Knowledge-Based Systems (2021)

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: https://www.journals.elsevier.com/knowledge-based-systems

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:

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

ABSTRACT

By integrating deep neural networks and hashing into cross-modal retrieval, deep cross-modal hashing (DCMHs) has achieved superior performance and thus has drew widespread attention. Nevertheless, there still remains two problems for existing DCMHs: (1) most existing DCMHs methods simply leverage single labels to compute the semantic similarity of cross-modal pairwise instances which neglects that many cross-modal datasets contain abundant semantic information with multi-labels. (2) a few DCMHs methods have utilized the multi-labels to supervise the learning of hash functions for DCMHs. Nevertheless, the feature space of multi-labels is too sparse to supervise the learning of DCMHs, which may lead to a suboptimal performance for DCMHs. To make full use of the multi-labels in cross-modal datasets and enhance the performance of DCMHs, we propose a multi-label modality enhanced attention based self-supervised deep cross-modal hashing (MMACH) method. Specifically, MMACH defines a multi-label modality enhanced attention module, which utilizes an attention mechanism to compensate the sparse feature vectors of multi-labels from multi-modal instances. MMACH also defines a multi-label cross-modal triplet loss to make sure that cross-modal instances with more common categories have more similar hash representations and vice versa. Afterwards, MMACH uses the enhanced multi-labels to supervise the learning of hash functions of other modalities with a self-supervised learning scheme, during which the defined multi-label cross-modal triplet loss is used to preserve the multi-label semantic relevance of cross-modal instances. Extensive experiments on four multi-label cross-modal datasets demonstrate that our MMACH can achieve prominent performance and outperform several baseline methods.

1. Introduction

With the advent and prevalence of Internet, more and more multi-modal data, such as graphics, texts, videos, images and so on, have been accumulated in social network. As data from distinct modalities may represent an identical object or event, it is plausible to bridge semantically relevant but modality differ-

*Corresponding author:
e-mail: songwuswu@swu.edu.cn (Song Wu)

ent data to implement massive cross-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 thus have dissimilar feature space, how to efficiently and effectively minimize the semantic gaps between these large-scale yet heterogeneous data and accurately calculate semantical similarity of cross-modal data still remains an intractable problem for cross-modal retrieval.

Generally, a large number of existing cross-modal retrieval methods, including topic models [3, 4, 5], subspace learning [6, 7, 8, 9, 10, 11], and deep models [12, 13, 14, 15, 16, 17, 18, 19, 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 scale of the multi-modal data, real-valued based cross-modal retrieval methods usually suffer from high computation cost and low retrieval accuracy. To solve this issue, hashing based cross-modal retrieval (also called cross-modal hashing (CMH)) methods would map high-dimensional original data from each modality into compact binary codes and calculate the semantic relevance of cross-modal pairwise instances with a fast XOR operation. As a result, it would meet low data storage requirement and ensure efficient similarity measurement and thus become a prevalent research topic in recent years.

Depending on whether category labels are leveraged during training stage, existing cross-modal hashing methods can be further divided into unsupervised methods [21, 22, 23, 24, 25, 26, 27] and supervised methods [28, 29, 30, 31, 32, 33, 34, 35]. Unsupervised cross-modal hashing methods transform the original cross-modal data to homogeneous binary codes by calculating the similarities of multi-modal data representations to protect their semantic relevance. By contrast, supervised cross-modal hashing methods encode the heterogeneous cross-modal instances into compact hash codes and preserve the cross-modal semantic similarities with the help of class labels. Compared to unsupervised ones, supervised cross-modal hashing methods can make full use of semantic relation of cross-modal instances by utilizing semantic labels and thus achieve remarkable boost of performance.

In the past few years, deep neural networks (DNN) are proposed and applied to many tasks such as sentence recognition, object detection, image caption and so on. Without exception, deep neural networks based cross-modal hashing are widely investigated. Due to the remarkable feature learning ability, deep cross-modal hashing method can capture the correlation across different modalities more effectively than hand-crafted methods.

In most of 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 instances, then the semantic similarity of the former should be higher than the semantic similarity of the latter. Therefore, many existing deep cross-modal hashing methods overlook the abundant semantic information in multiple-labels of practical cross-modal datasets and thus cannot accurately evaluate the semantic relevance of cross-modal pairwise instances, leading to the learned cross-modal hash projection functions suffered from suboptimal performance. To solve this problem, a few deep cross-modal hashing methods introduce self-supervised learning into deep cross-modal hashing and regard the multi-labels of original instances as a signal modality and learn a hash mapping 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 severely sparse, the multi-label based self-supervised learning strategy can make limited enhancement of performance of the learned cross-modal hash projection functions.

To further boost the performance of cross-modal hashing, we propose a multi-label modal enhanced attention based self-supervised deep cross-modal hashing (MMACH). Specifically, MMACH firstly defines a multi-label modal enhanced attention module (MMEA) to enrich the multi-label matrix to overcome the sparsity of multi-labels in self-supervised learning based deep cross-modal hashing, which utilizes three encoders to encode each original instance (including original image feature, original text feature and corresponding multi-label) into a common real-valued space, and then normalizes these real-valued features. Afterwards, the normalized real-valued features from the original image and the normalized real-valued features from the original text are fused into the normalized real-valued features by using a attention mechanism. Secondly, MMACH defines a multi-label cross-modal triplet loss (MCTL) to better evaluate the semantic similarity of multi-label cross-modal instances, i.e., MCTL constructs cross-modal triplets to keep the similarity of features of cross-modal pairwise instances with more common categories higher than the similarity of features of cross-modal pairwise instances with less common categories. Thirdly, MMACH introduces the multi-label modal enhanced attention module and the multi-label cross-modal triplet loss into self-supervised learning based deep cross-modal hashing to enhance the cross-modal retrieval. Fourthly, experiments on four standard cross-modal datasets are conducted and the experimental results demonstrate the enhancement of the performance of our proposed MMACH. The main contributions of our work include four-folds:

1. We propose a multi-label modal enhanced attention module. To solve the sparsity of multi-labels in self-supervised learning based deep cross-modal hashing, three encoders are pre-trained and transferred to map the original image-text pairs as well as their multi-labels into a common real-valued feature space. After normalization , the features of the original image and the features of the original text are fused into the features of the original multi-labels to enrich the sparse multi-labels.

2. We propose a multi-label cross-modal triplet loss. In multi-label learning, suppose that we have a triplet instance (a, b, c) and each instance has multi-labels, if instance a and instance b have more common categories than instance a and instance c, then a and b is more semantically relevant than a

and c, thus, the learned feature of a and the learned feature of b is more similar than the learned feature of a and the learned feature of c. Inspired by this, we define a multi-label cross-modal triplet loss, i.e,. if two cross-modal instances have more common categories than the other two cross-modal instances, then the similarity of the learned features of the former should be higher than the similarity of the learned features of the latter. To the best of our knowledge, this is the first time to introduce multi-label triplet loss into cross-modal retrieval.

3. We cooperate the multi-label modal enhanced attention module and the multi-label cross-modal triplet loss into cross-modal hashing and propose a multi-label modal enhanced attention based self-supervised deep cross-modal hashing (MMACH) method. MMACH leverages the multi-label modal enhanced attention module to generate an enhanced multi-label modality, with which a hash projection function is learned and used to supervise the training of hash mapping functions of other modalities. To preserve the semantic relevance of cross-modal instances unchanged during the hash representation stage, the multi-label cross-modal triplet loss is introduced.

4. We conduct extensive experiments to validate the efficiency of our proposed MMACH. Four well-known cross-modal benchmark datasets are utilized to conduct experiments to verify the effectiveness of MMACH and compare MMACH with several state-of-the-art cross-modal hashing methods.

The rest of the paper is organized as follows. Section 2 depicts the related work. Section 3 presents our multi-label modal enhanced attention based self-supervised deep cross-modal hashing (MMACH) method. Section 4 is the learning details of our MMACH. Section 5 shows the experiments on several datasets to validate the performance of MMACH. Section 6 concludes the paper.

