Knowledge-Based Systems 239 (2022) 107927

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Multi-label modality enhanced attention based self-supervised deep cross-modal hashing

Xitao Zou [a,b], Song Wu [a,*], Nian Zhang [c], Erwin M. Bakker [d]

[a] College of Computer and Information Science, Southwest University, Chongqing 400715, China
[b] Key Laboratory of Intelligent Information Processing and Control of Chongqing Municipal Institutions of Higher Education, Chongqing Three Gorges University, Wanzhou, Chongqing 404100, China
[c] Department of Electrical and Computer Engineering, University of the District of Columbia, Washington, D.C., 20008, USA
[d] LIACS Media Lab, Leiden University, Leiden, Netherlands

ARTICLE INFO

Article history:
Received 30 August 2021
Received in revised form 12 November 2021
Accepted 8 December 2021
Available online 28 December 2021

Keywords:
Deep cross-modal hashing
Attention mechanism
Multi-label semantic learning

ABSTRACT

The recent deep cross-modal hashing (DCMH) has achieved superior performance in effective and efficient cross-modal retrieval and thus has drawn increasing attention. Nevertheless, there are still two limitations for most existing DCMH methods: (1) single labels are usually leveraged to measure the semantic similarity of cross-modal pairwise instances while neglecting that many cross-modal datasets contain abundant semantic information among multi-labels. (2) several DCMH methods utilized the multi-labels to supervise the learning of hash functions. Nevertheless, the feature space of multi-labels suffers the weakness of sparse, resulting in sub-optimization for the hash functions learning. Thus, this paper proposed a multi-label modality enhanced attention-based self-supervised deep cross-modal hashing (MMACH) framework. Specifically, a multi-label modality enhanced attention module is designed to integrate the significant features from cross-modal data into multi-labels feature representations, aiming to improve its completion. Moreover, a multi-label cross-modal triplet loss is defined based on the criterion that the feature representations of cross-modal pairwise instances with more common categories should preserve higher semantic similarity than other instances. To the best of our knowledge, the multi-label cross-modal triplet loss is the first time designed for cross-modal retrieval. Extensive experiments on four multi-label cross-modal datasets demonstrate the effectiveness and efficiency of our proposed MMACH. Moreover, the MMACH also achieved superior performance and outperformed several state-of-the-art methods on the task of cross-modal retrieval. The source code of MMACH is available at https://github.com/SWU-CS-MediaLab/MMACH.

1. Introduction

With the advent and prevalence of Web 3.0, more and more multi-modal data, such as graphics, texts, videos, images, and so on, have been accumulated in the social network. As data from distinct modalities may represent an identical object or event, it is beneficial to bridge semantically relevant data from different modalities to implement massive multi-modal instances matching, fusing, and retrieval. Therefore, cross-modal retrieval [1,2] is proposed to retrieve semantically related data from one modality while the query data is from a distinct modality. Because data in different modalities have different distributions and dissimilar feature spaces, efficiently and effectively minimizing the semantic gaps between these large-scale yet heterogeneous data and accurately calculating the semantical similarity of cross-modal data are still big challenges for cross-modal retrieval.

Generally, a large number of existing cross-modal retrieval methods, including topic models [3–5], subspace learning [6–11], and deep models [12–20], project original features of cross-modal instances into a common real-valued subspace and measure the semantic similarities in the common real-valued subspace. However, due to the rapid increment of the amount and scale of the multi-modal data, real-valued-based cross-modal retrieval methods usually suffer the weakness of high computation costs and low retrieval accuracy. Thus, hashing-based cross-modal retrieval (also called cross-modal hashing (CMH)) methods are proposed to map high-dimensional data from each modality into compact binary codes and calculate the semantic relevance of cross-modal pairwise instances with an efficient XOR operation. Thus, CMH has been a prevalent research topic in recent years because of the significant strengths of low data storage and high similarity measurement.

Depending on whether category labels are leveraged during the training stage, existing cross-modal hashing methods can be further divided into unsupervised and supervised manners.

* Corresponding author.
E-mail address: songwuswu@swu.edu.cn (S. Wu).

https://doi.org/10.1016/j.knosys.2021.107927
0950-7051/

Unsupervised cross-modal hashing methods [21–27] transform the original modality data to homogeneous binary codes by calculating the similarities of different modality data representations while preserving the semantic relevance without the guidance of data labels. By contrast, supervised cross-modal hashing methods [28–35] encode the heterogeneous cross-modal instances into compact hash codes and keep the cross-modal semantic similarities with the supervised information of class labels. Compared to unsupervised manners, supervised cross-modal hashing methods can fully use semantic relations of cross-modal instances by utilizing semantic labels and thus significantly boost the performance of cross-modal retrieval.

In the past few years, deep neural networks (DNNs) have been proposed and applied to many tasks such as sentence recognition, object detection, image caption, etc. Without exception, deep neural networks based cross-modal hashing are widely investigated. Pairwise relationship guided deep hashing (PRDH) [36] integrates several different pairwise constraints to protect the semantic similarity of pairwise instances from both intra-modalities and inter-modalities. Deep cross-modal hashing (DCMH) [37] utilizes two deep neural networks to learn hash functions for image and text-modality data representations, respectively. Self-supervised adversarial hashing (SSAH) [38] regards the multi-labels of each image-text pair as a single modality and from which a hash projection function is learned to supervise the training of hash mapping functions for the image-modality as well as the text-modality. Due to the remarkable feature learning ability, deep cross-modal hashing methods can more effectively capture the correlation across different modalities than hand-crafted methods.

In most of the existing deep cross-modal hashing methods, two cross-modal pairwise instances are regarded as semantically similar only if they have at least one common category. They usually neglect the fact that if two cross-modal pairwise instances have more common labels than another cross-modal pairwise instance, then the semantic similarity of the former should be higher than the latter (As shown in Fig. 1). Therefore, most of the existing deep cross-modal hashing methods neglect the abundant semantic information in multiple-labels of cross-modal datasets£resulting in inaccurately evaluating the semantic relevance of cross-modal pairwise instances and weakly optimization of the learned cross-modal hash functions. Furthermore, a few deep cross-modal hashing methods introduce self-supervised learning into deep cross-modal hashing, which regard the multi-labels of original instances as a signal modality and learn a hash function to supervise the training of other modalities. This self-supervised-based deep cross-modal hashing can enhance the performance of cross-modal retrieval. However, as the original multi-label matrix is very sparse, the multi-label-based self-supervised learning strategy shows only a limited enhancement of the learned cross-modal hash projection functions.

To further boost the robustness of cross-modal hashing, we propose a multi-label modality enhanced attention-based self-supervised deep hashing (MMACH) for high-performance cross-modal retrieval. Specifically, a multi-label modality enhanced attention (MMEA) module is firstly defined to overcome the sparsity of the multi-label matrix in the self-supervised learning-based deep cross-modal hashing. The MMEA utilizes three encoders to transfer each original instance (including original image features, original text features, and corresponding multi-labels) into a latent feature space and then normalizes them to increase their discrimination. Afterward, the normalized feature representations from the original image and text modality are fused into the feature representations from the corresponding multi-labels by a self-attention mechanism, respectively. Secondly, a multi-label cross-modal triplet loss (MCTL) is designed to measure the semantic similarity of multi-label cross-modal instances. Suppose that we have a triplet of instances (a, b, c) and each instance has its corresponding multi-labels. If instance a and instance b have more common categories than instance a and instance c, thus a and b are more semantically relevant to each other than a and c, meanwhile, the learned features of a and b should be more similar than the learned features of a and c. Inspired by this, a multi-label cross-modal triplet loss is designed based on the fact that if two cross-modal instances have more categories in common than other instances, the similarity of the learned features should also be higher than others. The proposed modules of MMEA and MCTL are further integrated into a self-supervised learning-based deep cross-modal hashing framework for high-performance cross-modal retrieval. The main contributions of our work are three-fold:

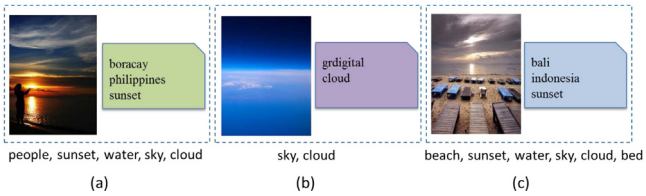


Fig. 1. This figure is the demonstration of three image-text instances with multiple labels. In previous deep cross-modal hashing approaches, the semantic similarity of the image-text instances in (a) and (b) is regarded as 1, because they have at least one common categories, i.e., *sky, cloud*. Analogously, the semantic similarity of the image-text pairs in (a) and (c) is regarded as 1, because they have several common categories *sunset, water, sky, cloud*. In fact, the semantic similarity of the image-text pairs in (a) and (c) is higher than that of the image-text pairs in (a) and (b), because the former pairs share more common categories than the latter pairs.

1. A novel multi-label modality enhanced attention (MMEA) module is designed to address the sparsity of the multi-labels-based similarity matrix in the self-supervised learning-based deep cross-modal hashing framework. Three encoders are firstly employed to transform the original image-text pairwise instances and their corresponding multi-labels into latent feature representations. The significantly useful semantic information of text and image feature representations are fused into their corresponding feature representations of multi-labels, respectively. The fusion process is based on a self-attention mechanism, which could effectively improve the completion of the multi-labels-based similarity matrix. 2. A robust multi-label cross-modal triplet loss (MCTL) is designed to measure the semantic similarity of multi-label cross-modal instances more correctly. The MCTL is constructed based on the observation that the feature representations of cross-modal pairwise instances with more common categories should also preserve higher semantic similarity than other cross-modal pairwise instances. To the best of our knowledge, the multi-label cross-modal triplet loss is the first time designed for the task of cross-modal retrieval.

3. The multi-label modality enhanced attention-based self-supervised deep cross-modal hashing (MMACH) is proposed. The MMACH integrated the designed multi-label modality enhanced attention (MMEA) module and the multi-label cross-modal triplet loss (MCTL) to improve the performance of cross-modal retrieval. Extensive experiments conducted on four well-known cross-modal datasets demonstrated the effectiveness of our MMACH. The comparison with several state-of-the-art baselines also shows the superiority of MMACH.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 presents details of our multi-label modality enhanced attention-based self-supervised deep cross-modal hashing (MMACH) framework. The learning produce

of MMACH is discussed in Section 4. Section 5 shows the evaluation as well as comparison experimental results on several datasets of MMACH. Section 6 concludes the MMACH in this paper.

2. Related work

2.1. Deep cross-modal hashing

Previous cross-modal hashing methods are shallow architecture-based methods that first extract hand-crafted features and then utilize these hand-crafted features to learn hash functions. These methods are based on a two-stage architecture where the two stages may not be optimally compatible, resulting in suboptimal performance. By contrast, deep cross-modal hashing methods benefit from the significant feature extraction capabilities of deep neural networks. Thus, they can better explore and exploit the correlations across different modalities in an end-to-end manner. As a result, deep cross-modal hashing retrieval has attracted increasing attention. Representative methods are deep cross-modal hashing (DCMH) [37], pairwise relation guided deep hashing (PRDH) [36], correlation hashing network (CHN) [39], cross-modal hamming hashing (CMHH) [40], and self-supervised adversarial hashing (SSAH) [38]. DCMH effectively projects image-text pairs into corresponding hash codes by using an end-to-end deep neural network framework. PRDH exploits intra-modal and inter-modal constraints of different pairwise instances to generate discriminative hash codes with a united deep learning framework. CHN defines a cosine max-margin loss to enhance the quality of the learned hash codes. CMHH uses an exponential focal loss to significantly penalize similar cross-modal pairs with Hamming distances larger than the Hamming radius threshold. SSAH introduces self-supervised learning to cross-modal hashing and learns the hash function (LabelNet) on the multi-label modality to supervise other modalities. Nonetheless, these methods either leverage single labels to calculate the semantic similarity of cross-modal pairwise instances or regard the semantic similarity of cross-modal pairwise instances with multiple labels as one when they have at least one common category. However, the fact that many cross-modal datasets have multiple labels containing abundant semantic information is neglected in these methods. Specifically, suppose two cross-modal instances have more common categories than some other cross-modal pairwise instances. In that case, the semantic similarity of the former pair is higher than the semantic similarity of the latter pair. Moreover, existing self-supervised-based deep cross-modal hashing methods often suffer from inferior performance because the hash function learned on the sparse multi-labels has a weak capacity to supervise the training of the hash functions of other modalities.

2.2. Attention mechanism

An attention mechanism [41–44] is first introduced and widely applied in natural language processing, which considers neighboring words when extracting features from one word. Subsequently, the attention mechanism is introduced to various computer vision tasks, where it is trained to identify what the model should concentrate on when performing a particular task. To date, only a few methods combine cross-modal hashing retrieval with an attention mechanism. Attention-aware deep adversarial hashing (DAH) [45] introduces an attention mechanism to cross-modal hashing and generates adaptive attention masks that divide the feature representations into attended and unattended feature representations. In our proposed method, the image and text modality feature representations are fused into the feature representations of multi-labels modality based on a novel self-attention mechanism. It could effectively improve the completion of a multi-label similarity matrix and supervise the training of hash functions for different modalities.

