368	0.6697	0.6830	0.7109	0.7141	0.6371	0.6592	0.6649	0.6763
	DOCH	0.7626	0.7716	0.7784	0.7827	0.7226	0.7280	0.7378	0.7412	0.6797	0.6877	0.6953	0.7006
	LEMON	0.7718	0.7805	0.7884	0.7930	0.7337	0.7495	0.7636	0.7696	0.6920	0.7040	0.7170	0.7249
	OSCMFH	0.7243	0.7317	0.7336	0.7403	0.7010	0.7073	0.7274	0.7211	0.6744	0.6878	0.6884	0.6912
	ROHLSE	**0.7889**	**0.8035**	**0.8079**	**0.8123**	**0.7573**	**0.7714**	**0.7783**	**0.7824**	**0.7117**	**0.7312**	**0.7380**	**0.7408**

Table 3
The mAP values of each approach on the NUS-WIDE dataset.

Task	Method	NR = 0.2				NR = 0.3				NR = 0.4			
		16	32	64	128	16	32	64	128	16	32	64	128
Image to Text	OLSH	0.4470	0.4698	0.4919	0.5103	0.4404	0.4631	0.4859	0.5083	0.4342	0.4522	0.4766	0.5001
	FOMH	0.5274	0.4651	0.5158	0.5290	0.5196	0.4779	0.5094	0.4986	0.5075	0.4807	0.4745	0.4697
	OASH	0.5355	0.5748	0.5683	0.5791	0.4994	0.5422	0.5445	0.5503	0.4696	0.4971	0.5004	0.5080
	DOCH	0.5786	0.5803	0.5899	0.5983	0.5423	0.5515	0.5592	0.5579	0.4927	0.5055	0.5103	0.5158
	LEMON	0.5827	0.5928	0.6009	0.6070	0.5404	0.5575	0.5604	0.5656	0.4959	0.5073	0.5169	0.5217
	OSCMFH	0.5374	0.5701	0.5761	0.5819	0.5119	0.5119	0.5223	0.5225	0.4750	0.4734	0.4777	0.4886
	ROHLSE	**0.6182**	**0.6189**	**0.6317**	**0.6325**	**0.5959**	**0.5994**	**0.6030**	**0.6046**	**0.5527**	**0.5517**	**0.5601**	**0.5645**
Text to Image	OLSH	0.4522	0.4870	0.5090	0.5355	0.4492	0.4806	0.5089	0.5310	0.4418	0.4773	0.5023	0.5201
	FOMH	0.5601	0.4941	0.5825	0.6003	0.5554	0.5090	0.5533	0.5487	0.5269	0.5066	0.5065	0.5048
	OASH	0.6378	0.6879	0.6812	0.6902	0.5909	0.6269	0.6416	0.6454	0.5415	0.5597	0.5723	0.5765
	DOCH	0.6841	0.6998	0.7019	0.7106	0.6264	0.6377	0.6477	0.6459	0.5532	0.5690	0.5763	0.5798
	LEMON	0.6949	0.7001	0.7126	0.7205	0.6275	0.6404	0.6489	0.6583	0.5581	0.5701	0.5845	0.5884
	OSCMFH	0.6105	0.6618	0.6679	0.6764	0.5811	0.5806	0.5909	0.5959	0.5178	0.5235	0.5287	0.5325
	ROHLSE	**0.7156**	**0.7237**	**0.7348**	**0.7423**	**0.6802**	**0.6873**	**0.6951**	**0.6969**	**0.6180**	**0.6215**	**0.6303**	**0.6316**

Table 4
The mAP values of each approach on the IAPR TC-12 dataset.

Task	Method	NR = 0.2				NR = 0.3				NR = 0.4			
		16	32	64	128	16	32	64	128	16	32	64	128
Image to Text	OLSH	0.3986	0.4036	0.4108	0.4157	0.3815	0.3986	0.4065	0.4102	0.3765	0.3903	0.4002	0.4057
	FOMH	0.3581	0.3650	0.3751	0.3752	0.3425	0.3510	0.3611	0.3607	0.3305	0.3425	0.3481	0.3509
	OASH	0.3308	0.3452	0.3508	0.3660	0.3252	0.3345	0.3400	0.3566	0.3204	0.3275	0.3320	0.3433
	DOCH	0.4287	0.4438	0.4556	0.4634	0.4033	0.4217	0.4304	0.4339	0.3822	0.3994	0.4155	0.4214
	LEMON	0.4392	0.4624	0.4797	0.4926	0.4243	0.4448	0.4621	0.4742	0.4092	0.4245	0.4428	0.4552
	OSCMFH	0.4168	0.4245	0.4347	0.4413	0.4091	0.4187	0.4264	0.4292	0.4019	0.4084	0.4162	0.4213
	ROHLSE	**0.4630**	**0.4860**	**0.5003**	**0.5099**	**0.4429**	**0.4658**	**0.4808**	**0.4910**	**0.4271**	**0.4463**	**0.4616**	**0.4736**
Text to Image	OLSH	0.4078	0.4256	0.4350	0.4445	0.4072	0.4198	0.4286	0.4401	0.4044	0.4124	0.4146	0.4201
	FOMH	0.3795	0.3893	0.4081	0.4112	0.3564	0.3680	0.3868	0.3915	0.3420	0.3567	0.3635	0.3742
	OASH	0.3514	0.3680	0.3980	0.4267	0.3409	0.3569	0.3764	0.4012	0.3306	0.3456	0.3579	0.3789
	DOCH	0.4755	0.5062	0.5247	0.5389	0.4435	0.4720	0.4912	0.4961	0.4126	0.4358	0.4636	0.4746
	LEMON	0.5307	0.5696	0.5985	0.6172	0.5023	0.5374	0.5641	0.5804	0.4778	0.5087	0.5256	0.5419
	OSCMFH	0.4668	0.4925	0.5135	0.5283	0.4518	0.4734	0.4954	0.5058	0.4373	0.4572	0.4725	0.4837
	ROHLSE	**0.5370**	**0.5788**	**0.6051**	**0.6211**	**0.5151**	**0.5502**	**0.5745**	**0.5931**	**0.4892**	**0.5252**	**0.5476**	**0.5626**

less semantic information than text features from the perspective of semantic information.