2. Related Work

2.1. Deep Cross-Modal Hashing

Among previous cross-modal hashing methods, shallow architecture based methods firstly extract hand-crafted features and then utilize these hand-crafted features to learn hash functions, which is a two-stage architecture and might not be optimally compatible with each other and may result in suboptimal performance. By contrast, deep cross-modal hashing methods make full use of the significant feature extraction capabilities of deep neural networks and thus can better explore the correlations across different modalities with an end-to-end style, therefore, deep cross-modal hashing retrieval has attracted increasing attention. Representative methods are deep cross-modal hashing (DCMH) [36], pairwise relation guided deep hashing (PRDH) [37], correlation hashing network (CHN) [38], cross-modal hamming hashing (CMHH) [39], and self-supervised adversarial hashing (SSAH) [40]. DCMH [36], which effectively projects image-text pairs into corresponding hash codes by using an end-to-end deep neural network framework. PRDH [37] exploits intra-modal and inter-modal constraints of different pairwise instances to generate accurate hash codes with a united deep learning framework. CHN [38] defines a cosine max-margin loss to enhance the quality of the learned hash codes. CMHH [39] uses an exponential focal loss to penalize significantly on similar cross-modal pairs with Hamming distances larger than the Hamming radius threshold. SSAH [40] introduces self-supervised learning into cross-modal hashing and trains a hash function (named LabelNet) on the multi-label modality to supervise the learning of other modalities. Nevertheless, these methods either leverage single labels to calculate the semantic similarity of cross-modal pairwise instances or simply regard the semantic similarity of cross-modal pairwise instances with multiple labels as 1 if these two cross-modal instances have at least one common categories, which overlooks the fact that many practical cross-modal datasets have multiple labels and there are abundant semantic information in multi-labels. Namely, if two cross-modal instances have more common categories than another cross-modal pairwise instances, then the semantic similarity of the former pairwise is higher than the semantic similarity of the latter pairwise. Moreover, the existing self-supervised based deep cross-modal hashing methods often suffer from inferior performance because hash function learned on the sparse multi-labels has a weak supervision capacity to train the hash functions of other modalities.

2.2. Attention Mechanism

Attention mechanism [41, 42, 43, 44] is firstly introduced and widely applied in natural language process which tries to consider the neighboring words when extracting features from one word. Subsequently, attention mechanism is introduced into the field of computer vision which is trained to identify where the model should concentrate on when performing a special task. To date, only a few methods combine cross-modal hashing retrieval with attention mechanism. Attention-aware deep adversarial hashing (DAH) [45] introduces attention mechanism into cross-modal hashing and generates the adaptive attention masks to divide the feature representations into the attended and the unattended feature representations. Different from this, we fuse the features of the image modality and the features of the text modality into the sparse multi-labels to train a strong hash function to supervise the learning of hash functions for the image modality and the text modality.

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 speaking, instances with multi-labels contain more semantic information than instances with signal labels. How to adequately mine the semantic information in multi-labels to accurately calculate the semantic similarities between instances with multi-labels remains a problem. To this end, [48] proposes a distance metric learning algorithm for multi-label classification, which integrates a pairwise multi-labels similarity constraint and a Jaccard Distance into multi-label learning and achieves competitive performance. To better explore the semantic information in multi-labels and further preserve the multi-labels similarity, especially preserve the multi-labels similarity of cross-modal instances, in this paper, we define a multi-label cross-modal triplet loss.

3. Proposed Method

In this section, we elaborate our proposed multi-label modal 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, we assume that each data in our method has three modalities (i.e., an image modality, a text modality and a multi-label modality). The framework of MMACH is shown in figure 1.

3.1. Notation and Problem Formulation

To help better understand this section, here we firstly give a formal definition of notations and problem formulations. We are provided a 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 feature, the original text feature and the multi-labels of the i-th training instance. If the i-th training instance is assigned to the j-th class, then the j-th 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 mapping functions to project the original images, the original multi-labels, the original texts into compact hash codes and preserve semantic similarities of these cross-modal instances unchanged. To achieve this goal, we first encode the original instances to 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 is pretty sparse, we utilize a multi-label modal enhanced attention to compensate it. The multi-label enhanced feature vectors are denoted as $\{H_i^A\}_{i=1}^n$. Afterwards, three deep neural networks are utilized to map the $\{F_i^E\}_{i=1}^n$, $\{G_i^E\}_{i=1}^n$ 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, we use a sign function to generate united hash codes matrix $B \in R^{n \times k}$ from the learned hash representations.

3.2. Modal Encoders

To extract features from the original instances, three encoders E_I, E_T and E_L are leveraged to encode each original image I_i, each original text T_i, and each original multi-label L_i 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

In cross-modal hashing field, many practical datasets (e. g., MIRFLICKR-25K [49] and NUS-WIDE [50]) contain multi-labels. Nevertheless, most previous methods merely regard two cross-modal instances as similar pairwise only if they have at least one common category, which overlooks the abundant semantic information in multi-labels and thus cannot accurately evaluate the semantic relevance of cross-modal pairwise 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 naturally come up to guide the learning of cross-modal hash projection functions. Nonetheless, as the original multi-label matrix is severely sparse, the multi-label based self-supervised learning strategy can make limited enhancement for the performance of the learned cross-modal hash projection functions. For this purpose, in this subsection, a multi-label modal enhanced attention module (MMEA) is proposed to enrich 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 subsection 3.2 to encode them into c-dimensional feature vectors F_i^E, G_i^E and H_i^E, then an attention mechanism is introduced to fuse relative semantic information of F_i^E and G_i^E to 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 $attention^{IL}$, $attention^{TL}$ are semantic affinity of F_i^E and H_i^E, G_i^E and H_i^E, respectively. $\| . \|$ is norm of a 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 equation 2 and 3, we can compensate the sparse multi-label L_i with the abundant semantic information in I_i and T_i, which with a self-supervising learning style can better guide the training of deep neural networks for the image modality and the text modality in return.

3.4. Multi-Label Cross-Modal Triplet Loss

Suppose that we have a cross-modal triplet (I_i, T_{p1}, T_{p2}), the image I_i is more semantically similar to the text T_{p1} than to the text T_{p2}. Their hash representations F_i^{hr}, G_{p1}^{hr} and G_{p2}^{hr} can easily learned with the corresponding hash mapping functions, i.e., $F_i^{hr} = f(F_i^E, \theta^I)$, $G_{p1}^{hr} = g(G_{p1}^E, \theta^T)$ and $G_{p2}^{hr} = g(G_{p2}^E, \theta^T)$. To preserve the semantic similarity unchanged 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, 52, 53], we define the multi-label cross-modal triplet loss (MCTL) as follows:

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

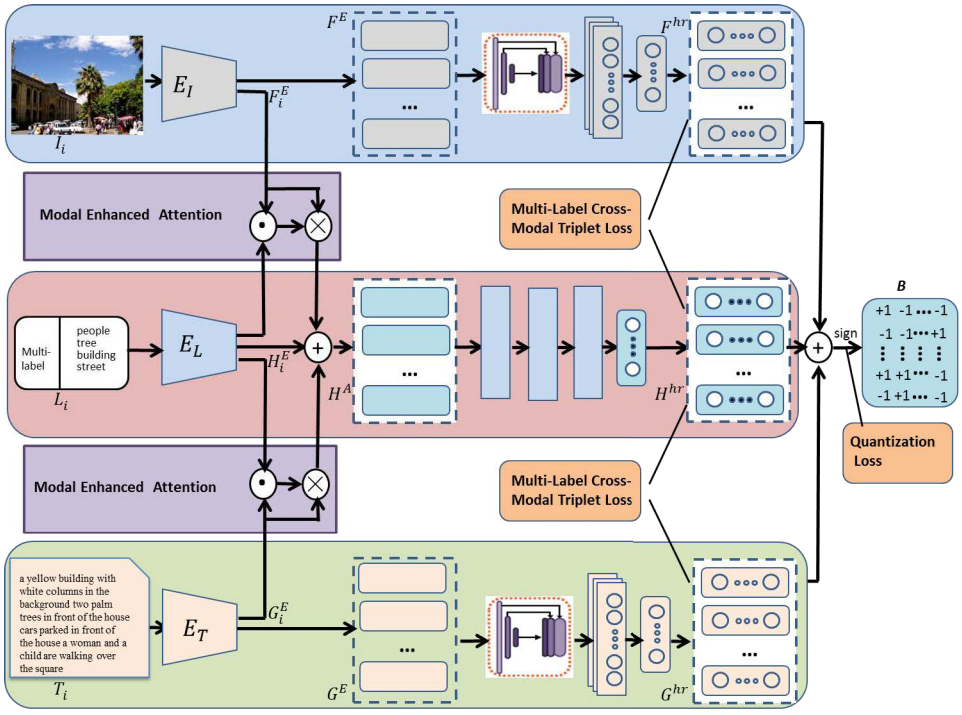


Fig. 1. The overall flowchart of our proposed MMACH method. MMACH includes three parts: (1) a modal encoder part (E_I, E_L and E_T), which composes of three deep neural networks to extract the features from the original instances of the image modality, the text modality, and the multi-label modality, respectively. (2) a multi-label modal enhanced attention part, which utilizes attention mechanism to extract semantically relevant information from the image modality and the text modality and then fuse them to the sparse multi-label modality. (3) a hash representation learning and hash codes generation part, which aims to make sure that semantically similar pairwise cross-modal instances have similar hash codes. The $\odot$ represents dot product, while the $\otimes$ represents element product, and the $\oplus$ denotes element add.