2.3. Multi-label learning

Multi-label learning pays attention to the issue that an instance is associated with several labels simultaneously [46,47]. Generally, instances with multi-labels contain more semantic information than instances with single labels. Adequately mining the semantic information in multi-labels to accurately calculate the semantic similarities between instances is still a challenge. To this end, [48] proposes a distance metric learning algorithm for multi-label classification, which integrates a pairwise multi-label similarity constraint and a Jaccard Distance into multi-label learning and achieves competitive performance. This paper defines a multi-label cross-modal triplet loss to better explore the semantic information in multi-labels and further preserve the multi-labels similarity, especially preserving the multi-label similarities of cross-modal instances.

3. Proposed method

In this section, we describe our proposed multi-label modality enhanced attention-based self-supervised deep cross-modal hashing (MMACH) method with the following subsections: notations and problem formulation, modal encoders, multi-label enhanced attention module, hash representations learning, and hash codes generation. For the sake of clarity, in the following, we always assume that each data instance has three modalities (i.e., an image-modality, a text modality, and a multi-label modality). The framework of MMACH is shown in Fig. 2.

3.1. Notation and problem formulation

To better understand the task of CMH, we firstly give a formal definition of notations and problem formulations. For a given training set of n instances $O = \{\{I_i\}_{i=1}^n, \{T_i\}_{i=1}^n, \{L_i\}_{i=1}^n\}$, where $I_i \in R^{d_I}$, $T_i \in R^{d_T}$ and $L_i \in R^{d_L}$ are the original image features, the original text features as well as the multi-labels of the ith training instance. If the ith training instance is assigned to the jth class, then the jth component of L_i equals 1 (i.e., $L_{ij} = 1$), otherwise $L_{ij} = 0$.

With the provided training set and semantic similarity matrices, the goal of cross-modal hashing is to learn three hash functions to project the original images, the original texts and the original multi-labels modality data into compact hash codes, meanwhile, effectively preserve semantic similarities of these cross-modal instances. To achieve this goal, the original instances of three modality data are encoded into c-dimensional feature vectors with pre-trained deep neural networks, i.e., $\{I_i\}_{i=1}^n$, $\{T_i\}_{i=1}^n$ and $\{L_i\}_{i=1}^n$ are projected into $\{F_i^E\}_{i=1}^n$, $\{G_i^E\}_{i=1}^n$ and $\{H_i^E\}_{i=1}^n$, respectively. As the original multi-labels are pretty sparse,a multi-label modality enhanced attention mechanism is designed to compensate for this weakness. The multi-label enhanced feature vectors are denoted as $\{H_i^A\}_{i=1}^n$. Afterwards, three deep neural networks are utilized to project $\{F_i^E\}_{i=1}^n$, $\{G_i^E\}_{i=1}^n$ and $\{H_i^A\}_{i=1}^n$ into k-dimensional hash representations $\{F_i^{hr}\}_{i=1}^n$, $\{G_i^{hr}\}_{i=1}^n$ and $\{H_i^{hr}\}_{i=1}^n$, respectively, i.e., $F_i^{hr} = f(F_i^E, \theta^I)$, $G_i^{hr} = g(G_i^E, \theta^T)$, $H_i^{hr} = h(H_i^A, \theta^L)$, where $f(., \theta^I)$, $g(., \theta^T)$ and $h(., \theta^L)$ are hash representation learning functions for the image-modality, the text-modality and the multi-label modality, respectively. θ^I, θ^T and θ^L are parameters of the three deep neural networks, respectively. Finally, a sign function is used to generate united hash codes matrix $B \in R^{n \times k}$ from the learned hash representations.

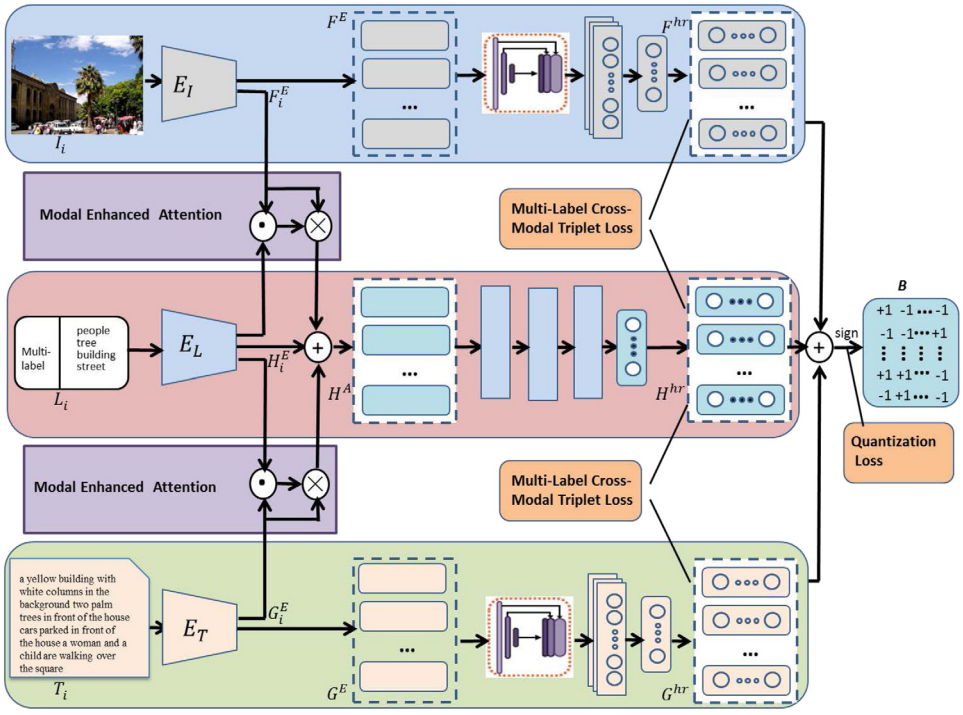


Fig. 2. This figure demonstrates the framework of our proposed MMACH method. The MMACH contains three modules: (1) The first module is a modal encoder part (E_I, E_L and E_T), it is composed of three deep neural networks to extract the features from the original instances of the image modality, the text modality, and the multi-label modality, respectively. (2) The second module is a multi-label modality enhanced attention module. It utilizes an attention mechanism to extract semantically relevant information from the image and text modality and subsequently fuses them to the sparse multi-label modality. (3) The third module is a hash representation learning and hash codes generation part. It aims to ensure that semantically similar pairs of cross-modal instances have similar hash codes. The $\odot$ represents the dot product, while $\otimes$ represents the element-wise product, and $\oplus$ denotes element-wise adds.

3.2. Modal encoders

In order to effectively extract discriminative features from the original instances, three encoders E_I, E_T and E_L are used to encode each original image I_i, text T_i, and multi-label L_i modality data into c-dimensional feature vectors F_i^E, G_i^E and H_i^E, respectively.

$$\begin{aligned} F_i^E &= E_I(I_i) \\ G_i^E &= E_T(T_i) \\ H_i^E &= E_L(L_i) \end{aligned} \tag{1}$$

3.3. Multi-label modality enhanced attention module

Many benchmark datasets for the task of cross-modal hashing retrieval (e.g., MIRFLICKR-25K [49] and NUS-WIDE [50], etc.) contain multi-labels. Nevertheless, most previous methods merely regard a pair of two cross-modal instances as similar if they share at least one common category. The abundant semantic information in multi-labels is neglected and thus cannot accurately evaluate the pairwise semantic relevance of cross-modal instances. As a result, the learned cross-modal hash projection functions have suboptimal performance. To solve this issue, a multi-label-based self-supervised learning strategy is designed to guide the learning of cross-modal hash projection functions. Because the original multi-label matrix suffers the weakness of sparse, a multi-label-based self-supervised learning strategy can only obtain a limited enhancement for the learned cross-modal hash projection functions. For this purpose, in this subsection, a multi-label modality enhanced attention module (MMEA) is proposed to improve the completion of the multi-label matrix. Specifically, for a given training image-text pair with multi-labels $\{I_i, T_i, L_i\}$, MMEA firstly utilizes the encoders in Section 3.2 to transfer them into c-dimensional feature vectors F_i^E, G_i^E and H_i^E, then an attention mechanism is introduced to fuse these relative semantic information of F_i^E and G_i^E into H_i^E. The corresponding formulations are as follows:

$$\begin{aligned} attention^{IL} &= \frac{F_i^E}{\| F_i^E \|} \cdot \frac{H_i^E}{\| H_i^E \|} \\ attention^{TL} &= \frac{G_i^E}{\| G_i^E \|} \cdot \frac{H_i^E}{\| H_i^E \|} \end{aligned} \tag{2}$$

Where $attention^{IL}$ and $attention^{TL}$ are semantic affinities of F_i^E and H_i^E, and G_i^E and H_i^E, respectively. $\| \cdot \|$ denotes a normalization on a feature vector.

$$H_i^A = H_i^E + attention^{IL} F_i^E + attention^{TL} G_i^E \tag{3}$$

Where H_i^A is the multi-label modal enhanced feature vector for the original multi-label L_i. By using Eqs. (2) and (3), we can compensate the sparsity of multi-label L_i with the abundant semantic information contained in I_i and T_i, and employ a self-supervising learning manner to better guide the training of deep neural networks for the image and the text modalities.

3.4. Multi-label cross-modal triplet loss

Suppose that we have a cross-modal triplet (I_i, T_{p1}, T_{p2}), where image I_i is more semantically similar to text T_{p1} than to text T_{p2}. Their respective hash representations F_i^{hr}, G_{p1}^{hr} and G_{p2}^{hr} can be easily learned with the respective hash mapping functions, $F_i^{hr} = f(F_i^E, \theta^I)$, $G_{p1}^{hr} = g(G_{p1}^E, \theta^T)$ and $G_{p2}^{hr} = g(G_{p2}^E, \theta^T)$. To preserve the semantic similarity during the hash representation learning procedure, the similarity of F_i^{hr} and G_{p1}^{hr} should be higher than the similarity of F_i^{hr} and G_{p2}^{hr}. Therefore, inspired by [51–53],

we define the multi-label cross-modal triplet loss (MCTL) as follows:

$$J^{IT}(I_i, T_{p1}, T_{p2}) = \sum_{I_i, T_{p1}, T_{p2}} max(0, \| F_i^{hr} - G_{p1}^{hr} \|_2^2 - \| F_i^{hr} - G_{p2}^{hr} \|_2^2 + \gamma) \quad (4)$$

Where $\| \cdot \|_2$ is the L_2 norm, and γ is a positive margin. Eq. (4) means that the L_2 distance of a more semantically similar multi-label cross-modal pair is smaller than the L_2 distance of a less semantically similar multi-label cross-modal pair by a margin of γ. By this manner, the multi-label cross-modal semantic similarity can be adequately protected during stage of hash representation learning.

3.5. Hash representations learning

During the stage of hash representation learning, the learned multi-label modal enhanced feature vectors $\{H_i^A\}_{i=1}^n$, the feature vectors for the image-modality $\{F_i^E\}_{i=1}^n$, and the feature vectors for the text-modality $\{G_i^E\}_{i=1}^n$ are forward into the deep neural network for the multi-label modality, the deep neural network for the image-modality, and the deep neural network for the text-modality, respectively. To preserve the semantic similarity of cross-modal instances during the hash representation learning stage, we introduce the multi-label cross-modal triplet loss in Section 3.4 into our method. Specifically, for cross-modal triplets $(H_i^A, F_{p1}^E, F_{p2}^E)$, $(F_i^E, H_{p1}^A, H_{p2}^A)$, $(H_i^A, G_{p1}^E, G_{p2}^E)$, and $(G_i^E, H_{p1}^A, H_{p2}^A)$, we define the following semantic similarity preserving loss functions:

$$\begin{aligned} J^{IL} &= J^{IL}(H_i^A, F_{p1}^E, F_{p2}^E) + J^{IL}(F_i^E, H_{p1}^A, H_{p2}^A) \\ &= \sum_{H_i^A, F_{p1}^E, F_{p2}^E} max(0, \| H_i^{hr} - F_{p1}^{hr} \|_2^2 - \| H_i^{hr} - F_{p2}^{hr} \|_2^2 + \gamma_1) \\ &+ \sum_{F_i^E, H_{p1}^A, H_{p2}^A} max(0, \| F_i^{hr} - H_{p1}^{hr} \|_2^2 - \| F_i^{hr} - H_{p2}^{hr} \|_2^2 + \gamma_2) \end{aligned} \quad (5)$$

Where J^{IL} is the cross-modal semantic similarity preserving loss for the image-modality and the multi-label modality. The multi-label semantic similarity of H_i^A and F_{p1}^E is higher than the multi-label semantic similarity of H_i^A and F_{p2}^E, and the multi-label semantic similarity of F_i^E and H_{p1}^A is higher than the multi-label semantic similarity of F_i^E and H_{p2}^A. And γ_1 and γ_2 are two positive margins.

$$\begin{aligned} J^{TL} &= J^{TL}(H_i^A, G_{p1}^E, G_{p2}^E) + J^{TL}(G_i^E, H_{p1}^A, H_{p2}^A) \\ &= \sum_{H_i^A, G_{p1}^E, G_{p2}^E} max(0, \| H_i^{hr} - G_{p1}^{hr} \|_2^2 - \| H_i^{hr} - G_{p2}^{hr} \|_2^2 + \gamma_3) \\ &+ \sum_{G_i^E, H_{p1}^A, H_{p2}^A} max(0, \| G_i^{hr} - H_{p1}^{hr} \|_2^2 - \| G_i^{hr} - H_{p2}^{hr} \|_2^2 + \gamma_4) \end{aligned} \quad (6)$$

Where J^{TL} is the cross-modal semantic similarity preserving loss for the text-modality and the multi-label modality, and the multi-label semantic similarity of H_i^A and G_{p1}^E is higher than the multi-label semantic similarity of H_i^A and G_{p2}^E, and the multi-label semantic similarity of G_i^E and H_{p1}^A is higher than the multi-label semantic similarity of G_i^E and H_{p2}^A. And γ_3 and γ_4 are two positive margins.

