Deep hashing with self-supervised asymmetric semantic excavation and margin-scalable constraint

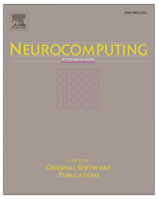
Yu, Z.; Wu, S.; Dou, Z.; Bakker, E.M.

Neurocomputing 483 (2022) 87–104

Contents lists available at ScienceDirect

Neurocomputing

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



Deep hashing with self-supervised asymmetric semantic excavation and margin-scalable constraint

Zhengyang Yu [a], Song Wu [a,*], Zhihao Dou [a], Erwin M. Bakker [b]

[a] College of Computer and Information Science, Southwest University, Chongqing, China
[b] LIACS Media Lab, Leiden University, Leiden, Netherlands

ARTICLE INFO

Article history:
Received 25 January 2021
Revised 15 December 2021
Accepted 23 January 2022
Available online 29 January 2022
Communicated by Zidong Wang

Keywords:
Deep supervised hashing
Asymmetric learning
Self-supervised learning

ABSTRACT

Due to its effectivity and efficiency, deep hashing approaches are widely used for large-scale visual search. However, it is still challenging to produce compact and discriminative hash codes for images associated with multiple semantics for two main reasons, 1) similarity constraints designed in most of the existing methods are based upon an oversimplified similarity assignment(i.e., 0 for instance pairs sharing no label, 1 for instance pairs sharing at least 1 label), 2) the exploration in multi-semantic relevance are insufficient or even neglected in many of the existing methods. These problems significantly limit the discrimination of generated hash codes. In this paper, we propose a novel Deep Hashing with Self-Supervised Asymmetric Semantic Excavation and Margin-Scalable Constraint(SADH) approach to cope with these problems. SADH implements a self-supervised network to sufficiently preserve semantic information in a semantic feature dictionary and a semantic code dictionary for the semantics of the given dataset, which efficiently and precisely guides a feature learning network to preserve multi-label semantic information using an asymmetric learning strategy. By further exploiting semantic dictionaries, a new margin-scalable constraint is employed for both precise similarity searching and robust hash code generation. Extensive empirical research on four popular benchmarks validates the proposed method and shows it outperforms several state-of-the-art approaches. The source codes URL of our SADH is: http://github.com/SWU-CS-MediaLab/SADH.

1. Introduction

The amount of image and video data in social networks and search engines are growing at an alarming rate. In order to effectively search large-scale high dimensional image data, Approximate Nearest Neighbor (ANN) search has been extensively studied by researchers [1,2]. Semantic hashing, first proposed in the pioneer work [3] is widely used in the field of large-scale image retrieval. It maps high-dimensional content features of pictures into Hamming space (binary space) to generate a low-dimensional hash sequence [1,2], which reflects the semantic similarity by distance between hash codes in the Hamming space. Hash algorithms can be broadly divided into data-dependent methods and data-independent methods [4] schemes. The most basic but representative data independent method is Locality Sensitive Hashing (LSH) [1], which generates embedding through random projections. However, these methods all require long binary code to achieve accuracy, which is not adapt to the processing of large-scale visual data. Recent research priorities have shifted to data-dependent approaches that can generate compact binary codes by learning large amount of data and information. This type of method embeds high-dimensional data into the Hamming space and performs bitwise operations to find similar objects. Recent data-dependent works such as [2,5–10] have shown better retrieval accuracy under smaller hash code length.

Although the above data-dependent hashing methods have certainly succeeded to some extent, they all use hand-crafted features, thereby limiting the retrieval accuracy of learning binary code. Recently, the deep-learning-based hashing methods have shown superior performance by combining the powerful feature extraction of deep learning [11–16]. Admitting significant progress achieved in large-scale image retrieval with deep hashing methods, there still remain crucial bottlenecks that limit the hashing retrieval accuracy for datasets like NUS-WIDE [17], MS-COCO [18], MIRFlickr-25K [19], where each image is annotated with multiple semantics. Firstly, to the best of our knowledge, most of the existing supervised hashing methods use semantic-level labels to examine the similarity between instance pairs following a common experimental protocol. That is, the similarity score will be

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

https://doi.org/10.1016/j.neucom.2022.01.082
0925-2312/© 2022 Elsevier B.V. All rights reserved.

assigned as ‘1’ if the item pair shares at least one semantic label and ‘0’ if none of the semantic labels are shared. Based upon this coarsely defined similarity metric, in many of the existing methods [20,11,21], the exact degree of similarity (i.e., how many exact semantics are shared) cannot be quantified, therefore they fail to search for similarity information at a fine-grained level. Additionally, by further utilizing semantic labels, exploring semantic relevance to facilitate the similarity searching process can bring numerous merits for hashing function learning, e.g., the interclass instance pairs can be better separated which can provide better efficiency and robustness in the training process [22]; the shared image representations can be learned which is beneficial for hashing function learning [23]. Many existing deep hashing methods ignore to leverage such valuable semantic information [11–13,15,16], leading to inferior retrieval performance. A few of the existing methods [24–26,23] solve this problem by adding an auxiliary classifier to enhance the preservation of global semantic information. However, the complex semantic correlations under mentioned multi-label scenarios are still insufficiently discovered and cannot be effectively embedded into hash codes.

To tackle the mentioned flaws, we proposed a novel Deep Hashing with Self-Supervised Asymmetric Semantic Excavation and Margin-Scalable Constraint (SADH) approach to improve the accuracy and efficiency of multi-label image retrieval. Holding the motivation of thoroughly discover semantic relevance, as shown in Fig. 1, in our work, in spite of using an auxiliary classifier following methods like [24–26,23], semantic relevance from multi-label annotations are thoroughly excavated through a self-supervised Semantic-Network. While a convolutional neural network namely Image-Network, projects original image inputs into semantic features and hash codes. Inspired by methods like [27–30], we propose a novel asymmetric guidance mechanism to efficiently and effectively transfer semantic information from Semantic-Network to Image-Network, firstly we refine the abstract semantic features and binary codes of the entire training set labels generated by Semantic-Network into two semantic dictionaries by removing the duplications, by which the global knowledge stored in semantic dictionaries can seamlessly supervise the feature learning and hashing generation of Image-Network for each sampled mini-batch of input images with asymmetric association. Additionally, we are also motivated to search pairwise similarity at a fine-grained level. To this end, a well-defined margin-scalable pairwise constraint is proposed. Unlike conventional similarity constraint used in many existing methods [20,11,21] with which all the similarity instance pairs are penalized with the same strength, by looking up the semantic dictionaries, our margin-scalable constraint can dynamically penalize instance pairs with respect to their corresponding semantic similarity in fine-grained level (i.e., for a given similarity score of one instance pair, the more identical semantics they share, the larger penalty would be given on them), with which our SADH is empowered to search for discriminative visual feature representations and corresponding combat hashing representations. The main contributions of this paper are as follows:

- [1] We propose a novel end-to-end deep hashing framework which consists of Image-Network and Semantic-Network. With a novel asymmetric guidance mechanism, rich semantic information preserved by Semantic-Network can be seamlessly transferred to Image-Network, which can ensure that the global semantic relevance can be sufficiently discovered and utilized from multi-label annotations of the entire training set.
- [2] We devise a novel margin-scalable pairwise constraint based upon the semantic dictionaries, which can effectively search for precise pairwise similarity information in a semantically fine-grained level to facilitate the discrimination of generated hash codes.

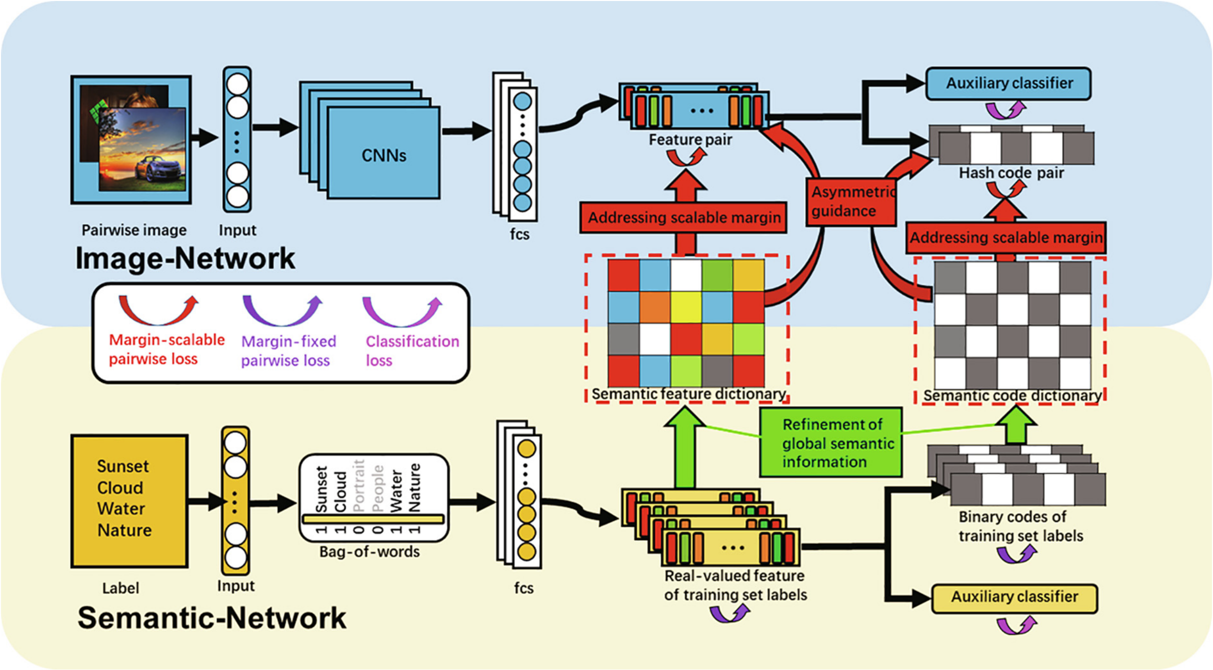


Fig. 1. The overall framework of our proposed SADH, Image-Network plotted in blue background is comprised of CNN layers for deep image representations, while Semantic-Network plotted in yellow background is a self-supervised MLP network which abstracts semantic features from one-hot annotations as inputs. Both networks embeds deep features into a semantic space through a semantic layer, and independently obtain classification outputs and binary codes using multi-task learning framework. Semantic-Network is first trained until convergence, then global semantic information of the entire training set labels is refined by Semantic-Network into two semantic dictionaries, such refined semantic information is transferred to Image-Network by asymmetric guidance on both feature learning and hash code generation. The semantic dictionaries are further utilized to dynamically assign each instance pairs of Image-Network with a scalable margin in the pairwise constraint.