Where $\|\cdot\|_{L_2}$ is the L_2 norm, γ is a positive margin. Equation 4 means that the L_2 distance of a more multi-label semantic similar cross-modal pairwise is smaller than the L_2 distance of a less multi-label semantic similar cross-modal pairwise with a margin γ. By this way, the multi-label cross-modal semantic similarity can be adequately protected during hash representation learning stage.

3.5. Hash Representations Learning

In hash representation learning stage, 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 as input of the deep neural network for the multi-label modality, deep neural network for the image-modality, and deep neural network for the text-modality, respectively. To preserve semantic similarity of cross-modal instances during hash representation learning procedure, we introduce the multi-label cross-modal triplet loss in subsection 3.4. Namely, 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:

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

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. 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. γ_1 and γ_2 are two positive margins.

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

Where J^{TL} is the cross-modal semantic similarity preserving loss for the text-modality and the multi-label modality. 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. 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. γ_3 and γ_4 are two positive margins.

3.6. Hash Codes Generation

By subsection 3.5, 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 original images $\{I_i\}_{i=1}^n$, original texts $\{T_i\}_{i=1}^n$, and original multi-labels $\{L_i\}_{i=1}^n$, respectively. However, the goal of cross-modal hashing is to map multimodal data into compact hash codes. To this end, we utilize a sign function to approximatively generate the hash codes from the learned hash representations:

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

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

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

Moreover, to further decrease the information loss in equation 7, a quantization loss is introduced as follows:

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

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

Merging the cross-modal semantic similarity preserving losses and the quantization loss together, the whole loss function is depicted as follows:

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

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

3.7. Hash Representations Learning Networks

For the image-modality, we fine-tune the TxtNet in SSAH [40] ($c \rightarrow MS \rightarrow 4096 \rightarrow 512 \rightarrow k$) 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 also 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 when keeping the other three fixed. The detailed execution and optimization for MMACH are depicted 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 DNN for the multi-label modality can be learned by stochastic gradient descent (SGD) and back-propagation (BP). Detailedly, in each iteration, four batches of training cross-modal triplets are randomly selected to execute our algorithm, i.e., for each selected multi-label enhanced feature vector H_i^A, the gradient is computed as follows:

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

Afterwards, the $\frac{\partial J}{\partial \theta^L}$ can be calculated from $\frac{\partial J}{\partial H_{\cdot}^{hr}}$ with the chain rule. Finally, the θ^L can be optimized with $\frac{\partial J}{\partial \theta^L}$ and the 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 DNN for the image modality can be optimized by SGD and BP. Concretely, in each epoch, two batches of training cross-modal triplets are randomly selected to run our method, i.e., for each selected image feature vector F_i^E, the gradient is calculated as follows:

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

Further, the $\frac{\partial J}{\partial \theta^I}$ can be calculated from $\frac{\partial J}{\partial F_{\cdot}^{hr}}$ with the chain rule. Finally, the θ^I can be optimized with $\frac{\partial J}{\partial \theta^I}$ and the 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 DNN for the text modality can be optimized by SGD and BP. Concretely, in each epoch, two batches of training cross-modal triplets are randomly selected to execute our algorithm, i.e., for

each selected text feature vector G_i^E, the gradient is calculated as follows:

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

Afterwards, the $\frac{\partial J}{\partial \theta^T}$ can be calculated from $\frac{\partial J}{\partial G^{hr}}$ with the chain rule. Finally, the θ^T can be optimized with $\frac{\partial J}{\partial \theta^T}$ and the 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 equation 7.

Algorithm 1 MMACH: Multi-Label Modal 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 epoches of the algorithm is *max_epoch*.
mini-batch size $n_{batch} = 128$.
Output:
Deep neural networks parameters θ^I, θ^T and θ^L for hash representation learning, and 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 equation 1.
Learning the multi-label enhanced feature vectors $\{H_i^A\}_{i=1}^n$ from $\{H_i^E\}_{i=1}^n$ with equation 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 $Triplet_{IL}$ to construct a mini-batch and randomly select n_{batch} triplets from $Triplet_{LI}$ to construct a mini-batch. Randomly select n_{batch} triplets from $Triplet_{TL}$ to construct a mini-batch and randomly select n_{batch} triplets from $Triplet_{LT}$ to construct a mini-batch.
    For each feature vector H_i^A in the mini-batches, calculate $H_j^{hr} = h(H_i^A, \theta^L)$ by forward propagation.
    Update $\{H_i^{hr}\}_{i=1}^n$.
    Compute the derivative of θ^L in equation 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 $Triplet_{IL}$ to construct a mini-batch and randomly select n_{batch} triplets from $Triplet_{LI}$ to construct a mini-batch.
    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 in equation 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 $Triplet_{TL}$ to construct a mini-batch and randomly select n_{batch} triplets from $Triplet_{LT}$ to construct a mini-batch.
    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 in equation 13.
    Utilize back-propagation to update the network parameters θ^T.
  end for
  Optimize B by utilizing equation 7.
until the max epoch number *max_epoch*

4.5. Complexity Analysis

We analyze the algorithm complexity of MMACH in this part. For the overall loss function (equation 10) of MMACH, its complexity can be calculated as follows: $O(n_{IL})+O(n_{TL})+O(n\times k)\approx O(n^2)$, 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, in this section, we implement 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 Flickr website. In our experiment, instances that have at least 20 textual tags are selected and thus finally 20, 015 image-text pairs with multi-labels are remained, and each of the selected instances is assigned to at least one of the 24 given labels. We encode each textual tag into a 1386-dimensional BOW (bag-of-words) feature in our experiment.

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. We encode each textual tag into a 1000-dimensional BOW feature in our experiment.

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 80 classes. Moreover, 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 when giving a query instance. The hash lookup protocol returns retrieval instances within a certain Hamming radius from the

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

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** [36], **PRDH** [37], **CMHH** [39], **CHN** [38], **SSAH** [40] and **MLSPH** [56] are chose as baseline methods in our experiment. Concretely, 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 X-P GPU server. During 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 and 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. The length of hash codes is set to 64 to find out the best value of hyper-parameter α. The MAPs of our proposed MESDCH method under different α are recorded and then depicted in Figure 2. 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 conduct the experiment to examine the effectiveness of our proposed multi-label enhanced attention module. Concretely, we firstly remove the modal attention attention in our proposed MMACH (i.e., we set $H^A = H^E$ in Figure 1) and keep other parts unchanged, we name this variation as MLSDCH. Afterwards, we compare MLSDCH with MMACH on the four datasets MIRFLICKR-25K, NUS-WIDE, Microsoft COCO2014 and IAPRTC-12. The corresponding MAPs under different hash code lengths of 16, 32 and 64 are shown in Table 2.

From 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 enhance the performance of cross-modal hashing retrieval, which is partly because multi-label enhanced attention module compensates the sparse feature space. In addition, Figure 3 presents the top 4 cross-modal retrieval results by MMACH and MLSDCH on four datasets, 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 of our proposed multi-label cross-modal triplet loss in our proposed MMACH method and keep other parts fixed, and 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 distinct hash code lengths 16, 32 and 64 are shown in Table 3.

From Table 3, we can see that the MAPs of MMACH is 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 compare to MSE loss. Furthermore, Figure 4 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 to further 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,

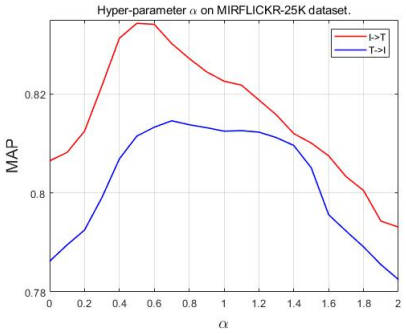


Fig. 2. Sensitivity analysis of the hyper-parameter α on MIRFLICKR25K dataset.

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**

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. This demonstrates that MMACH can utilize 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 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 modalities. Among deep neural networks based baseline methods, MLSPH has the highest MAP values in most cases which is partly because MLSPH introduces a multi-label semantic preserving module and it can compute the semantic relevance of original data more precisely.