To evaluate the performances of hashing approaches in online learning scenarios, we give their mAP values after each data chunk arrives. We conduct some retrieval experiments on the MIRFlickr, NUS-WIDE, and IAPR TC-12 datasets, where the hash code length is set to 64 bits and the noise ratio is set to 0.2, 0.3, and 0.4, respectively. The mAP curves of all approaches on three datasets are shown in Figs. 2–4. We can see that the mAP value of most methods rises with the increase of new data chunks. It indicates that the retrieval performance of the online hashing approach is improved continuously with the continuous arrival of training data. Therefore, online hashing is more suitable for the retrieval tasks of large-scale streaming data. In addition, we can see that our proposed ROHLSE approach outperforms other online hashing approaches on three datasets, which verifies the superiority of our ROHLSE approach.

4.5. Parameter sensitivity analysis

To evaluate the influence of parameters in the ROHLSE method on retrieval performance, we conduct some experiments with different noise ratio settings on the NUS-WIDE dataset. Here, the hash code length is set to 16 bits. In this experiment, we change the values of two parameters by fixing other parameters. Figs. 5–7 show the mAP values of our ROHLSE approach on the NUS-WIDE dataset with different parameter settings.

Fig. 5 plots the retrieval results of ROHLSE with varying the value of the parameters α and β under different noise ratio settings. It is

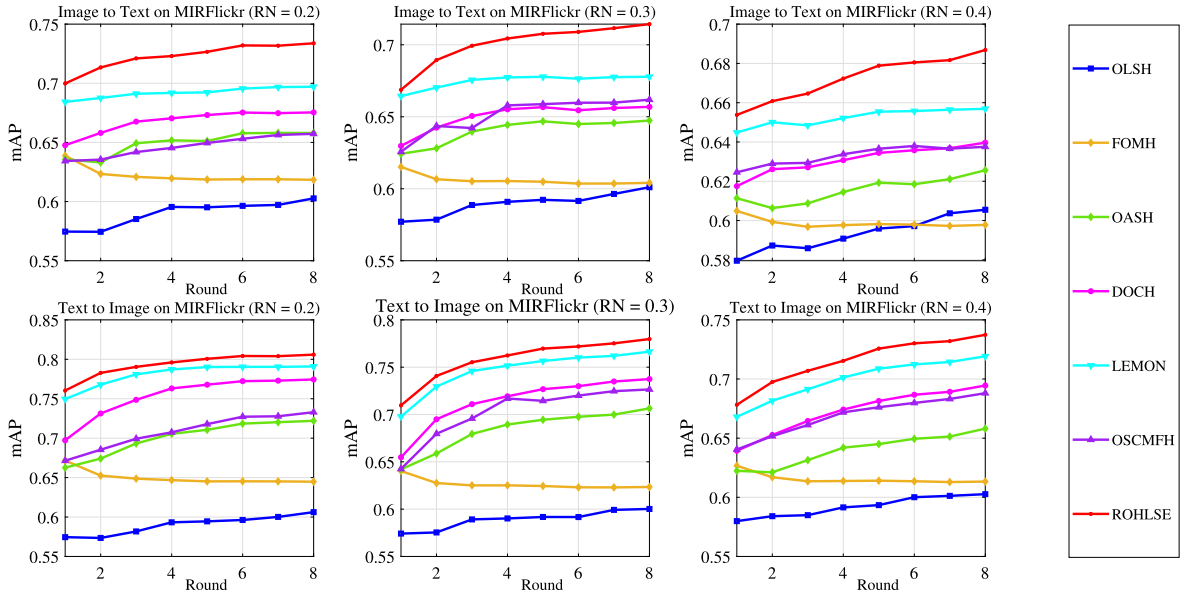


Fig. 2. The mAP values in each round on the MIRFlickr (64 bits) dataset.