3.6. Hash codes generation

In Section 3.5, we described how we can acquire the hash representations $\{F_i^{hr}\}_{i=1}^n$, $\{G_i^{hr}\}_{i=1}^n$ and $\{H_i^{hr}\}_{i=1}^n$ for the original images $\{I_i\}_{i=1}^n$, texts $\{T_i\}_{i=1}^n$, and multi-labels $\{L_i\}_{i=1}^n$, respectively. However, the goal of cross-modal hashing is to map multi-modal data into compact hash codes. To this end, we utilize a sign function to approximately generate the hash codes from the learned hash representations:

$$B_i = sign(\frac{F_i^{hr} + G_i^{hr} + H_i^{hr}}{3}) \quad (7)$$

Where $B_i \in R^k$ is the hash codes for the ith instance. To minimize the information loss in Eq. (7), we firstly squeeze the hash representations from a real-valued space into $[-1, 1]$ with the following *tanh* function:

$$\begin{aligned} F_i^{hr} &= tanh(F_i^{hr}) \\ G_i^{hr} &= tanh(G_i^{hr}) \\ H_i^{hr} &= tanh(H_i^{hr}) \end{aligned} \quad (8)$$

Where $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Moreover, to further decrease the information loss in Eq. (7), the following quantization loss is also introduced:

$$J_{quantization} = \frac{\sum_{i=1}^n (\| B_i - F_i^{hr} \|_2^2 + \| B_i - G_i^{hr} \|_2^2 + \| B_i - H_i^{hr} \|_2^2)}{3nk} \quad (9)$$

Where n and k are the number of training instances and the length of hash codes, respectively.

Combining the cross-modal semantic similarity preserving losses with the quantization loss together, the complete loss function can be obtained as follows:

$$J = \frac{1}{n_{IL}^2 k} J^{IL} + \frac{1}{n_{TL}^2 k} J^{TL} + \alpha J_{quantization} \quad (10)$$

Where α is a hyper-parameter to balance the cross-modal semantic similarity preserving losses and the quantization loss. n_{IL} is the number of cross-modal triplets from the image-modality and the multi-label modality, and n_{TL} is the number of cross-modal triplets from the text-modality and the multi-label modality.

3.7. Hash representations learning networks

For the image-modality, we fine-tune the multi-scale (MS) fusion based TxtNet in SSAH [38] ($c \rightarrow MS \rightarrow 4096 \rightarrow 512 \rightarrow k$) to learn the corresponding hash representations from the encoded features.

For the text-modality, the TxtNet in SSAH is fine-tuned ($c \rightarrow MS \rightarrow 4096 \rightarrow 512 \rightarrow k$) and utilized to learn the corresponding hash representations from the encoded features.

For the multi-label modality, a deep neural network with three fully-connected layers ($c \rightarrow 8192 \rightarrow k$) is introduced to learn the hash representations from the encoded features.

4. Learning algorithm of MMACH

To learn the optimized θ^I, θ^T, θ^L and B, an alternating strategy is introduced to update one of θ^I, θ^T, θ^L and B, while keeping the other three fixed. The detailed execution and optimization schema for MMACH are given in Algorithm 1.

4.1. Optimize θ^L with θ^I, θ^T and B unchanged

While we keep θ^I, θ^T and B unchanged, the parameters θ^L of the DNN for the multi-label modality can be learned by stochastic gradient descent (SGD) and back-propagation (BP). Detailedly, in each iteration, four training batches of cross-modal triplets are randomly selected to execute our algorithm. For each selected multi-label enhanced feature vector H_i^A, the gradient is computed as follows:

$$\begin{aligned}\frac{\partial J}{\partial H_{\cdot}^{hr}} =& \frac{2}{n_{IL}^2 k}\Big(\sum_{H_i^A,F_{p1}^E,F_{p2}^E}(F_{p2}^{hr}-F_{p1}^{hr})+\sum_{F_i^E,H_{p1}^A,H_{p2}^A}(H_{p1}^{hr}-H_{p2}^{hr})\Big)\\ &+\frac{2}{n_{TL}^2 k}\Big(\sum_{H_i^A,G_{p1}^E,G_{p2}^E}(G_{p2}^{hr}-G_{p1}^{hr})+\sum_{G_i^E,H_{p1}^A,H_{p2}^A}(H_{p1}^{hr}-H_{p2}^{hr})\Big)\\ &-\frac{2\alpha\sum_{i=1}^{n}(B_i-H_i^{hr})}{3nk}\end{aligned} \tag{11}$$

Afterwards, $\frac{\partial J}{\partial \theta^L}$ can be calculated from $\frac{\partial J}{\partial H_{\cdot}^{hr}}$ by applying the chain rule. Finally, the θ^L can be optimized using $\frac{\partial J}{\partial \theta^L}$ and back-propagation.

4.2. Optimize θ^I with θ^L, θ^T and B unchanged

While we keep θ^T, θ^L and B unchanged, the parameters θ^I of the DNN for the image modality can be optimized by SGD and BP. During each epoch, two training batches of cross-modal triplets are randomly selected to run our method. For each selected image feature vector F_i^E, the gradient is calculated as follows:

$$\begin{aligned}\frac{\partial J}{\partial F_{\cdot}^{hr}} =& \frac{2}{n_{IL}^2 k}\Big(\sum_{H_i^A,F_{p1}^E,F_{p2}^E}(F_{p1}^{hr}-F_{p2}^{hr})+\sum_{F_i^E,H_{p1}^A,H_{p2}^A}(H_{p2}^{hr}-H_{p1}^{hr})\Big)\\ &-\frac{2\alpha\sum_{i=1}^{n}(B_i-F_i^{hr})}{3nk}\end{aligned} \tag{12}$$

Furthermore, $\frac{\partial J}{\partial \theta^I}$ can be calculated from $\frac{\partial J}{\partial F_{\cdot}^{hr}}$ by applying the chain rule. Finally, the θ^I can be optimized by using $\frac{\partial J}{\partial \theta^I}$ and back-propagation.

4.3. Optimize θ^T with θ^I, θ^L and B unchanged

When we keep θ^I, θ^L and B unchanged, the parameters θ^T of the DNN for the text modality can be optimized by SGD and BP. During each epoch, two training batches of cross-modal triplets are randomly selected to execute our algorithm. For each selected text feature vector G_i^E, the gradient is calculated as follows:

$$\begin{aligned}\frac{\partial J}{\partial G_{\cdot}^{hr}} =& \frac{2}{n_{IL}^2 k}\Big(\sum_{H_i^A,G_{p1}^E,G_{p2}^E}(G_{p1}^{hr}-G_{p2}^{hr})+\sum_{G_i^E,H_{p1}^A,H_{p2}^A}(H_{p2}^{hr}-H_{p1}^{hr})\Big)\\ &-\frac{2\alpha\sum_{i=1}^{n}(B_i-G_i^{hr})}{3nk}\end{aligned} \tag{13}$$

Afterwards, $\frac{\partial J}{\partial \theta^T}$ can be calculated from $\frac{\partial J}{\partial G_{\cdot}^{hr}}$ by using the chain rule. Finally, the θ^T can be optimized by using $\frac{\partial J}{\partial \theta^T}$ and back-propagation.

4.4. Optimize B with θ^I, θ^T and θ^L unchanged

When we keep θ^I, θ^T and θ^L unchanged, the hash codes B can be optimized with Eq. (7).

Algorithm 1 MMACH: Multi-Label Modality Enhanced Attention based Self-Supervised Deep Cross-Modal Hashing.

Input:
training instances: $O = \{\{I_i\}_{i=1}^n, \{T_i\}_{i=1}^n, \{L_i\}_{i=1}^n\}$.
the maximal epochs of the algorithm is *max_epoch*.
mini-batch size $n_{batch} = 128$.

Output:
Deep neural networks parameters are θ^I, θ^T and θ^L for hash representation learning, and the hash codes matrix B.

Encoding the original instances $\{I_i\}_{i=1}^n$, $\{T_i\}_{i=1}^n$, $\{L_i\}_{i=1}^n$ to c-dimensional features $\{F_i^E\}_{i=1}^n$, $\{G_i^E\}_{i=1}^n$ and $\{H_i^E\}_{i=1}^n$ with Eq. (1).
Learning the multi-label enhanced feature vectors $\{H_i^A\}_{i=1}^n$ from $\{H_i^E\}_{i=1}^n$ with Eqs. (2) and (3).
Generating n_{IL} $(H_i^A, F_{p1}^E, F_{p2}^E)$ (the triplets set is named $Triplet_{IL}$) and n_{IL} $(F_i^E, H_{p1}^A, H_{p2}^A)$ (the triplets set is named $Triplet_{LI}$) from $\{H_i^A\}_{i=1}^n$ and $\{F_i^E\}_{i=1}^n$, generating n_{TL} $(H_i^A, G_{p1}^E, G_{p2}^E)$ (the triplets set is named $Triplet_{TL}$) and n_{TL} $(G_i^E, H_{p1}^A, H_{p2}^A)$ (the triplets set is named $Triplet_{LT}$) from $\{H_i^A\}_{i=1}^n$ and $\{G_i^E\}_{i=1}^n$.
Initialize the deep neural network parameters θ^I, θ^T, θ^L, hash representations $\{F_i^{hr}\}_{i=1}^n$, $\{G_i^{hr}\}_{i=1}^n$, $\{H_i^{hr}\}_{i=1}^n$, hash codes matrix B, and the epoch numbers $batchnum_L = \lceil (n_{IL}+n_{TL})/(2n_{batch}) \rceil$, $batchnum_I = \lceil n_{IL}/n_{batch} \rceil$, $batchnum_T = \lceil n_{TL}/n_{batch} \rceil$.
repeat
  for $j = 1$ *to* $batchnum_L$ **do**
    Randomly select n_{batch} triplets from $Triplet_{IL}$, n_{batch} triplets from $Triplet_{LI}$, n_{batch} triplets from $Triplet_{TL}$, and n_{batch} triplets from $Triplet_{LT}$ to construct the respective four mini-batches.
    For each feature vector H_i^A in the mini-batches, calculate $H_i^{hr} = h(H_i^A, \theta^L)$ by forward propagation.
    Update $\{H_i^{hr}\}_{i=1}^n$.
    Compute the derivative of θ^L using Eq. (11).
    Utilize back-propagation to update the network parameters θ^L.
  end for
  for $j = 1$ *to* $batchnum_I$ **do**
    Randomly select n_{batch} triplets from $Triplet_{IL}$ and n_{batch} triplets from $Triplet_{LI}$ to construct the respective two mini-batches.
    For each feature vector F_i^E in the mini-batches, calculate $F_i^{hr} = f(F_i^E, \theta^I)$ by forward propagation.
    Update $\{F_i^{hr}\}_{i=1}^n$.
    Compute the derivative of θ^I using Eq. (12).
    Utilize back-propagation to update the network parameters θ^I.
  end for
  for $j = 1$ *to* $batchnum_T$ **do**
    Randomly select n_{batch} triplets from $Triplet_{TL}$ and n_{batch} triplets from $Triplet_{LT}$ to construct the respective two mini-batches.
    For each feature vector G_i^E in the mini-batches, calculate $G_i^{hr} = g(G_i^E, \theta^T)$ by forward propagation.
    Update $\{G_i^{hr}\}_{i=1}^n$.
    Compute the derivative of θ^T using Eq. (13).
    Utilize back-propagation to update the network parameters θ^T.
  end for
  Optimize B by utilizing Eq. (7).
until the max epoch number *max_epoch*

4.5. Complexity analysis

The time complexity of the overall loss function (Eq. (10)) of MMACH can be calculated as follows: $O(n_{IL})+O(n_{TL})+O(n \times k) \approx O(n)$, as $k \ll n$ and k, n_{IL}, n_{TL} are of the same magnitude as n.