- [3] Without losing generality, we comprehensively evaluate our proposed method on CIFAR-10, NUS-WIDE, MS-COCO, and MIRFlickr-25K to cope with image retrieval task, the effectiveness of proposed modules in our method is endorsed by exhaustive ablation studies. Additionally, we show how to seamlessly extend our SADH algorithm from single-modal scenario to multi-modal scenario. Extensive experiments demonstrate the superiority of our SADH in both image retrieval and cross-modal retrieval, as compared with several state-of-the-art hashing methods.

2. Related work

In this section, we discuss works that are inspiring for our SADH or relevant to four popular research topics in learning to hash.

2.1. Unsupervised hashing methods

The unsupervised hashing methods endeavors to learn a set of hashing functions without any supervised information, they preserve the geometric structure (e.g., the similarity between neighboring samples) of the original data space, by which instance pairs that are close in the original data space are projected into similar hash codes, while the separated pairs in the original data space are projected into dissimilar hash codes. Locality sensitive hashing is the pioneer work of unsupervised hashing, which is first proposed in [31,32], the basic idea of LSH is to learn a family of hashing functions that assigns similar item pairs with a higher probability of being mapped into the same hash code than dissimilar ones. Following [31,32], many variants of LSH has been proposed, e.g., [33–35] extends LSH from the traditional vector-to-vector nearest neighbor search to subspace-to-subspace nearest neighbor search with angular distance as subspace similarity metric. Although LSH can effectively balance computational cost and retrieval accuracy, but it has no exploration on the specific data distributions and often reveals inferior performance. In this paper, we focus on the data-dependent(learning to hash methods). The representative unsupervised learning to hash method includes ITQ [36] which is the first method that learns relaxed hash codes with principal component analysis and iteratively minimize the quantization loss. SH [8] proves the problem of finding good binary code for a given dataset is equivalent to the NP-hard graph partitioning problem, then the spectral relaxation scheme of the original problem is solved by identify the eigenvector solution. LSMH [37] utilizes matrix decomposition to refine the original feature space into a latent feature space which makes both the latent features and binary codes more discriminative, this simultaneous feature learning and hashing learning scheme is followed by many latter methods. JSH [38] collaboratively optimize a regression term which regress high-dimensional features into clustering centroids that are viewed as anchors and a sparse feature selection term. Unlike JSH which utilize anchor-based correlation, our SADH correlate global semantic label features with instance features. Meanwhile, using a deep hashing framework, the proper projection from latent features to hash codes can be adaptively learned.

2.2. Supervised hashing methods

The supervised hashing methods can use the available supervised information such as labels or semantic affinities to guide feature extraction and hash code generation, which can achieve more robust retrieval performance than unsupervised methods. Supervised hashing with kernel (KSH) [6] and supervised discrete hashing (SDH) [39] generate binary hash codes by minimizing the Hamming distance through similar data point pairs. Distortion Minimization Hashing (DMS) [9], Minimum Loss Hashing (MLH) [40]. Binary Reconstruction Embedding(BRE) [9] learns hashing function by minimizing the reconstruction loss to similarities in the original feature space and Hamming space. In [41,42], Support Vector Machine(SVM) is used to learn a set of hyperplanes as a hash function family, by which the margin between the selected support vectors belonging to similar and dissimilar pairs are maximized to generate discriminative binary codes. [43] utilize Binary Matrix Factorization to extract semantic correlations from partially provided social tags. Although the above hashing methods have certainly succeeded to some extent, they all use hand-crafted features that do not fully capture the semantic information and cannot search for similarity information in latent feature space and Hamming space simultaneously, thereby causing suboptimal problem. Recently, the deep learning-based hashing methods have shown superior performance by exploiting the powerful feature extraction of deep learning [40,44–54]. In particular, Convolutional Neural Network Hash (CNNH) [23] is a two-stage hashing method, where the pairwise similarity matrix is decomposed to approximate the optimal hash code representations which can directly guide hash function learning. However, in the two-stage framework of CNNH, the generation of latent features are not participated in the generation of approximate hash codes, so it fails to perform simultaneous feature extraction and hash code learning which limit the discrimination of hash codes. To solve this limitation, Yan et al. [37] improved [23] by equally dividing the latent features into pieces then projecting the pieces of features into the bit-wise representations of hash codes under a one-stage framework. Similarly DSPH[11] performs joint hash code learning and feature learning under a one-stage framework. DDSH [20] adopt an alternative training strategy to optimize the continuous features and binary codes individually. DAGH [55] simultaneously optimize an anchor regression term with the metric loss term, which is beneficial for capturing the global anchor-graph similarity information. Unlike DAGH which focus on tackling the anchor-graph-based scenarios, our proposed SADH mainly focus on capturing more global semantic correlation from multi-label annotations.

Although these methods have obtained satisfactory retrieval performance, they are still suboptimal for multi-label datasets, as they fail to sufficiently discover semantic relevance from multi-label annotations, additionally they only utilize the earlier mentioned coarsely defined similarity supervision(either 0 or 1), which fails to construct more precise pairwise correlations between pairs of hash codes and deep features, significantly downgrading retrieval accuracy. As stated by [56], multi-label images are widely involved in many large-scaled image retrieval systems, so it is valuable to improve the retrieval performance under this scenario. Many recent works are proposed which aim to fully exploit semantic labels in hash function learning. One natural and popular strategy used in a number of recent methods like [24,57–62] is to add an auxiliary classifier in the hashing network to learn the hashing task and classification task simultaneously, which can provide more robust hash function learning by preserving semantic-specific features. A novel and effective methods DSEH [63] utilizes a self-supervised semantic network to capture rich semantic information from semantic labels to guide the feature learning network which learns hash function for images. In comparison with auxiliary classifiers used in [24,57–62], the Semantic-Network used in DESH [63] can capture more complex semantic correlations and can directly supervise the hash code generation, which significantly improves the retrieval performance in multi-label scenarios, however DSEH uses a conventional negative log-likelihood objective function which still cannot search for similarity information in a fine-grained level. Several methods design weighted ranking loss to solve this problem, e.g., HashNet [14] tackle the ill-posed gradient problem of learning discrete hash function by changing the widely used negative log-likelihood objective function [63,11] into a Weighted Maximum Likelihood(WML) estimation. Yan et al. propose an instance-aware hashing framework for multi-label image retrieval in [56], where a weighted triplet loss is included based upon multi-label annotations. Similarly, DSRH [64] designs a Surrogate Loss, in which the NDCG score is calculated which is related to the instance pairs' shared number of labels. However, since both [56,64] design their weighted ranking loss in triplet form, they only consider preserving correct ranking of instances, instead of directly optimizing the multi-level pairwise semantic similarity. IDHN [65] calculate a soft semantic similarity score (i.e., the cosine similarity between label pairs) to replace the hard-assigned semantic similarity metric, which directly perform as the supervision of negative log-likelihood pairwise loss. Although the soft semantic similarity score used in IDHN and the weight factor used in [56,14,64] can reflect multi-level semantic similarity between labels, but they cannot guarantee that the predefined similarity measurement such as NDCG and cosine similarity is the optimal choice for supervising similarity searching of hash codes.

Unlike these methods, we design a new similarity constraint in a contrastive form [66], which contains a margin parameter which can reflect the strength of supervision given on instance pairs. Inspired by DSEH [63], we observe that, using a self-supervised training scheme and taking semantic labels as inputs, Semantic-Network can generate highly discriminative hash codes and its retrieval performance is not sensitive to the selection of hyper-parameter. Taking advantage of these characteristics of Semantic-Network, we consider the pairwise similarity preserved by Semantic-Net- work as the optimum of an ideal hash function, by calculating a scalable margin factor for each item pairs with respect to the corresponding semantic information stored by Semantic-Network, our new similarity constraint can dynamically and accurately penalize the item pairs with respect to multi-level semantic similarity to learn combat hash codes. Note that the margin used in our method is originated form [66], this is different from the hyperplane margin used in SVM-based methods like [41,42], which is maximized between negative and positive support vectors. Additionally, a similar form of contrastive loss function can be also seen in MMHH [67], which also contains a margin value. However different from our SADH, which is mainly focus on multi-label image retrieval, MMHH is focused on alleviating the vulnerability to noisy data. In comparison with our scalable margin, the margin used in MMHH is fixed based on manual selection, which is viewed as Hamming radius to truncate the contrastive loss, preventing it from being excessively large for noisy data.

2.3. Asymmetric hashing methods

Most classical hashing methods build pairwise interaction in symmetric form, recently asymmetric hashing methods have shown the power of learning distinct hash functions and building asymmetric interactions in similarity search. Asymmetric LSH [27] extends LSH to solve the approximate Maximum Inner Product Search (MIPS) problem by generalizing the MIPS problem to an ANN problem with asymmetric transformation. However, asymmetric LSH is data-independent and can hardly achieve satisfactory result. SSAH [68] directly solve the MIPS problem by approximating the full similarity matrix using asymmetric learning structure. [29] theoretically interprets that there is an exponential gap between the minimal binary code length of symmetric and asymmetric hashing. NAMVH [62] learns a real-valued non-linear embedding for novel query data and a multi-integer embedding for the entire database and correlate two distinct embedding asymmetrically. In the deep hashing framework ADSH [30], only query points are engaged in the stage of updating deep network parameters, while the hash codes for database are directly learned as a auxiliary variable, the hash codes generated by the query and database are correlated through asymmetric pairwise constraints, such that the dataset points can be efficiently utilized during the hash function learning procedure. In comparison with [30] building asymmetric association between query and database, notably the cross-modal hashing framework AGAH [69] is devoted to use the asymmetric learning strategy to fully preserve semantic relevance between multi-modal feature representations and their corresponding label information to eliminate modality gap. It constructs asymmetric interaction between binary codes belonging to heterogeneous modalities and semantic labels. Different from AGAH, which separately learns hash function for each single semantics to build asymmetric interaction with modalities, our method leverage a self-supervised network to directly learn hash function for multi-label annotations, which can indicate more fine-grained similarity information. We preserve semantic information from labels of the entire training set, which in turn being refined in form of two semantic dictionaries. Comparing to DSEH [63] which utilize an alternative training strategy and point-to-point symmetric supervision, with the asymmetric guidance of two dictionaries in our method, the global semantic relevance can be more powerfully and efficiently transferred to hash codes and latent feature generated by each sampled mini-batch of images.