(3) Compared to hand-crafted methods, deep neural networks based methods usually achieve higher MAP values, which is partly because deep neural networks based methods make full use of the superior features learning capability of 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, which is partly because MMACH defines a multi-label modal enhanced attention module to compensate the sparse features of multi-labels. Moreover, MMACH utilize multi-label cross-modal triplet loss to select multi-label semantic similar triplets, while SSAH regards the semantic similarity of two instances as 1 only if there are at least one common categories, which neglects the difference of multi-labels.

To further compare MMACH with baseline CMH methods, we compare the precision-recall curves of MMACH and all baseline methods on four experimental datasets with different hash codes length. Figures 5, 6, 7 and 8 are the precision-recall curves of all methods with different datasets and hash code length. From these figures, we can observe that our proposed MMACH outperforms most baseline methods in most cases. Meanwhile, the precision-recall curves are approximately identical to our observations on the MAP scores.

Moreover, top$N$-precision curves of MMACH and baseline methods on datasets MIRFLICKR-25K, NUS-WIDE, Microsoft COCO2014 and iaprtc12 with hash codes length of 16, 32 and 64 are drew and presented in Figure 9, 10, 11 and 12. From these results, we can see that, in most cases, MMACH can achieve better performance than baseline methods, which is 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. Figure 13 and Figure 14 illustrate the corresponding heatmaps. From these heatmaps, it is obvious that our MMACH can more accurately correlate the corresponding semantic categories compared to DCMH and SSAH in most cases, which demonstrates the powerful multi-label semantic preserving capability of our proposed MMACH.

6. Conclusion

In this paper, we introduce a prominent cross-modal hashing method termed multi-label modal enhanced attention based

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.

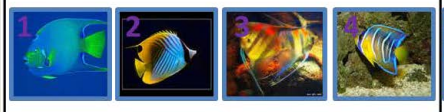
Task	Method	MIRFlickr-25K			NUS-WIDE			MS COCO			IAPRTC-12		
		16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I→T	MMACH-MSE	0.8006	0.8158	0.8282	0.6215	0.6533	0.6692	0.6912	0.7168	0.7364	0.5286	0.5450	0.5795
	MMACH	**0.8085**	**0.8235**	**0.8348**	**0.6489**	**0.6679**	**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 [36]	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 [37]	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 [39]	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 [38]	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 [40]	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 [36]	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 [37]	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 [39]	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 [38]	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 [40]	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

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. 3. 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 frame are the matching image.

self-supervised deep cross-modal hashing (MMACH). A novel multi-label modal enhanced attention module is designed in MMACH to compensate the sparse feature vectors of multi-labels from multi-modal instances and based on this enhanced multi-labels, self-supervised learning is introduced to train a multi-label hash function to supervise the training of hash functions of other modalities. Furthermore, a multi-label cross-modal triplet loss is defined in MMACH to ensure that hash repre-

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. 4. 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 frame are the matching image.

sentations of pairwise cross-modal instances with more common labels should be more similar than that of pairwise cross-modal instances with less common labels. Massive experiments on several renowned cross-modal benchmark datasets indicated that MMACH method surpasses baseline methods and acquires competitive cross-modal retrieval performance.

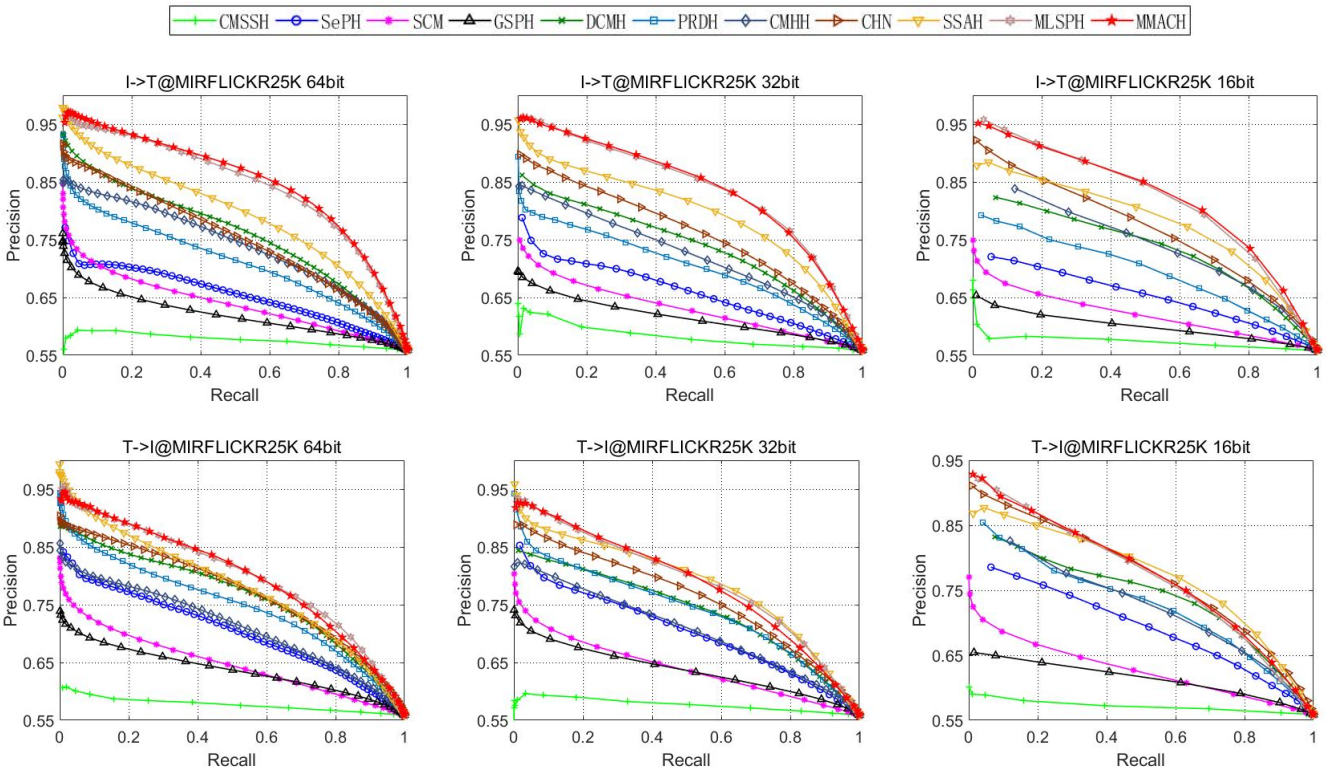


Fig. 5. Precision-Recall Curves on MIRFLICKR-25K.