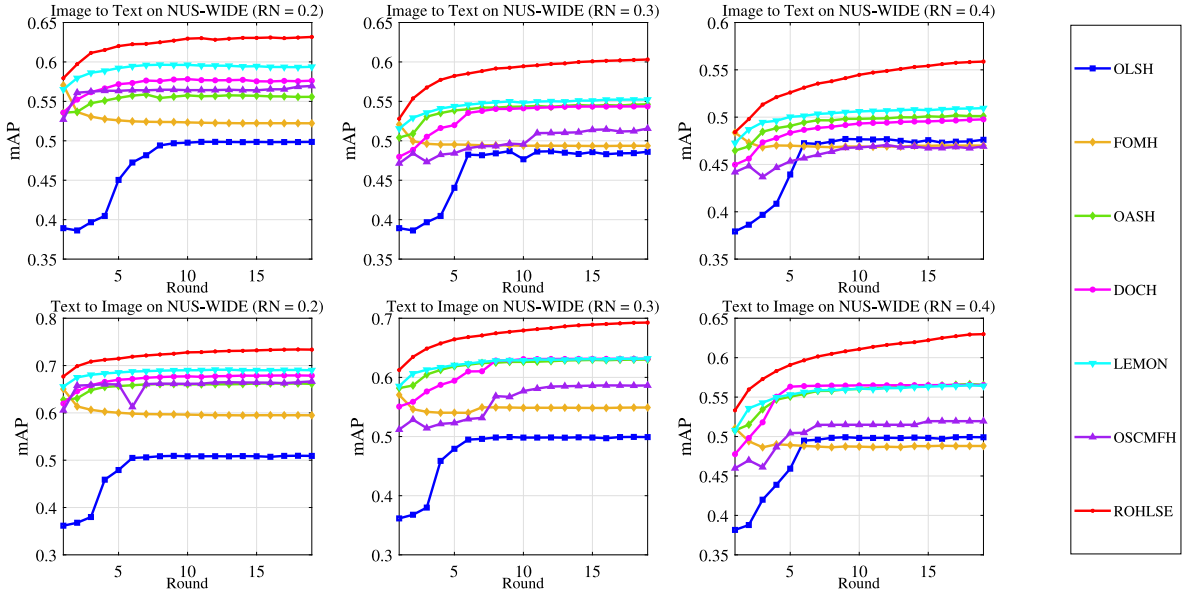


Fig. 3. The mAP values in each round on the NUS-WIDE (64 bits) dataset.

clear to see that α is used to control the low-rank and sparse constraint term, and β aims to control the sparse constraint term. From Fig. 5, we can find that the ROHLSE approach can achieve stable retrieval performance in a wide range of parameter values. In addition, it can be seen from Fig. 6 that the mAP values of the proposed ROHLSE approach are also relatively stable with different settings of the parameters μ and γ in most cases. However, it is noteworthy that the retrieval performance of ROHLSE declines significantly when the parameter γ is assigned a large value and the parameter μ is assigned a small value. Fig. 7 shows the results of ROHLSE with different values of the parameters δ and λ under different noise ratio settings. It can be found that ROHLSE in the two retrieval tasks achieves relatively stable results with only slight fluctuations. Therefore, we can know that the parameters δ and λ are insensitive to retrieval performance. From Figs. 5–7, we can see that our ROHLSE approach achieves relatively stable results with a wide range of parameter values.

4.6. Ablation studies

To evaluate the effectiveness of each component in the proposed model, we conduct ablation experiments with different hash code lengths on the MIRFlickr dataset. ROHLSE considers the real case that there are missing and wrong labels in the observed labels. Meanwhile, it fully exploits the latent semantic correlations between labels, thereby generating more discriminative hash codes. Therefore, four variants are designed based on ROHLSE, namely ROHLSE-F, ROHLSE-L, ROHLSE-R and ROHLSE-D. To be specific, ROHLSE-F has removed the data

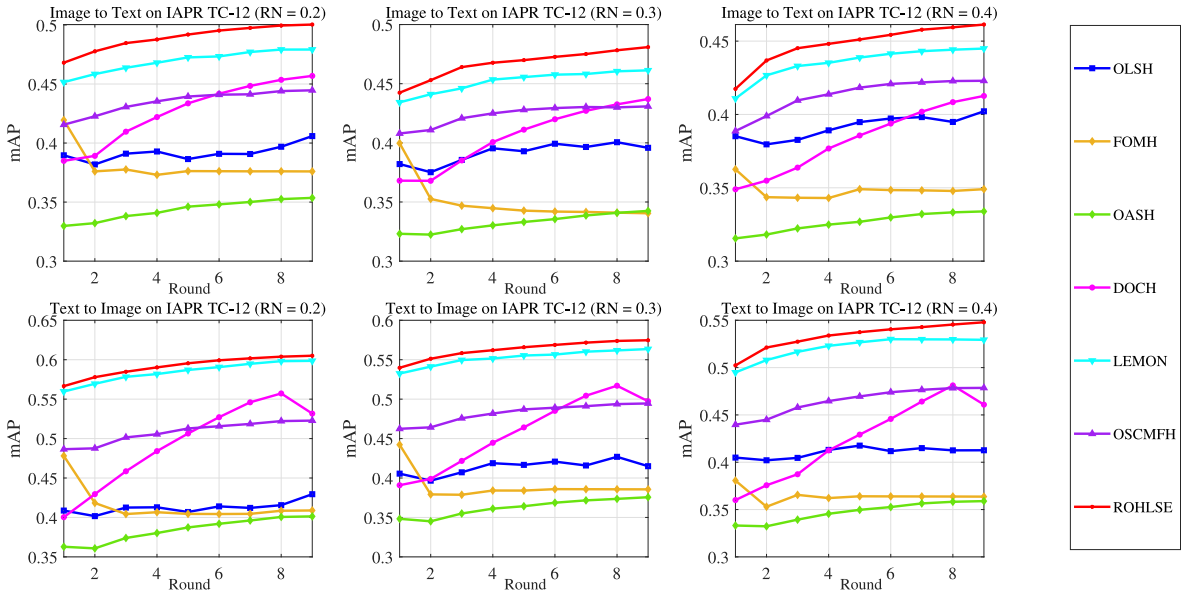


Fig. 4. The mAP values in each round on the IAPR TC-12 (64 bits) dataset.