5. Experiments

In order to validate the performance of our proposed MMACH method and compare it with several state-of-the-art cross-modal hashing methods, we conducted extensive experiments on four benchmark datasets.

5.1. Datasets

MIRFLICKR-25K [49]: the original MIRFLICKR-25K dataset is made up of 25,000 image-text pairs from the Flickr website. In our experiment, instances that have at least 20 textual tags are selected and thus 20,015 image-text pairs with multi-labels remain, where each of the selected instances is assigned to at

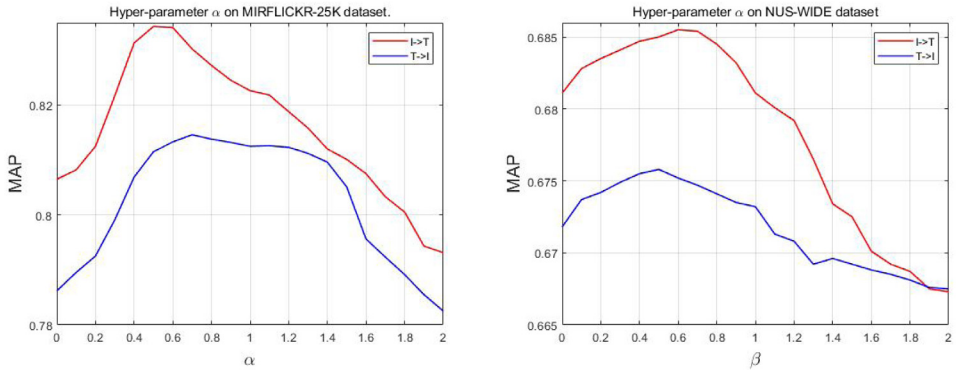


Fig. 3. Sensitivity analysis of the hyper-parameter α on MIRFLICKR25K and NUS-WIDE datasets.

Table 1
Detailed settings of experimental datasets.

Dataset	Used	Train	Query	Retrieve	Tag dimension	Labels
MIRFLICKR-25K	20,015	10,000	2,000	18,015	1,386	24
NUS-WIDE	190,421	10,500	2,100	188,321	1,000	21
MS COCO2014	122,218	10,000	5,000	117,218	2,026	80
IAPRTC-12	19,999	10,000	2,000	17,999	1,251	275

least one of the 24 given labels. For our experiments, we encode each textual tag into a 1386-dimensional BOW (bag-of-words) feature.

NUS-WIDE [50]: the original NUS-WIDE dataset contains 269,468 image-text pairs. We first abandon the data without categories, then choose data classified by the 21 most-frequent categories to construct a subset, which has 190,421 image-text pairs. For our experiments, we encode each textual tag into a 1000-dimensional BOW feature.

Microsoft COCO2014 [54]: the original Microsoft COCO2014 dataset comprises two parts: training set with 82,785 images, and validation set with 40,504 images. Each image contains 5 captions (which is regarded as a text modality). We first abandon instances that have no captions, then we combine the training set and validation set together to construct a subset with 122,218 image-text pairs, and each instance is annotated with at least one of the 80 classes. The text of each instance is represented as a 2026-dimensional BOW feature.

IAPRTC-12 [55]: the original IAPRTC-12 dataset is composed of 20,000 image-text pairs. In our experiment, we first eliminate instances without tags and then construct a subset of 19,999 image-text pairs with 275 categories. The text of each instance is encoded into a 1251-dimensional BOW feature.

Furthermore, the detailed information, including number of used instances, number of training set, number of query set, number of retrieval set, dimension of tags for each instance, and categories for the four experimental datasets are listed in Table 1. [56] provides more detailed information for experimental settings.

5.2. Evaluation metrics

For cross-modal hashing retrieval, two of the most prevalent leveraged retrieval protocols are Hamming ranking and hash lookup. Specifically, the Hamming ranking protocol ranks the retrieval results in ascending order of the Hamming distance for given a query instance. The hash lookup protocol returns retrieval instances within a certain Hamming radius from the query instance. In practical applications, Mean Average Precision (MAP), topN precision curves (topN Curves) and precision recall curves are three substitutions of the above two retrieval protocols. Thus, Mean Average Precision, Mean Average Precision and precision–recall curves are used as evaluation metrics to validate the performance of our proposed MMACH method and in the comparison with several state-of-the-art baseline methods.

5.3. Baselines and implementation details

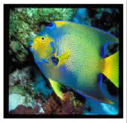
Several CMH methods, including hand-crafted based CMH methods **CMSSH** [57], **SePH** [58], **SCM** [31] and **GSPH** [20] and deep feature based CMH methods **DCMH** [37], **PRDH** [36], **CMHH** [40], **CHN** [39], **SSAH** [38] and **MLSPH** [56] are chosen as baseline methods in our experiments. The source codes of GSPH, SePH, SCM, CMSSH, SSAH, DCMH and MLSPH have been released and we cautiously implement them. For other methods, we cautiously implement them by ourselves.

By using the open source deep learning framework pytorch, our experiments are executing on an NVIDIA GTX Titan XP GPU server. During the training stage, each multi-label cross-modal triplet (a, b, c) is generated by using the following rule: a and b are instances from the first modality, while instance c is from another modality. Moreover, a and b have more common categories than a and c. In our experiments, the modal encoders E_L and E_T employ the universal sentence encoder [59] to encode each original text or original multi-label text into 512-dimensional feature vectors, and the modal encoder E_I utilizes ResNet34 [60] to extract the features of each original image. We acquire the output of the global average pool and resize it to a 512-dimensional feature vector. In our experiments, the maximum training epoch is set to 200, the learning rate is initialized to $10^{-1.5}$ and gradually lowered to 10^{-6} in 200 epochs. For all experiments, $I \rightarrow T$ represents the cases when using a querying image while returning text, while $T \rightarrow I$ represents the cases when using a querying text while returning an image. Source code will be released at: https://github.com/SWU-CS-MediaLab/MMACH.

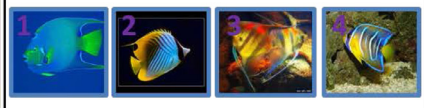
5.4. Performance comparisons and discussion

5.4.1. Hyper-parameters experiment

In this subsection, experiments are conducted on two datasets, i.e., MIRFLICKR-25K and NUS-WIDE. The length of hash codes is

Dataset	Query Image	MMACH: Retrieval texts	MLSDCH: Retrieval texts
MIRFLICKR-25K		1. maldives fuvahmulah kulhi mangrove sunset sunrise atoll gnaviyani pond wetland land swim reflation nikon red sky blue millzero d300 boy 2. trees sunset naturesfinest 3. crane gru sunset hdr tramonto cielo sky ray raggi light luci chdk milano soe flickrsbest Damniwishidtakenthat 4. okmulgee oklahoma sunset red drippingspringslake tree water reflection blueribbonwinner abigfave explore	1. maldives fuvahmulah kulhi mangrove sunset sunrise atoll gnaviyani pond wetland land swim reflation nikon red sky blue millzero d300 boy 2. trees sunset naturesfinest 3. boracay philippines sunset 4. roady photo photograph digital jlbrown jumpinjimmyjava canon40d roadart darksky thefunhouse
NUS-WIDE		1. fish angelfish 2. tropicalfish cichlid angelfish 3. fish yellow zoo angelfish 4. fish aquarium blue angelfish	1. fish angel boat ship tank angelfish 2. tropicalfish cichlid angelfish 3. 2005 beauty rock mexico angelfish 4. pink woman girl lady female bed pattern dress legs polkadots mauve knees shins angelfish lowcontrast patterned cocktaildress lowbrightness heartbreaktohate
Microsoft COCO		1. a man on a horse in a flat pasture; a second horse behind him on the left; 2. three people are riding on brown horses in the foreground; three red houses with a brown thatched roof and lila flowers with green leaves behind it; a white sky in the background; 3. a dark and a light brown horse with red saddles are standing on a path in the foreground; high grass and a wooded hill behind it; 4. a group of people is riding on brown horses on a green meadow; grey clouds in the background;	1. a man on a horse in a flat pasture; a second horse behind him on the left; 2. a grey statue of a man on a horse on a base made of marmol, with a fence in front of it and trees behind it; 3. four tourists are riding on brown horses on a gravel road; a green slope with a few bushes in the background; 4. four people riding on horses; two foals next to the horses; a creek with a brown rock face and forest in the background;
IAPRTC-12		1. a fountain and cobbled walkway in the foreground, a pink and white buidling with many arches in the background; trees on the right 2. a white building with lots of columns and arches, a neat lawn and neatly cut trees and bushes in the foreground; the flag of Paraguay is waving at the top of the building; there is a flower bed on the left; 3. a very modern building; stairs are leading up to the entrance; the walls are entirely made of glass; one red huge column is supporting the big roof; rails in the foreground; a green tree on the left; 4. Several flagpoles with waving flags on a green lawn in the foreground; a large grey and black building behind it; a huge column with a football on top on the left; a blue sky with white clouds in the background;	1. a fountain and cobbled walkway in the foreground, a pink and white buidling with many arches in the background; trees on the right 2. a white building with lots of columns and arches, a neat lawn and neatly cut trees and bushes in the foreground; the flag of Paraguay is waving at the top of the building; there is a flower bed on the left; 3. a large building on the left, a palm tree in centre of picture, (mostly) white cars in the street at a junction, some of them turning left, others going straight; there are red umbrellas in a park on the right; people are walking through the park, others are crossing the road in the foreground; 4. Front view of a huge dam; water is flowing through one tiny spot; backwater is flowing off on the left; green reed in the foreground;

(a) Image-to-text retrieval

Dataset	Query Text	MMACH: Retrieval images	MLSDCH: Retrieval images
MIRFLICKR-25K	maldives fuvahmulah kulhi mangrove sunset sunrise atoll gnaviyani pond wetland land swim reflation nikon red sky blue millzero d300 boy		
NUS-WIDE	. fish angelfish		
Microsoft COCO2014	a man on a horse in a flat pasture; a second horse behind him on the left;		
IAPRTC-12	. a fountain and cobbled walkway in the foreground, a pink and white buidling with many arches in the background; trees on the right		

(b) Text-to-image retrieval

Fig. 4. Examples of top 4 cross-modal retrieval results by MMACH and MLSDCH on four datasets. For (a) using images to retrieve texts, the matching texts are in blue. For (b) using texts to retrieve images, the purple number in each image is the ranking order, and the blue frames indicate the matching image.

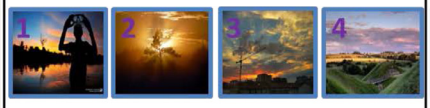
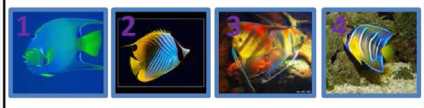
set to 64 to find out the best value of hyper-parameter α. The MAPs of our proposed MMACH method under different α are recorded and then depicted in Fig. 3. From this figure, it is obvious that our proposed MMACH method can achieve better performance when $\alpha = 0.6$. Therefore, in the subsequent experiments, we set $\alpha = 0.6$ for MMACH.

