



When CLIP meets cross-modal hashing retrieval: A new strong baseline

Xinyu Xia^{a,b}, Guohua Dong^{a,*}, Fengling Li^c, Lei Zhu^b, Xiaomin Ying^{a,*}

^a Center for Computational Biology, Beijing Institute of Basic Medical Sciences, Beijing 100850, China

^b School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China

^c Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia

ARTICLE INFO

Keywords:

Cross-modal retrieval

Hashing

CLIP

Modality fusion

Contrastive learning

ABSTRACT

Recent days witness significant progress in various multi-modal tasks made by Contrastive Language-Image Pre-training (CLIP), a multi-modal large-scale model that learns visual representations from natural language supervision. However, the potential effects of CLIP on cross-modal hashing retrieval has not been investigated yet. In this paper, we for the first time explore the effects of CLIP on cross-modal hashing retrieval performance and propose a simple but strong baseline Unsupervised Contrastive Multi-modal Fusion Hashing network (UCMFH). We first extract the off-the-shelf visual and linguistic features from the CLIP model, as the input sources for cross-modal hashing functions. To further mitigate the semantic gap between the image and text features, we design an effective contrastive multi-modal learning module that leverages a multi-modal fusion transformer encoder supervising by a contrastive loss, to enhance modality interaction while improving the semantic representation of each modality. Furthermore, we design a contrastive hash learning module to produce high-quality modal-correlated hash codes. Experiments show that significant performance improvement can be made by our simple new unsupervised baseline UCMFH compared with state-of-the-art supervised and unsupervised cross-modal hashing methods. Also, our experiments demonstrate the remarkable performance of CLIP features on cross-modal hashing retrieval task compared to deep visual and linguistic features used in existing state-of-the-art methods. The source codes for our approach is publicly available at: <https://github.com/XinyuXia97/UCMFH>.

1. Introduction

Fast and accurate cross-modal retrieval system has become an urgent demand for searching the semantic relevant instances among heterogeneous modalities, e.g., taking an image as the query to retrieve the relevant text descriptions or expecting relevant images when giving a query text description [1–4]. With this trend, cross-modal hashing retrieval methods are gaining increasing importance due to their great gains of both computation and storage in massive multimedia data [5–9].

In cross-modal hashing retrieval, the main objective is to learn a common binary code representation for data samples from different modalities (such as images and texts). This task poses two important challenges: (1) How to effectively mitigate the semantic gap between the heterogeneous modalities [10,11] during the hash code learning. Different modalities may have different levels of abstraction and semantics. For instance, an image may contain high-level visual features, while a text description may contain more abstract concepts. The cross-modal hashing model needs to identify a space to bridge the

semantic gap and learn a common representation that captures the shared information between different modalities. (2) How to preserve the heterogeneous data correlation into binary codes. The cross-modal hashing model needs to learn a representation that preserves the data correlation of inter- and intra-modalities. For example, images with similar semantic labels and their corresponding text descriptions should have similar binary codes.

Various techniques [12–15] have been proposed to address the above challenges. According to whether semantic labels are used or not, those methods can be roughly divided into two categories: supervised and unsupervised. Supervised methods [13–17] leverage pre-annotated labels to preserve correlations of samples among different modalities, thus learning more discriminative representations and yielding more promising results. However, labeling large-scale datasets is a challenging and time-consuming task in real-world applications. In contrast, unsupervised approaches [12,18–21] could generate hash codes by mining the correlation structure from the data itself, without relying

* Corresponding authors.

E-mail addresses: xiaxyu97@gmail.com (X. Xia), dgh1991.learn@gmail.com (G. Dong), fenglingli2023@gmail.com (F. Li), leizhu0608@gmail.com (L. Zhu), yingxmbio@foxmail.com (X. Ying).

<https://doi.org/10.1016/j.inffus.2023.101968>

Received 11 April 2023; Received in revised form 29 June 2023; Accepted 2 August 2023

Available online 6 August 2023

1566-2535/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

on semantic tags. Such unsupervised methods are more competitive for large-scale deployment in real-world retrieval scenarios.

Over the past few years, numerous unsupervised hashing methods have been developed for efficient cross-modal retrieval. These methods are based on a range of learning strategies, including matrix factorization [18,22], graph regularization [23,24] and deep learning [19, 23–25], to preserve semantic correlations while projecting features from heterogeneous modalities into a shared Hamming space. Among those works, deep cross-modal hashing methods [12,19,22–25] have shown impressive performance due to the powerful feature representation abilities of deep neural networks. However, despite their success, there are still several weaknesses that limit their performance: (1) weak semantic interaction of different modalities, which may lead to incomplete or inaccurate cross-modal correlations. (2) these methods rely heavily on training data, which can restrict their generalization ability to new and unseen data.

Recently, Contrastive Language-Image Pre-training (CLIP) [26], a large-scale model that learns to associate images and text via pre-training on large-scale image-text pairs, has demonstrated impressive semantic comprehension abilities and remarkable zero-shot or few-shot learning capabilities. CLIP's success has significantly transformed the cross-modal field, and its superiority in multi-modal tasks is increasingly being recognized [27–29]. Despite the numerous successful trials of CLIP, a comprehensive evaluation of its impact and performance on cross-modal hashing retrieval has not yet been explored.

In this paper, we conduct extensive research to explore the potential of CLIP on cross-modal hashing retrieval. The main contributions of our method are summarized as follows:

- We propose to explore the effects of the pre-trained large-scale multimodal CLIP model on the cross-modal hashing retrieval task. To the best of our knowledge, there is still no similar work.
- Building upon the highly-performing embeddings captured by CLIP, we propose a simple yet effective unsupervised contrastive multi-modal fusion hashing network (UCMFH) for cross-modal retrieval. We propose a multi-modal fusion Transformer encoder to capture rich multi-modal semantics and enhance the interaction between different modalities. These semantics are then transferred to hash codes through a contrastive hash learning module, resulting in highly informative and compact representations suitable for cross-modal retrieval.
- Extensive experiments on several public benchmark datasets demonstrate the superior performance of the proposed method on cross-modal hashing retrieval compared with both unsupervised and supervised SOTA methods, even for retrieving unseen data. Moreover, we compare the CNN/BoW features used in existing cross-modal hashing methods with pre-trained CLIP embeddings by integrating several state-of-the-art hashing works on cross-modal retrieval. The experimental results demonstrate that significant improvement can indeed be achieved with pre-trained CLIP features.

The reminder of this paper is organized as follows. We will in turn introduce the related work (Section 2), the proposed new baseline method (Section 3) and the validating experiments (Section 4). Finally, the conclusion will be given in Section 5.

2. Related work

In this section, we briefly review the related cross-modal hashing works according to whether supervised information, such as semantic labels or matrices, is utilized. We divide the literature into supervised and unsupervised methods.

2.1. Supervised cross-modal hashing

In general, supervised cross-modal hashing methods rely on the use of semantic labels or similarity matrices to supervise the learning of intra- and inter-modality semantics. These methods can be further classified as either shallow or deep cross-modal hashing approaches, depending on the depth and complexity of the model structure used in the hashing process.

Shallow Cross-modal Hashing. At the early stage, several supervised cross-modal hashing methods are proposed to bridge the semantic gap in cross-modal retrieval with shallow learning models. Supervised Matrix Factorization Hashing (SMFH) [30] employs collective matrix factorization to generate unified hash codes that consider both label consistency across different modalities and local geometric consistency in each modality. Matrix Tri-Factorization Hashing (MTFH) [31] designs an efficient objective function to simultaneously learn the modality-specific hash codes with different length settings, as well as two semantic correlation matrices to semantically correlate the different hash representations for heterogeneous data. Scalable Asymmetric Discrete Cross-modal Hashing (BATCH) [32] fully exploits semantic labels through semantic latent space learning and distance-distance difference minimization. BATCH is one of the early works to use distance-distance difference minimization to embed semantic information into hash codes.

Deep Cross-modal Hashing. Motivated by the powerful deep neural networks (DNN) [33–35], many deep cross-modal hashing methods have been proposed and they achieve superior performance on cross-modal retrieval. Deep Cross-modal Hashing (DCMH) [36] designs an end-to-end deep learning framework which learns multi-modal features and generates hash codes simultaneously. Consistency-Preserving Adversarial Hashing (CPAH) [37] introduces a consistency refined module and a multi-task adversarial learning module to address the challenges of modality representation separation and preservation of multimodal semantic consistency. Graph Convolutional Hashing (GCH) [38] designs a GCN based hashing network that leverages a novel semantic encoder to maintain semantic consistency in the multi-modal feature encoding process. Deep Cross-Modal Hashing with Hashing Functions and Unified Hash Codes Jointly Learning (DCHUC) [39] proposes an iterative optimization algorithm to jointly learn hash codes and hash functions. Deep Adversarial Discrete Hashing (DADH) [40] employs adversarial training to learn features across modalities and ensure the distribution consistency of feature representations across modalities. Efficient Hierarchical Message Aggregation Hashing (HMAH) [41] proposes a correlation knowledge distillation strategy to transfer fine-grained multi-modal semantic correlations from a teacher module to lightweight student hashing modules.

2.2. Unsupervised cross-modal hashing

Unsupervised cross-modal hashing methods aim to learn hash codes and functions without the supervision of pre-annotated semantic labels or matrices. In the absence of such supervision, these methods capture the inherent inter-modal and intra-modal data correlations [42,43] to learn common representations. They can also be categorized into shallow and deep methods according to their basic learning frameworks.