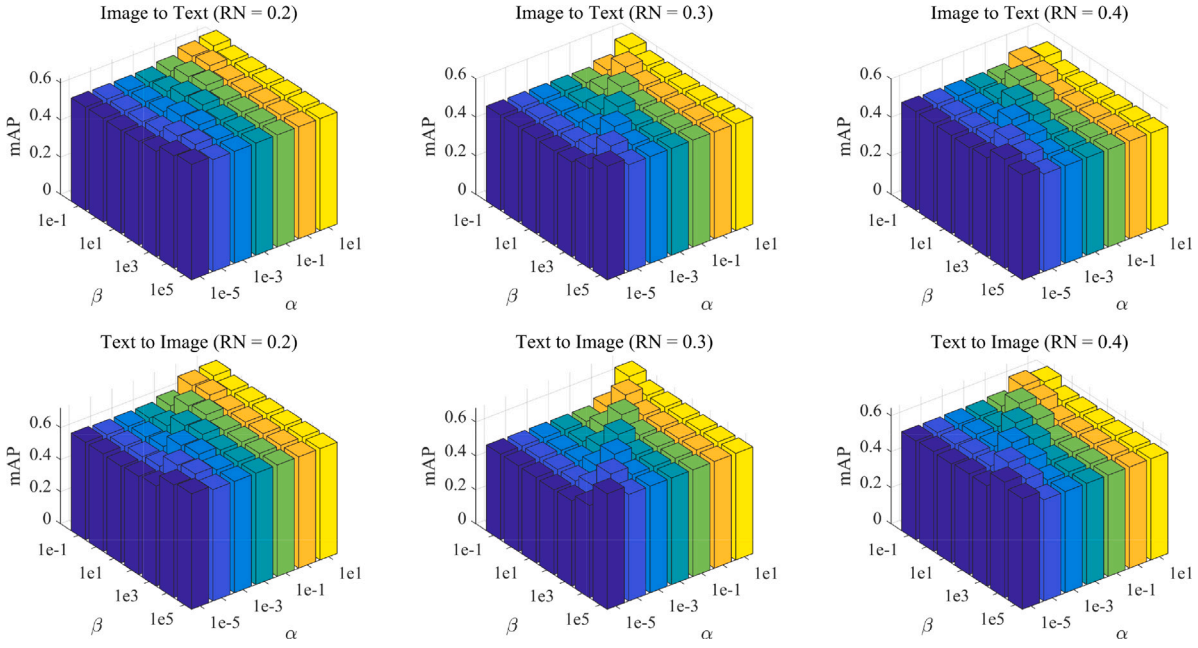


Fig. 5. The mAP values of ROHLSE with different values of parameters α and β on the NUS-WIDE dataset.

Table 5
The ablation study on different hash code lengths on the MIRFlickr dataset.

Task	Method	NR = 0.2				NR = 0.3				NR = 0.4			
		16	32	64	128	16	32	64	128	16	32	64	128
Image to Text	ROHLSE-F	0.7155	0.7228	0.7284	0.7307	0.6900	0.7010	0.7060	0.7104	0.6590	0.6696	0.6790	0.6825
	ROHLSE-L	0.6875	0.6891	0.6893	0.6901	0.6790	0.6832	0.6833	0.6836	0.6649	0.6738	0.6783	0.6782
	ROHLSE-R	0.6757	0.6801	0.6801	0.6794	0.6691	0.6728	0.6748	0.6744	0.6631	0.6688	0.6718	0.6723
	ROHLSE-D	0.6544	0.6725	0.6734	0.6721	0.6314	0.6419	0.6669	0.6578	0.6099	0.6248	0.6423	0.6439
	ROHLSE	**0.7227**	**0.7308**	**0.7337**	**0.7366**	**0.6983**	**0.7088**	**0.7143**	**0.7168**	**0.6689**	**0.6789**	**0.6883**	**0.6887**
Text to Image	ROHLSE-F	0.7809	0.7927	0.7993	0.8025	0.7481	0.7592	0.7670	0.7711	0.6982	0.7162	0.7258	0.7290
	ROHLSE-L	0.7324	0.7411	0.7437	0.7453	0.7253	0.7288	0.7323	0.7318	0.7069	0.7148	0.7215	0.7205
	ROHLSE-R	0.7165	0.7257	0.7264	0.7270	0.7103	0.7156	0.7195	0.7190	0.6975	0.7036	0.7091	0.7091
	ROHLSE-D	0.6950	0.7208	0.7225	0.7256	0.6621	0.6755	0.7103	0.7047	0.6256	0.6489	0.6723	0.6747
	ROHLSE	**0.7889**	**0.8035**	**0.8079**	**0.8123**	**0.7573**	**0.7714**	**0.7783**	**0.7824**	**0.7117**	**0.7312**	**0.7380**	**0.7408**