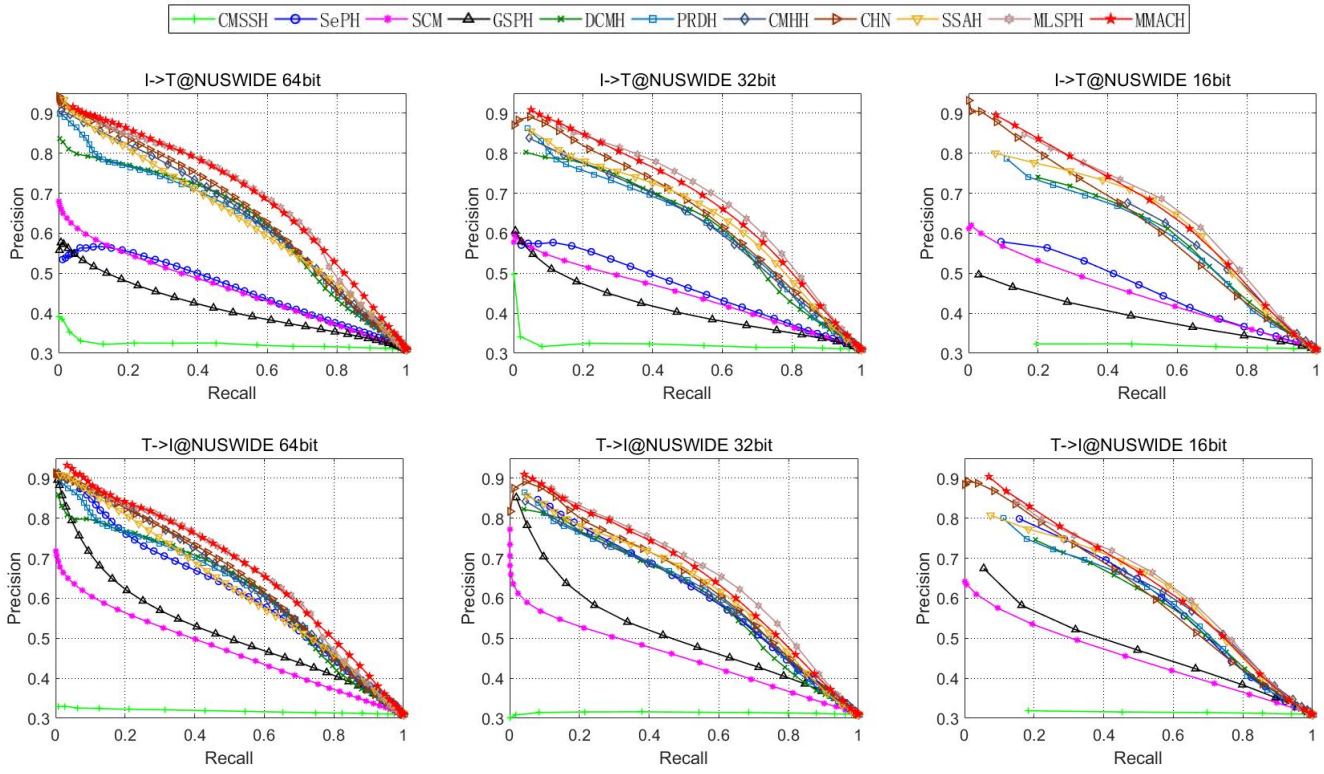


Fig. 6. Precision-Recall Curves on NUS-WIDE.

Acknowledgments

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

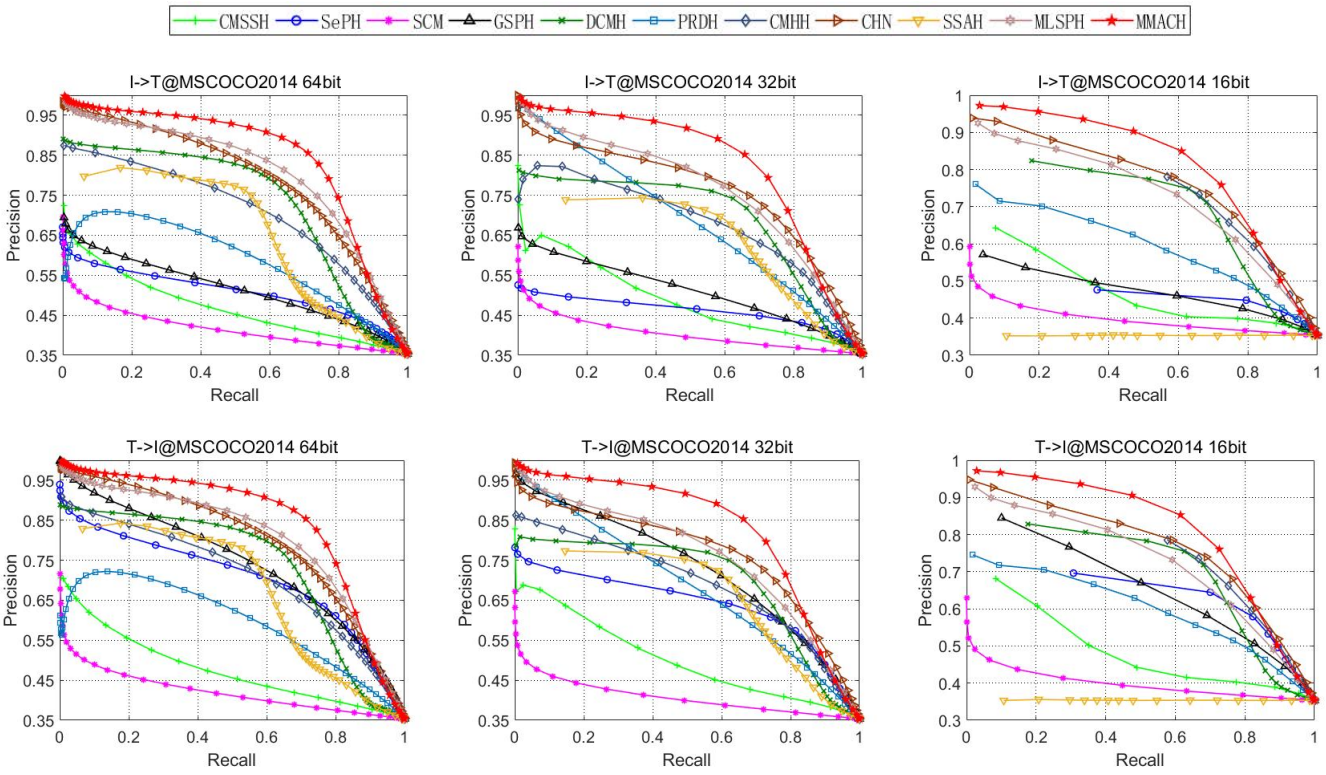


Fig. 7. Precision-Recall Curves on Microsoft COCO2014.

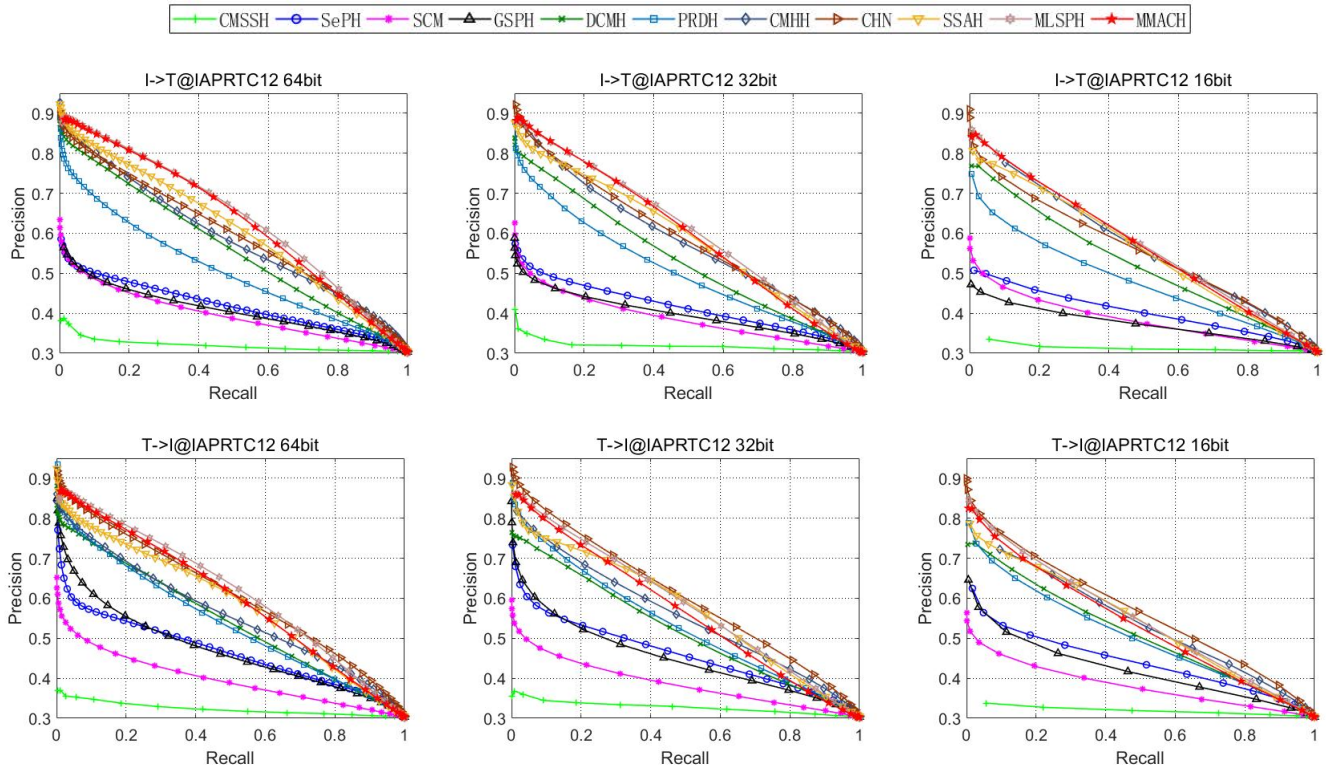


Fig. 8. Precision-Recall Curves on IAPRTC-12.

grant #2011927 and DoD grant #W911NF1810475.