Shallow Cross-modal Hashing. Early unsupervised cross-modal hashing methods develop several strategies to tackle the challenge of learning hash functions without supervision information. Cross-View Hashing (CVH) [7] formulates the problem of learning hash functions as an NP hard minimization problem and further transforms it into a tractable eigenvalue problem through a novel relaxation. Inter-Media Hashing (IMH) [44] discovers a common hamming space by exploiting inter-media and intra-media consistency to ensure consistent representation of heterogeneous multi-modal data. Latent Semantic Sparse Hashing (LSSH) [45] utilizes sparse coding to capture visual

salient structures and adopts matrix factorization to learn textual latent concepts. Collective Matrix Factorization Hashing (CMFH) [18] designs unified hash codes using collective matrix factorization to enable searching for different modalities. Semantic-Rebased Cross-modal Hashing (SRCH) [46] proposes sparse graph structures to exploit similarity information, which can effectively address the degradation problem, and leverages similarity-preserving and quantization strategies to generate hash codes.

Deep Cross-modal Hashing. Deep cross-modal hashing methods have shown remarkable performance in cross-modal retrieval tasks without any supervision information. Deep Binary Reconstruction (DBRC) [12] proposes a deep binary reconstruction network to achieve heterogeneous modal relation modeling and hash code learning. Unsupervised Deep Cross Modal Hashing (UDCMH) [22] solves the discrete constrained objective function by optimizing unified binary codes in an alternating manner, and assigning weights for different modalities dynamically during optimization. Deep Joint Semantic Reconstructing Hashing (DJSRH) [19] reconstructs joint-semantics structure to learn hash codes, which explicitly integrates the original neighborhood information from multi-modal data. Joint-modal Distribution-based Similarity Hashing (JDSH) [25] constructs a joint-modal similarity matrix and proposes a sampling and weighting scheme, which is able to generate more discriminative hash codes. Aggregation-based Graph Convolutional Hashing (AGCH) [23] aggregates structural information and applies various similarity measures to obtain a joint similarity matrix. Correlation-Identity Reconstruction Hashing (CIRH) [24] designs a multi-modal collaborated graph to construct heterogeneous multi-modal correlations and performs the semantic aggregation on graph networks to generate a multi-modal complementary representation.

3. The proposed method

Artificial intelligence has been revolutionized by the advent of large-scale models trained on large-scale data, and one of the most promising models is the Contrastive Language Image Pre-training (CLIP) model developed by OpenAI [26]. It is pre-trained on approximately 400 million image-text pairs from the internet and is capable of understanding and processing both text and images by maximizing their agreement through contrastive learning.

Specially, CLIP learns to represent both images and texts in a shared embedding space, where the similarity between a given image and text can be measured as the cosine similarity between their respective embeddings. One of the major advantages of CLIP is its ability to generalize effectively to a wide range of downstream tasks, without requiring fine-tuning on specific datasets. This is attributable to the fact that CLIP is trained on a diverse set of image and text pairs, encompassing a wide range of concepts and domains.

CLIP has demonstrated state-of-the-art performance on a variety of benchmark multi-modal tasks, including text-to-image synthesis [27], 3D scene comprehension [28], and scene text detection [29]. Drawing inspiration from these studies, we are motivated to explore the effects of CLIP's linguistic and visual features on cross-modal hashing retrieval.

3.1. Framework overview

Fig. 1 shows the basic learning framework of the proposed new baseline. Firstly, we extract features of both image and text from the pre-trained CLIP backbone. Then, fine-grained semantic interaction is performed through a multi-modal fusion transformer encoder. It simultaneously extracts multi-modal fusion semantics and constructs a common cross-modal subspace, while enhances the semantic interaction of heterogeneous modalities. We further present a contrastive hash learning module to generate semantic preserved hash codes. In the following parts, we will give a detailed introduction of the proposed method.

Table 1

Main notations in our method.

Notation	Description
$\mathcal{O} = \{(x_i, y_i) i \in [1, n]\}$	Training set
$x_i \in \mathbb{R}^{1 \times d_v}$	Visual data of the i th sample
$y_i \in \mathbb{R}^{1 \times d_t}$	Textual data of the i th sample
$b_i \in \{-1, 1\}^{1 \times k}$	Hash code of the i th sample
$\mathbf{B} = \{b_i i \in [1, n]\}$	Hash code matrix
n	Number of samples
k	Length of hash code
d_v/d_t	Dimension of visual feature/textual feature

3.2. Notation and problem definition

To facilitate our presentation, we first introduce the following notations used in this paper. For clarity, we use visual and textual modalities as an example to explain our method. In our paper, the boldface uppercase letters, e.g., \mathbf{A} , and the boldface lowercase letters, e.g., \mathbf{a} , represent matrices and vectors, respectively. Suppose we have a training set: $\mathcal{O} = \{(x_i, y_i) | i \in [1, n]\}$, where $x_i \in \mathbb{R}^{1 \times d_v}$ denotes the visual data of the i th sample, $y_i \in \mathbb{R}^{1 \times d_t}$ denotes the textual data of the i th sample, and n , d_v and d_t indicate the number of samples, the dimension of visual feature and the dimension of textual feature, respectively. Our final goal is to learn a compact binary representation $b_i \in \{-1, 1\}^{1 \times k}$ of each sample, where k is the length of hash code. So hash code matrix can be defined as $\mathbf{B} = \{b_i | i \in [1, n]\}$. The above notations are summarized in Table 1.

3.3. Contrastive multi-modal learning

We propose a contrastive multi-modal fusion learning module, which is based on the standard Transformer Encoder architecture, to model the correspondence between image-text pairs and enhance semantic-interaction between these two modalities. To achieve this, we first concatenate the output of the CLIP feature encoder, $f_v \in \mathbb{R}^{d_v}$ and $f_t \in \mathbb{R}^{d_t}$, where d_v and d_t denote the dimensions of the visual and textual features, respectively. This results in a multi-modal feature vector, $f \in \mathbb{R}^d$, where $d = d_v + d_t$. The multi-modal fusion Transformer encoder then takes the multi-modal feature vector f as input and leverages the self-attention mechanism to capture the intra-modal and inter-modal dependencies and semantic relevance among the features. Although the self-attention mechanism requires more computational resources when the input dimension d increases, utilizing the pre-trained CLIP to extract features make the computational complexity controllable in this paper.

In the self-attention mechanism, the multi-modal features f are utilized to construct queries, keys, and values. This can be mathematically formulated as follows:

$$\begin{aligned} f_v &= CLIP_{visual}(x_i), \\ f_t &= CLIP_{textual}(y_i), \\ f &= \text{concat}[f_v, f_t], \end{aligned} \quad (1)$$

$$q_i = f_i W^Q, k_j = f_j W^K, v_j = f_j W^V, \quad (2)$$

where $W^Q \in \mathbb{R}^{d \times d_k}$, $W^K \in \mathbb{R}^{d \times d_k}$ and $W^V \in \mathbb{R}^{d \times d_v}$ are trainable parameters, respectively. d_k , d_v are the dimension of keys and values, respectively. For any image-text pair, multi-modal Transformer Encoder can generate more representational multi-modal features f' :

$$f'_i = SAtt(q_i, K, V) = \sum_{j=1}^m softmax\left(\frac{q_i k_j^T}{\sqrt{d_k}}\right) v_j, \quad (3)$$

where m represents batch size, $SAtt$ represents self-attention mechanism. K and V are respective the keys and values matrices. Further, we use Feed-Forward Network (FFN) and Shortcut to generate the

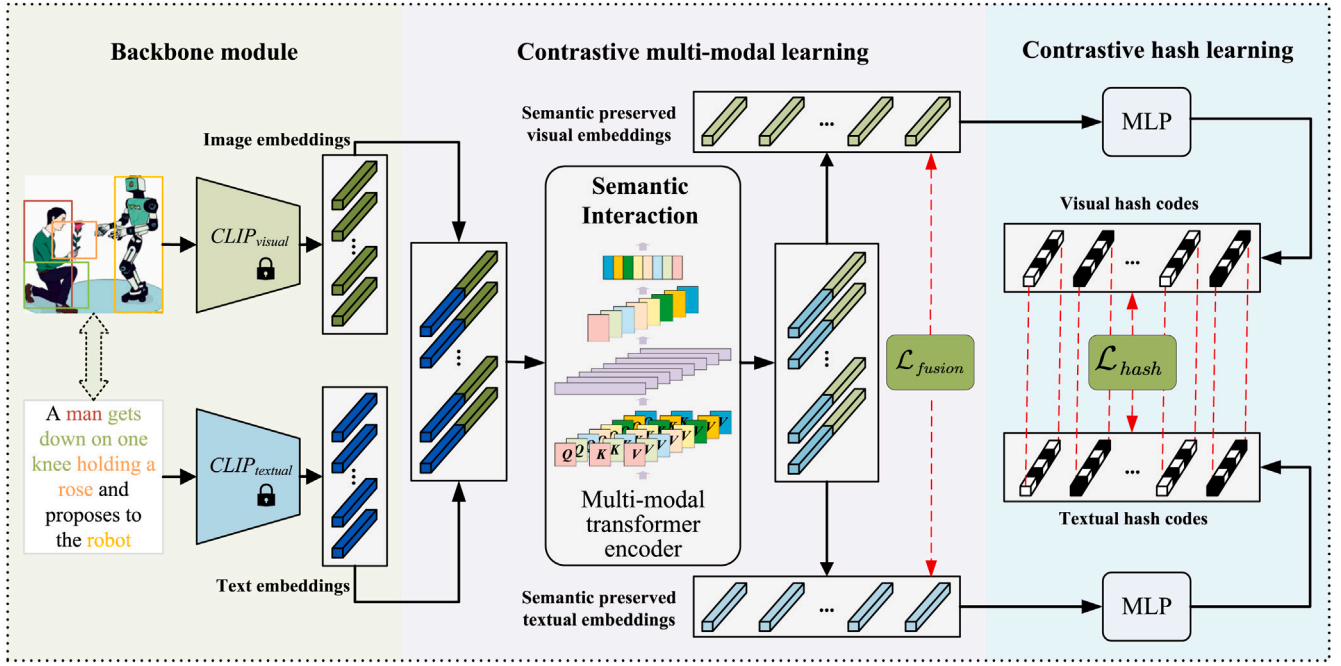


Fig. 1. The basic learning framework of the proposed unsupervised contrastive multi-modal fusion hashing network (UCMFH). UCMFH first extracts features of image and text from the pre-trained CLIP backbone. Then, it performs semantic interaction through a multi-modal transformer encoder. It further presents a contrastive hashing learning module to generate hash codes that preserve semantics. More details can be found in Section 3.