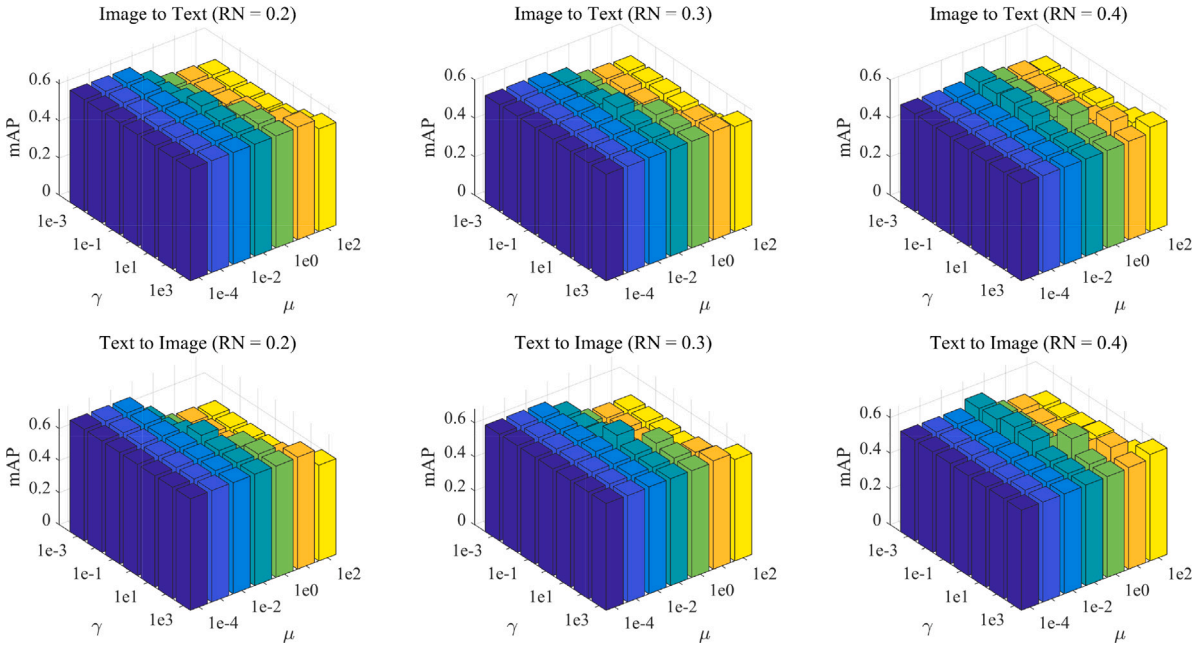


Fig. 6. The mAP values of ROHLSE with different values of parameters μ and γ on the NUS-WIDE dataset.

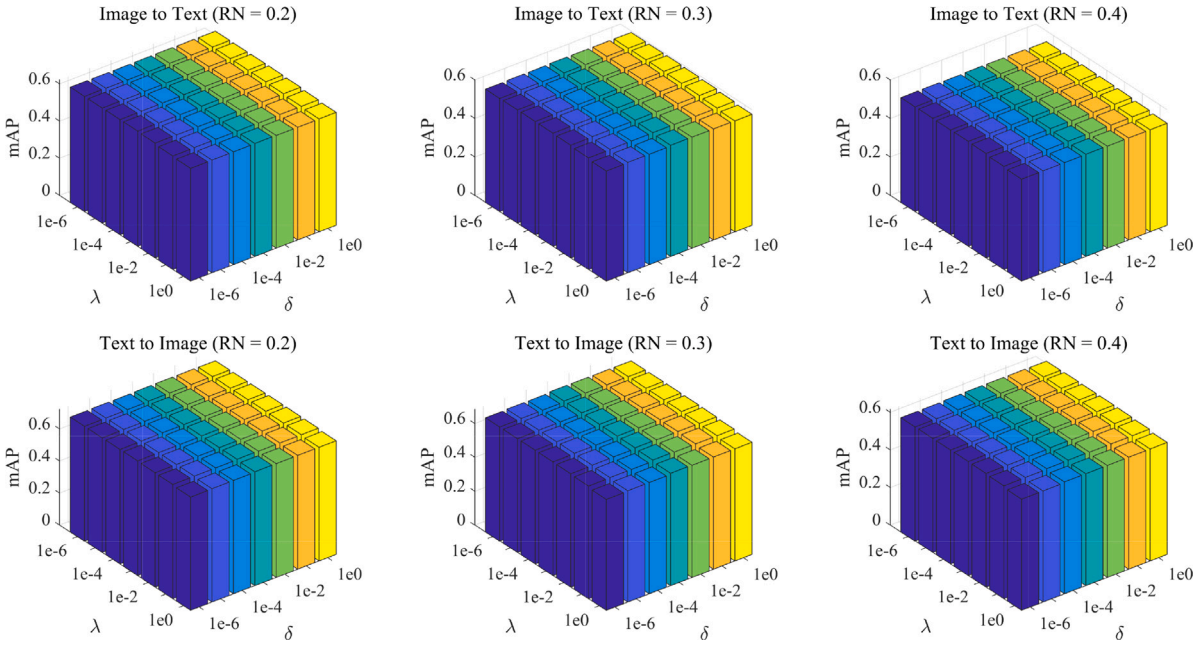
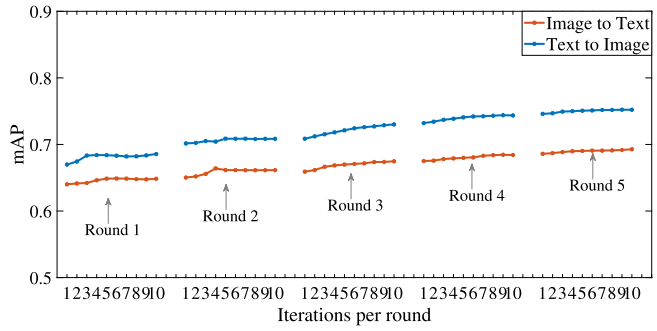
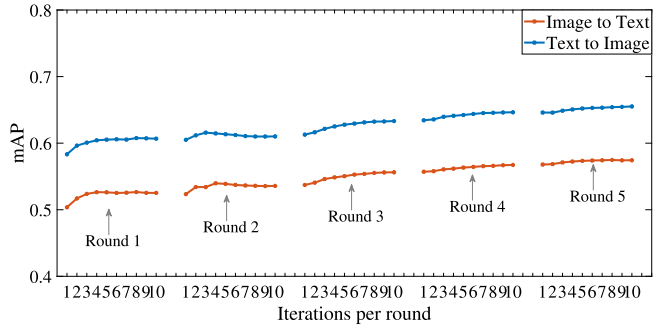


Fig. 7. The mAP values of ROHLSE with different values of parameter δ and λ on the NUS-WIDE dataset.