5.4.2. Validation of the effectiveness of multi-label modality enhanced attention

In this subsection, we examine the effectiveness of our proposed multi-label modality enhanced attention module. Concretely, we first remove the multi-label modality enhanced module in our proposed MMACH (i.e., we set $H^A = H^E$ in Fig. 2)

Dataset	Query Image	MMACH: Retrieval texts	MMACH-MSE: Retrieval texts
MIRFLICKR-25K		1. maldives fuvahmulah kulhi mangrove sunset sunrise atoll gnaviyani pond wetland land swim reflation nikon red sky blue millzero d300 boy 2. trees sunset naturesfinest 3. crane gru sunset hdr tramonto cielo sky ray raggi light luci chdk milano soe flickrsbest Damniwishidtakenthat 4. okmulgee oklahoma sunset red drippingspringslake tree water reflection blueribbonwinner abigfave explore	1. contraluz pandorga perico playa puestassol puntaumbria fab amazingcolors 2. ravenelle second life torley solo piano kenny bumby sweet mermaids romance moonlight craig altman animations dancing cats explore enjoy love youguys areverylucky 3. kelowna bc canada ubcokanagan 4. aldoaldoz fochi fuochi san giovanni firenze florence italia italy toscana tuscany ialia florencia
NUS-WIDE		1. fish angelfish 2. tropicalfish cichlid angelfish 3. fish yellow zoo angelfish 4. fish aquarium blue angelfish	1. fish angelfish tropicalfish denmarksaquarium 2. fish animal angelfish bermudaaquariumandzoo 3. philippines scuba diving underwater angelfish 4. ocean school sea fish water georgia aquarium scales angelfish striped
Microsoft COCO		1. a man on a horse in a flat pasture; a second horse behind him on the left; 2. three people are riding on brown horses in the foreground; three red houses with a brown thatched roof and lila flowers with green leaves behind it; a white sky in the background; 3. a dark and a light brown horse with red saddles are standing on a path in the foreground; high grass and a wooded hill behind it; 4. a group of people is riding on brown horses on a green meadow; grey clouds in the background;	1. people is riding on brown horses on a green meadow; grey clouds in the background; 2. a woman and other people are riding on horses on a grey, deep sandy trail through a forest with green trees 3. many people are riding on brown horses on a light brown dune in the shade; dark bushes behind them; a light blue sky in the background; 4. a cattle herd on a pasture with mainly white cows and two black ones
IAPRTC-12		1. a fountain and cobbled walkway in the foreground, a pink and white buidling with many arches in the background; trees on the right 2. a white building with lots of columns and arches, a neat lawn and neatly cut trees and bushes in the foreground; the flag of Paraguay is waving at the top of the building; there is a flower bed on the left; 3. a very modern building; stairs are leading up to the entrance; the walls are entirely made of glass; one red huge column is supporting the big roof; rails in the foreground; a green tree on the left; 4. Several flagpoles with waving flags on a green lawn in the foreground; a large grey and black building behind it; a huge column with a football on top on the left; a blue sky with white clouds in the background;	1. an inner courtyard with a fountain and flower pots in the centre; several arches surround the courtyard on two levels in front of the red building with a blue entrance; more flower pots below the arches 2. . a fountain and cobbled walkway in the foreground, a pink and white building with many arches in the background; trees on the right 3. a swimming pool in the foreground; behind it a bar with chairs and two people, and a bench with one person lying on it; upper level with doors and a blue rail 4. a large building on the left, a palm tree in centre of picture, (mostly) white cars in the street at a junction, some of them turning left, others going straight; there are red umbrellas in a park on the right; people are walking through the park, others are crossing the road in the foreground

(a) Image-to-text retrieval

Dataset	Query Text	MMACH: Retrieval images	MMACH-MSE: Retrieval images
MIRFLICKR-25K	maldives fuvahmulah kulhi mangrove sunset sunrise atoll gnaviyani pond wetland land swim reflation nikon red sky blue millzero d300 boy		
NUS-WIDE	. fish angelfish		
Microsoft COCO2014	a man on a horse in a flat pasture; a second horse behind him on the left;		
IAPRTC-12	. a fountain and cobbled walkway in the foreground, a pink and white buidling with many arches in the background; trees on the right		

(b) Text-to-image retrieval

Fig. 5. Examples of top 4 cross-modal retrieval results by MMACH and MMACH-MSE on four datasets. For (a) using images to retrieve texts, the matching texts are in blue. For (b) using texts to retrieve images, the purple number in each image is the ranking order, and the blue frames indicate the matching image.

and keep other parts unchanged, and we name this variation as MLSDCH. Afterward, we compare MLSDCH with MMACH on the four datasets MIRFLICKR-25K, NUS-WIDE, Microsoft COCO2014, and IAPRTC-12. The corresponding MAPs under the different hash code lengths of 16, 32, and 64 are shown in Table 2.

From the MAPs in Table 2, it demonstrates that in most cases, the MAPs of MMACH is higher than that of MLSDCH, showing that our proposed multi-label enhanced attention module can improve the performance of cross-modal hashing retrieval, which is partly because the multi-label modality enhanced attention module compensates for the sparse feature space. In addition,

Table 2
Performance of MMACH compared to MLSDCH in terms of MAPs on four datasets: MIRFLICKR-25K, NUS-WIDE, Microsoft COCO2014 and IAPRTC-12. The best MAP scores are shown in boldface.

Task	Method	MIRFlickr-25K			NUS-WIDE			MS COCO			IAPRTC-12		
		16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I→T	MLSDCH	0.8024	0.8186	0.8278	0.6330	0.6577	**0.6851**	0.6826	0.7182	0.7306	0.5218	0.5433	0.5730
	MMACH	**0.8085**	**0.8235**	**0.8348**	**0.6489**	**0.6679**	0.6847	**0.6989**	**0.7322**	**0.7540**	**0.5421**	**0.5752**	**0.6031**
T→I	MLSDCH	0.7796	0.8010	0.8115	0.6371	0.6613	0.6718	**0.6989**	0.7164	0.7280	0.4962	0.5297	0.5501
	MMACH	**0.7872**	**0.8011**	**0.8162**	**0.6450**	**0.6653**	**0.6758**	0.6913	**0.7245**	**0.7515**	**0.5316**	**0.5619**	**0.5866**

Table 3
Performance of MMACH compared to MMACH-MSE in terms of MAPs on four datasets: MIRFLICKR-25K, NUS-WIDE, Microsoft COCO2014 and IAPRTC-12. The best MAP scores are shown in boldface.

Task	Method	MIRFlickr-25K			NUS-WIDE			MS COCO			IAPRTC-12		
		16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I→T	MMACH-MSE	0.8006	0.8158	0.8282	0.6215	0.6533	0.6692	0.6912	0.7168	0.7364	0.5286	0.5450	0.5795
	MMACH	**0.8085**	**0.8235**	**0.8348**	**0.6489**	**0.6679**	**0.6847**	**0.6989**	**0.7322**	**0.7540**	**0.5421**	**0.5752**	**0.6031**
T→I	MMACH-MSE	0.7714	0.7952	0.8065	0.6362	0.6573	0.6698	0.6531	0.6882	0.6971	0.5026	0.5190	0.5485
	MMACH	**0.7872**	**0.8011**	**0.8162**	**0.6450**	**0.6653**	**0.6758**	**0.6913**	**0.7245**	**0.7515**	**0.5316**	**0.5619**	**0.5866**

Table 4
Comparison to baselines in terms of MAP on four datasets: MIRFLICKR-25K, NUS-WIDE, Microsoft COCO2014, IAPRTC-12, respectively. The best accuracy is shown in boldface.

Task	Method		MIRFlickr-25K			NUS-WIDE			MS COCO			IAPR TC-12		
			16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I→T	Hand-crafted methods	CMSSH [57]	0.5600	0.5709	0.5836	0.3092	0.3099	0.3396	0.5439	0.5450	0.5410	0.3049	0.3074	0.3130
		SePH [58]	0.6740	0.6813	0.6803	0.4797	0.4859	0.4906	0.4295	0.4353	0.4726	0.4186	0.4298	0.4315
		SCM [31]	0.6354	0.6407	0.6556	0.4626	0.4792	0.4886	0.4252	0.4344	0.4574	0.3887	0.3945	0.4068
		GSPH [20]	0.6068	0.6191	0.6230	0.4015	0.4151	0.4214	0.4427	0.4733	0.4840	0.3716	0.3921	0.4015
	Deep methods	DCMH [37]	0.7316	0.7343	0.7446	0.5445	0.5597	0.5803	0.5228	0.5438	0.5419	0.4536	0.4727	0.4919
		PRDH [36]	0.6952	0.7072	0.7108	0.5919	0.6059	0.6116	0.5238	0.5521	0.5572	0.4761	0.4883	0.4925
		CMHH [40]	0.7334	0.7281	0.7444	0.5530	0.5698	0.5559	0.5463	0.5676	0.5674	0.4903	0.5074	0.5152
		CHN [39]	0.7504	0.7495	0.7461	0.5754	0.5966	0.6015	0.5763	0.5822	0.5805	0.4962	0.5070	0.5241
		SSAH [38]	0.7745	0.7882	0.7990	0.6163	0.6278	0.6140	0.5127	0.5256	0.5067	0.5348	0.5619	0.5781
		MLSPH [56]	0.8076	0.8235	0.8337	0.6405	0.6604	0.6734	0.6557	0.7011	0.7271	0.5342	0.5721	0.5994
		MMACH	**0.8085**	**0.8235**	**0.8348**	**0.6489**	**0.6679**	**0.6847**	**0.6989**	**0.7322**	**0.7540**	**0.5421**	**0.5752**	**0.6031**
T→I	Hand-crafted methods	CMSSH [57]	0.5726	0.5776	0.5753	0.3167	0.3171	0.3179	0.3793	0.3876	0.3899	0.3189	0.3282	0.3229
		SePH [58]	0.7139	0.7258	0.7294	0.6072	0.6280	0.6291	0.4348	0.4606	0.5195	0.4667	0.4857	0.4936
		SCM [31]	0.6340	0.6458	0.6541	0.4261	0.4372	0.4478	0.4118	0.4183	0.4345	0.3824	0.3897	0.4002
		GSPH [20]	0.6282	0.6458	0.6503	0.4995	0.5233	0.5351	0.5435	0.6039	0.6461	0.4177	0.4452	0.4641
	Deep methods	DCMH [37]	0.7607	0.7737	0.7805	0.5793	0.5922	0.6014	0.4883	0.4942	0.5145	0.4851	0.4976	0.5171
		PRDH [36]	0.7626	0.7718	0.7755	0.6155	0.6286	0.6349	0.5122	0.5190	0.5404	0.5112	0.5283	0.5403
		CMHH [40]	0.7320	0.7183	0.7279	0.5739	0.5786	0.5639	0.4884	0.4554	0.4846	0.4790	0.4951	0.4963
		CHN [39]	0.7776	0.7775	0.7798	0.5816	0.5967	0.5992	0.5198	0.5320	0.5409	0.4994	0.5370	0.5397
		SSAH [38]	0.7860	0.7974	0.7910	0.6204	0.6251	0.6215	0.4832	0.4831	0.4922	0.5265	0.5594	0.5726
		MLSPH [56]	0.7852	**0.8041**	0.8146	0.6433	0.6633	0.6724	0.6494	0.6955	0.7193	0.5252	**0.5624**	**0.5938**
		MMACH	**0.7872**	0.8011	**0.8162**	**0.6450**	**0.6653**	**0.6758**	**0.6913**	**0.7245**	**0.7515**	**0.5316**	0.5619	0.5866

Fig. 4 presents the top 4 cross-modal retrieval results by MMACH and MLSDCH on four datasets, and it can be observed that in most cases, MMACH can retrieve more accurate candidates than MLSDCH.

5.4.3. Validation of the effectiveness of multi-label cross-modal triplet loss

In this part, we conduct experiments to verify the performance of our proposed multi-label cross-modal triplet loss. Specifically, we firstly utilize MSE (Mean Square Error) loss to replace our proposed multi-label cross-modal triplet loss in our proposed MMACH method and keep other parts fixed. We name this variation as MMACH-MSE. Subsequently, we compare MMACH with MMACH-MSE on the four datasets MIRFLICKR-25K, NUS-WIDE, Microsoft CO-CO2014, and IAPRTC-12. The corresponding MAPs under the distinct hash code lengths 16, 32, and 64 are shown in Table 3.

From Table 3, we can see that the MAPs of MMACH are always higher than that of MMACH-MSE. This demonstrates the effectiveness of our proposed multi-label cross-modal triplet loss, which is partly because multi-label cross-modal triplet loss can better preserve the multi-label semantic relevance compared to MSE loss. Furthermore, Fig. 5 lists the top 4 cross-modal retrieval results by MMACH and MMACH-MSE on four datasets. It can be observed that in most cases, MMACH can retrieve more accurate candidates than MMACH-MSE.

5.4.4. Comparison with state-of-the-art CMH methods

In this subsection, experiments are conducted further to investigate the performance of our proposed MMACH method. Specifically, we compare MMACH with several state-of-the-art cross-modal hashing methods in terms of MAP scores, precision–recall curves, and top$N$-precision curves on four datasets (i.e., MIRFLICKR-25K, NUS-WIDE, IAPRTC-12, and Microsoft COCO2014).

The MAPs of MMACH and baseline methods under distinct hash code lengths 16, 32, and 64 are listed in Table 4. Based on the experimental results, we have the following findings:

(1) Compared to both hand-crafted baseline methods and deep neural networks-based baseline methods, our proposed MMACH method can achieve higher MAP values in most cases.