multi-modal representations. It can be formulated as

$$\begin{aligned} \mathbf{z}_i &= RS_2(RS_1(f'_i)), \\ RS_1(\cdot) &= Norm(Drop(\cdot) + f'_i), \\ RS_2(\cdot) &= Norm(Drop(FFN(\cdot)) + (\cdot)), \end{aligned} \quad (4)$$

where $Norm(\cdot)$ and $Drop(\cdot)$ represent Layer Normalization and Dropout function, respectively. $\mathbf{z}_i = \text{concat}[\mathbf{z}_i^v, \mathbf{z}_i^t]$ represents the output of multi-modal fusion Transformer encoder.

To ensure that the data representation of the same category within the same modality contains consistent category semantics, we define the modality contrastive loss as follows:

$$\begin{aligned} \mathcal{L}_{fusion}^v &= - \sum_{i=1}^m \log \frac{\exp(\langle \mathbf{z}_i^v, \mathbf{z}_i^{v+} \rangle / \tau)}{\sum_{j=1}^m \exp(\langle \mathbf{z}_i^v, \mathbf{z}_j^v \rangle / \tau)}, \\ \mathcal{L}_{fusion}^t &= - \sum_{i=1}^m \log \frac{\exp(\langle \mathbf{z}_i^t, \mathbf{z}_i^{t+} \rangle / \tau)}{\sum_{j=1}^m \exp(\langle \mathbf{z}_i^t, \mathbf{z}_j^t \rangle / \tau)}, \end{aligned} \quad (5)$$

where m denote the batch size, τ denote the temperature coefficient, and $\langle \cdot, \cdot \rangle$ denote the similarity function. It is important to note that both $\mathbf{z}_i^v, \mathbf{z}_i^{v+}$ and $\mathbf{z}_i^t, \mathbf{z}_i^{t+}$ represent the representations of the truly aligned image-text pairs. Therefore, the multi-modal contrastive loss can be defined as follows:

$$\mathcal{L}_{fusion} = \mathcal{L}_{fusion}^v + \mathcal{L}_{fusion}^t. \quad (6)$$

3.4. Contrastive hash learning

In this section, we introduce a private hash function \mathcal{H} implemented by MLP(Multilayer Perceptron) for each modality to learn a common subspace, which can be formulated as

$$\mathbf{h}_i^v = \mathcal{H}(\mathbf{z}_i^v; \theta_i), \mathbf{h}_i^t = \mathcal{H}(\mathbf{z}_i^t; \theta_t). \quad (7)$$

Then we introduce the contrastive loss and generate the final hash code as follows:

$$\begin{aligned} \mathcal{L}_{hash}^v &= - \sum_{i=1}^m \log \frac{\exp(\langle \mathbf{h}_i^v, \mathbf{h}_i^{v+} \rangle / \tau)}{\sum_{j=1}^m \exp(\langle \mathbf{h}_i^v, \mathbf{h}_j^v \rangle / \tau)}, \\ \mathcal{L}_{hash}^t &= - \sum_{i=1}^m \log \frac{\exp(\langle \mathbf{h}_i^t, \mathbf{h}_i^{t+} \rangle / \tau)}{\sum_{j=1}^m \exp(\langle \mathbf{h}_i^t, \mathbf{h}_j^t \rangle / \tau)}, \end{aligned} \quad (8)$$

$$\mathbf{b}_i^v = \text{sign}(\mathbf{h}_i^v), \mathbf{b}_i^t = \text{sign}(\mathbf{h}_i^t), \quad (9)$$

where $\text{sign}(\cdot)$ is a quantization operation. $\mathbf{h}_i^v, \mathbf{h}_i^{v+}$ and $\mathbf{h}_i^t, \mathbf{h}_i^{t+}$ are truly aligned image-text pairs, respectively. Thus, the total hashing loss is defined as

$$\mathcal{L}_{hash} = \mathcal{L}_{hash}^v + \mathcal{L}_{hash}^t. \quad (10)$$

3.5. Objective function

By integrating Eqs. (6) and (10) into a unified learning framework, we derive the overall objective function of the proposed method as

$$\min_{\theta_m, \theta_i, \theta_t} \mathcal{L} = \lambda_1 \mathcal{L}_{fusion} + \lambda_2 \mathcal{L}_{hash}, \quad (11)$$

where λ_1 and λ_2 are two trade-off hyper-parameters to adjust the importance of the first and the second regularization items. We summarize the basic learning process of the proposed method in Algorithm 1.

Algorithm 1 Optimization process for UCMFH

Input: n aligned image-text pairs as the training set; the batch size m ; the learning rate of multi-modal Transformer encoder l_{rm} and hash functions l_{ri}, l_{rt} ; the number of epochs t ; hyper-parameters λ_1 and λ_2 .

Output: Network parameters of the to-be-learned multi-modal fusion Transformer encoder θ_m and hash functions: θ_i and θ_t .

- 1: Initialize the parameters of multi-modal fusion and hash learning networks: θ_m, θ_i and θ_t .
- 2: **for** $i \in [1, \frac{n}{m}]$ **do**
- 3: Randomly select m training instance pairs;
- 4: Generate \mathbf{z}_i^v and \mathbf{z}_i^t by forward propagation;
- 5: Calculate loss with Eq.(6);
- 6: Generate \mathbf{h}_i^v and \mathbf{h}_i^t by forward propagation;
- 7: Calculate loss with Eq.(10);
- 8: Update the parameters θ_m, θ_i and θ_t by back propagation;

4. Experiments

In this section, we conduct various experiments to verify the effects of CLIP on cross-modal hashing retrieval task.

Table 2
Basic statistics of the experimental datasets.

Datasets	Categories	Training	Query	Retrieval
MIR Flickr	24	5,000	2,000	18,015
NUS-WIDE	10	5,000	2,000	184,577
MS COCO	80	5,000	2,000	121,287

Firstly, we conduct retrieval accuracy comparison from two aspects: (1) we conduct a comprehensive comparison of the proposed method with 15 state-of-the-art supervised and unsupervised cross-modal hashing methods. (2) we select six representative state-of-the-arts to further compare the performance of CLIP features with existing CNN/BoW features, which are the most common backbone that SOTA cross-modal hashing methods used.

We also design external experiments to further demonstrate the effectiveness of the proposed method, containing ablation experiment, parameter sensitivity analysis, convergence analysis and real cross-modal retrieval visualization.

4.1. Evaluation datasets

To ensure a fair comparison of our proposed method with the state-of-the-art cross-modal hashing methods, we employ identical experimental datasets and settings as those in state-of-the-art unsupervised cross-modal hashing retrieval method [24], and these datasets and settings are also the most commonly used ones in cross-modal retrieval. The specific statistics are presented in Table 2.

MIR Flickr[47] dataset comprises 25,000 pairs of images and tags, which correspond to 24 distinct concepts. We have excluded pairs with fewer than 20 tags, resulting in a final set of 20,015 image-tag pairs. In order to maintain consistency with the experimental protocols of other cross-modal hashing methods [19,22–25], we have randomly selected 2,000 pairs from this set as the query set, while the remaining 18,015 pairs form the retrieval set. We have also selected 5,000 pairs from the retrieval set as the training set.

NUS-WIDE[48] dataset contains 269,648 images and 81 concepts. Following the experimental setup of previous methods [19,22–25], we have chosen 186,577 image-text pairs corresponding to the top 10 most common concepts to create the experimental dataset. We have randomly selected 2,000 image-tag pairs as the query set, while the remaining 184,577 image-text pairs form the retrieval set. From the retrieval set, we have selected 5,000 image-text pairs to form the training set.

MS COCO[49] dataset comprises 123,287 image-text pairs categorized into 80 independent categories. The dataset partition settings are similar to those of MIR Flickr and NUS-WIDE, with the training set and query set consisting of 5,000 and 2,000 randomly selected instances, respectively. The retrieval set comprises a total of 121,287 instances.

4.2. Baselines and evaluation criterion

We compare the proposed method with 5 state-of-the-art supervised baselines and 10 state-of-the-art unsupervised baselines: (1) **Supervised baselines**. DLFH [50], LEMON [51], BATCH [32], DADH [40] and HMAH [41]. (2) **Unsupervised baselines**. CVH [7], IMH [44], LSSH [45], CMFH [18], DBRC [12], UDCMH [22], DJSRH [19], JDSH [25], AGCH [23] and CIRH [24].

Among those methods, CVH, IMH, LSSH, CMFH, DLFH, LEMON and BATCH are shallow methods, while DBRC, UDCMH, DJSRH, JDSH, AGCH, CIRH, DADH and HMAH are deep methods.