References

[1] Yuxin Peng, Xin Huang, and Yunzhen Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE*

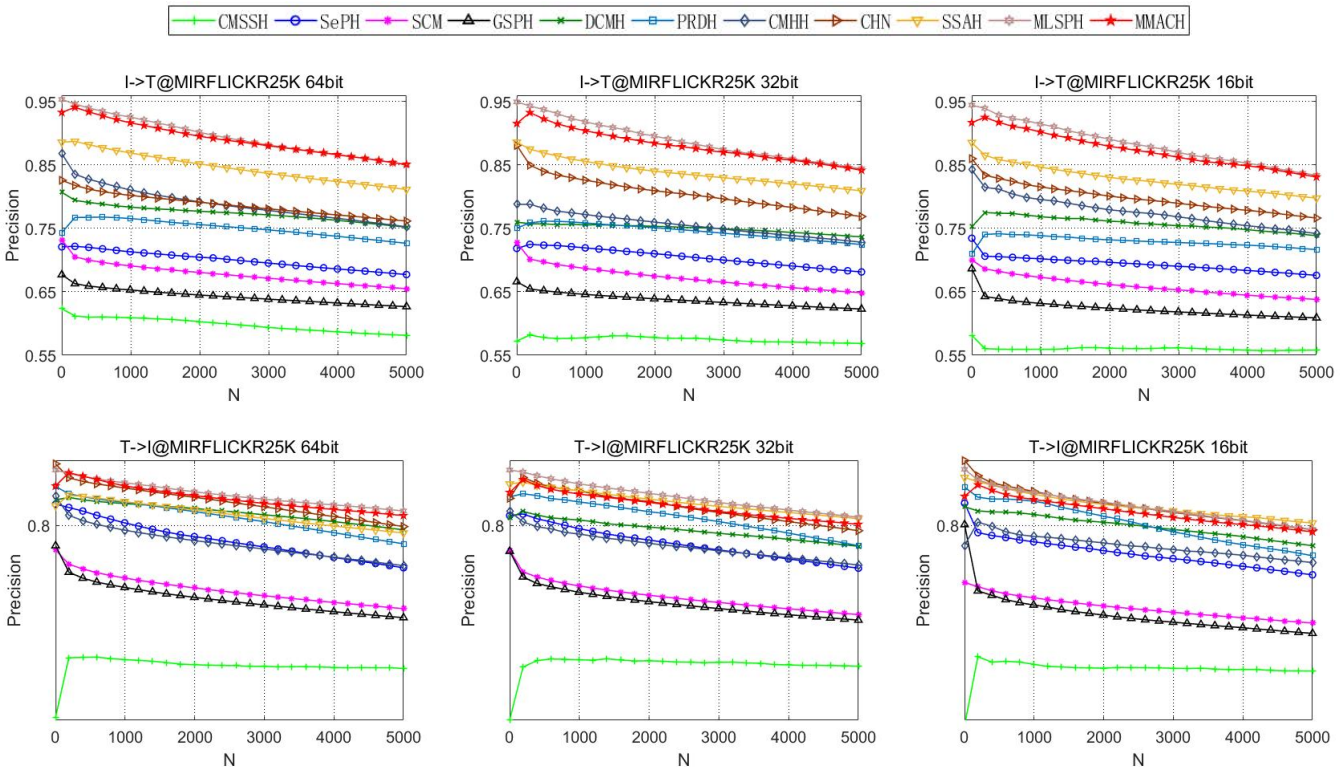


Fig. 9. top*N*-precision curves on MIRFLICKR-25K

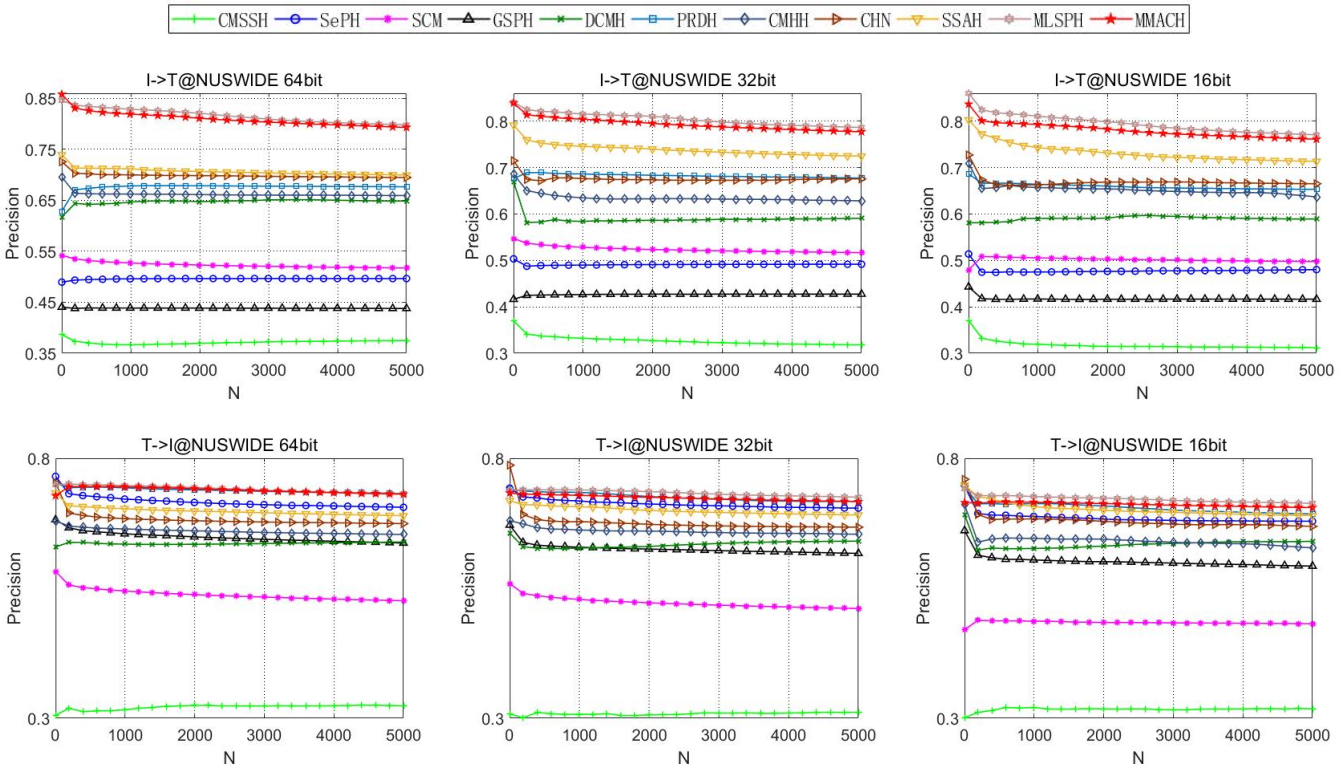


Fig. 10. top*N*-precision curves on NUS-WIDE

Transactions on circuits and systems for video technology, 28(9):2372–2385, 2017.