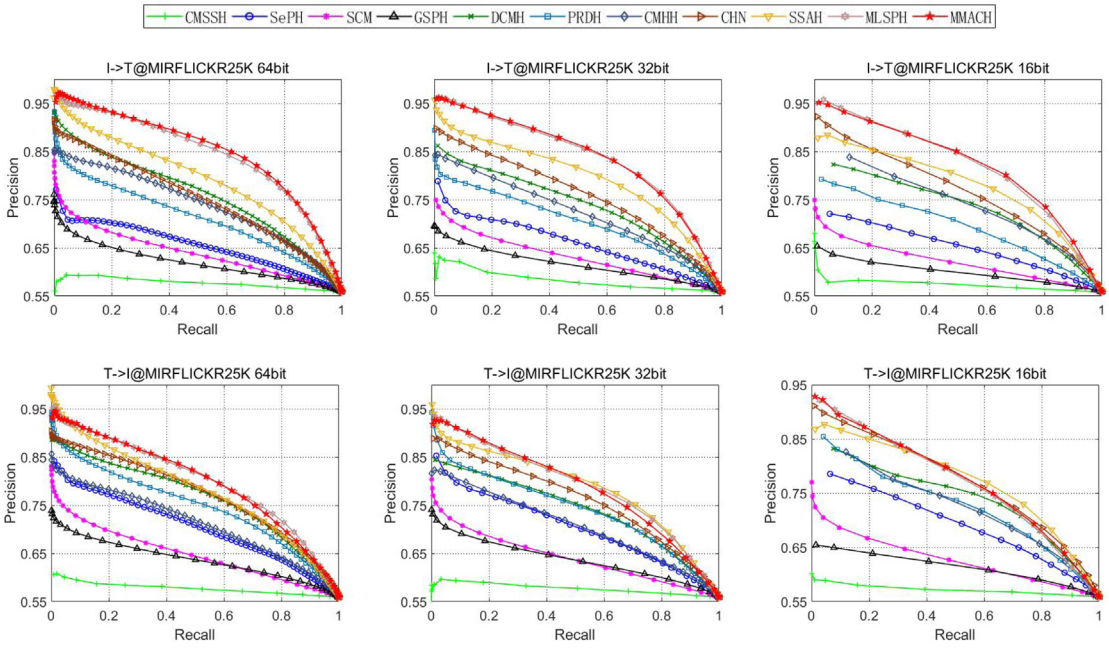


Fig. 6. Precision-Recall Curves on MIRFLICKR-25K.

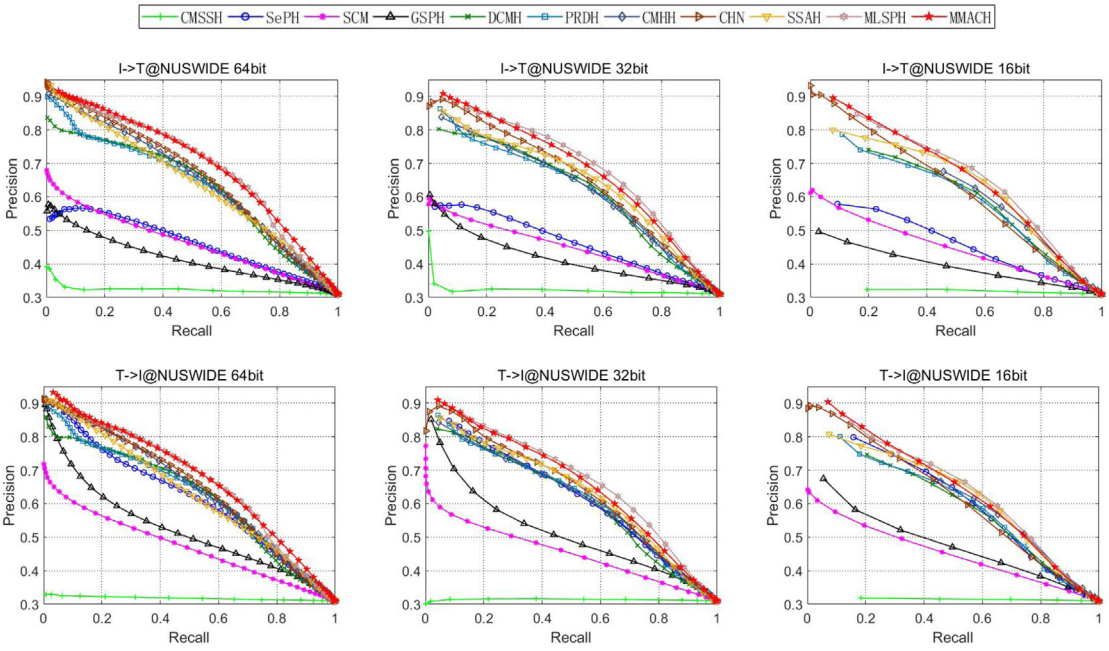


Fig. 7. Precision-Recall Curves on NUS-WIDE.

This demonstrates that MMACH can utilize the multi-label modality enhanced attention module, multi-label cross-modal triplet loss, and self-supervised learning strategy to enhance the performance of deep cross-modal hashing retrieval.

(2) Among the hand-crafted baseline methods, SePH has the highest MAP values in most cases, which is partly because SePH utilizes kernel logistic regression to learn hash projection functions for each modality. Among deep neural network-based baseline methods, MLSPH has the highest MAP values in most cases, partly because MLSPH introduces a multi-label semantic preserving module and can compute the semantic relevance of original data more precisely.

(3) Compared to hand-crafted methods, deep neural network-based methods usually achieve higher MAP values, partly because deep neural network-based methods make full use of the excellent features learning capability of these deep neural networks.

(4) Both SSAH and MMACH leverage self-supervised learning to supervise the training of hash projection functions for all modalities. However, MMACH outperforms SSAH in all cases, partly because MMACH defines a multi-label modality enhanced attention module to compensate for the sparsity of multi-label features. Moreover, MMACH utilizes multi-label cross-modal triplet loss to select multi-label semantic similar triplets. Meanwhile, SSAH regards the semantic similarity of two instances as 1, if there is at least one common category, neglecting the differences between multi-labels.

To further compare MMACH with the baseline CMH methods, we compare the precision–recall curves of MMACH and all

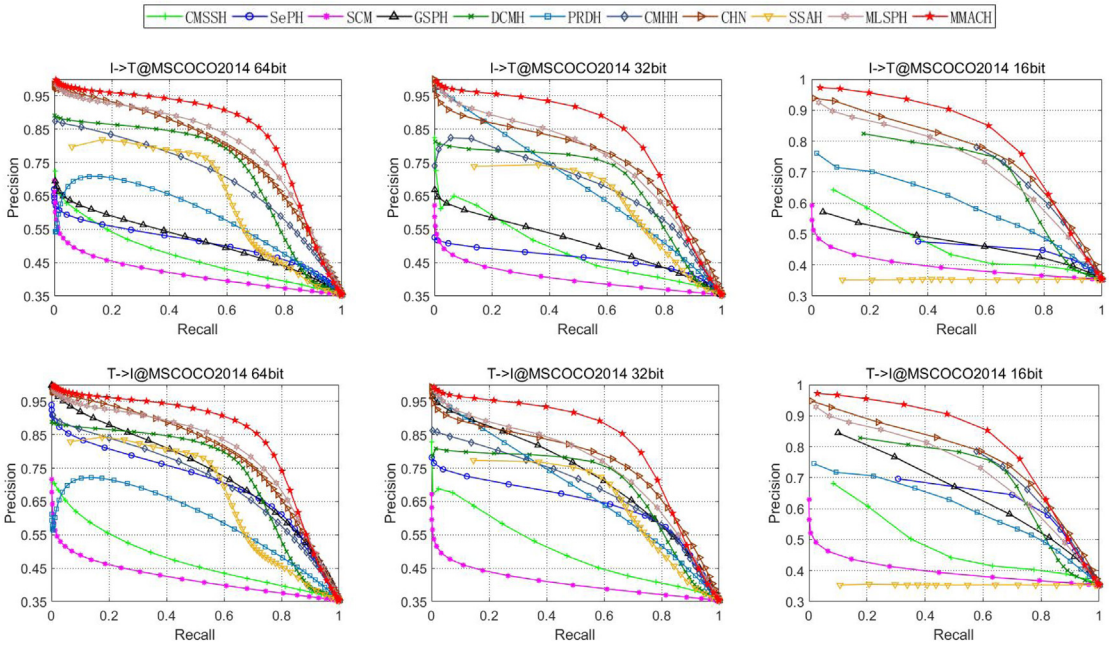


Fig. 8. Precision-Recall Curves on Microsoft COCO2014.

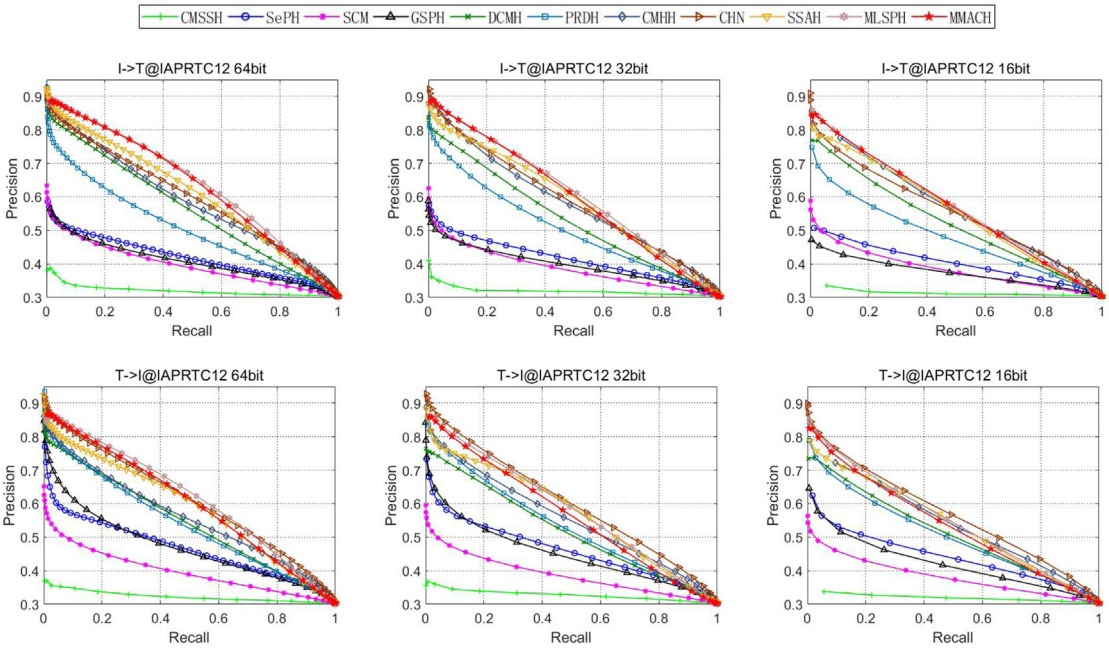


Fig. 9. Precision-Recall Curves on IAPRTC-12.

baseline methods on four experimental datasets with different hash codes length. Figs. 6–9 are the precision–recall curves of all methods with different datasets and hash code length. From these figures, we have the following observations:

(1) In most cases, the precision–recall curves of our proposed MMACH method are higher than that of most baseline methods. This demonstrates that MMACH can achieve better cross-modal retrieval performance than most baseline methods.

(2) The precision–recall curves of all methods are approximately identical to the corresponding observations on the MAP scores.

(3) In some cases, the precision–recall curves of MLSPH are higher than that of MMACH. In contrast, in other cases, the precision–recall curves of MMACH are higher than that of MLSPH. This is partly because MLSPH and MMACH both consider multi-label semantic similarity. Meanwhile, MLSPH utilizes a ResNet to extract the features of images. At the same time, MMACH defines a multi-labels modality enhanced attention module to supervise better the learning of hash projection functions with a self-supervised style.

Moreover, top$N$-precision curves of MMACH and baseline methods on datasets MIRFLICKR-25K, NUS-WIDE, Microsoft COCO2014 and IAPRTC-12 with hash codes lengths of 16, 32, and 64 are depicted in Figs. 10–13. From these results, we can draw the following conclusions:

(1) In most cases, the top-N precision curves of our proposed MMACH method are higher than that of most baseline methods.

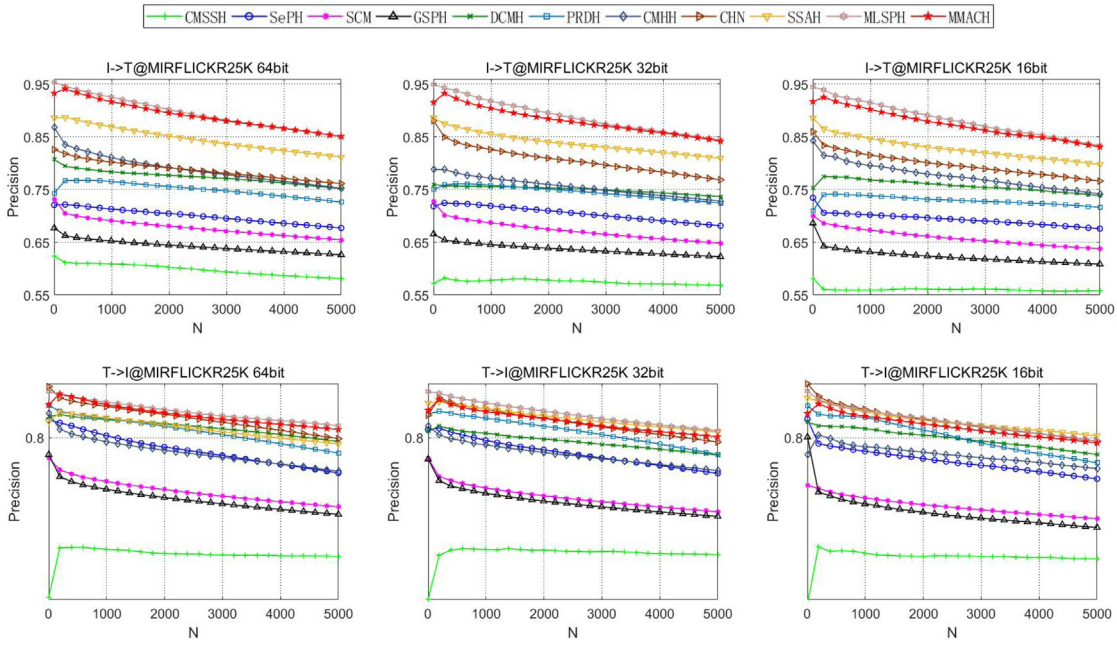


Fig. 10. top*N*-precision curves on MIRFLICKR-25K.