Similar to the previous works [19,22–25], we use the commonly adopted mean Average Precision (mAP) [52,53] to evaluate the retrieval performance of all methods. Specifically, we compute the mAP scores for two different cross-modal retrieval tasks: retrieving text samples using image queries and retrieving image samples using text

queries. Given a set of queries $Q = [q_1, q_2, \dots, q_{n_i}]$, we compute mAP as follows:

$$mAP = \frac{1}{n_i} \sum_{i=1}^{n_i} AP_i, \quad (12)$$

where n_i is the size of Q , and AP_i (Average Precision) can be calculated as

$$AP_i = \frac{1}{P_i} \sum_{k=1}^n \frac{P_{ik}}{k} \times \phi_{ik}, \quad (13)$$

where P_i is the number of similar samples of the query q_i in the database, n is the number of samples in the database, P_{ik} is the number of similar samples among the top k retrieved samples of the query q_i . ϕ_{ik} is a binary indicator function, where $\phi_{ik} = 1$ if the k th retrieved sample is similar to query q_i and $\phi_{ik} = 0$ otherwise. We regard two samples as similar if they share at least one consistent semantic label.

4.3. Implementation details

We first introduce the implementation details of the sub-modules in our method as follows:

- **Contrastive Multi-modal Learning.** In our design, we have incorporated two standard transformer encoder layers. The self-attention mechanism in our model has a latent dimension of 1,024.
- **Contrastive Hash Learning.** Our Contrastive Hash Learning module is implemented based on Fully-Connected (FC) layers. Specifically, these FC layers consist of linear projection layers, layer-normalization layers, and activation function $Tanh(\cdot)$.
- **Contrastive Hash Learning.** Our Contrastive Hash Learning module is implemented based on Fully-Connected (FC) layers. Specifically, the visual module is a fully connected layer ($512 \rightarrow 4096 \rightarrow ReLU \rightarrow Dropout \rightarrow k \rightarrow Tanh(\cdot)$), where $ReLU$ is an activation function, $Tanh(\cdot)$ is an activation function, and k is the code length. The textual module structure is consistent with the visual module structure.

The hyper-parameters used in our experiments are consistent across all datasets, including MIR Flickr, NUS-WIDE, and MS COCO, with values of $\lambda_1 = 1.0$ and $\lambda_2 = 0.1$. We adopt ADAM [54] optimizer and the learning rate is empirically set to 0.001. The batch size is set to 128. We conduct our experiments on a server with single NVIDIA RTX 2080Ti GPU and two 2.10 GHz Intel(R) Xeon(R) Silver 4110 CPUs.

4.4. Retrieval accuracy comparison

To comprehensively evaluate the efficacy of our proposed method, we perform a cross-modal retrieval comparison on three datasets with the hash code length of 16 bits, 32 bits, 64 bits and 128 bits in Table 3, respectively. Most experimental results are directly brought from original paper [24]. The other results DLFH [50], LEMON [51], BATCH [32], DADH [40], HMAH [41] are implemented carefully based on the released source codes.

As shown in Table 3, the proposed method achieves the best retrieval accuracy with hash code length varying from 16 bits to 128 bits. Specifically, on I2T and T2I retrieval tasks, our method outperforms the best deep baseline CIRH [24] by 2.2% ~ 4.0% and 5.9% ~ 6.3% on MIR Flickr, by 3.4% ~ 4.3% and 8.1% ~ 8.9% on NUS-WIDE, and by 8.9% ~ 10.8% and 4.7% ~ 7.6% on MS COCO, respectively.

We then select six best-performed cross-modal hashing methods containing JSRH [19], JDSH [25], CIRH [24], DADH [40], HMAH [41] and our proposed method, to further verify the superiority of CLIP features compared with traditional CNN/BoW features. Among the above six methods, the first three methods and our method are unsupervised, while DADH and HMAH are supervised. We use two different ways to extract features as backbone to evaluate the retrieval performance of all six methods:

Table 3

Comparison results of mAP@50 on I2T and T2I retrieval tasks on three datasets at different code lengths.

Task	Methods	Reference	MIR Flickr				NUS-WIDE				MS COCO				
			16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	
I2T	Supervised	DLFH [50]	TIP19	0.547	0.604	0.64	0.668	0.555	0.607	0.64	0.669	0.552	0.606	0.655	0.671
		LEMON [51]	MM20	0.735	0.767	0.795	0.809	0.724	0.767	0.776	0.812	0.67	0.762	0.786	0.808
		BATCH [32]	TKDE21	0.813	0.845	0.858	0.866	0.818	0.837	0.861	0.868	0.819	0.846	0.86	0.865
		DADH [40]	ICMR20	0.877	0.896	0.892	0.899	0.777	0.787	0.808	0.825	-	-	-	-
		HMAH [41]	TMM22	0.895	0.908	0.924	0.936	0.818	0.841	0.857	0.868	0.729	0.785	0.824	0.837
	Unsupervised	CVH [7]	IJCAI11	0.606	0.599	0.596	0.589	0.372	0.363	0.404	0.39	0.505	0.509	0.519	0.51
		IMH [44]	SIGMOD13	0.612	0.601	0.592	0.579	0.47	0.473	0.476	0.459	0.57	0.615	0.613	0.587
		LSSH [45]	SIGIR14	0.584	0.599	0.602	0.614	0.481	0.489	0.507	0.507	0.652	0.707	0.746	0.773
		CMFH [18]	TIP16	0.621	0.624	0.625	0.627	0.455	0.459	0.465	0.467	0.621	0.669	0.525	0.562
		DBRC [12]	TMM18	0.617	0.619	0.62	0.621	0.424	0.459	0.447	0.447	0.567	0.591	0.617	0.627
		UDCMH [22]	IJCAI18	0.689	0.698	0.714	0.717	0.511	0.519	0.524	0.558	-	-	-	-
		DJSRH [19]	ICCV19	0.81	0.843	0.862	0.876	0.724	0.773	0.798	0.817	0.678	0.724	0.743	0.768
		JDSH [25]	SIGIR20	0.832	0.853	0.882	0.892	0.736	0.793	0.832	0.835	0.694	0.738	0.769	0.788
		AGCH [23]	TMM21	0.865	0.887	0.892	0.912	0.809	0.83	0.831	0.852	0.741	0.772	0.789	0.806
		CIRH [24]	TKDE22	0.901	0.913	0.929	0.937	0.815	0.836	0.854	0.862	0.797	0.819	0.83	0.849
		UCMFH	Our Method	0.923	0.953	0.964	0.965	0.849	0.879	0.894	0.899	0.892	0.919	0.938	0.938
		T2I	Supervised	DLFH [50]	TIP19	0.614	0.699	0.72	0.74	0.636	0.693	0.719	0.746	0.615	0.68
LEMON [51]	MM20			0.732	0.753	0.767	0.781	0.732	0.762	0.769	0.772	0.724	0.743	0.764	0.781
BATCH [32]	TKDE21			0.752	0.772	0.779	0.789	0.746	0.763	0.784	0.79	0.745	0.766	0.785	0.788
DADH [40]	ICMR20			0.871	0.866	0.869	0.882	0.74	0.754	0.752	0.76	-	-	-	-
HMAH [41]	TMM22			0.819	0.837	0.846	0.859	0.722	0.757	0.763	0.775	0.787	0.862	0.911	0.928
Unsupervised	CVH [7]		IJCAI11	0.591	0.583	0.576	0.576	0.401	0.384	0.442	0.432	0.543	0.553	0.56	0.542
	IMH [44]		SIGMOD13	0.603	0.595	0.589	0.58	0.478	0.483	0.472	0.462	0.641	0.709	0.705	0.652
	LSSH [45]		SIGIR14	0.637	0.659	0.659	0.672	0.577	0.617	0.642	0.663	0.612	0.682	0.742	0.795
	CMFH [18]		TIP16	0.642	0.662	0.676	0.685	0.529	0.577	0.614	0.645	0.627	0.667	0.554	0.595
	DBRC [12]		TMM18	0.618	0.622	0.626	0.628	0.455	0.459	0.468	0.473	0.635	0.671	0.697	0.735
	UDCMH [22]		IJCAI18	0.692	0.704	0.718	0.733	0.637	0.653	0.695	0.716	-	-	-	-
	DJSRH [19]		ICCV19	0.786	0.822	0.835	0.847	0.712	0.744	0.771	0.789	0.65	0.753	0.805	0.823
	JDSH [25]		SIGIR20	0.825	0.864	0.878	0.88	0.721	0.785	0.794	0.804	0.703	0.759	0.793	0.825
	AGCH [23]		TMM21	0.829	0.849	0.852	0.88	0.769	0.78	0.798	0.802	0.746	0.774	0.797	0.817
	CIRH [24]		TKDE22	0.867	0.885	0.9	0.901	0.774	0.803	0.81	0.817	0.811	0.847	0.872	0.895
	UCMFH		Our Method	0.92	0.947	0.959	0.961	0.859	0.884	0.899	0.903	0.887	0.921	0.939	0.942

"–" denotes unavailable results, because the source paper did not report corresponding results, and the source code was not released.

- **CNN/BoW:** The extraction of image and text features can be accomplished through the utilization of a Convolutional Neural Network (CNN) [55] network and Bag-of-Words (BoW) [56] technique, respectively.
- **CLIP:** We utilize the Contrastive Language-Image Pre-training (CLIP) [26] visual encoder and textual encoder as feature extractors for images and text, respectively.

Detailed experimental results and corresponding analyses are presented in the following subsections. For a fair comparison with all baselines, we use the same backbone and the same data splitting. Our experimental results are meticulously implemented based on the released source codes.