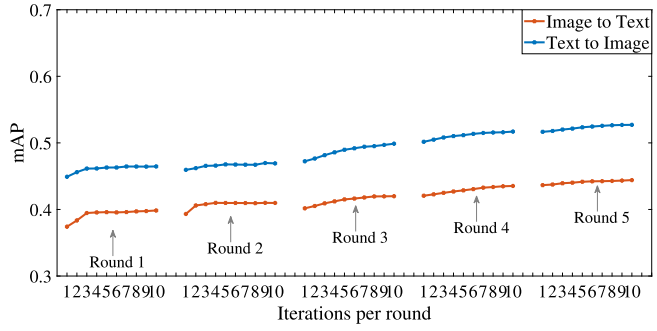
feature matrix to guide the learning of the clean labels compared with ROHLSE (i.e. $\mu = 0$); ROHLSE-L is to remove the latent correlation constraint between multiple labels; ROHLSE-R means to remove the low-rank and sparse constraints from the original model (i.e. $\alpha = 0$); ROHLSE-D is to ignore the semantic relationship between old and new data established by chunk similarity. Table 5 shows the mAP values of ROHLSE and its variant on the MIRFlickr dataset. It can be found that ROHLSE outperforms ROHLSE-F under different noise ratio settings. The main reason is that ROHLSE employs the data feature matrix to guide the learning of the clean labels. We can see that the retrieval results of the ROHLSE-L method are inferior to that of ROHLSE. This is because ROHLSE fully exploits the latent correlation between multiple labels. In addition, ROHLSE-R and ROHLSE-D outperform other variants with different hash code lengths. It indicates that the low-rank



(a) MIRFlickr



(b) NUS-WIDE



(c) IAPR TC-12

Fig. 8. Convergence analysis of ROHLSE on three datasets.

and sparse constraints can suppress the impact of noisy labels in real scenes. Furthermore, the results of ROHLSE-D show that the semantic relationship between old and new data can improve the retrieval results of the model by constructing chunk similarity.

4.7. Convergence analysis

The proposed ROHLSE approach develops an iterative optimization strategy to solve the proposed model. In this subsection, its convergence analysis is given on the MIRFlickr, NUS-WIDE, and IAPR TC-12 datasets. Here, the hash code length is set to 32 bits, and the noise ratio is set to 0.3. Fig. 8 shows the mAP values of the first five rounds with different iterations on three datasets. It can be seen that our ROHLSE approach can converge within ten iterations for each round on different datasets. This verifies the efficiency of the developed optimization scheme.

4.8. Time cost analysis

In this subsection, we have conducted some experiments to verify the efficiency of different methods on the MIRFlickr dataset. Here, we set the hash code length to 16, 32, 64, and 128 bits, respectively. The training time of all online hashing methods is shown in Table 6. We can see from Table 6 that the time cost of DOCH increases as the number of data chunks increases. This is because the number of anchor points is increasing cumulatively and thus its training time gradually increases among all online hashing methods. In addition, it can be found that all methods also consume more time as the length of the hash code increases. This indicates that the training cost of all methods is influenced by the hash code length. It is worth noting that our proposed method cannot achieve the highest efficiency among all hashing methods. The main reason is that our ROHLSE method needs additional time to perform low-rank and sparse decomposition on a noisy label matrix.

5. Conclusions

In this paper, we propose a novelty online hashing approach for cross-modal retrieval, i.e. robust online hashing with label semantic enhancement (ROHLSE). This approach not only employs online learning to deal with streaming data, but also takes into account the missing and wrong labels in the real scene. To be specific, ROHLSE seeks to recover the clean labels from the provided noisy labels by imposing low-rank and sparse constraints, thus enhancing the robustness of the proposed model. In addition, it can use the representation of samples to predict the labels by fully exploiting the dependency between the instances and their labels. To effectively avoid semantic forgetting in online learning, our ROHLSE approach establishes the relationship between new and old data through chunk similarity. Finally, our ROHLSE approach fully explores the semantic correlations between labels. Experiments on several benchmark datasets show that the proposed ROHLSE approach outperforms the state-of-the-art online hashing methods in cross-modal retrieval.

The proposed ROHLSE approach only deals with the fully-paired multi-modal streaming data with noisy labels. However, it cannot effectively solve the semi-paired cross-modal retrieval problem. In future work, we will further explore how to apply our proposed method to this scenario.

Declaration of competing interest

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted, and manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously.

Data availability

Data will be made available on request

Acknowledgments

This work was supported by the National Natural Science Foundation of China [Grant No. 62162033, U21B2027], Yunnan Foundation Research Projects, China [Grant No. 202201AT070154, 202101BE070001-056], Yunnan Provincial Major Science and Technology Special Plan Projects, China [Grant No. 202002AD080001, 202103AA080015], Yunnan Xingdian Talent Support Plan Project, China.

Table 6
Training time (seconds) of all methods on the MIRFlickr dataset.