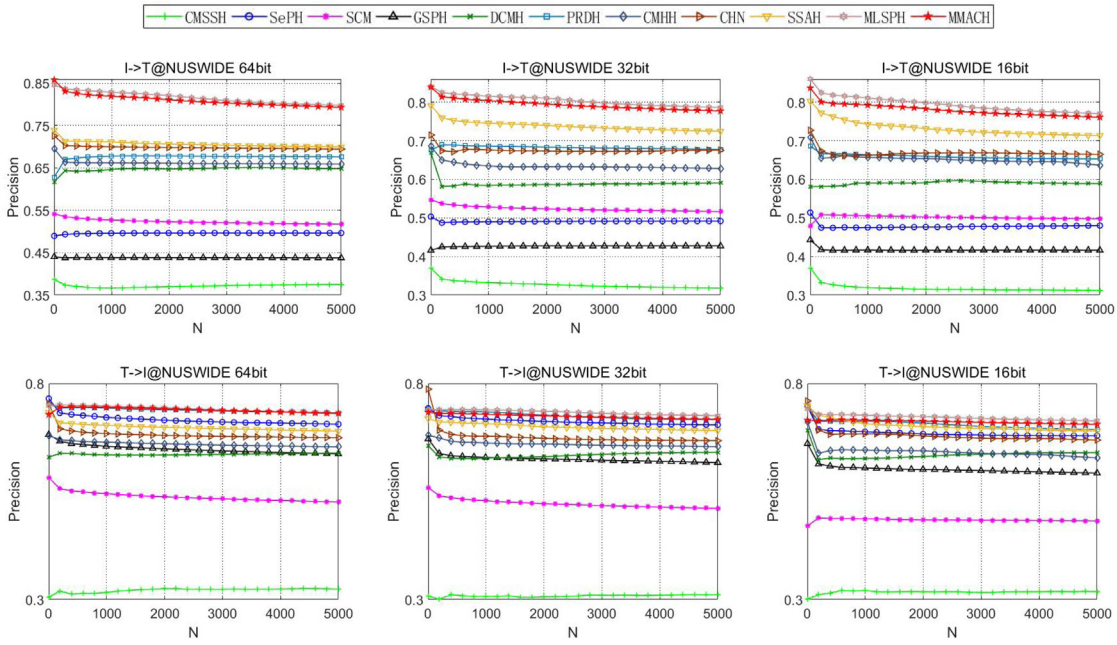


Fig. 11. top*N*-precision curves on NUS-WIDE.

It shows that MMACH outperforms most baseline methods on cross-modal retrieval.

(2) In all cases, MMACH achieves competitive performance with MLSPH, partly because these two methods are both multi-labels semantic protection-based methods.

(3) The top-*N* precision curves of all methods are nearly consistent with the observed MAP values and precision–recall curves.

5.5. Heatmap visualization of the image modality

To verify the robustness of features extracted by the deep convolutional neural networks, we utilize the GRAD-CAM [61] to visualize the heatmaps of input images for our proposed MMACH as well as DCMH and SSAH on datasets IAPRTC-12 and MIRFLICKR-25K. Figs. 14 and 15 show the corresponding heatmaps. From these heatmaps, it is obvious that in most cases, our proposed MMACH can more accurately correlate the corresponding semantic categories compared to DCMH and SSAH, which demonstrates the powerful multi-label semantic preserving capability of our proposed MMACH.

5.6. Running time analysis

We further evaluate the running time of our proposed MMACH method. Specifically, we record the running time of both MMACH and three representative baseline methods (DCMH [37], PRDH

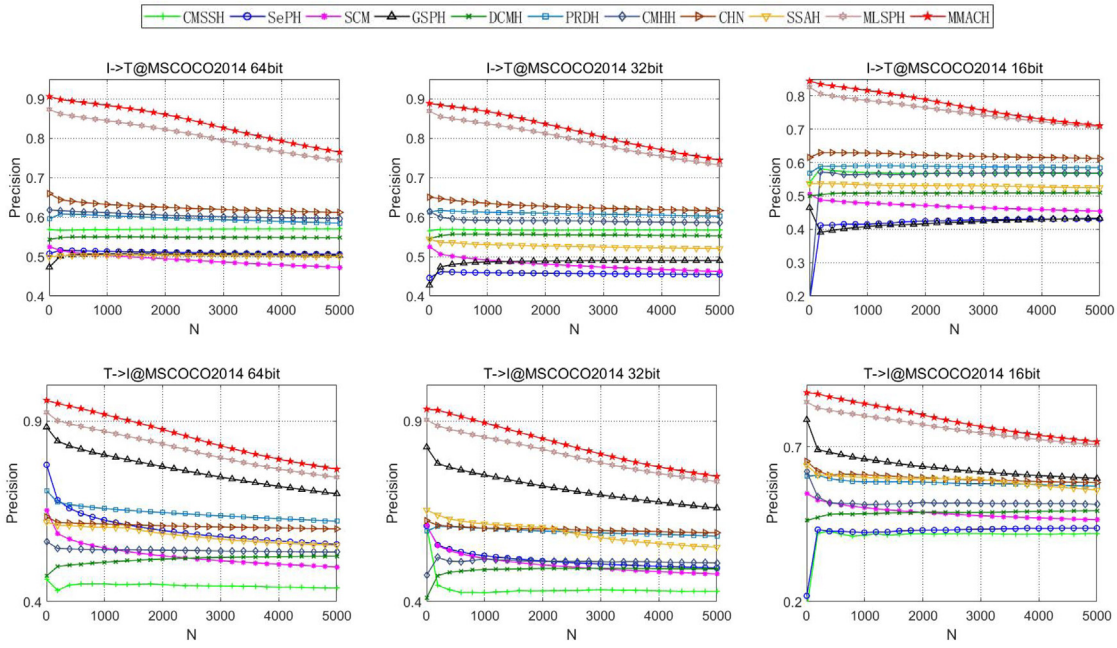


Fig. 12. top*N*-precision curves on Microsoft COCO2014.

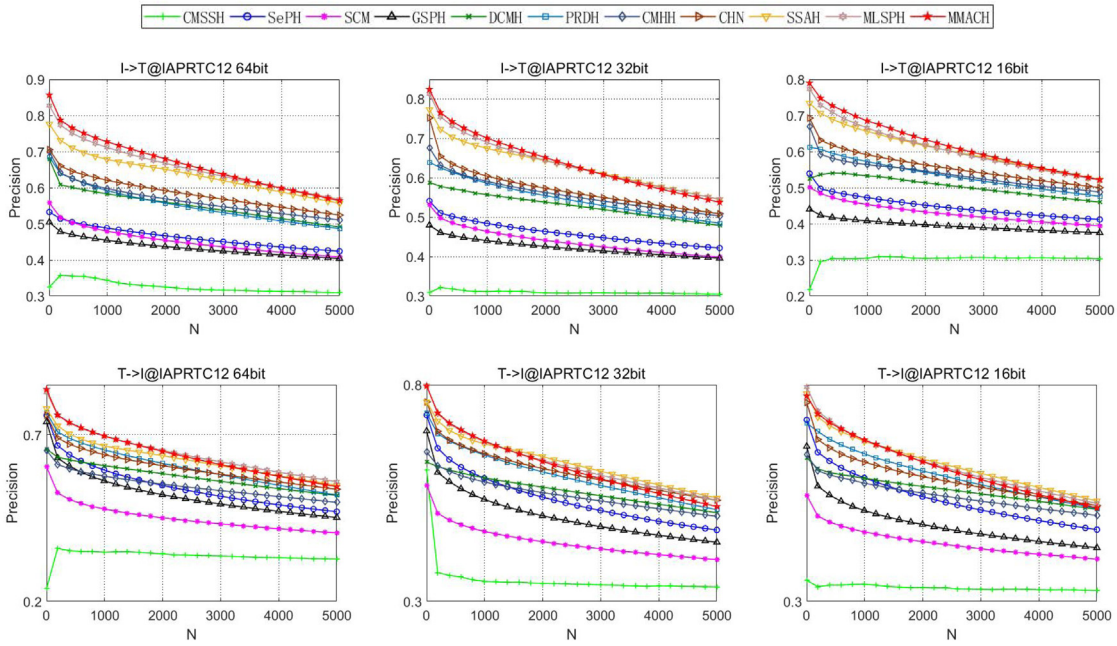


Fig. 13. top*N*-precision curves on IAPRTC-12.

[36], SSAH [38]). These methods are executed on an NVIDIA GTX Titan XP GPU server with the maximum epoch $max\_epoch = 200$, and the experimental dataset is MIRFLICKR-25K, and the length of hash code is 64. The results are presented in Table 5. In Table 5, it can be observed that the running time of MMACH is higher than that of DCMH and PRDH, which is partly because that, compared with PRDH and DCMH, MMACH utilizes a self-supervised learning style, which introduces a deep neural network for the label modality. Meanwhile, the running time of MMACH is lower than that of SSAH, which is partly because that SSAH further introduces generative adversarial networks, which needs more running time.

Table 5
Comparison of running time to some baseline methods.

	Running time
DCMH	28 h
PRDH	31 h
SSAH	42 h
MMACH	38 h

6. Conclusion

This paper introduces a prominent cross-modal hashing method termed multi-label modality enhanced attention-based

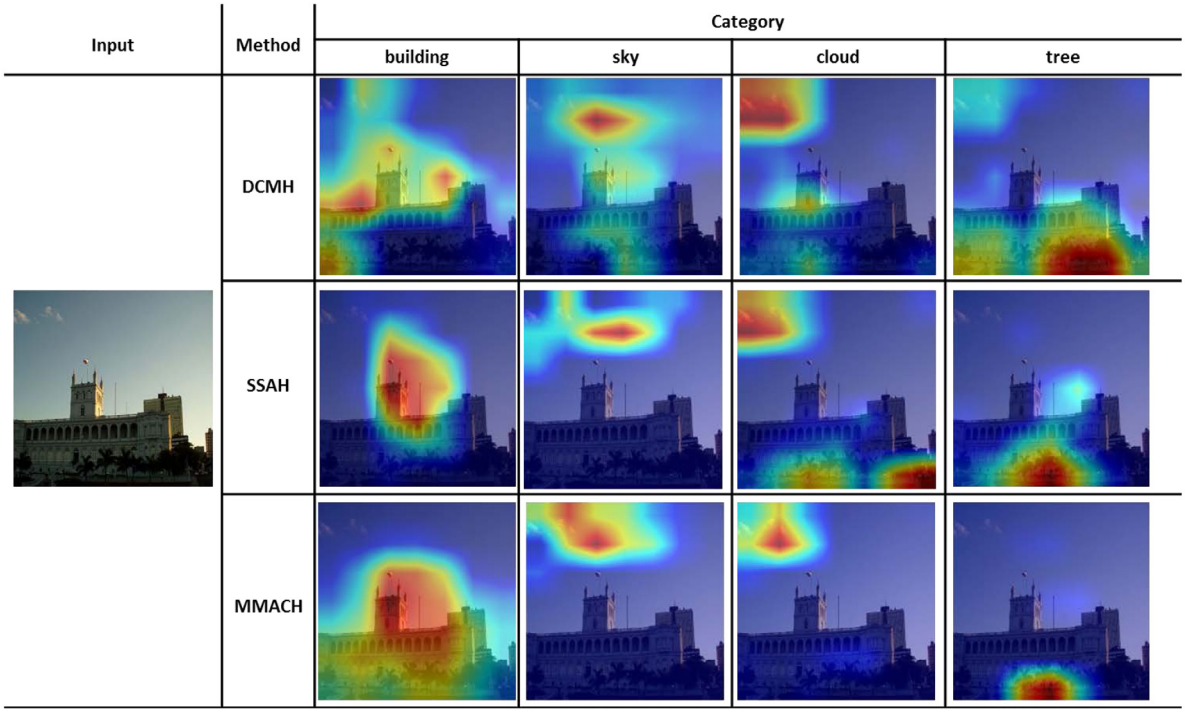


Fig. 14. Grad-CAM visualization of MMACH compared to SSAH and DCMH for a randomly selected image from multi-label dataset IAPRTC-12 with respect to different ground-truth categories.