4.4.1. Cross-modal retrieval on MIR Flickr

Table 4 reports the experimental results on MIR Flickr dataset over DJSRH [19], JDSH [25], CIRH [24], DADH [40], HMAH [41] and our proposed method. Based on CNN/BoW features, our proposed method respectively outperforms the best deep unsupervised baseline CIRH [24] by 0.4% ~ 2.2% and 4.2% ~ 6.3% on I2T and T2I retrieval tasks. Based on CLIP features, our proposed method outperforms the best deep baseline CIRH by 2.7% ~ 3.8% and 3.9% ~ 4.9% on I2T and T2I retrieval tasks. As shown in Table 4, CLIP features yield a great improvement compared with CNN/BoW features.

Furthermore, we observed that whether adopt CNN/BoW or CLIP as backbone, our proposed method achieves similar performance compared with previous unsupervised work DJSRH, JDSH and state-of-the-art method CIRH. Specifically, in the case of 16 bits, 32 bits, 64 bits and 128 bits retrieval tasks with CLIP features, the gap between I2T and T2I are only 0.3%, 0.5%, 0.5% and 0.4%, while the gap are 1.5%, 1.3%, 2.4% and 2.3% when using CIRH method. Based on CNN/BoW features, in the case of 16 bits, 32 bits, 64 bits and 128 bits retrieval tasks, the gap are only 0.0%, 0.3%, 0.3% and 0.2%. The experimental results shows that our proposed method also achieves the same performance steadily based on CNN/BoW features.

In order to observe the performance of retrieving topk samples with CLIP features in unsupervised cross-modal hashing, we also compare the mAP-topk results on MIR Flickr with CLIP features at 128 bits, and the results in Figs. 2 (a) and (d) confirm the superiority of the proposed method again.

Table 4

Comparison results of mAP@50 with different features on I2T and T2I retrieval tasks on MIR Flickr at different code lengths.

Task	Features	Methods	Unsupervised	MIR Flickr			
				16 bits	32 bits	64 bits	128 bits
I2T	CNN/BoW	DJSRH	✓	0.81	0.843	0.862	0.872
		JDSH	✓	0.832	0.853	0.882	0.892
		CIRH	✓	0.91	0.913	0.929	0.937
		DADH	✓	0.877	0.896	0.892	0.899
		HMAH	✓	0.895	0.908	0.924	0.936
		UCMFH	✓	0.93	0.935	0.94	0.941
	CLIP	DJSRH	✓	0.832	0.879	0.89	0.901
		JDSH	✓	0.898	0.92	0.933	0.941
		CIRH	✓	0.896	0.915	0.931	0.935
		DADH	✓	0.922	0.926	0.936	0.939
T2I	CNN/BoW	HMAH	✓	0.94	0.964	0.975	0.982
		UCMFH	✓	0.923	0.953	0.964	0.965
	CLIP	DJSRH	✓	0.817	0.839	0.849	0.862
		JDSH	✓	0.872	0.895	0.902	0.906
		CIRH	✓	0.881	0.902	0.907	0.912
		DADH	✓	0.909	0.917	0.922	0.919
	CLIP	HMAH	✓	0.853	0.884	0.897	0.912
		UCMFH	✓	0.92	0.947	0.959	0.961

To further evaluate the performance of our proposed method, we compare it with supervised SOAT methods DADH [40] and HMAH [41]. We observed that our method outperforms the supervised methods with CNN/BoW features. When we adopt the CLIP features, our method also demonstrates satisfactory performance. Specially, in the case of 16 bits, 32 bits, 64 bits and 128 bits T2I tasks, our method outperforms the best result of supervised method by 1.1%, 3.0%, 3.7% and 4.2%. These results suggest that our approach is an effective and viable alternative for cross-modal retrieval tasks, particularly in scenarios where labeled data may be limited or unavailable.

4.4.2. Cross-modal retrieval on NUS-WIDE

Based on CNN/BoW features, our proposed method outperforms the best deep unsupervised baseline CIRH [24] by 0.9% ~ 2.6% and 6.1% ~ 7.0% on I2T and T2I retrieval tasks, respectively. As for CLIP features, our proposed method respectively outperforms the best deep unsupervised baseline CIRH by 4.3% ~ 6.2% and 7.0% ~ 8.4% on I2T and

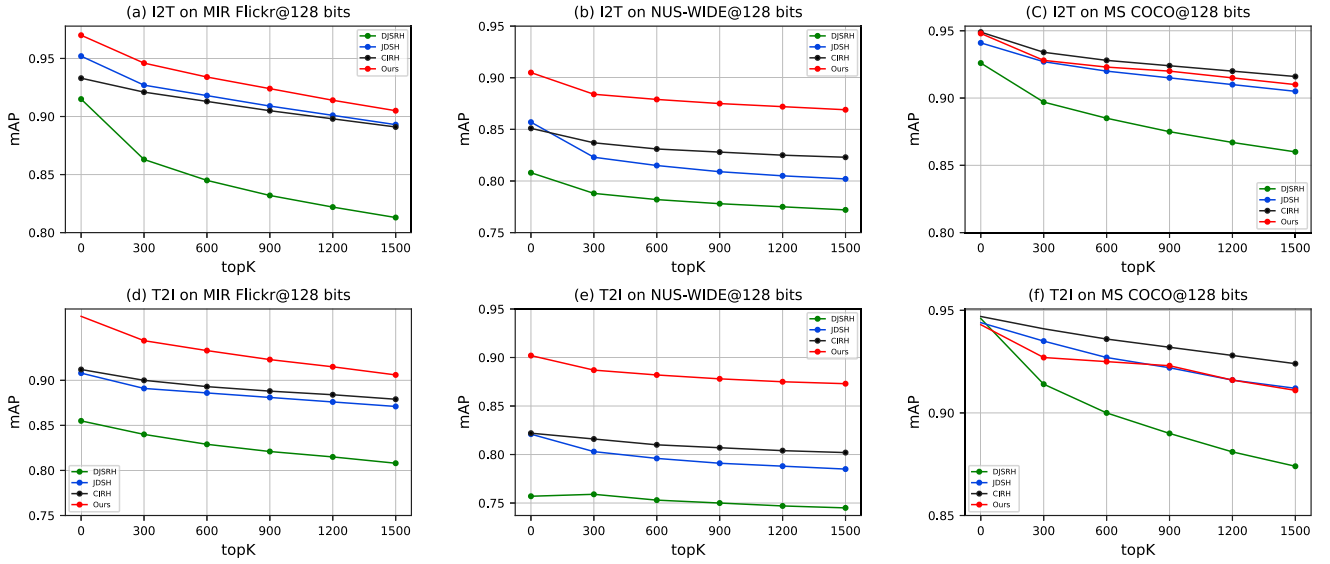


Fig. 2. The mAP-topK curves of all compared approaches on MIR Flickr, NUS-WIDE and MS COCO at code length of 128 bits.

Table 5

Comparison results of mAP@50 with different features on I2T and T2I retrieval tasks on NUS-WIDE at different code lengths.

Task	Features	Methods	Unsupervised	NUS-WIDE			
				16 bits	32 bits	64 bits	128 bits
I2T	CNN/BoW	DJSRH	✓	0.724	0.773	0.798	0.817
		JDSH	✓	0.736	0.793	0.832	0.835
		CIRH	✓	0.815	0.836	0.854	0.862
		DADH	✓	0.777	0.787	0.808	0.825
		HMAH	✓	0.818	0.857	0.868	0.868
		UCMFH	✓	0.841	0.859	0.874	0.871
	CLIP	DJSRH	✓	0.73	0.776	0.801	0.805
		JDSH	✓	0.751	0.786	0.838	0.851
		CIRH	✓	0.795	0.82	0.838	0.854
		DADH	✓	0.812	0.84	0.854	0.863
T2I	CNN/BoW	HMAH	✓	0.867	0.893	0.901	0.913
		UCMFH	✓	0.849	0.882	0.894	0.897
	CLIP	DJSRH	✓	0.712	0.744	0.771	0.789
		JDSH	✓	0.721	0.785	0.794	0.804
		CIRH	✓	0.774	0.803	0.81	0.817
		DADH	✓	0.74	0.754	0.752	0.760
	CLIP	HMAH	✓	0.722	0.757	0.763	0.775
		Ours	✓	0.844	0.866	0.877	0.878
	CLIP	DJSRH	✓	0.728	0.744	0.757	0.772
		JDSH	✓	0.73	0.773	0.812	0.829
		CIRH	✓	0.786	0.806	0.817	0.83
		DADH	✓	0.821	0.821	0.834	0.835
		HMAH	✓	0.786	0.81	0.815	0.827
		UCMFH	✓	0.859	0.886	0.901	0.9

Table 6

Comparison results of mAP@50 with different features on I2T and T2I retrieval tasks on MS COCO at different code lengths.