2.4. Cross-modal hashing methods

Cross-modal hashing (CMH) has become an active research area since IMH [70] extends the scenario of hashing from similarity search of traditional homogeneous data to heterogeneous data by exploring inter-and-intra consistency and projecting the multi-modality data to a common hamming space. Followed by which a number of CMH methods are proposed, representative unsupervised methods include LSSH [71] which is the first CMH method that simultaneous do similarity search in latent feature space and Hamming space, CMFH [72] uses collective matrix factorization to correlate different modalities and CVH [73]which is the extension of SH for solving cross-view retrieval. Similar to single modal hashing, CMH can achieve more powerful performance with supervised information. SCM [74] is the first attempt to integrate semantic labels into a CMH framework. SePH [75] minimize the Kullback–Leibler (KL) divergence between the pairwise similarity of labels and hash codes. Recently, due to the powerful ability of deep learning in feature extraction, more and more efforts have been devoted to deep cross-modal hashing. Similar to DSPH [11], DCMH [76] and PRDH [77] performs simultaneous feature learning and hash learning under and end-to-end framework. The preservation of semantic relevance is also beneficial for bridging heterogeneous data. CPAH [78] devise an adversarial module and classification module to align the feature distribution and semantic consistency between different modality data. DCE [79] propose a collaborative latent space framework that is capable of dealing with both single-modal and cross-modal hashing tasks. Like DCE, our method SADH is also capable of dealing with both scenarios. DSMHN [80] propose a deep multi-task learning framework with auxiliary classifier, intra-modality and inter-modality similarity constraint. SSAH [68] utilize the self-supervised semantic network in a way that is similar to DSEH, to learn a common semantic space for different modalities. In this paper, although we mainly focus on the single-modal scenario, the core components of our SADH algorithm can be seamlessly integrated in a cross-modal hashing framework. The extension of our method from single-modal to multi-modal scenarios is discussed, and we demonstrate that our SADH can achieve state-of-the-art experimental performance in both scenarios.

3. The proposed method

We elaborate our proposed SADH in details. Firstly, the problem formulation for hash function learning is presented. Afterwards, each module as well as the optimization strategy in the Semantic-Network and Image-Network are explicitly described. As can be seen in the overall framework Fig. 1, SADH consists of two networks, where Semantic-Network is a pure MLP network for semantic preservation with labels in form of bag-of-words as inputs. Image-Network utilizes convolutional neural network to extract high-dimensional visual feature from images, which in turn being projected into binary hash codes, with both deep features (generated by semantic layer) and hash codes (generated by hash layer) under asymmetric guidance of Semantic-Network as shown in Fig. 1.

3.1. Problem definition

First the notations used in the rest of the paper are introduced. Following methods like [63,24,39,64], we consider the common image retrieval scenario where images are annotated by semantic labels, let $O = \{o_l\}_{i=1}^{m}$ denote a dataset with m instances, and $o_l = (v_l, l_l)$ where $v_l \in \mathbb{R}^{1\times d_v}$ is the original image feature from the l-th sample. Assuming that there are C classes in this dataset, o_i will be annotated with multi-label semantic $l_l = [l_{i1}, \ldots, l_{ic}]$, where $l_{ij} = 1$ indicates that o_l belongs to the j-th class, and $l_{ij} = 0$ if not. The image-feature matrix is noted as V, and the label matrix as L for all instances. The pairwise multi-label similarity matrix S is used to describe semantic similarities between each of the two instances, where $S_{i,j} = 1$ means that O_i is semantically similar to O_j, otherwise $S_{i,j} = 0$. In a multi-label setting, two instances $(O_l and O_j)$ are annotated by multiple labels. Thus, we define $S_{i,j} = 1$, if O_l and O_j share at least one label, otherwise $S_{i,j} = 0$. The main goal in deep hashing retrieval is to identify a nonlinear hash function, i.e., $H : o \rightarrow h \in \{-1, 1\}^K$, where K is the length of each hash codes, to encode each item o_l into a K-bit hash code $H_i \in \{-1, 1\}$, whereby the correlation of all item pairs are maintained. The similarity between a hash code pair H_i, H_j are evaluated by their Hamming distance $dis_H(H_i, H_j)$, which might be a challenging and costly calculation [81]. The inner-product $\langle H_i, H_j \rangle$ can be used as a surrogate which relates to hamming distance as follows:

$$dis_H = \frac{1}{2}\left(K - \langle H_i, H_j \rangle\right). \tag{1}$$

3.2. Self-supervised semantic network

To enrich the semantic information in generated hash codes, we designed a self-supervised MLP network namely Semantic-Network to leverage abundant semantic correlations from multi-label annotations, the semantic information preserved by Semantic-Network will be further refined to perform as the guidance of the hash function learning process of Image-Network.

Semantic-Network extracts high-dimensional semantic features through fully-connected layers with multi-label annotations as inputs $\left(\text{i.e.}, H_i^l = f^l\left(l_l, \theta^l\right)\right)$, where f^l is the nonlinear hash function for Semantic-Network, while θ^l denotes the parameters for Semantic-Network. With a sign function the learned H^l can be discretized into binary codes:

$$B^l = \text{sign}\left(H^l\right) \in \{-1, 1\}^K. \tag{2}$$

For comprehensive preservation of semantic information especially in multi-label scenarios, the abstract semantic features $F^l = \left[F_l^l, \ldots, F_n^l\right]$ of Semantic-Network are also exploited to supervise the semantic learning of Image-Network.

3.2.1. Cosine-distance-based similarity evaluation

In Hamming space, the similarity of two hash codes H_i, H_j can be defined by the Hamming distance $dist_H(*, *)$. To preserve the similarity of item pairs, whereby similar pairs are clustered and dissimilar pairs scattered, a similarity loss function of Semantic-Network is defined as follows:

$$J_s = \sum\nolimits_{i,j=1}^{n}\left(s_{i,j}\text{dis}_H(H_i, H_j) + (1 - s_{i,j})\max\left(m - \text{dis}_H(H_i, H_j), 0\right)\right) \tag{3}$$

Where J_s denotes the similarity loss function, by which the similarity of two generated hash codes H_i and H_j can be preserved. $\text{dis}_H(H_i, H_j)$ represents the Hamming distance between H_i and H_j. To avoid the collapsed scenario [21], a contrastive form of loss function is applied with a margin parameter m, with which the hamming distance of generated hash code pairs are expected to be less than m. With the mentioned relationship Eq. (1) between Hamming distance and inner-product, the similarity loss can be redefined as:

$$J_s = \frac{1}{2}\sum\nolimits_{i,j=1}^{n}\left(s_{i,j}\max\left(m - \langle H_i, H_j\rangle, 0\right) + (1 - s_{i,j})\max\left(m + \langle H_i, H_j\rangle, 0\right)\right) \tag{4}$$

where the margin parameter induce the inner-product of dissimilar pairs to be less than $-m$, while that of similar ones to be larger than m, note that this form of contrastive similarity constraint derives from [66] where margin is a hyper-parameter which is different from the hyper-plane margin used in SVM-based methods [41,42]. For enhancement of similarity preservation, we expect the similarity constraint to be extended by ensuring the discrimination of deep semantic features. However because of the difference between the distributions of features from Semantic-Network and Image-Network, the inner-product $\langle ., . \rangle \in (-\infty, \infty)$ will no longer be a plausible choice for the similarity evaluation between the semantic features of the two networks. As the choice of margin parameter m is ambiguous. One way to resolve this flaw is to equip the two networks with the same activate function, for example a sigmoid or tanh, at the output of the semantic layer to limit the scale of output features to a fixed range, nevertheless we expect both of the networks to maintain their own scale of feature representations. Considering the fact that hash codes are discretized to either -1 or 1 at each bit, meanwhile all generated hash codes have the same length K, therefore in the similarity evaluation in Hamming space, we choose to focus more on the angles between hash codes, instead of the absolute distance between them. Hence we adopt the cosine distance $\cos(., .)$ as a replacement:

$$\cos\left(H_i, H_j\right) = \frac{< H_i, H_j >}{\|H_i\|\|H_j\|} \tag{5}$$

where $\cos\left(H_i, H_j\right) \in (-1, 1)$. Although pairwise label information is adopted to store the semantic similarity of hash codes, the label information is not fully exploit. Thus Semantic-Network will further exploit semantic information with an auxiliary classifier as shown in Fig. 1. Many recent works directly map the learned binary codes into classification predictions by using a linear classifier [24,63]. To prevent the interference between the classification stream and hashing stream, and to avoid the classification performance being too sensitive to the length of hash codes, we jointly learn the classification task and hashing task under a multi-task learning scheme without mutual interference [82,83].

The final object function of Semantic-Network can be formulated as:

$$\begin{aligned}&\min_{B^l,\theta^l,\hat{L}^l} J_{Lab}\\ =\ & \alpha J_1+\lambda J_2+\eta J_3+\beta J_4\\ =\ & \tfrac{\alpha}{2}\sum_{i,j=1}^{n} S_{i,j}\max\left(m-\Delta_{i,j}^{l},0\right)+\left(1-S_{i,j}\right)\max\left(m+\Delta_{i,j}^{l},0\right)\\ +\ & \tfrac{\lambda}{2}\sum_{i,j=1}^{n} S_{i,j}\max\left(m-\Gamma_{i,j}^{l},0\right)+\left(1-S_{i,j}\right)\max\left(m+\Gamma_{i,j}^{l},0\right)\\ +\ & \eta\left|\left|\hat{L}^l-L\right|\right|_2^2+\beta\left|\left|H^l-B^l\right|\right|_2^2\end{aligned} \tag{6}$$

where the margin is a manually-selected hyper-parameter $m \in (0,1)$. Taking semantic labels as inputs and being trained in self-supervised manner, it's relatively easy for Semantic-Network to achieve robust retrieval accuracy, and it's performance is not sensitive to the selection of margin value, with respect to the sensitivity analysis latter in 4.3.2., it can consistently achieve robust performance when m is relatively small, so we directly set it as 0 in experiments. J_l and J_2 are the similarity loss for the learned semantic features and hash codes respectively with $\Delta_{i,j}^{l}=\cos\left(F_i^l,F_j^l\right),\Gamma_{i,j}^{l}=\cos\left(H_i^l,H_j^l\right)$. The classification loss J_3 calculates the difference between input labels and predicted labels. J_4 is the quantization loss for the discretization of learned hash codes.