[2] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

[3] Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. Learning cross-

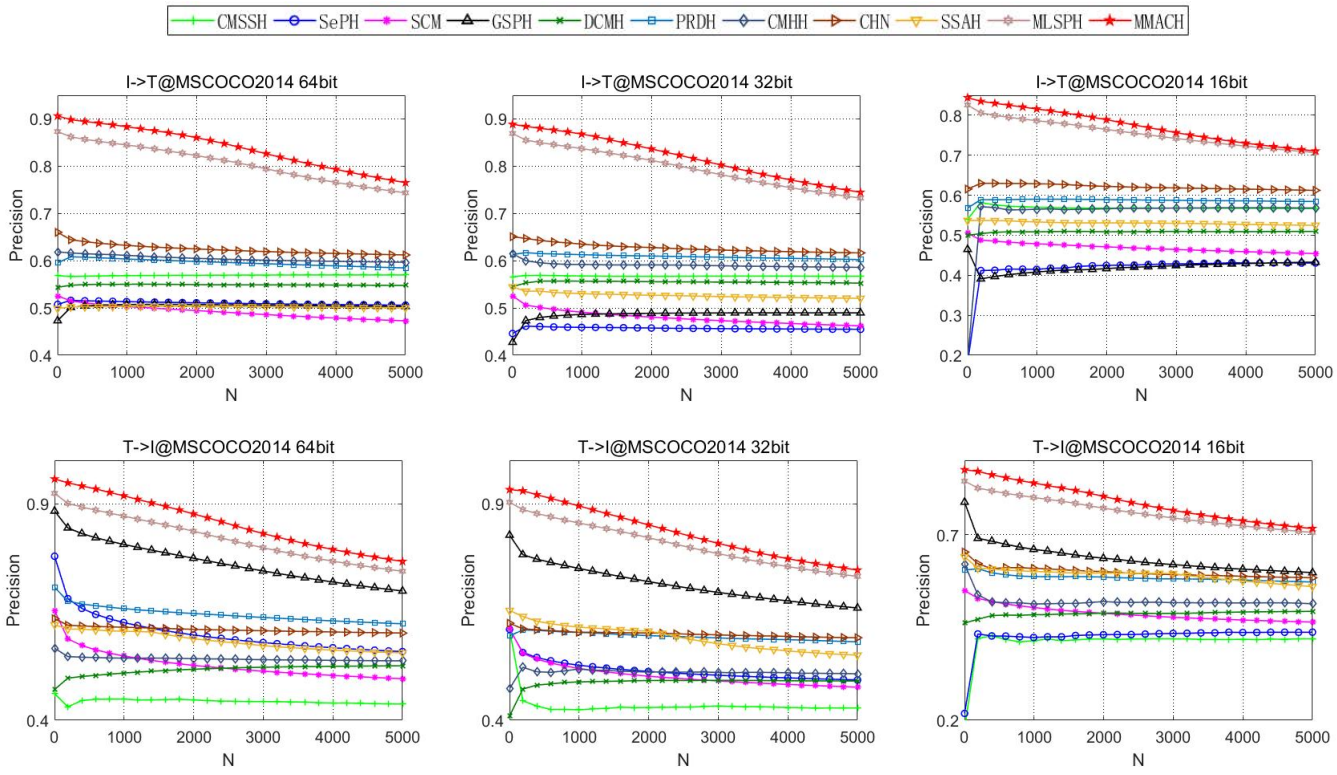


Fig. 11. top*N*-precision curves on Microsoft COCO2014

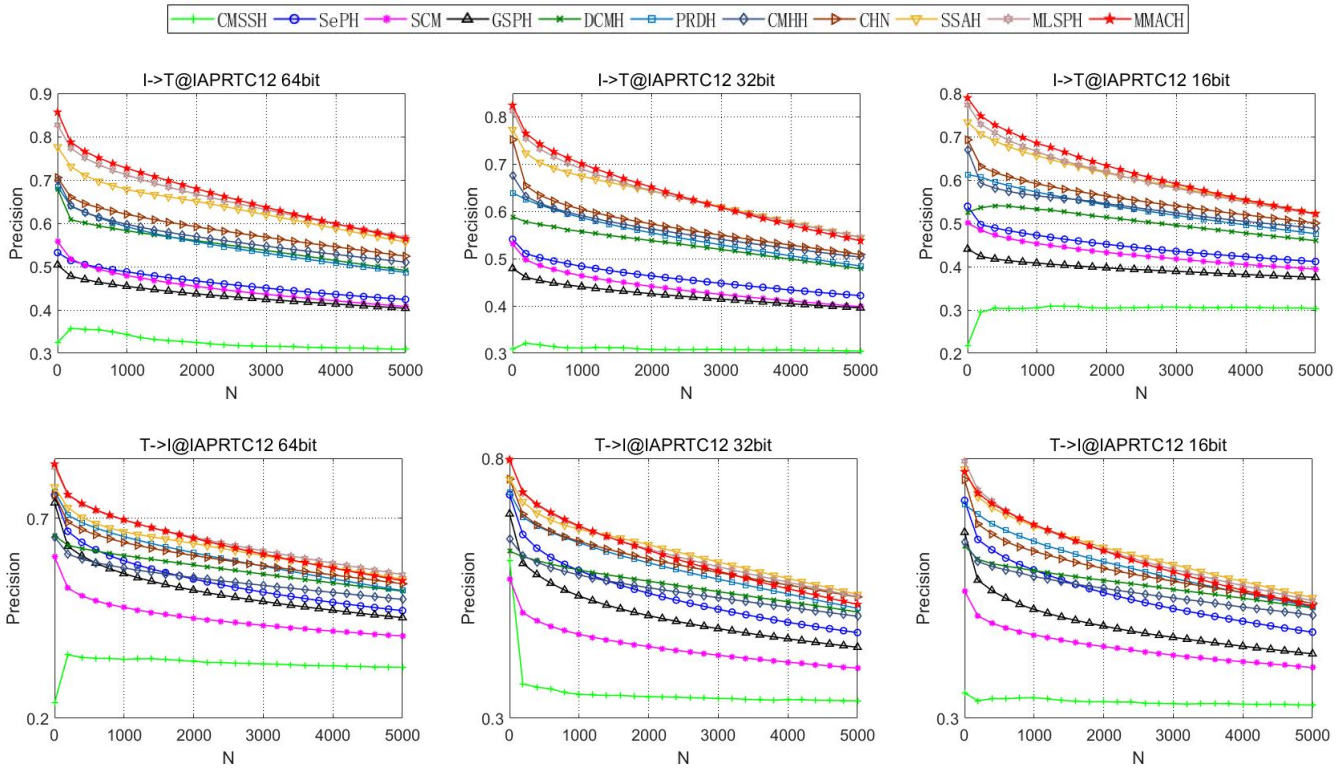


Fig. 12. top*N*-precision curves on IAPRTC-12

modality similarity for multinomial data. In *2011 International Conference on Computer Vision*, pages 2407–2414. IEEE, 2011.

[4] Yin Zheng, Yu-Jin Zhang, and Hugo Larochelle. Topic modeling of multimodal data: an autoregressive approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1370–1377, 2014.

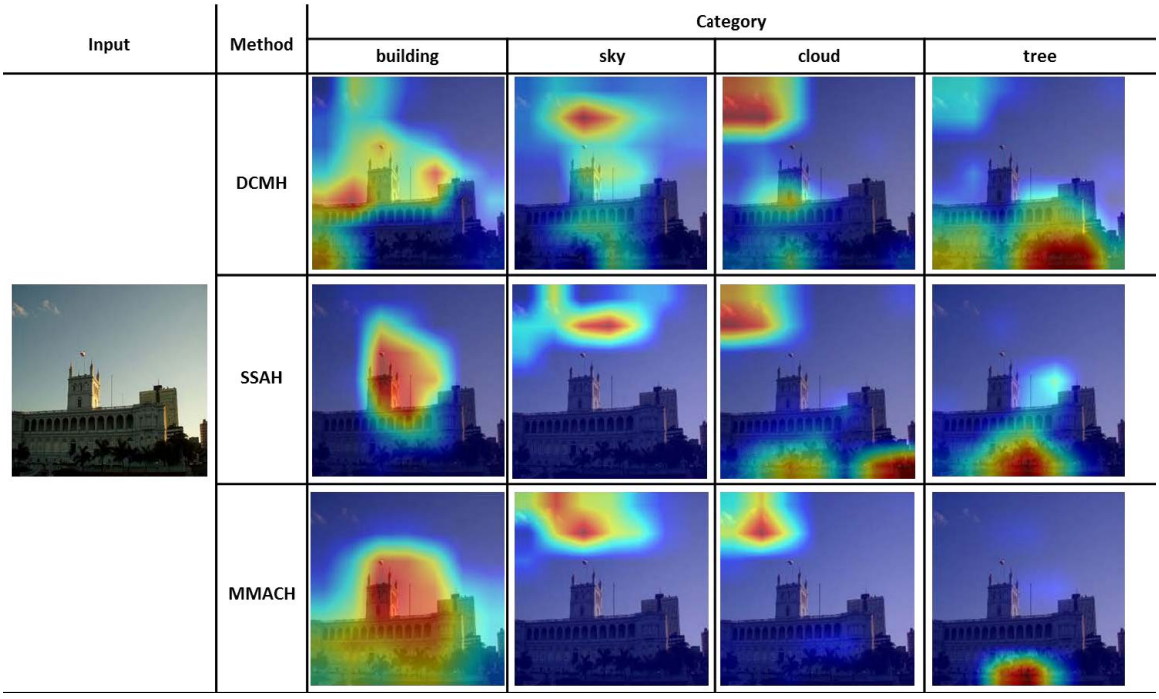


Fig. 13. Grad-CAM visualization of MMACH compared to SSAH and DCMH for a randomly selected image from multi-label dataset IAPRTC-12 with respect to different ground-truth categories.