Task	Features	Methods	Unsupervised	MSCOCO			
				16 bits	32 bits	64 bits	128 bits
I2T	CNN/BoW	DJSRH	✓	0.678	0.724	0.743	0.768
		JDSH	✓	0.694	0.738	0.769	0.788
		CIRH	✓	0.797	0.819	0.83	0.849
		DADH	✓	–	–	–	–
		HMAH	✓	0.729	0.785	0.824	0.837
		UCMFH	✓	0.822	0.843	0.858	0.858
	CLIP	DJSRH	✓	0.835	0.87	0.909	0.924
		JDSH	✓	0.86	0.923	0.935	0.94
		CIRH	✓	0.88	0.916	0.941	0.942
		DADH	✓	0.859	0.887	0.899	0.895
T2I	CNN/BoW	HMAH	✓	0.818	0.903	0.928	0.946
		UCMFH	✓	0.892	0.919	0.938	0.938
	CLIP	DJSRH	✓	0.65	0.753	0.805	0.823
		JDSH	✓	0.703	0.759	0.793	0.825
		CIRH	✓	0.811	0.847	0.872	0.895
		DADH	✓	–	–	–	–
	CLIP	HMAH	✓	0.787	0.862	0.911	0.928
		UCMFH	✓	0.826	0.852	0.864	0.874
	CLIP	DJSRH	✓	0.841	0.863	0.914	0.931
		JDSH	✓	0.86	0.928	0.941	0.948
		CIRH	✓	0.885	0.925	0.944	0.952
		DADH	✓	0.879	0.902	0.905	0.905
		HMAH	✓	0.835	0.927	0.907	0.963
		UCMFH	✓	0.887	0.921	0.939	0.942

T2I retrieval tasks. Based on the comprehensive evaluation results, it is evident that CLIP features significantly enhance cross-modal retrieval performance compared to traditional CNN/BoW features (see Table 5).

Under CLIP features, the mAP-topK results in Figs. 2 (d) and (e) also verify the superiority of the proposed method compared with other unsupervised cross-modal hashing baselines. We guess that with CLIP features on NUS-WIDE dataset, our method could learn more consistent features of the heterogeneous modalities, thus making a great promotion of cross-modal retrieval than other unsupervised methods.

Compared to supervised cross-modal methods utilizing CNN/BoW features, our method yields superior performance across all tested methods. Additionally, when employing CLIP features, our method outperforms the best result of the supervised method by 3.8%, 6.5%, 6.7% and 6.5% for T2I tasks using 16 bits, 32 bits, 64 bits, and 128 bits representations, respectively. The results obtained in our experiments provide compelling evidence of the effectiveness and potential of our unsupervised approach in cross-modal retrieval tasks. Remarkably, our method even obtaining competitive results with the supervised methods, demonstrating its superiority in capturing the complex correlations between heterogeneous modalities.

4.4.3. Cross-modal retrieval on MS COCO

On the MS COCO dataset, the performance of the proposed UCMFH method is not as good as that on the MIR Flickr and NUS-WIDE datasets. Nevertheless, our results show that the UCMFH method still achieves competitive performance compared to other unsupervised or supervised methods. This could potentially be attributed to the dataset's complexity in terms of semantic correlations. Specifically, MS-COCO contains a larger number of diverse semantic concepts, with each image or text labeled with more than one tag from a set of 80 categories. This greater complexity in the dataset could have made it harder for the model to always learn meaningful and consistent representations across modalities, and could explain the performance of all methods that not always at their best.

As shown in Table 6, We can observe that when adopting CNN/BoW as backbone, our method outperforms the best baseline CIRH by 2.5%, 2.4%, 2.8% and 0.9% on I2T retrieval task, respectively. On T2I task, our UCMFH method obtains top-2 performance upon 16 bits, 32 bits and 64 bits, and also has competitive performance comparing with other methods on 128 bits.

According to the results presented in Table 6, it is evident that the CLIP features exhibit superior performance compared to the traditional

Table 7

The mAP performance of ablation experiments on MIR Flickr and NUS-WIDE.

Task	Variant	MIR Flickr				NUS-WIDE			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
I2T	w/o TEs	0.896	0.913	0.913	0.921	0.803	0.821	0.829	0.829
	w/o Dis	0.730	0.770	0.783	0.793	0.723	0.764	0.696	0.679
	w/o Hash	0.644	0.560	0.628	0.652	0.513	0.364	0.492	0.495
	UCMFH	0.923	0.953	0.964	0.965	0.849	0.882	0.894	0.897
T2I	w/o TEs	0.876	0.887	0.893	0.894	0.800	0.801	0.812	0.820
	w/o Dis	0.82	0.792	0.785	0.809	0.753	0.753	0.703	0.678
	w/o Hash	0.531	0.643	0.664	0.623	0.490	0.517	0.370	0.436
	UCMFH	0.920	0.947	0.959	0.961	0.859	0.886	0.901	0.900

CNN/Bow features. Our approach demonstrates that the average retrieval accuracy of CLIP features is 7.24% higher than that of CNN/BoW features. Furthermore, in the context of cross-modal retrieval, the state-of-the-art method CIRH reveals that the average retrieval performance of CLIP features is 8.31% higher than that of CNN/BoW features.

We plot the mAP-topk curves for the I2T and T2I retrieval tasks on the MS COCO dataset in Figs. 2 (d) and (f), respectively. As shown in the figures, as the number of query instances increases, the retrieval mAP criterion gradually decreases.

4.4.4. Conclusion of the retrieval accuracy comparison

The experimental results reported on three datasets demonstrate the superiority of CLIP features over traditional CNN/BoW features in cross-modal retrieval tasks. The proposed UCMFH method shows great potential in unsupervised cross-modal retrieval, outperforming several baseline methods, including supervised ones, on various evaluation metrics. These results suggest that the combination of unsupervised learning and CLIP features can be a promising approach for cross-modal retrieval tasks.

4.5. Ablation experiments

In order to evaluate the performance of key components in our model, we design three variants for comparison purposes. These variants are as follows: (1) **w/o TEs**: In this setting, we remove the multi-modal fusion Transformer encoder and instead directly adopt the original features to guide the learning of the deep hash function. (2) **w/o Dis**: In this variant, we remove the multi-modal contrastive loss. (3) **w/o Hash**: In this variant, we directly remove the hashing contrastive loss.

The experimental results for these three variants are reported in Table 7. Based on the results presented in this table, we have arrived at the following analyses:

- The removal of the multi-modal fusion Transformer encoder leads to a decline in performance. When compared to using the original features as supervision, the multi-modal fusion Transformer provides superior guidance for deep hash function learning. Furthermore, it is obvious that in the absence of the Transformer encoders, our method exhibits a greater gap between the I2T and T2I retrieval tasks. This serves as evidence of the effectiveness of enhancing multi-modal interaction.
- The second variant also experiences a substantial reduction in performance. It is worth noting that the decline in performance is more pronounced, indicating that multi-modal fusion semantic information is beneficial for the entire model.
- The experimental results of the third setting show that the contrastive loss can assist in generating higher quality hash code.
- In comparison to the first and second variants, the third variant exhibits a more significant decrease in retrieval accuracy. This implies that the elimination of the hash function loss has a more detrimental effect on the retrieval results.

Table 8

Comparison results of mAP@50 on I2T and T2I real-value retrieval task on three datasets.

	MIR Flickr		NUS-WIDE		MSCOCO	
	I2T	T2I	I2T	T2I	I2T	T2I
CLIP	0.88	0.841	0.805	0.783	0.917	0.931
UCMFH	0.971	0.971	0.916	0.923	0.973	0.971

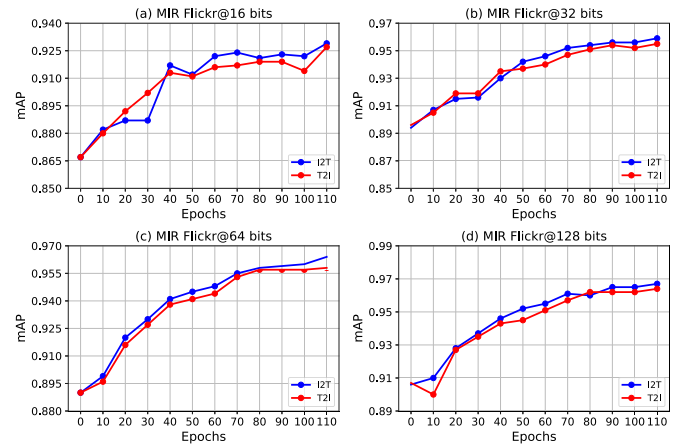


Fig. 3. The mAP variations with the number of epochs on MIR Flickr at various code lengths.

In order to evaluate the efficiency of the multi-modal fusion Transformer encoder, we exclusively employed it to generate retrieval features and compared its performance with that of CLIP features on MIR Flickr, NUS-WIDE and MS COCO datasets. The results, presented in Table 8, demonstrate a statistically significant improvement of over 10% in retrieval performance achieved by our method compared to CLIP's features on MIR Flickr and MS COCO.

4.6. Convergence analysis

Fig. 3 depicts the convergence curve versus different epochs on MIR Flickr. It can be observed that as the number of epochs increases, the mAP value shows an initial rapid increase, followed by a gradual smoothing before reaching a stable state. These results demonstrate the swift and effective training of our model.