3.2.2. Global semantic preservation with semantic dictionary construction

In existing self-supervised hashing methods [63,68], the self-supervised network normally guides the deep hashing network with a symmetric point-to-point strategy, hash codes generated by one mini-batch of image are directly associated with the hash codes generated by the corresponding mini-batch of labels. Under such mechanism, the global semantic information is insufficiently transferred to deep hashing network, meanwhile the similarity search process excessively focus on the semantics that frequently appear, whereas the semantics with lower frequency of occurrence are relatively neglected. In this paper, we are motivated to alleviate the mentioned drawbacks of existing guidance mechanism. Inspired by methods like [30,69], we seek for constructing an asymmetric interaction between the output features of deep hashing network and the global semantic information to significantly empowered the effectiveness of similarity search. In this section we first discuss how such global semantic information can be extracted and refined into two semantic dictionaries.

As illustrated in Fig. 1, we first train Semantic-Network parameters until convergence to minimize J_{lab} w.r.t. the objective function Eq. (6), i.e., $\tilde{\theta}^l=\arg\min_{\theta^l} J_{lab}$. Next we fix the network parameter $\tilde{\theta}^l$ of Semantic-Network and use it to refine the global semantic information from the multi-label annotations of the entire dataset into a semantic code dictionary U and a corresponding semantic feature dictionary Q, this process can be formulated as follows:

$$U=sign\left(\sigma\left(E_U^l\left(\tilde{\theta}^l,unique\left(\tilde{L}\right)\right)\right)\right) \tag{7}$$

$$Q=E_Q^l\left(\tilde{\theta}^l,unique\left(\tilde{L}\right)\right) \tag{8}$$

where σ is the tanh function. E_U^l and E_Q^l are binary code encoder and semantic feature encoder respectively using Semantic-Network. $\tilde{L}$ represents the multi-label annotations of the entire training set. $unique(\cdot)$ is the deduplication operation which summarizes the multi-label annotations into unique representations. Note that $U=\{u_i\}_{i=1}^{C}$ where $u_i \in [-1,1]$ and $Q=\{q_i\}_{i=1}^{C}$, where C is the total number of deduplicated training set labels.

3.3. Deep feature learning network with asymmetric guidance

We apply an end-to-end convolutional neural network namely Image-Network for image feature learning, which can extract and embed deep visual features from images into high dimensional semantic features and simultaneously project them into output representations for multi-label classification task and hashing task, similar to Semantic-Network, two tasks are learned simultaneously under a multi-task learning framework. The semantic feature extraction and hash function learning of Image-Network will be guided by the semantic dictionaries U and Q using an asymmetric learning strategy, the asymmetric similarity constraint can be formulated as follows:

$$\begin{aligned}J_s=\sum_{i=1}^{n}\sum_{j=1}^{c}\frac{1}{2}&\left(s_{i,j}\max\left(m-\cos\left(H_i,u_j\right),0\right)\right.\\ &\left.+\left(1-s_{i,j}\right)\max\left(m+\cos\left(H_i,u_j\right),0\right)\right)\end{aligned} \tag{9}$$

where $s_{i,j}$ is an asymmetric affinity matrix.

3.3.1. Margin-scalable constraint

In most contrastive or triplet similarity constraints used in deep hash methods[84,85,30], the choice of the margin parameter mainly relies on manual tuning. As demonstrated in Section 4.3.2, we observe that, in comparison with the self-supervised Semantic-Network, the deep Image-Network is fairly sensitive to the choice of margin, which means that a good selection of margin is valuable for robust hash function learning. Additionally, in multi-label scenarios, it would be more desirable if the margin can be scaled to be larger for item pairs that share more semantic similarities than those less semantically similar pairs, in this case the scale of margin can be equivalent to the strength of constraint. Thus setting a single fixed margin value may downgrade the storage of similarity information. Holding the motivation of dynamically selecting optimized margin for each sampled instance pairs with respect to their exact degree of semantic similarity, we propose a margin-scalable similarity constraint based on the semantic maps generated by Semantic-Network. Relying on the insensitivity of Semantic-Network to selection of margin, we leverage information in semantic dictionaries to calculate scalable margin and to indicate relative semantic similarity, i.e., for two hash codes H_i^v and H_j^v generated by Image-Network, a pair of corresponding binary codes $u_{H_i^v}$ and $u_{H_j^v}$ are represented by addressing the semantic code map U with their semantic labels as index. The scalable margin M_{H_i,H_j} for H_i^v and H_j^v is calculated by:

$$M_{H_i,H_j}=\max\left(0,\cos\left(u_{H_i^v},u_{H_j^v}\right)\right) \tag{10}$$

As $\cos(.,.) \in (-1,1)$, a positive cosine distance between item pairs in the semantic code dictionary will be assigned to similar item pairs and will be used by Image-Network to calculate their scalable margin, while the negative cosine distances will scale the margin to 0. This is due to the nature of multi-label tasks, where the 'dissimilar' situation only refers to item pairs with none identical label. While for a similar item pair, the number of shared labels may come from a wide range. Thus in similarity preservation, dissimilar items are given a weaker constraint, whereas the similar pairs are constrained in a more precise and strict way. For two sampled sets of hash codes or semantic features G_1 and G_2 with size of n_1 and n_2, the margin-scalable constraint J_{ms} can be given by:

$$\begin{aligned} & J_{ms}(G_1, G_2) \\ & = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \tfrac{1}{2}\left(S_{i,j} max\left(M_{G_i,G_j} - \cos\left(G_1^i, G_2^j\right), 0\right)\right. \\ & \left. + (1 - S_{i,j}) max\left(M_{G_i,G_j} - \cos\left(G_1^i, G_2^j\right), 0\right)\right) \end{aligned} \tag{11}$$

The final object function of Image-Network can be formulated as:

$$\begin{aligned} & \min_{B^v, \theta^v, \widehat{L}^v} J_{\text{Img}} \\ = \; & \alpha J_1 + \lambda J_2 + \alpha J_3 + \lambda J_4 + \eta J_5 + \beta J_6 \\ = \; & \alpha J_{ms}(F^v, F^v) + \lambda J_{ms}(H^v, H^v) + \alpha J_{ms}(F^v, Q) \\ + \; & \lambda J_{ms}(H^v, U) + \eta \|\widehat{L}^v - L\|_2^2 + \beta \|H^v - B^v\|_2^2 \end{aligned} \tag{12}$$

where J_l and J_2 are margin-scalable losses for semantic features and hash codes generated by Image-Network, with symmetric association between instance pairs. J_3 and J_4 are margin-scalable losses with asymmetric guidance of semantic dictionaries U and Q on hash codes and semantic Features generated by Image-Network. J_5 and J_6 are classification loss and quantization loss similarly defined in Semantic-Network.

3.4. Optimization

It is noteworthy to mention that, the Image-Network is trained after the convergence of Semantic-Network is obtained. First we iteratively optimize the objective function Eq. (6) by exploring multi-label information to learn θ^l, H^l and $\widehat{L}^l$. With the finally trained Semantic-Network we obtain U and Q. Then the parameters of Semantic-Network will be fixed, and L_{img} will be optimized through θ^v, H^v and $\widehat{L}^v$ with the guidance of U and Q. Finally, we obtain binary hash codes $B = \text{sign}(H^v)$. The entire learning algorithm is summarized in Algorithm 1 in more detail.

3.4.1. Optimization of semantic-network

The gradient of J_{Lab} w.r.t each Hash code H_i^l in sampled mini-batch is

$$\frac{\partial J_{Lab}}{\partial H_i^l} = \begin{cases} \sum_{j=1 s_{i,j}=1}^{n} \frac{\lambda}{2}\left(m - \frac{H_j^l}{\|H_i^l\|\|H_j^l\|} + \frac{H_i^l \Gamma_{i,j}^l}{\|H_i^l\|_2^2}\right) + 2\beta\left(H_i^l - B_i^l\right) \\ \qquad \text{if } s_{i,j} = 1, \; \Gamma_{i,j}^l < m \\ \sum_{j=1 s_{i,j}=0}^{n} \frac{\lambda}{2}\left(m + \frac{H_j^l}{\|H_i^l\|\|H_j^l\|} - \frac{H_i^l \Gamma_{i,j}^l}{\|H_i^l\|_2^2}\right) + 2\beta\left(H_i^l - B_i^l\right) \\ \qquad \text{if } s_{i,j} = 0, \; \Gamma_{i,j}^l > -m \end{cases} \tag{13}$$

where $\Gamma_{i,j}^l = \cos\left(H_i^l, H_j^l\right)$. $\frac{\partial J_{Lab}}{\partial F_j^l}$ can be obtained similarly, $\frac{\partial J_{Lab}}{\partial \theta^l}$ can be computed by using the chain rule, then θ^l can be updated for each iteration using Adam with back propagation.