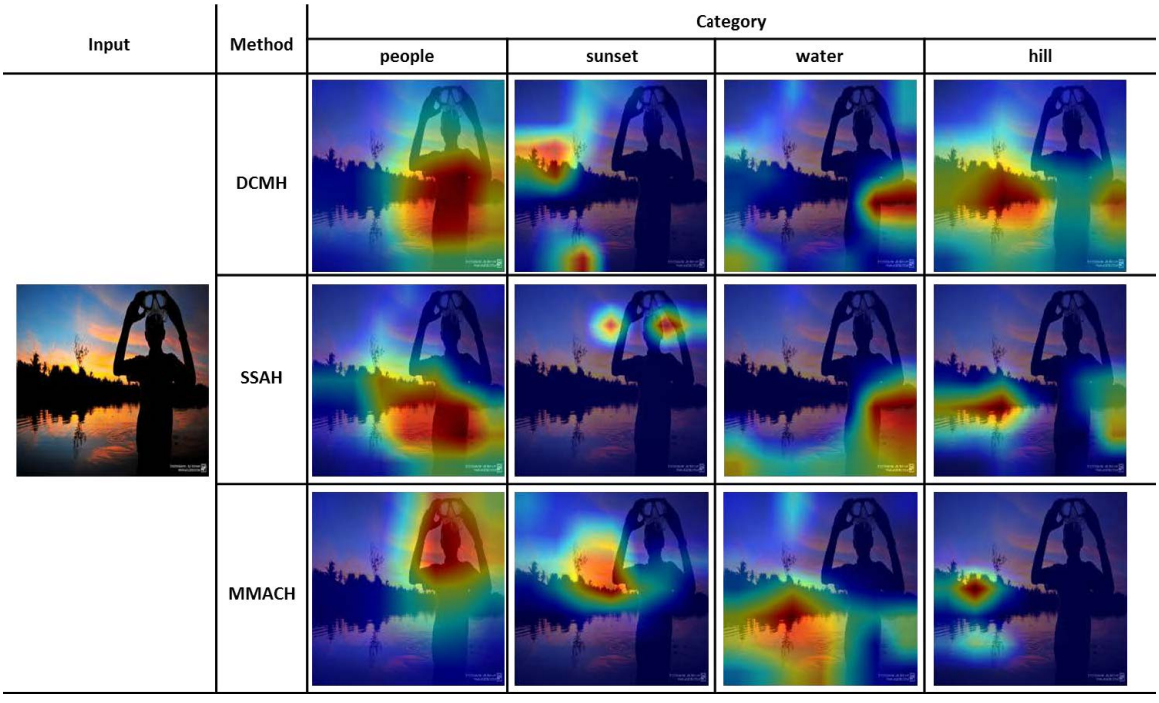


Fig. 14. Grad-CAM visualization of MMACH compared to SSAH and DCMH for a randomly selected image from multi-label dataset MIRFLICKR-25K with respect to different ground-truth categories.

[5] Yanfei Wang, Fei Wu, Jun Song, Xi Li, and Yueting Zhuang. Multi-modal mutual topic reinforce modeling for cross-media retrieval. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 307–316. ACM, 2014.

[6] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2160–2167. IEEE, 2012.

[7] Xiao-Yuan Jing, Rui-Min Hu, Yang-Ping Zhu, Shan-Shan Wu, Chao Liang, and 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, and Jian Pei. Parallel field alignment for cross media retrieval. In *Proceedings of the 21st ACM*

international conference on Multimedia, pages 897–906. ACM, 2013.

[9] Yue Ting Zhuang, Yan Fei Wang, Fei Wu, Yin Zhang, and 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, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2):210–233, 2014.

[11] Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2010–2023, 2015.

[12] Jian Wang, Yonghao He, Cuicui Kang, Shiming Xiang, and Chunhong Pan. Image-text cross-modal retrieval via modality-specific feature learning. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 347–354. ACM, 2015.

[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[14] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.

[15] Xinyang Jiang, Fei Wu, Xi Li, Zhou Zhao, Weiming Lu, Siliang Tang, and Yueting Zhuang. Deep compositional cross-modal learning to rank via local-global alignment. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 69–78. ACM, 2015.

[16] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE transactions on cybernetics*, 47(2):449–460, 2016.

[17] Yuxin Peng and Jinwei Qi. Cm-gans: cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1):22, 2019.

[18] Fangming Zhong, Zhikui Chen, and Geyong Min. Deep discrete cross-modal hashing for cross-media retrieval. *Pattern Recognition*, 83:64–77, 2018.

[19] Lin Wu, Yang Wang, and Ling Shao. Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 28(4):1602–1612, 2018.

[20] Devraj Mandal, Kunal N Chaudhury, and Soma Biswas. Generalized semantic preserving hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 28(1):102–112, 2018.

[21] Guiguang Ding, Yuchen Guo, and Jile Zhou. Collective matrix factorization hashing for multimodal data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2075–2082, 2014.

[22] Yixian Fang, Bin Li, Xiaozhou Li, and Yuwei Ren. Unsupervised cross-modal similarity via latent structure discrete hashing factorization. *Knowledge-Based Systems*, 218:106857, 2021.

[23] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[24] Shupeng Su, Zhisheng Zhong, and 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*, pages 3027–3035, 2019.

[25] Chao Li, Cheng Deng, Lei Wang, De Xie, and Xianglong Liu. Coupled cyclegan: Unsupervised hashing network for cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 176–183, 2019.

[26] Gengshen Wu, Zijia Lin, Jungong Han, Li Liu, Guiguang Ding, Baochang Zhang, and Jialie Shen. Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval. In *IJCAI*, pages 2854–2860, 2018.

[27] Yixian Fang, Huaxiang Zhang, and Yuwei Ren. Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing. *Knowledge-Based Systems*, 171:69–80, 2019.

[28] Min Meng, Haitao Wang, Jun Yu, Hui Chen, and Jigang Wu. Asymmetric supervised consistent and specific hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 30:986–1000, 2020.

[29] Haopeng Qiang, Yuan Wan, Ziyi Liu, Lun Xiang, and Xiaojing Meng. Discriminative deep asymmetric supervised hashing for cross-modal retrieval. *Knowledge-Based Systems*, 204:106188, 2020.

[30] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

[31] Dongqing Zhang and 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, and Kris M Kitani. Deep supervised hashing with triplet labels. In *Asian conference on computer vision*, pages 70–84. Springer, 2016.

[33] Zhan Yang, Liu Yang, Osolo Ian Raymond, Lei Zhu, Wenti Huang, Zhifang Liao, and Jun Long. Nsdh: A nonlinear supervised discrete hashing framework for large-scale cross-modal retrieval. *Knowledge-Based Systems*, 217:106818, 2021.

[34] Song Wang, Huan Zhao, and Kei Nai. Learning a maximized shared latent factor for cross-modal hashing. *Knowledge-Based Systems*, 228:107252, 2021.

[35] Fengling Li, Tong Wang, Lei Zhu, Zheng Zhang, and Xinhua Wang. Task-adaptive asymmetric deep cross-modal hashing. *Knowledge-Based Systems*, 219:106851, 2021.

[36] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3232–3240, 2017.

[37] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao. Pairwise relationship guided deep hashing for cross-modal retrieval. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[38] Yue Cao, Mingsheng Long, Jianmin Wang, and Philip S Yu. Correlation hashing network for efficient cross-modal retrieval. *arXiv preprint arXiv:1602.06697*, 2016.

[39] Yue Cao, Bin Liu, Mingsheng Long, and Jianmin Wang. Cross-modal hamming hashing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 202–218, 2018.

[40] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4242–4251, 2018.

[41] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[42] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[43] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and 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*, pages 1480–1489, 2016.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[45] Xi Zhang, Hanjiang Lai, and Jiashi Feng. Attention-aware deep adversarial hashing for cross-modal retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pages 591–606, 2018.

[46] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

[47] Mohammad S Sorower. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18:1–25, 2010.

[48] Henry Gouk, Bernhard Pfahringer, and Michael Cree. Learning distance metrics for multi-label classification. In *Asian Conference on Machine Learning*, pages 318–333. PMLR, 2016.

[49] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43. ACM, 2008.

[50] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and 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*, page 48. ACM, 2009.

[51] Weiwei Liu and Ivor W Tsang. Large margin metric learning for multi-label prediction. In *AAAI*, 2015.

[52] Yi Zhang and Jeff Schneider. Maximum margin output coding. *Computer*

Science, pages 1575–1582, 2012.

[53] Yashaswi Verma and C. V. Jawahar. Image annotation by propagating labels from semantic neighbourhoods. *International Journal of Computer Vision*, 121(1):126–148, 2017.

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[55] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villaseñor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer vision and image understanding*, 114(4):419–428, 2010.

[56] Xitao Zou, Xinzhi Wang, Erwin M Bakker, and Song Wu. Multi-label semantics preserving based deep cross-modal hashing. *Signal Processing: Image Communication*, 93:116131, 2021.

[57] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and 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*, pages 3594–3601. IEEE, 2010.

[58] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Semantics-preserving hashing for cross-view retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3864–3872, 2015.

[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. *arXiv preprint arXiv:1803.11175*, 2018.

[60] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[61] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.