Bits	Methods	1st	2st	3st	4st	5st	6st	7st	8st
16	OLSH	3.45	2.17	2.68	2.93	3.45	3.89	4.36	4.63
	FOMH	0.71	0.52	0.51	0.48	0.53	0.51	0.48	0.48
	OASH	0.91	0.52	0.52	0.51	0.53	0.53	0.51	0.50
	DOCH	1.45	4.53	12.09	19.23	25.95	32.06	37.55	41.50
	LEMON	0.15	0.15	0.12	0.11	0.12	0.10	0.13	0.12
	OSCMFH	0.45	0.21	0.45	0.26	0.23	0.30	0.26	0.21
	ROHLSE	3.45	2.17	2.68	2.93	2.45	2.89	2.36	2.63
32	OLSH	4.15	2.67	3.15	3.82	4.07	4.76	5.06	5.33
	FOMH	0.79	0.81	0.80	0.83	0.82	0.82	0.89	0.81
	OASH	1.08	0.75	0.77	0.78	0.75	0.72	0.73	0.65
	DOCH	5.08	19.39	45.28	76.04	90.33	100.04	136.53	147.62
	LEMON	0.19	0.18	0.18	0.16	0.17	0.17	0.19	0.18
	OSCMFH	0.47	0.32	0.30	0.28	0.26	0.26	0.23	0.28
	ROHLSE	4.14	3.08	3.01	3.05	3.04	3.07	3.04	2.89
64	OLSH	4.36	3.29	3.15	3.61	4.46	4.86	5.25	5.75
	FOMH	0.89	0.94	0.90	0.93	0.92	0.92	0.89	0.91
	OASH	1.28	0.85	0.87	0.88	0.86	0.84	0.88	0.86
	DOCH	8.08	25.39	60.28	89.04	111.33	136.04	159.53	175.62
	LEMON	0.36	0.34	0.35	0.36	0.34	0.32	0.35	0.32
	OSCMFH	0.56	0.42	0.43	0.42	0.46	0.43	0.43	0.42
	ROHLSE	5.40	3.40	3.23	3.40	3.17	3.22	3.12	3.06
128	OLSH	4.61	3.67	3.79	3.82	4.59	5.41	5.92	6.31
	FOMH	1.11	1.55	1.58	1.48	1.56	1.54	1.58	1.48
	OASH	1.82	1.75	1.23	1.25	1.42	1.35	1.41	1.36
	DOCH	41.89	125.39	200.28	258.04	303.33	354.04	404.53	431.62
	LEMON	0.48	0.43	0.43	0.42	0.47	0.43	0.43	0.42
	OSCMFH	0.67	0.52	0.50	0.52	0.56	0.56	0.53	0.52
	ROHLSE	5.72	3.95	3.91	3.91	3.98	4.03	3.87	3.97

References

[1] Xiao Lin, Shuzhou Sun, Wei Huang, Bin Sheng, Ping Li, David Dagan Feng, EAPT: Efficient attention pyramid transformer for image processing, IEEE Trans. Multimed. 25 (2021) 3120873.

[2] Shuzhou Sun, Shuaifeng Zhi, Janne Heikkilä, Li Liu, Evidential uncertainty and diversity guided active learning for scene graph generation, in: The Eleventh International Conference on Learning Representations, 2023.

[3] Youxiang Duan, Ning Chen, Peiying Zhang, Neeraj Kumar, Lunjie Chang, Wu Wen, MS2GAH: Multi-label semantic supervised graph attention hashing for robust cross-modal retrieval, Pattern Recognit. 128 (2022) 108676.

[4] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, Samuel Albanie, Cross modal retrieval with querybank normalisation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5194–5205.

[5] Donglin Zhang, Xiao-Jun Wu, Scalable discrete matrix factorization and semantic autoencoder for cross-media retrieval, IEEE Trans. Cybern. (2020).

[6] Ruoqi Su, Di Wang, Zhen Huang, Yuan Liu, Yaqiang An, Online adaptive supervised hashing for large-scale cross-modal retrieval, IEEE Access 8 (2020) 206360–206370.

[7] Tao Yao, Gang Wang, Lianshan Yan, Xiangwei Kong, Qingtang Su, Caiming Zhang, Qi Tian, Online latent semantic hashing for cross-media retrieval, Pattern Recognit. 89 (2019) 1–11.

[8] Yu-Wei Zhan, Yongxin Wang, Yu Sun, Xiao-Ming Wu, Xin Luo, Xin-Shun Xu, Discrete online cross-modal hashing, Pattern Recognit. 122 (2022) 108262.

[9] Liang Xie, Jialie Shen, Lei Zhu, Online cross-modal hashing for web image retrieval, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30, 2016.

[10] Di Wang, Quan Wang, Yaqiang An, Xinbo Gao, Yumin Tian, Online collective matrix factorization hashing for large-scale cross-media retrieval, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1409–1418.

[11] Xiaofeng Zhu, Zi Huang, Heng Tao Shen, Xin Zhao, Linear cross-modal hashing for efficient multimedia search, in: Proceedings of the 21st ACM International Conference on Multimedia, 2013, pp. 143–152.

[12] Jile Zhou, Guiguang Ding, Yuchen Guo, Latent semantic sparse hashing for cross-modal similarity search, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, 2014, pp. 415–424.

[13] Guiguang Ding, Yuchen Guo, Jile Zhou, Collective matrix factorization hashing for multimodal data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2075–2082.

[14] Hui Cui, Lei Zhu, Chaoran Cui, Xiushan Nie, Huaxiang Zhang, Efficient weakly-supervised discrete hashing for large-scale social image retrieval, Pattern Recognit. Lett. 130 (2020) 174–181.

[15] Xuanwu Liu, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Guoqiang Xiao, Maozu Guo, Weakly supervised cross-modal hashing, IEEE Trans. Big Data 8 (2) (2019) 552–563.

[16] Chao Zhang, Huaxiong Li, Yang Gao, Chunlin Chen, Weakly-supervised enhanced semantic-aware hashing for cross-modal retrieval, IEEE Trans. Knowl. Data Eng. (2022).

[17] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, Dacheng Tao, Self-supervised adversarial hashing networks for cross-modal retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4242–4251.

[18] Yue Cao, Mingsheng Long, Jianmin Wang, Han Zhu, Correlation autoencoder hashing for supervised cross-modal search, in: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, 2016, pp. 197–204.

[19] Hua-Junjie Huang, Rui Yang, Chuan-Xiang Li, Yuliang Shi, Shanqing Guo, Xin-Shun Xu, Supervised cross-modal hashing without relaxation, in: 2017 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2017, pp. 1159–1164.