3.4.2. Optimization of image-network

The gradient of J_{Img} w.r.t each Hash code H_i^v in sampled mini-batch is

$$\frac{\partial J_{\text{Img}}}{\partial H_i^v} = \lambda \frac{\partial J_2}{\partial H_i^v} + \lambda \frac{\partial J_4}{\partial H_i^v} + \beta \frac{\partial J_6}{\partial H_i^v} \tag{14}$$

where

$$\frac{\partial J_2}{\partial H_i^v} = \begin{cases} \sum_{j=1 s_{i,j}=1}^{n} \frac{1}{2}\left(M_{H_i,H_j} - \frac{H_j^v}{\|H_i^v\|\|H_j^v\|} + \frac{H_i^v \Gamma_{i,j}^v}{\|H_i^v\|_2^2}\right) \\ \qquad \text{if } s_{i,j} = 1, \; M_{H_i,H_j} > \Gamma_{i,j}^v \\ \sum_{j=1 s_{i,j}=0}^{n} \frac{1}{2}\left(\frac{H_i^v \Gamma_{i,j}^v}{\|H_i^v\|_2^2} - \frac{H_j^v}{\|H_i^v\|\|H_j^v\|} - M_{H_i,H_j}\right) \\ \qquad \text{if } s_{i,j} = 0, \; M_{H_i,H_j} > \Gamma_{i,j}^v \end{cases}$$

where $\Gamma_{i,j}^v = \cos\left(H_i^v, H_j^v\right)$. $\frac{\partial J_6}{\partial H_i^v} = 2\left(H_i^l - B_i^l\right)$, the calculation of $\frac{\partial J_4}{\partial H_i^v}$ resembles $\frac{\partial J_2}{\partial H_i^v}$, $\frac{\partial J_{\text{Img}}}{\partial F_i^v}$ can be obtained similarly to $\frac{\partial J_{\text{Img}}}{\partial H_i^v}$, $\frac{\partial J_{\text{Img}}}{\partial \theta^v}$ can be computed by using the chain rule, then θ^v can be updated for each iteration using SGD with back propagation.

Algorithm 1: The learning algorithm of our *SADH*

```
Input:
  Image set $V$, Label set $L$
Output:
  semantic feature map $Q$, and semantic code map $U$,
  parameters $\theta^v$ for Image-Network,
  Optimal code matrix for Image-Network $B^v$
Initialization:
  Initialize network parameters $\theta^l$ and $\theta^v$
  Hyper-parameters: $\alpha, \lambda, \eta, \beta, m$
  Mini-batch size $M$, learning rate: $lr$
  maximum iteration numbers $t^l, t^v$
Stage1: Hash learning for the self-supervised network
  (Semantic-Network)
  for $t^l$ iteration do
    Calculate derivative using Eq. (13)
    Update $\theta^l$ by using Adam and back propagation
  end for
  Update semantic feature map $Q$ and semantic code map
U by Semantic-Network for each semantic as input
Stage2: Hash learning for the feature learning network
  (Image-Network)
  for $t^v$ iteration do
    Calculate derivative using Eq. (14)
    Update $\theta^v$ by using SGD and back propagation
  end for
  Update the parameter $B^v$ by $B^v = \text{sign}(H^v)$
```

3.5. Extension to cross-modal hashing

As mentioned in Section 2.4, hashing in Cross-modal scenarios has arouse extensive attention of many researchers, in which a common Hamming space is expected to be learned to perform mutual retrieval between data of heterogeneous modalities. In this paper, we mainly consider the single-modal retrieval of image data, but the flexibility of margin-scalable constraint and asymmetric guidance mechanism allows us to readily extend our SADH algorithm to achieve cross-modal hashing. Suppose the training instances consists of N different modalities, with corresponding hash codes $H^j, j = 1, \ldots, N$, and semantic features $F^j, j = 1, \ldots, N$. Then the extension of our proposed method in Eq. (4) can be formulated as:

$$\begin{aligned}&\min_{B^j,\theta^j,\widehat{L^j}}\sum_{j=1}^{N}\alpha J_{ms}\left(F^j,F^j\right)+\lambda J_{ms}\left(H^j,H^j\right)\\&+\alpha J_{ms}\left(F^j,Q\right)+\lambda J_{ms}\left(H^j,U\right)\\&+\eta\|\widehat{L^j}-L^j\|_2^2+\beta\|H^j-B^j\|_2^2\end{aligned}\tag{15}$$

Without loss of generality, following methods like [76–78,28], we focus on cross-modal retrieval for bi-modal data (i.e., image and text) in experimental analysis.

4. Experiments and analysis

In this section, we conducted extensive experiments to verify three main issues of our proposed SADH method: (1) To illustrate the retrieval performance of SADH compared to existing state-of-the-art methods. (2) To evaluate the improvements of efficiency in our method compared to other methods. (3) To verify the effectiveness of different modules proposed in our method.

4.1. Datasets and experimental settings

The evaluation is based on four mainstream image retrieval datasets: CIFAR-10 [86], NUS-WIDE [17], MIRFlickr-25K [19], MS-COCO [87].

CIFAR-10: CIFAR-10 contains 60,000 images with a resolution of 32 × 32. These images are divided into 10 different categories, each with 6,000 images. In the CIFAR-10 experiments, following [88], we select 100 images per category as testing set(a total of 1000) and query set, the remaining as database(a total of 59,000), 500 images per category are selected from the database as a training set(a total of 5000).

NUS-WIDE: NUS-WIDE contains 269,648 image-text pairs. This data set is a multi-label image set with 81 ground truth concepts. Following a similar protocol as in [24,88], we use the subset of 195,834 images which are annotated by the 21 most frequent classes (each category contains at least 5,000 images). Among them, 100 image-text pairs and 500 image-text pairs are randomly selected in each class as the query set (2100 in total) and the training set (10,500 in total), respectively. The remaining 193,734 image-text pairs are selected as database.

MIRFlickr-25K: The MIRFlickr25K dataset consists of 25,000 images collected from the Flickr website. Each instance is annotated by one or more labels selected from 38 categories. We randomly selected 1,000 images for the query set, 4,000 images for the training set and the remaining images as the retrieval database.

MS-COCO: The MS-COCO dataset consists of 82,783 training images and 40,504 validation images, each image is annotated with at least one of the 80 semantics, we combine the training set and validation set and prune the images with no categories, which gives us 122,218 images. We randomly selected 5,000 images for the query set, 10,000 images for the training set and the remaining images as the retrieval database. For cross-modal retrieval, the text instances are presented in form of 2028-dimensional Bag-of-Word vectors.

For image retrieval, we compare our proposed SADH with several state-of-the-art approaches including LSH [1], SH [8], ITQ [2], LFH [89], DSDH [24], HashNet [14], DPSH [11], DBDH [90], CSQ [91] and DSEH [63] on all the four datasets. For cross-modal retrieval, we compare our SADH with 3 state-of-the-art deep cross-modal hashing frameworks including DCMH[76], PRDH[77], SSA-H[68]. These methods are briefly introduced as follows:

1. Locality-Sensitive Hashing (LSH) [1] is a data-inde-pendent hashing method that employs random projections as hash function.
2. Spectral Hashing (SH) [8]is a spectral method which transfers the original problem of finding the best hash codes for a given dataset into the task of graph partitioning.
3. Iterative quantization (ITQ) [2] is a classical unsupervised hashing method. It projects data points into a low dimensional space by using principal component analysis (PCA), then minimize the quantization error for hash code learning.
4. Latent Factor Hashing (LFH) [89] is a supervised method based on latent hashing models with convergence guarantee and linear-time variant.
5. Deep Supervised Discrete Hashing (DSDH) [24] is the first supervised deep hashing method that simultaneously utilize both semantic labels and pairwise supervised information, the hash layer in DSDH is constrained to be binary codes.
6. HashNet [14] is a supervised deep architecture for hash code learning, which includes a smooth activation function to resolve the ill-posed gradient problem during training.
7. Deep pairwise-supervised hashing (DPSH) [11] is a representative deep supervised hashing method that jointly performs feature learning and hash code learning for pairwise application.
8. Deep balanced discrete hashing for image retrieval (DBDH) [90] is a recent supervised deep hashing method which uses a straight-through estimator to actualize discrete gradient propagation.
9. Central Similarity Quantization for Efficient Image and Video Retrieval (CSQ) [91] defines the correlation of hash codes through a global similarity metric, to identify a common center for each hash code pairs.
10. Deep Joint Semantic-Embedding Hashing (DSEH) [63] is a supervised deep hashing method that employs a self-supervised network to capture abundant semantic information as guidance of a feature learning network.
11. Deep cross modal hashing (DCMH)[76] is a supervised deep hashing method that integrates feature learning and hash code learning in an end-to-end framework.
12. Pairwise Relationship Guided Deep Hashing (PRDH) [77] is a supervised deep hashing method that utilize both intra-modal and inter-modal pairwise constraints to search for similarity information.
13. Self-supervised adversarial hashing networks for cross-modal retrieval(SSAH) [68] is a deep supervised cross-modal method that utilize a self-supervised network to constitute a common semantic space to bridge data from image modality and text modality.

Among the above approaches, LSH [1], SH [8], ITQ [2], LFH [89] are non-deep hashing methods, for these methods, 4096-dimentional deep features extracted from AlexNet [44] and 2048-dimentio-nal deep features extracted from ResNet50 [40] are utilized for two datasets: NUS-WIDE and CIFAR-10 as inputs, when AlexNet features are used the baseline is named as 'method name-A'. When ResNet50 features are used, the baseline is named as 'method name-R'. The other six baselines (i.e., DSDH, HashNet, DPSH, DBDH and DSEH) are deep hashing methods, for which images on three dataset (i.e., NUS-WIDE, CIFAR-10 and MIRFlickr-25 k) are resized to 224 × 224 and used as inputs. LSH, SH, ITQ, LFH, DSDH, HashNet, DPSH, DCMH and SSAH are carefully carried out based on the source codes provided by the authors, while for the rest of the methods, they are carefully implemented by ourselves using parameters as suggested in the original papers.

We evaluate the retrieval quality by three widely used evaluating metrics: Mean Average Precision (MAP), Precision-Recall curve, and Precision curve with the number of top returned results as variable (topK-Precision).

Specifically, given a query instance q, the Average Precision (AP) is given by: $AP(q) = \frac{1}{n_q}\sum_{i=1}^{n_{database}} P_{q_i} I(i)$, where $n_{database}$ is the total number of instances in the database, n_q is the number of similar samples, P_{q_i} is the probability of instances of retrieval results being similar to the query instance at cut-off i, And $I(i)$ is the indicator function that indicates the i-th retrieval instance I is similar to query image to q, if $I(i) = 1$, and $I(i) = 0$ otherwise.

The larger the MAP is, the better the retrieval performance. Since NUS-WIDE is relatively large, we only consider the top 5,000 neighbors (MAP@5000), when computing MAP for NUS-WIDE, while for CIFAR-10 and MIRFlickr-25 K, we calculate MAP for the entire retrieval database (MAP-@ALL).