4.7. Parameter sensitivity analysis

We conduct sensitivity experiments to explore the variations of mAP when hyper-parameter values changing. As shown in Eq. (11), there are two hyper-parameters: λ_1 and λ_2 . Specifically, the λ_1 is used to balance the performance of semantic expression, while λ_2 assigns the weight

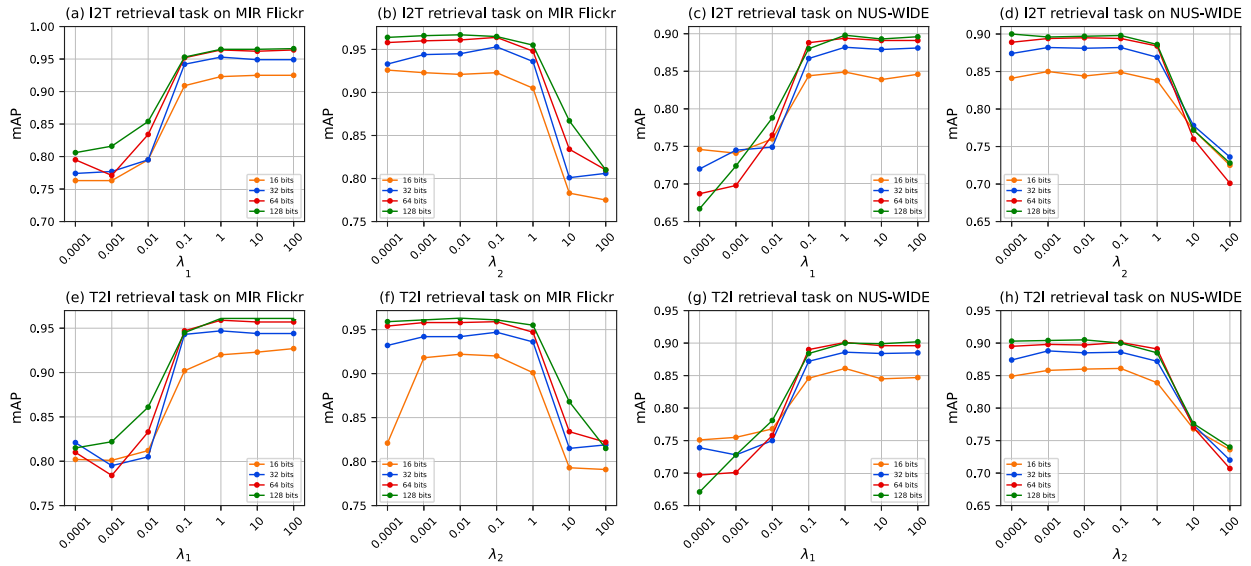


Fig. 4. Parameter sensitivity curves.




Query	Retrieved samples
<p>I2T</p> 	<p>Retrieved Candidates for our model:</p> <ul style="list-style-type: none"> A batter getting ready to bunt while the catcher waits with his glove in the air for the pitch.(19.97%) A baseball player getting ready to bunt the ball.(19.97%) A baseball batter and catcher anticipating a baseball pitch.(19.97%) A baseball batter holds his bat and waits for the pitch.(19.97%) A couple of young people are playing a baseball game.(19.97%) <p>Retrieved Candidates for CIRH model:</p> <ul style="list-style-type: none"> A big fighter jet sits on the te(100.0%) A man begins a snowboard trip while two people watch him from behind.(0.0%) A group of people riding skate boards down a hill.(0.0%) A person attempting to pitch a baseball from the mound.(0.0%) A batter getting ready to bunt while the catcher waits with his glove in the air for the pitch.(0.0%)
<p>T2I</p> <p>Query sentence: A group of people sitting around a table with a pizza on it.</p>	<p>Retrieved Candidates for our model:</p>  <p>Retrieved Candidates for CIRH model:</p> 

Fig. 5. The visualization shows the top-5 retrieval results for image-to-text (I2T) and text-to-image (T2I) tasks. We use cosine distance as a similarity metric to compare instances and apply the softmax function to predict the probability of each candidate outcome.

of learning hash codes. The possible values of the hyper-parameters belong to $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$.

We display the performance of all hyper-parameters in Fig. 4. The four colors represent four different hash code lengths. According to these experimental results, we can find that, (1) with λ_1 goes up, the performances show quickly increase trends, then flatten out. (2) When λ_2 is less than 1.0, the performances keep stable.

In fact, the mAP performance on different datasets shows the same trend, showing the generalization ability of our method. And as we all know, this ability is a crucial factor in achieving robust cross-modal retrieval performance, as it enables the model to effectively learn and represent the underlying relationships between heterogeneous modalities across different datasets.

4.8. Visualization

To evaluate the effectiveness of the proposed method in real-world cross-modal retrieval, we compare it with the state-of-the-art unsupervised cross-modal hashing retrieval method CIRH on both the image-to-text (I2T) and text-to-image (T2I) tasks.

Our model is trained on the MIR Flickr dataset, and we randomly select 100 images and 500 sentences from the MS COCO dataset (each image corresponds to 1–5 sentences) as the test set. For the I2T task, we randomly select an image from the test set and present the top-5 textual results based on their similarity to the image. For the T2I task, we randomly select a sentence (“a group of people sitting around a table with a pizza on it”), and the top-5 retrieval results are retrieved. We display the results in Fig. 5.

Obviously, our method retrieves all ground-truth textual descriptions, while CIRH only retrieve two correct descriptions. For the T2I task, we observe that the corresponding image appears in the first position of the candidate images in our method, while it appears in the fourth position when using CIRH. These results indicate that our method can enhance the semantic interaction of different modalities better than the compared methods.

Furthermore, our method demonstrates good retrieval performance on the MS COCO dataset, even though it is trained on MIR Flickr. This confirms the generalization ability of our method to perform well on unseen data.

5. Conclusion

In this paper, we explore the potential of large-scale models for cross-modal hashing retrieval tasks, focusing on the multi-modal pre-trained model CLIP. The experimental results demonstrate that using CLIP features in cross-modal hashing retrieval outperforms existing features in state-of-the-art methods. Moreover, we propose an unsupervised contrastive multi-modal fusion hashing network to provide a new strong baseline for the research field of cross-modal hashing retrieval. The extensive experimental results further support the effectiveness of our proposed model. The main challenge of our method and other large-scale models is the long encoding time when retrieving. However, there exist many optimizing strategies to overcome this, such as parallelizing computation and utilizing high computing cards. These strategies could significantly reduce the encoding time and improve the retrieval efficiency of our method. Nevertheless, our study highlights the emerging trend of using large-scale models for multimodal learning, as demonstrated by the success of CLIP and its ability to process multi-modal data. This trend has opened up new avenues for developing more efficient and effective models in various tasks such as cross-modal retrieval, image captioning, and visual question answering. Overall, our findings suggest that large-scale models like CLIP hold great potential for advancing the field of multimodal learning and its applications.

CRediT authorship contribution statement

Xinyu Xia: Conceptualization, Methodology, Software, Writing – original draft. **Guohua Dong:** Visualization, Investigation, Writing – original draft, Writing – review & editing, Supervision. **Fengling Li:** Visualization, Validation. **Lei Zhu:** Conceptualization, Writing – original draft. **Xiaomin Ying:** Investigation, Supervision.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

Funding: This work was supported by the National Key Research and Development Program of China under Grant No. 2022YFF1202400.

References

- [1] Shiyuan He, Weiyang Wang, Zheng Wang, Xing Xu, Yang Yang, Xiaoming Wang, Heng Tao Shen, Category alignment adversarial learning for cross-modal retrieval, *IEEE Trans. Knowl. data Eng.* (2022).
- [2] Yong Tian, Lian Zhou, Yuejie Zhang, Tao Zhang, Weiguo Fan, Deep cross-modal face naming for people news retrieval, *IEEE Trans. Knowl. data Eng.* 33 (5) (2019) 1891–1905.
- [3] Fangcen Liu, Chenqiang Gao, Yongqing Sun, Yue Zhao, Feng Yang, Anyong Qin, Deyu Meng, Infrared and visible cross-modal image retrieval through shared features, *IEEE Trans. Circuits Syst. Video Technol.* 31 (11) (2021) 4485–4496.
- [4] Liang Xie, Jialie Shen, Lei Zhu, Online cross-modal hashing for web image retrieval, in: *AAAI Conference on Artificial Intelligence*, Vol. 30, no. 1, 2016.
- [5] 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).
- [6] Zheng Zhang, Haoyang Luo, Lei Zhu, Guangming Lu, Heng Tao Shen, Modality-invariant asymmetric networks for cross-modal hashing, *IEEE Trans. Knowl. data Eng.* (2022).
- [7] Shaishav Kumar, Raghavendra Udapa, Learning hash functions for cross-view similarity search, in: *International Joint Conference on Artificial Intelligence*, 2011.
- [8] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, Dacheng Tao, Triplet-based deep hashing network for cross-modal retrieval, *IEEE Trans. image Process.* 27 (8) (2018) 3893–3903.
- [9] Yudong Chen, Sen Wang, Jianglin Lu, Zhi Chen, Zheng Zhang, Zi Huang, Local graph convolutional networks for cross-modal hashing, in: *ACM International Conference on Multimedia*, 2021, pp. 1921–1928.
- [10] Chang Tang, Xinwang Liu, Xiao Zheng, Wanqing Li, Jian Xiong, Lizhe Wang, Albert Y Zomaya, Antonella Longo, DeFusionNET: Defocus blur detection via recurrently fusing and refining discriminative multi-scale deep features, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (2) (2020) 955–968.
- [11] Chang Tang, Xinwang Liu, Xinzhou Zhu, En Zhu, Zhigang Luo, Lizhe Wang, Wen Gao, CGD: Multi-view clustering via cross-view graph diffusion, in: *AAAI Conference on Artificial Intelligence*, 2020, pp. 5924–5931.
- [12] Xuelong Li, Di Hu, Feiping Nie, Deep binary reconstruction for cross-modal hashing, in: *ACM International Conference on Multimedia*, 2017, pp. 1398–1406.
- [13] Lin Wu, Yang Wang, Ling Shao, Cycle-consistent deep generative hashing for cross-modal retrieval, *IEEE Trans. image Process.* 28 (4) (2018) 1602–1612.
- [14] Xuanwu Liu, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Yazhou Ren, Maozu Guo, Ranking-based deep cross-modal hashing, in: *AAAI Conference on Artificial Intelligence*, Vol. 33, no. 01, 2019, pp. 4400–4407.
- [15] Heng Tao Shen, Luchen Liu, Yang Yang, Xing Xu, Zi Huang, Fumin Shen, Richang Hong, Exploiting subspace relation in semantic labels for cross-modal hashing, *IEEE Trans. Knowl. data Eng.* 33 (10) (2020) 3351–3365.
- [16] Feng Zheng, Yi Tang, Ling Shao, Hetero-manifold regularisation for cross-modal hashing, *IEEE Trans. Pattern Anal. Mach. Intell.* 40 (5) (2016) 1059–1071.
- [17] Kai Li, Guo-Jun Qi, Jun Ye, Kien A. Hua, Linear subspace ranking hashing for cross-modal retrieval, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (9) (2016) 1825–1838.
- [18] Guiguang Ding, Yuchen Guo, Jile Zhou, Collective matrix factorization hashing for multimodal data, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2075–2082.