[20] Jun Tang, Ke Wang, Ling Shao, Supervised matrix factorization hashing for cross-modal retrieval, IEEE Trans. Image Process. 25 (7) (2016) 3157–3166.

[21] Zijia Lin, Guiguang Ding, Mingqing Hu, Jianmin Wang, Semantics-preserving hashing for cross-view retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3864–3872.

[22] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, Xuelong Li, Learning discriminative binary codes for large-scale cross-modal retrieval, IEEE Trans. Image Process. 26 (5) (2017) 2494–2507.

[23] Xiaozhao Fang, Zhihu Liu, Na Han, Lin Jiang, Shaohua Teng, Discrete matrix factorization hashing for cross-modal retrieval, Int. J. Mach. Learn. Cybern. 12 (10) (2021) 3023–3036.

[24] Fatih Cakir, Sarah Adel Bargal, Stan Sclaroff, Online supervised hashing, Comput. Vis. Image Underst. 156 (2017) 162–173.

[25] Long-Kai Huang, Qiang Yang, Wei-Shi Zheng, Online hashing, in: IJCAI, 2013, pp. 1422–1428.

[26] Mingbao Lin, Rongrong Ji, Hong Liu, Xiaoshuai Sun, Shen Chen, Qi Tian, Hadamard matrix guided online hashing, Int. J. Comput. Vis. 128 (8) (2020) 2279–2306.

[27] Chen-Lu Ding, Xin Luo, Xiao-Ming Wu, Yu-Wei Zhan, Rui Li, Hui Zhang, Xin-Shun Xu, Weakly-supervised online hashing with refined pseudo tags, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 375–385.

[28] Xu Lu, Lei Zhu, Zhiyong Cheng, Jingjing Li, Xiushan Nie, Huaxiang Zhang, Flexible online multi-modal hashing for large-scale multimedia retrieval, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1129–1137.

[29] Xin Liu, Jinhan Yi, Yiu-ming Cheung, Xing Xu, Zhen Cui, OMGH: Online manifold-guided hashing for flexible cross-modal retrieval, IEEE Trans. Multimed. (2022).

[30] Yongxin Wang, Xin Luo, Xin-Shun Xu, Label embedding online hashing for cross-modal retrieval, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 871–879.
[31] Zhenqiu Shu, Li Li, Jun Yu, Donglin Zhang, Zhengtao Yu, Xiao-Jun Wu, Online supervised collective matrix factorization hashing for cross-modal retrieval, Appl. Intell. (2022) 1–18.
[32] Emmanuel J. Candès, Xiaodong Li, Yi Ma, John Wright, Robust principal component analysis? J. ACM 58 (3) (2011) 1–37.
[33] Huaxiong Li, Chao Zhang, Xiuyi Jia, Yang Gao, Chunlin Chen, Adaptive label correlation based asymmetric discrete hashing for cross-modal retrieval, IEEE Trans. Knowl. Data Eng. (2021).
[34] Albert Gordo, Florent Perronnin, Yunchao Gong, Svetlana Lazebnik, Asymmetric distances for binary embeddings, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2013) 33–47.
[35] Junfeng Yang, Yin Zhang, Alternating direction algorithms for \ell_1-problems in compressive sensing, SIAM J. Sci. Comput. 33 (1) (2011) 250–278.
[36] Jian-Feng Cai, Emmanuel J. Candès, Zuowei Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. 20 (4) (2010) 1956–1982.
[37] Wei Liu, Cun Mu, Sanjiv Kumar, Shih-Fu Chang, Discrete graph hashing, Neural Inform. Process. Syst. 27 (2014).
[38] Mark J. Huiskes, Michael S. Lew, The mir flickr retrieval evaluation, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, 2008, pp. 39–43.
[39] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, Yantao Zheng, Nus-wide: A real-world web image database from national university of singapore, in: Proceedings of the ACM International Conference on Image and Video Retrieval, 2009, pp. 1–9.
[40] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villasenor, Michael Grubinger, The segmented and annotated IAPR TC-12 benchmark, Comput. Vis. Image Underst. 114 (4) (2010) 419–428.

Li Li is currently pursuing toward Master degree at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. Her current research interests include multimedia information retrieval and machine learning.

Zhenqiu Shu received the Ph.D. degree in computer applications at Nanjing University of Science and Technology. In February 2021, he joined the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, where he is currently an associate professor. Before joining in Kunming University of Science and Technology University, he had been a postdoctoral in Jiangnan University for four years. His research interests include image processing, computer vision and machine learning.

Zhengtao Yu received the Ph.D. degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2005. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. His current research interests include natural language processing, information retrieval, and machine learning.

Xiao-Jun Wu received the B.Sc. degree in mathematics from Nanjing Normal University, Nanjing, China, in 1991, and the M.S. and Ph.D. degrees in pattern recognition and intelligent systems from the Nanjing University of Science and Technology, Nanjing, in 1996 and 2002, respectively. From 1996 to 2006, he taught at the School of Electronics and Information, Jiangsu University of Science and Technology, where he was promoted to Professor. He has been with the School of AI & CS, Jiangnan University, since 2006, where he is a Professor of Computer Science and Technology. He was a Visiting Researcher with the CVSSP, University of Surrey, U.K., from 2003 to 2004. He has published over 300 research papers in refereed international journals and conferences. His current research interests include pattern recognition and computational intelligence. He was a Fellow of the International Institute for Software Technology, United Nations University, from 1999 to 2000. He was a recipient of the Most Outstanding Postgraduate Award from the Nanjing University of Science and Technology.