4.2. Implementation details

Semantic-Network is built with four fully-connected layers, with which the input labels are transformed into hash codes ($L \rightarrow 4096 \rightarrow 2048 \rightarrow N$). Here the output includes both the K-dimensional hash code and the C-dimensional multi-label predictions, $N = K + C$.

We built ImageNet based on ResNet50 [40], the extracted visual features of ResNet are embedded into 2048-dimensional semantic features, which is followed by the two extra layers (i.e., Hash layer and Classification layer) with K nodes for hash code generation and C nodes for classification. It is noted that except for output layers, the network is pre-trained on ImageNet dataset. The implementation of our method is based on the Pytorch framework and executed on NVIDIA TITAN X GPUs for 120 epochs of training.

The Adam optimizer [92] is applied to Semantic-Network, while the stochastic Gradient descent (SGD) method is applied to Image-Network. The batch size is set to 64. The learning rates are chosen from 10^{-3} to 10^{-8} with a momentum of 0.9.

4.3. Performance evaluation

4.3.1. Comparison to state of the art

To validate the retrieval performance of our method for image retrieval, we compare the experimental results of SAD-H with other state-of-the-art methods including LSH [1], SH [8], ITQ [2], LFH [89], DSDH [24], HashNet [14], DPSH [11], DBDH [90], CSQ [91] and DSEH [63] on CIFAR-10, NUS-WIDE, MIRFlickr-25 K and MS-COCO. Table 1 shows the top 10 retrieved images in database for 3 sampled images in MIRFlickr-25 K, it can be observed that in difficult cases, SADH reveals better semantic consistency than HashNet. Table 2 to Table 5 report the MAP results of different methods, note that for NUS-WIDE, MAP is calculated for the top 5000 returned neighbors. Fig. 2–7 show the overall retrieval performance of SADH compared to other baselines in terms of

Table 1
Examples of top 10 retrieved images by SADH and DSDH on MIRFlickr-25 K for 48 bits. The semantically incorrect images are marked with a red border.

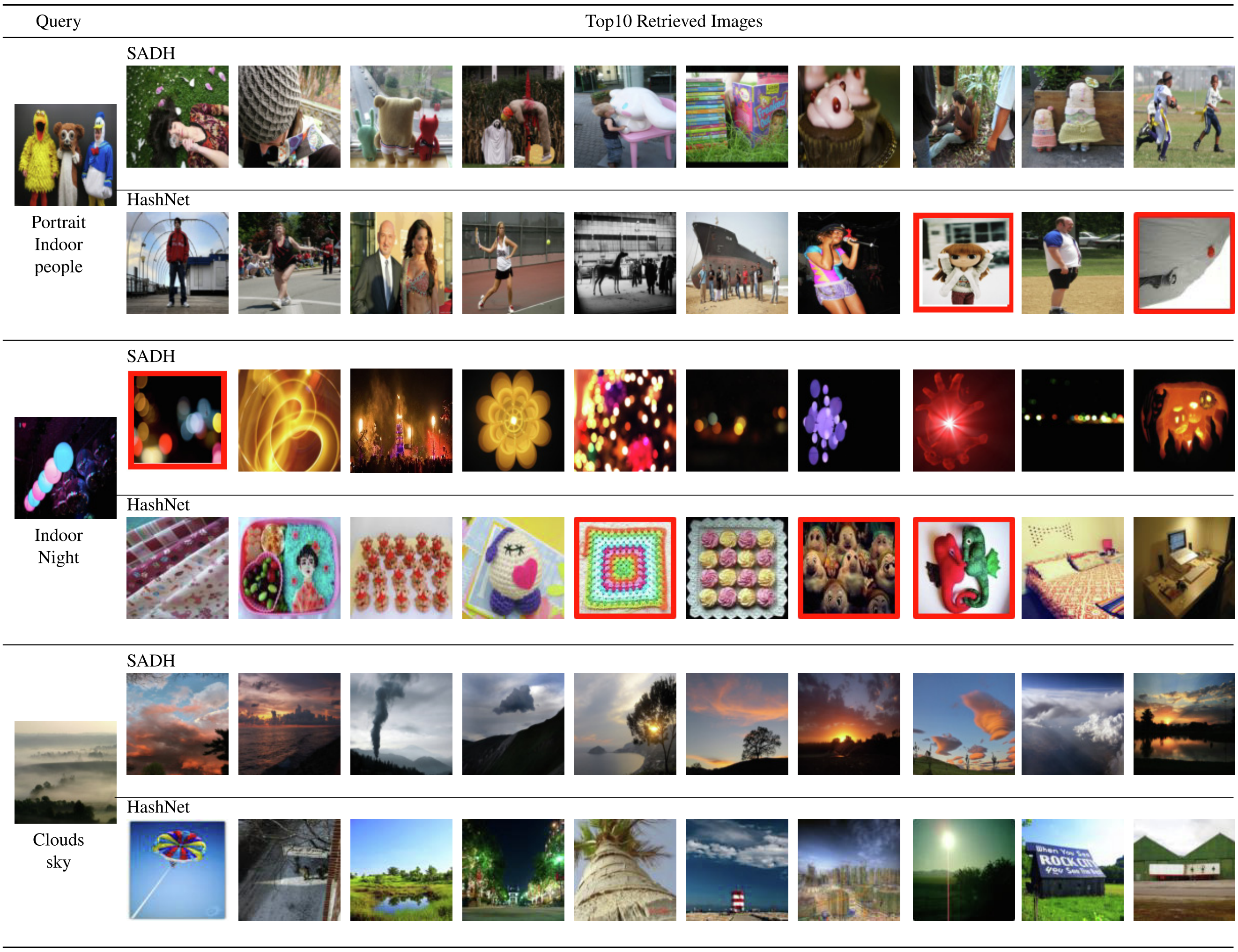


Table 2
MAP@ALL on CIFAR-10 for image retrieval.

Method	CIFAR-10 (MAP@ALL)			
	16 bits	32 bits	48 bits	64 bits
LSH-A [1]	0.4443	0.5302	0.5839	0.6326
ITQ-A [2]	0.2094	0.2355	0.2424	0.2535
SH-A [8]	0.1866	0.1900	0.2044	0.2020
LFH-A [89]	0.1599	0.1608	0.1705	0.1693
LSH-R [1]	0.1017	0.1018	0.1017	0.1020
ITQ-R [2]	0.1038	0.1040	0.1041	0.1043
SH-R [8]	0.1036	0.1031	0.1028	0.1029
LFH-R [89]	0.1041	0.1049	0.1048	0.1046
DSDH [24]	0.7514	0.7579	0.7808	0.7690
HashNet [14]	0.6975	0.7821	0.8045	0.8128
DPSH [11]	0.7870	0.7807	0.7982	0.8003
DBDH [90]	0.7892	0.7803	0.7797	0.7914
CSQ [91]	0.7761	0.7775	–	0.7741
DSEH [63]	0.8025	0.8130	0.8214	0.8301
SADH	**0.8755**	**0.8832**	**0.8913**	**0.8783**

precision-recall curve and precision curves by varying the number of top returned images, shown from 1 to 1000, on NUS-WIDE, CIFAR-10, MS-COCO and MIRFlickr-25K respectively. SADH substantially outperforms all other state-of-the-art methods. It can be noticed that SADH outperforms other methods for almost all the leng-ths of hash bits with a steady performance on both datasets. This is due to the multi-task learning structure in our method with which the classification output and hashing output are obtained independently, and the two tasks are not mutually interfered. It is also noteworthy that, with abundant semantic information leveraged from the self-supervised network and the pairwise information derived from the margin-scalable constraint, SADH obtained an impressive retrieval performance on both single-label datasets and multi-label datasets as shown in Table 3,Table 4.

4.3.2. Sensitivity analysis

Four hyper-parameters $\alpha, \lambda, \gamma, \beta$ are selected in the objective functions Eq. (6) and Eq. (12). Here we examine the effect of different selections of these hyper-parameters on the performance of SADH in a range between $1e-2$ and $1e+2$. As shown in Fig. 8, the performance of SADH is relatively robust to the selection of α, λ and γ, better performance can be obtained when the discretization strength β is smaller than 1. The best performance can be achieved when $\alpha = 0.1, \lambda = 1, \gamma = 0.1, \beta = 0.01$.

To illustrate the earlier mentioned difference of two networks' sensitivity to margin parameter in contrastive loss, we replace the scalable margin module in Image-Network by margin constant m in Semantic-Network and report their MAP with 48-bit length under different choices of m on CIFAR-10 and MIRFlicker-25 K. As shown in Fig. 9, we can see that under different choices of margin, Semantic-Network reveals relatively slight changes in MAP, and it's performance is consistently robust when m is relatively small, so we set m as 0 for all the scenarios. While Image-Network is highly sensitive to the choice of margin with a largest MAP gap of roughly 0.14 at margin = 0 and margin = 0.2. Which to some extend reveals the significance of proper selection of margin and the feasibility of calculating margin for different item pairs rely on the hash codes generated by Semantic-Network based on the insensitivity of its performance to the selection of margin parameter.

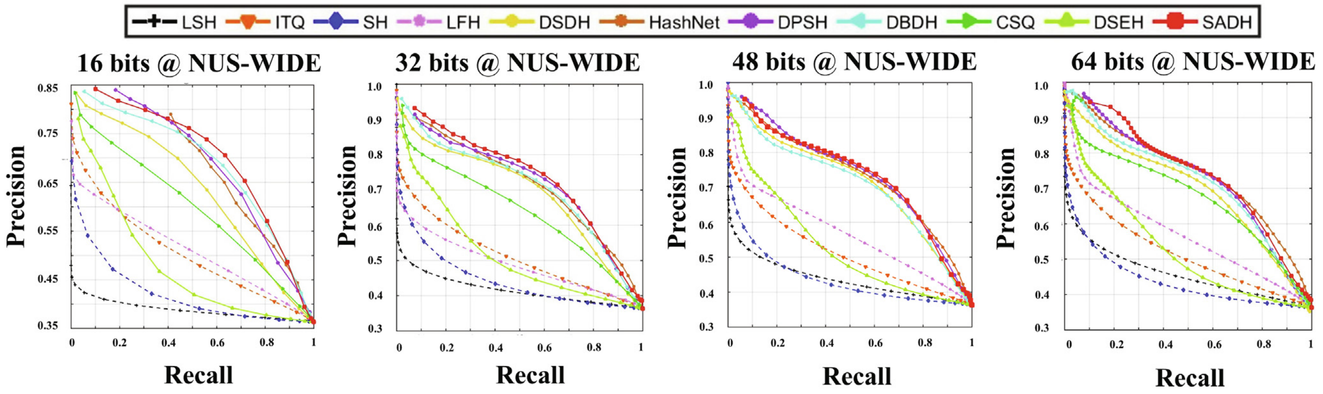


Fig. 2. precision-recall curves on NUS-WIDE.