- [19] Shupeng Su, Zhisheng Zhong, Chao Zhang, Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval, in: IEEE/CVF International Conference on Computer Vision, 2019, pp. 3027–3035.
- [20] Hengtong Hu, Lingxi Xie, Richang Hong, Qi Tian, Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3123–3132.
- [21] Di Wang, Quan Wang, Xinbo Gao, Robust and flexible discrete hashing for cross-modal similarity search, IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2017) 2703–2715.
- [22] Gengshen Wu, Zijia Lin, Jungong Han, Li Liu, Guiguang Ding, Baochang Zhang, Jialie Shen, Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval, in: International Joint Conference on Artificial Intelligence, Vol. 5, 2018.
- [23] Peng-Fei Zhang, Yang Li, Zi Huang, Xin-Shun Xu, Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval, IEEE Trans. Multimedia 24 (2021) 466–479.
- [24] Lei Zhu, Xize Wu, Jingjing Li, Zheng Zhang, Weili Guan, Heng Tao Shen, Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval, IEEE Trans. Knowl. data Eng. (2022).
- [25] Song Liu, Shengsheng Qian, Yang Guan, Jiawei Zhan, Long Ying, Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval, in: International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1379–1388.
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
- [27] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, Timo Aila, Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis, 2023, arXiv preprint [arXiv:2301.09515](https://arxiv.org/abs/2301.09515).
- [28] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, Wenping Wang, CLIP2scene: Towards label-efficient 3D scene understanding by CLIP, 2023, arXiv preprint [arXiv:2301.04926](https://arxiv.org/abs/2301.04926).
- [29] Wenwen Yu, Yuliang Liu, Wei Hua, Deqiang Jiang, Bo Ren, Xiang Bai, Turning a CLIP model into a scene text detector, 2023, arXiv preprint [arXiv:2302.14338](https://arxiv.org/abs/2302.14338).
- [30] Jun Tang, Ke Wang, Ling Shao, Supervised matrix factorization hashing for cross-modal retrieval, IEEE Trans. image Process. 25 (7) (2016) 3157–3166.
- [31] Xin Liu, Zhikai Hu, Haibin Ling, Yiu-ming Cheung, MTFH: A matrix tri-factorization hashing framework for efficient cross-modal retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 43 (3) (2019) 964–981.
- [32] Yongxin Wang, Xin Luo, Liqiang Nie, Jingkuan Song, Wei Zhang, Xin-Shun Xu, BATC: A scalable asymmetric discrete cross-modal hashing, IEEE Trans. Knowl. data Eng. 33 (11) (2020) 3507–3519.
- [33] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
- [34] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
- [35] Sepp Hochreiter, Jürgen Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
- [36] Qing-Yuan Jiang, Wu-Jun Li, Deep cross-modal hashing, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3232–3240.
- [37] De Xie, Cheng Deng, Chao Li, Xianglong Liu, Dacheng Tao, Multi-task consistency-preserving adversarial hashing for cross-modal retrieval, IEEE Trans. image Process. 29 (2020) 3626–3637.
- [38] Ruiqing Xu, Chao Li, Junchi Yan, Cheng Deng, Xianglong Liu, Graph convolutional network hashing for cross-modal retrieval, in: International Joint Conference on Artificial Intelligence, Vol. 2019, 2019, pp. 982–988.
- [39] Rong-Cheng Tu, Xian-Ling Mao, Bing Ma, Yong Hu, Tan Yan, Wei Wei, Heyan Huang, Deep cross-modal hashing with hashing functions and unified hash codes jointly learning, IEEE Trans. Knowl. data Eng. 34 (2) (2020) 560–572.
- [40] Cong Bai, Chao Zeng, Qing Ma, Jinglin Zhang, Shengyong Chen, Deep adversarial discrete hashing for cross-modal retrieval, in: International Conference on Multimedia Retrieval, 2020, pp. 525–531.
- [41] Wentao Tan, Lei Zhu, Jingjing Li, Huaxiang Zhang, Junwei Han, Teacher-student learning: Efficient hierarchical message aggregation hashing for cross-modal retrieval, IEEE Trans. Multimedia (2022).
- [42] Xinwang Liu, Li Liu, Qing Liao, Siwei Wang, Yi Zhang, Wenxuan Tu, Chang Tang, Jiyuan Liu, En Zhu, One pass late fusion multi-view clustering, in: International Conference on Machine Learning, 2021, pp. 6850–6859.
- [43] Xinwang Liu, Xinzhou Zhu, Miaomiao Li, Lei Wang, Chang Tang, Jianping Yin, Dinggang Shen, Huaimin Wang, Wen Gao, Late fusion incomplete multi-view clustering, IEEE Trans. Pattern Anal. Mach. Intell. 41 (10) (2018) 2410–2423.
- [44] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, Heng Tao Shen, Inter-media hashing for large-scale retrieval from heterogeneous data sources, in: ACM SIGMOD International Conference on Management of Data, 2013, pp. 785–796.
- [45] Jile Zhou, Guiguang Ding, Yuchen Guo, Latent semantic sparse hashing for cross-modal similarity search, in: International ACM SIGIR Conference on Research & Development in Information Retrieval, 2014, pp. 415–424.
- [46] Weiwei Wang, Yuming Shen, Haofeng Zhang, Li Liu, Semantic-rebased cross-modal hashing for scalable unsupervised text-visual retrieval, Inform. Process. Manag. 57 (6) (2020) 102374.
- [47] Mark J. Huiskes, Michael S. Lew, The mir flickr retrieval evaluation, in: ACM International Conference on Multimedia Information Retrieval, 2008, pp. 39–43.
- [48] 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: ACM International Conference on Image and Video Retrieval, 2009, pp. 1–9.
- [49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C Lawrence Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.
- [50] Qing-Yuan Jiang, Wu-Jun Li, Discrete latent factor model for cross-modal hashing, IEEE Trans. image Process. 28 (7) (2019) 3490–3501.
- [51] Yongxin Wang, Xin Luo, Xin-Shun Xu, Label embedding online hashing for cross-modal retrieval, in: ACM International Conference on Multimedia, 2020, pp. 871–879.
- [52] Lei Zhu, Xu Lu, Zhiyong Cheng, Jingjing Li, Huaxiang Zhang, Deep collaborative multi-view hashing for large-scale image search, IEEE Trans. image Process. 29 (2020) 4643–4655.
- [53] Xu Lu, Lei Zhu, Zhiyong Cheng, Liqiang Nie, Huaxiang Zhang, Online multi-modal hashing with dynamic query-adaption, in: International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 715–724.
- [54] Diederik P. Kingma, Jimmy Ba, Adam: A method for stochastic optimization, 2015, arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [55] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, Ruslan R Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, 2012, arXiv preprint [arXiv:1207.0580](https://arxiv.org/abs/1207.0580).
- [56] Youngjoong Ko, A study of term weighting schemes using class information for text classification, in: International ACM SIGIR Conference on Research and Development in Information Retrieval, 2012, pp. 1029–1030.



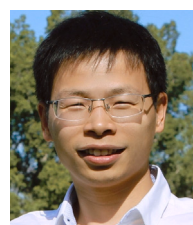
Xinyu Xia is currently a postgraduate with both Beijing Institute of Basic Medical Sciences and Shandong Normal University, China. His current research interests include information retrieval, hashing and multimodal learning.



Guohua Dong is currently an associate professor with Beijing Institute of Basic Medical Sciences, China. She received the Ph.D. degree from College of Computer, National University of Defense Technology in 2019. Her current research interests include information retrieval, hashing, multi-modal learning and brain-inspired artificial intelligence.



Fengling Li is currently a Ph.D. student with Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney. Her current research interest mainly focuses on the multi-modal analysis and retrieval.



Lei Zhu is currently a professor with the School of Information Science and Engineering, Shandong Normal University. He received his B.Eng. and Ph.D. degrees from Wuhan University of Technology in 2009 and Huazhong University Science and Technology in 2015, respectively. He was a Research Fellow at the University of Queensland (2016–2017). His research interests are in the area of large-scale multimedia content analysis and retrieval. Zhu has co-authored more than 100 peer-reviewed papers, such as ACM SIGIR, ACM MM, IEEE TPAMI, IEEE TIP, IEEE TKDE, and ACM TOIS. His publications have attracted more than 6,000 Google citations. At present, he serves as the Associate

Editor of IEEE TBD, ACM TOMM, and Information Sciences. He has served as the Area Chair, Senior Program Committee or reviewer for more than 40 well-known international journals and conferences. He won ACM SIGIR 2019 Best Paper Honorable Mention Award, ADMA 2020 Best Paper Award, ChinaMM 2022 Best Student Paper Award, ACM China SIGMM Rising Star Award, Shandong Provincial Entrepreneurship Award for Returned Students, and Shandong Provincial AI Outstanding Youth Award.



Xiaomin Ying received the B.S. and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 1997, and 2003, respectively. She is currently a professor in Bioinformatics. Her current research interests include machine learning and interdisciplinary researches of AI and biology.