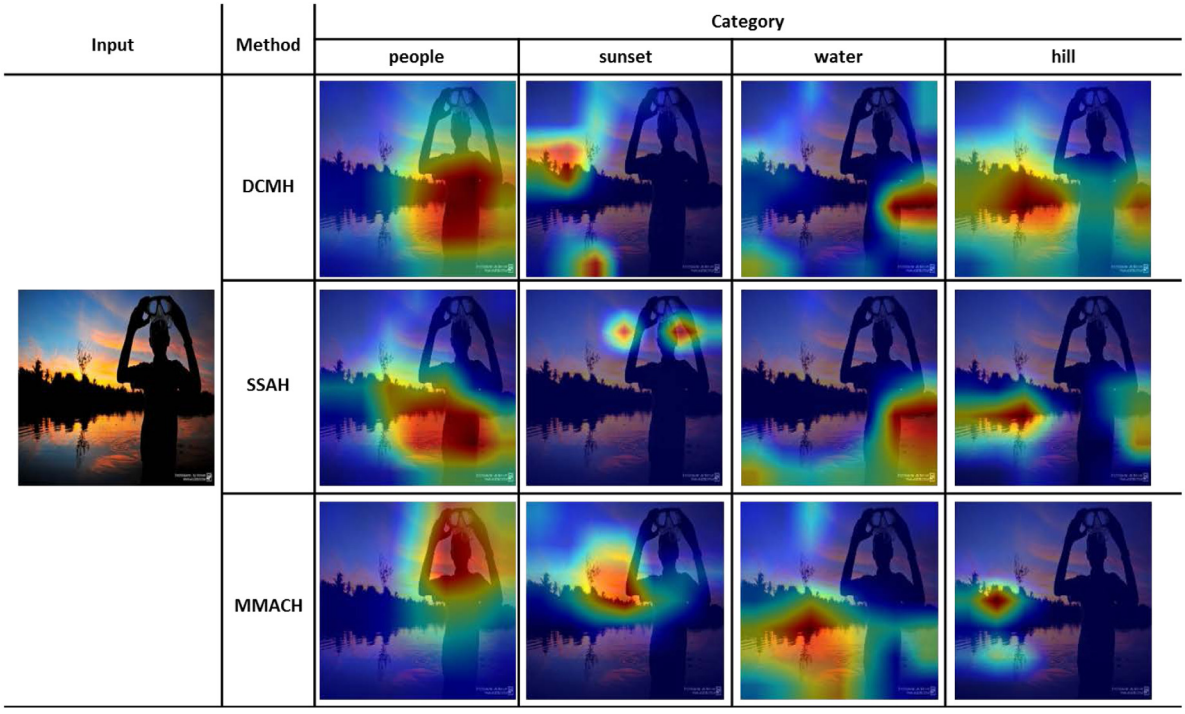


Fig. 15. Grad-CAM visualization of MMACH compared to SSAH and DCMH for a randomly selected image from multi-label dataset MIRFLICKR-25K with respect to different ground-truth categories.

self-supervised deep cross-modal hashing (MMACH). A novel multi-label modality enhanced attention (MMEA) module is designed in MMACH to compensate for the sparse feature representations of multi-labels from multi-modal instances. Based on these enhanced multi-labels, self-supervised learning is introduced to supervise the training of hash functions of other modalities. Furthermore, a multi-label cross-modal triplet loss (MCTL) is defined in MMACH to ensure that the feature representations of cross-modal pairwise instances with more common categories should preserve higher semantic similarity than other instances. Extensive experiments on several well-known cross-modal benchmark datasets indicated the effectiveness of the

proposed MMEA and MCTL. Meanwhile, the MMACH method surpasses the performance of the baseline methods and acquires competitive cross-modal retrieval performance.

CRediT authorship contribution statement

Xitao Zou: Conceptualization, Methodology, Experiments, Writing – original draft. **Song Wu:** Conceptualization, Revise manuscript. **Nian Zhang:** Revise manuscript. **Erwin M. Bakker:** Revise manuscript.

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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61806168), Fundamental Research Funds for the Central Universities, China (SWU117059), Venture & Innovation Support Program for Chongqing Overseas Returnees, China (CX2018075), National Science Foundation (NSF), USA grant #2011927 and DoD, USA grant #W911NF1810475.

References

[1] Yuxin Peng, Xin Huang, Yunzhen Zhao, An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2017) 2372–2385.
[2] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, Liang Wang, A comprehensive survey on cross-modal retrieval, 2016, arXiv preprint arXiv:1607.06215.
[3] Yangqing Jia, Mathieu Salzmann, Trevor Darrell, Learning cross-modality similarity for multinomial data, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2407–2414.
[4] Yin Zheng, Yu-Jin Zhang, Hugo Larochelle, Topic modeling of multimodal data: an autoregressive approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1370–1377.
[5] Yanfei Wang, Fei Wu, Jun Song, Xi Li, Yueting Zhuang, Multi-modal mutual topic reinforce modeling for cross-media retrieval, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 307–316.
[6] Abhishek Sharma, Abhishek Kumar, Hal Daume, David W Jacobs, Generalized multiview analysis: A discriminative latent space, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2160–2167.
[7] Xiao-Yuan Jing, Rui-Min Hu, Yang-Ping Zhu, Shan-Shan Wu, Chao Liang, Jing-Yu Yang, Intra-view and inter-view supervised correlation analysis for multi-view feature learning, in: Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.
[8] Xiangbo Mao, Binbin Lin, Deng Cai, Xiaofei He, Jian Pei, Parallel field alignment for cross media retrieval, in: Proceedings of the 21st ACM International Conference on Multimedia, ACM, 2013, pp. 897–906.
[9] Yue Ting Zhuang, Yan Fei Wang, Fei Wu, Yin Zhang, Wei Ming Lu, Supervised coupled dictionary learning with group structures for multi-modal retrieval, in: Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013.
[10] Yunchao Gong, Qifa Ke, Michael Isard, Svetlana Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, Int. J. Comput. Vis. 106 (2) (2014) 210–233.
[11] Kaiye Wang, Ran He, Liang Wang, Wei Wang, Tieniu Tan, Joint feature selection and subspace learning for cross-modal retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 38 (10) (2015) 2010–2023.
[12] Jian Wang, Yonghao He, Cuicui Kang, Shiming Xiang, Chunhong Pan, Image-text cross-modal retrieval via modality-specific feature learning, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM, 2015, pp. 347–354.
[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, Tomas Mikolov, Devise: A deep visual-semantic embedding model, in: Advances in Neural Information Processing Systems, 2013, pp. 2121–2129.
[14] Andrej Karpathy, Armand Joulin, Li F. Fei-Fei, Deep fragment embeddings for bidirectional image sentence mapping, in: Advances in Neural Information Processing Systems, 2014, pp. 1889–1897.
[15] Xinyang Jiang, Fei Wu, Xi Li, Zhou Zhao, Weiming Lu, Siliang Tang, Yueting Zhuang, Deep compositional cross-modal learning to rank via local-global alignment, in: Proceedings of the 23rd ACM International Conference on Multimedia, ACM, 2015, pp. 69–78.
[16] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, Shuicheng Yan, Cross-modal retrieval with CNN visual features: A new baseline, IEEE Trans. Cybern. 47 (2) (2016) 449–460.
[17] Yuxin Peng, Jinwei Qi, CM-GANs: cross-modal generative adversarial networks for common representation learning, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 15 (1) (2019) 22.
[18] Fangming Zhong, Zhikui Chen, Geyong Min, Deep discrete cross-modal hashing for cross-media retrieval, Pattern Recognit. 83 (2018) 64–77.
[19] Lin Wu, Yang Wang, Ling Shao, Cycle-consistent deep generative hashing for cross-modal retrieval, IEEE Trans. Image Process. 28 (4) (2018) 1602–1612.
[20] Devraj Mandal, Kunal N. Chaudhury, Soma Biswas, Generalized semantic preserving hashing for cross-modal retrieval, IEEE Trans. Image Process. 28 (1) (2018) 102–112.
[21] 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.
[22] Yixian Fang, Bin Li, Xiaozhou Li, Yuwei Ren, Unsupervised cross-modal similarity via latent structure discrete hashing factorization, Knowl.-Based Syst. 218 (2021) 106857.
[23] Jian Zhang, Yuxin Peng, Mingkuan Yuan, Unsupervised generative adversarial cross-modal hashing, pp. Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[24] Shupeng Su, Zhisheng Zhong, Chao Zhang, Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3027–3035.
[25] Chao Li, Cheng Deng, Lei Wang, De Xie, Xianglong Liu, Coupled cyclegan: Unsupervised hashing network for cross-modal retrieval, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 33, No. 01, 2019, pp. 176–183.
[26] 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: IJCAI, pp. 2854–2860.
[27] Yixian Fang, Huaxiang Zhang, Yuwei Ren, Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing, Knowl.-Based Syst. 171 (2019) 69–80.
[28] Min Meng, Haitao Wang, Jun Yu, Hui Chen, Jigang Wu, Asymmetric supervised consistent and specific hashing for cross-modal retrieval, IEEE Trans. Image Process. 30 (2020) 986–1000.
[29] Haopeng Qiang, Yuan Wan, Ziyi Liu, Lun Xiang, Xiaojing Meng, Discriminative deep asymmetric supervised hashing for cross-modal retrieval, Knowl.-Based Syst. 204 (2020) 106188.
[30] Liangli Zhen, Peng Hu, Xu Wang, Dezhong Peng, Deep supervised cross-modal retrieval, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[31] Dongqing Zhang, Wu-Jun Li, Large-scale supervised multimodal hashing with semantic correlation maximization, in: Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.
[32] Xiaofang Wang, Yi Shi, Kris M. Kitani, Deep supervised hashing with triplet labels, in: Asian Conference on Computer Vision, Springer, 2016, pp. 70–84.
[33] Zhan Yang, Liu Yang, Osolo Ian Raymond, Lei Zhu, Wenti Huang, Zhifang Liao, Jun Long, NSDH: A nonlinear supervised discrete hashing framework for large-scale cross-modal retrieval, Knowl.-Based Syst. 217 (2021) 106818.
[34] Song Wang, Huan Zhao, Kei Nai, Learning a maximized shared latent factor for cross-modal hashing, Knowl.-Based Syst. 228 (2021) 107252.
[35] Fengling Li, Tong Wang, Lei Zhu, Zheng Zhang, Xinhua Wang, Task-adaptive asymmetric deep cross-modal hashing, Knowl.-Based Syst. 219 (2021) 106851.
[36] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, Xinbo Gao, Pairwise relationship guided deep hashing for cross-modal retrieval, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[37] Qing-Yuan Jiang, Wu-Jun Li, Deep cross-modal hashing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3232–3240.
[38] 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.
[39] Yue Cao, Mingsheng Long, Jianmin Wang, Philip S. Yu, Correlation hashing network for efficient cross-modal retrieval, 2016, arXiv preprint arXiv:1602.06697.
[40] Yue Cao, Bin Liu, Mingsheng Long, Jianmin Wang, Cross-modal Hamming hashing, in: Proceedings of the European Conference on Computer Vision ECCV, 2018, pp. 202–218.

[41] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio, Neural machine translation by jointly learning to align and translate, 2014, arXiv preprint arXiv:1409.0473.
[42] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.
[43] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, Eduard Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.
[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[45] Xi Zhang, Hanjiang Lai, Jiashi Feng, Attention-aware deep adversarial hashing for cross-modal retrieval, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 591–606.
[46] Min-Ling Zhang, Zhi-Hua Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2013) 1819–1837.
[47] Mohammad S. Sorower, A Literature Survey on Algorithms for Multi-Label Learning, Vol. 18, Oregon State University, Corvallis, 2010, pp. 1–25.
[48] Henry Gouk, Bernhard Pfahringer, Michael Cree, Learning distance metrics for multi-label classification, in: Asian Conference on Machine Learning, PMLR, 2016, pp. 318–333.
[49] Mark J. Huiskes, Michael S. Lew, The MIR flickr retrieval evaluation, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, ACM, 2008, pp. 39–43.
[50] 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, ACM, 2009, p. 48.
[51] Weiwei Liu, Ivor W. Tsang, Large margin metric learning for multi-label prediction, in: AAAI, 2015.
[52] Yi Zhang, Jeff Schneider, Maximum margin output coding, Comput. Sci. (2012) 1575–1582.
[53] Yashaswi Verma, C.V. Jawahar, Image annotation by propagating labels from semantic neighbourhoods, Int. J. Comput. Vis. 121 (1) (2017) 126–148.
[54] 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.
[55] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villaseñor, Michael Grubinger, The segmented and annotated IAPR TC-12 benchmark, Comput. Vis. Image Underst. 114 (4) (2010) 419–428.
[56] Xitao Zou, Xinzhi Wang, Erwin M. Bakker, Song Wu, Multi-label semantics preserving based deep cross-modal hashing, Signal Process., Image Commun. 93 (2021) 116131.
[57] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, Nikos Paragios, Data fusion through cross-modality metric learning using similarity-sensitive hashing, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 3594–3601.
[58] 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.
[59] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al., Universal sentence encoder, 2018, arXiv preprint arXiv:1803.11175.
[60] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, pp. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[61] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.