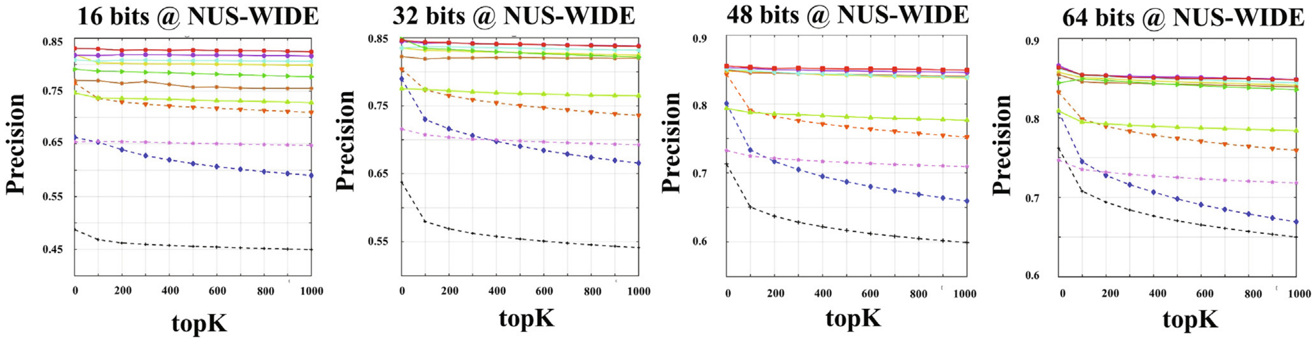


Fig. 3. TopK-precision curves on NUS-WIDE.

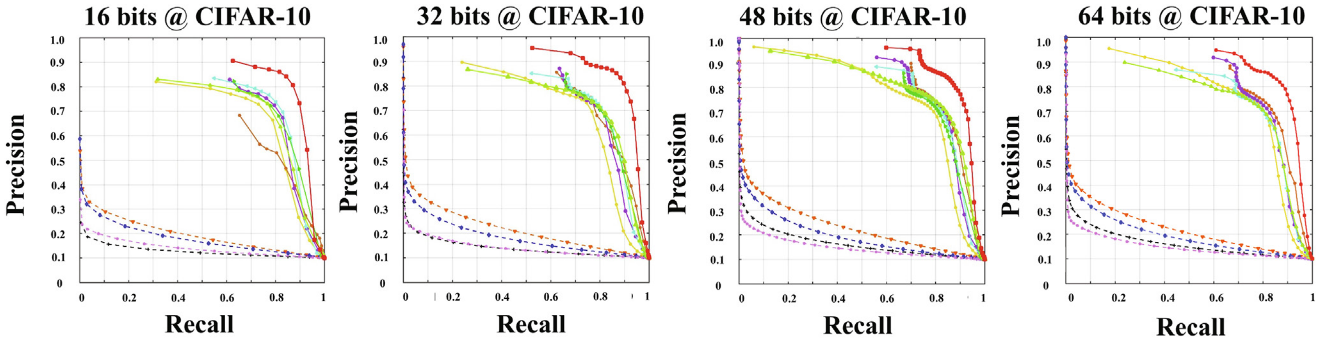


Fig. 4. precision-recall curves on CIFAR-10.

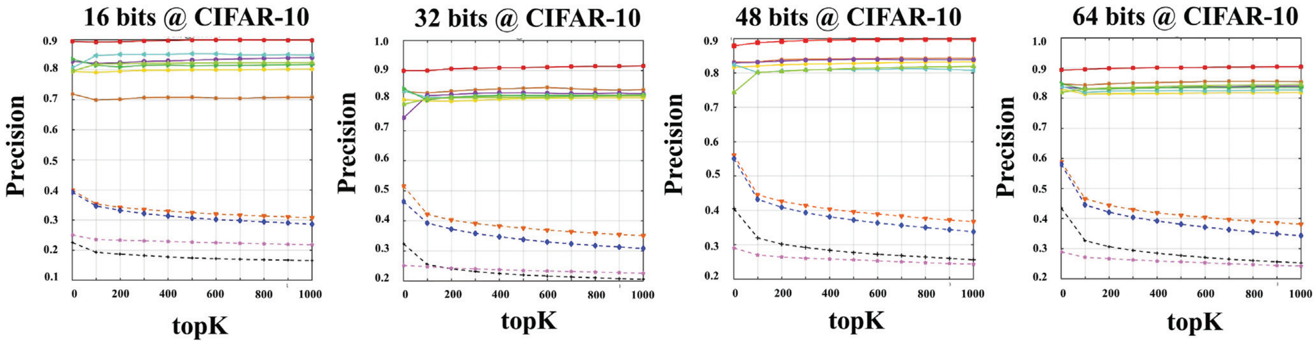


Fig. 5. TopK-precision curves on CIFAR-10.

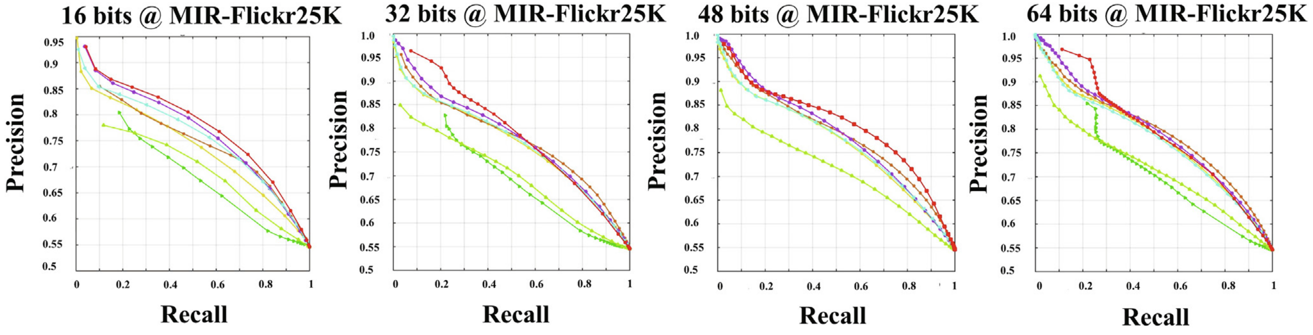


Fig. 6. precision-recall curves on MIR-Flickr25K.

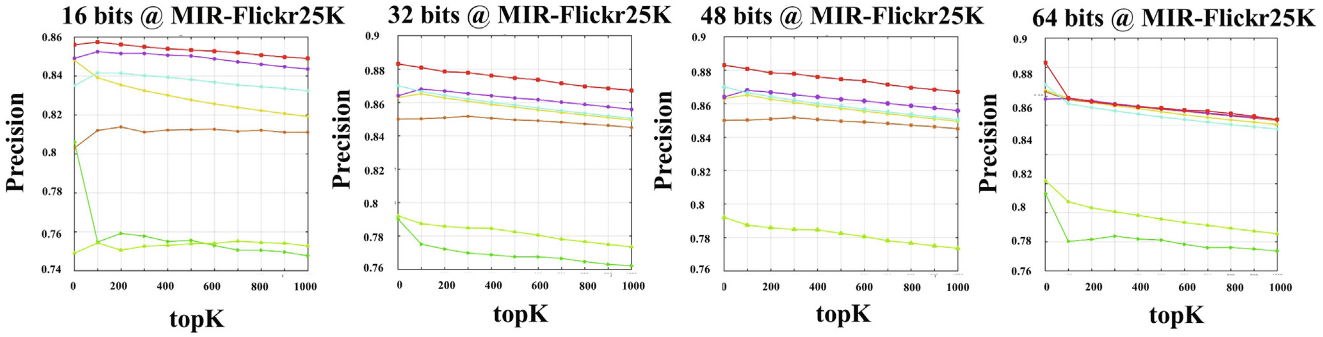


Fig. 7. TopK-precision curves on MIR-Flickr25K.

Table 3
MAP@5000 on NUS-WIDE for image retrieval.

Method	NUS-WIDE (MAP@5000))			
	16 bits	32 bits	48 bits	64 bits
LSH-A [1]	0.4443	0.5302	0.5839	0.6326
ITQ-A [2]	0.2094	0.2355	0.2424	0.2535
SH-A [8]	0.1866	0.1900	0.2044	0.2020
LFH-A [89]	0.1599	0.1608	0.1705	0.1693
LSH-R [1]	0.4627	0.5180	0.5481	0.5750
ITQ-R [2]	0.2476	0.2722	0.2813	0.2895
SH-R [8]	0.2244	0.1938	0.1806	0.1922
LFH-R [89]	0.2347	0.2849	0.2997	0.2955
DSDH [24]	0.7941	0.8076	0.8318	0.8297
HashNet [14]	0.7554	0.8163	0.8340	0.8439
DPSH [11]	0.8094	0.8325	0.8441	0.8520
DBDH [90]	0.8052	0.8107	0.8277	0.8324
CSQ [91]	0.7853	0.8213	–	0.8316
DSEH [63]	0.7319	0.7466	0.7602	0.7721
SADH	**0.8352**	**0.8454**	**0.8487**	**0.8646**

Table 4
MAP@ALL on MIRFLICKR-25 K for image retrieval.

Method	MIRFlickr-25 K (MAP@ALL)			
	16 bits	32 bits	48 bits	64 bits
DSDH [24]	0.7541	0.7574	0.7616	0.7680
HashNet [14]	0.7440	0.7685	0.7757	0.7815
DPSH [11]	0.7672	0.7694	0.7722	0.7772
DBDH [90]	0.7530	0.7615	0.7634	0.7653
CSQ [91]	0.6702	0.6735	–	0.6843
DSEH [63]	0.6832	0.6863	0.6974	0.6970
SADH	**0.7731**	**0.7698**	**0.7993**	**0.7873**

4.4. Empirical analysis

Three additional experimental settings are designed and used to further analyse SADH.

Table 5
MAP@ALL on MS-COCO for image retrieval.

Method	MS-COCO (MAP@ALL)			
	16 bits	32 bits	48 bits	64 bits
DSDH [24]	0.6093	0.6482	0.6615	0.6740
HashNet [14]	0.6873	0.7184	0.7301	0.7362
DPSH [11]	0.6610	0.6825	0.6887	0.6850
DSEH [63]	0.5897	0.6048	0.6133	0.6188
SADH	**0.7176**	**0.7507**	**0.7558**	**0.7736**

4.4.1. Ablation study

We investigate the impact of the different proposed modules on the retrieval performance of SADH. SADH-sym refers is built by replacing the asymmetric association between Image-Network and Semantic-Network by conventional point-to-point symmetric learning strategy, SADH-mars is built by removing the margin-scalable constraint from Image-Network, SADH-cos refers to replacing the cosine similarity module by the logarithm Maximum a Posterior (MAP) estimation of pairwise similarity loss which is used in many deep hashing approaches [63,24]:

$$J_s = -\sum_{i,j=1}^{m}\left(S_{i,j}H_i^TH_j - \log\left(1+\exp\left(H_i^TH_j\right)\right)\right) \tag{16}$$

Results are shown in Table 6 for both NUS-WIDE and CIFAR-10 for hash codes of 32 bits. Considering the results, we can see that the asymmetric guidance from Semantic-Network with rich semantic information plays an essential role on the performance of our method, meanwhile the margin-scalable constraint from Image-Network itself also significantly improves retrieval accuracy. It can also be observed that when using the cosine similarity, better performance is achieved than using the MAP estimation of pairwise similarity.

As a further demonstration of the effectiveness of the margin-scalable constraint, we compare it with several choices of single constants on our SADH. For 50 epochs, the top 5000 MAP results on MIR-Flickr25K and CIFAR-10 are given for every 10 epochs

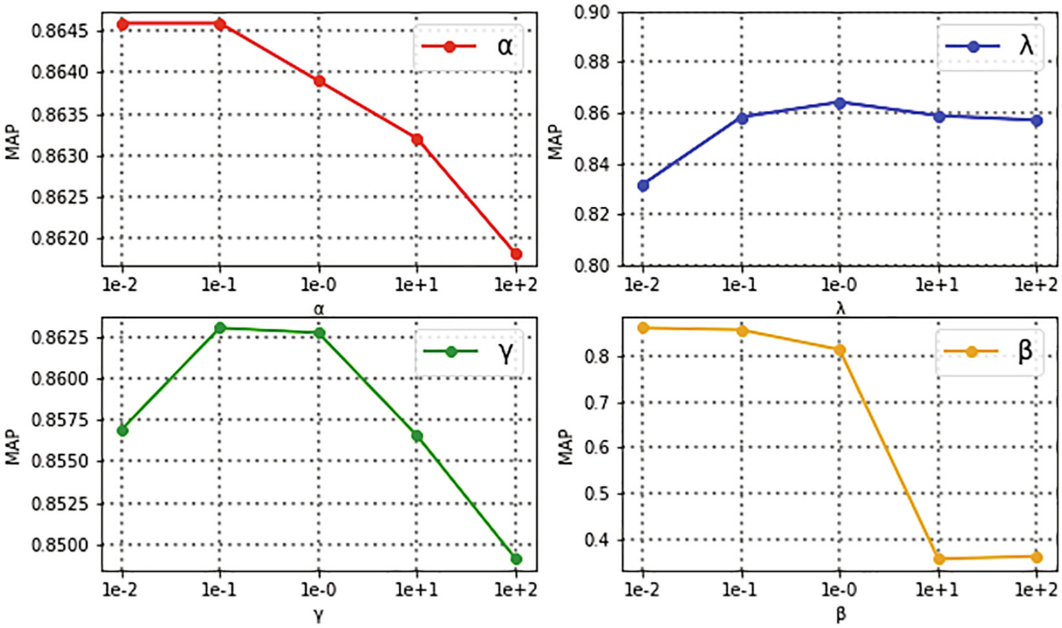
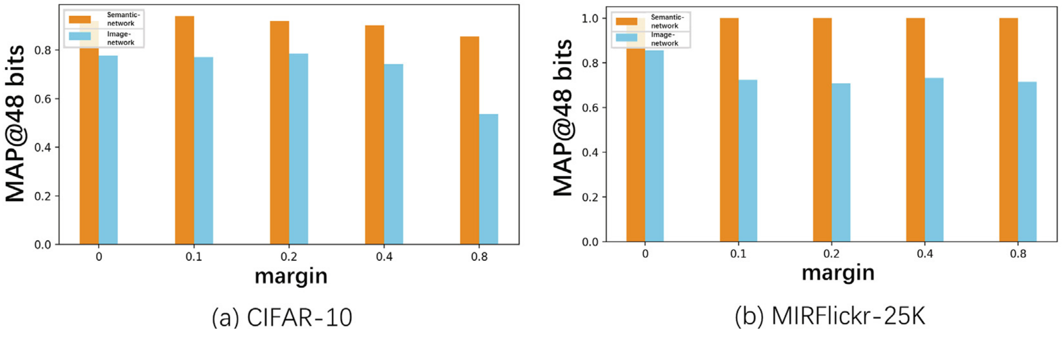


Fig. 8. Sensitivity analysis of four hyper-parameters. The dataset is NUS-WIDE and the code length is 64-bit.



(a) CIFAR-10 (b) MIRFlickr-25K

Fig. 9. Sensitivity analysis on the margin parameter, the orange bars corresponds to the result of Semantic-Network, the blue bars corresponds to the result of Image-Network.

Table 6
Ablation study on several modules in SADH, with MAP on NUS-WIDE and CIFAR-10 at hash length 32 bits.

Methods	NUS-WIDE (MAP@5000)	CIFAR-10 (MAP@ALL)
SADH-sym	0.8031	0.8152
SADH-mars	0.8174	0.8249
SADH-cos	0.8168	0.8502
SADH	**0.8454**	**0.8832**

respectively. As illustrated in Fig. 10, it is clear that in both the single-labeled and multi-labeled scenario, a scalable margin achieves better retrieval accuracy than using fixed margin constants. Furthermore, it is observed that on CIFAR-10, scalable margin result in faster convergence of SADH during training.

4.4.2. Training efficiency analysis

Fig. 11 shows the change of MAP using 64-bit hash codes during training time of 1000 s, with a comparison of SADH with DSEH, DSDH, DBDH, and DPSH on NUS-WIDE. It is distinct that, SADH costs significantly less time than the other methods to achieve a similar MAP. In comparison with DPSH, SADH reduces training time by approximately two times to achieve a MAP of 0.85. Furthermore, SADH displays the tendency of convergence much earlier than DSEH, with much high MAP. This is because Image-Network and Semantic-Network are trained alternatively for multiple rounds in DSEH, with the generated hash codes and semantic features of Image-Network being supervised by same number of those generated by Semantic-Network. Whereas in SADH Semantic-Network will cease to train after one round of convergence. And the converged Semantic-Network will be utilized to produce semantic dictionaries for each cases of semantic label. The semantic dictionaries directly supervise Image-Network with asymmetric pairwise correlation without further use of Semantic-Network.

4.4.3. Visualization of hash codes

Fig. 12 is the t-SNE [93] visualization of hash codes generated by DSDH and SADH on CIFAR-10, hash codes that belong to 10 different classes. Each class is assigned a different color. It can be observed that hash codes in different categories are discriminatively separated by SADH, while the hash codes generated by DSDH do not show such a clear characteristic. This is because the cosine similarity and scalable margin mechanism used in SADH can provide a more accurate inter-and-intra-class similarity preservation resulting in more discriminative hash codes in comparison to the mentioned form of pairwise similarity loss (16) used in DSDH.

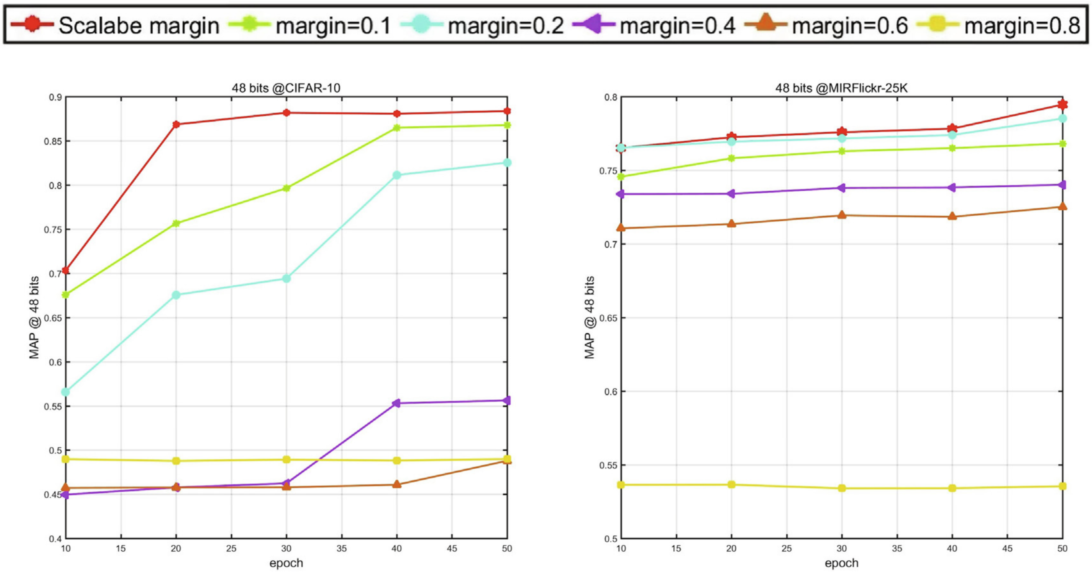


Fig. 10. Map during 50 epochs on CIFAR-10 and MIRFlickr-25 K with different choice of margins.

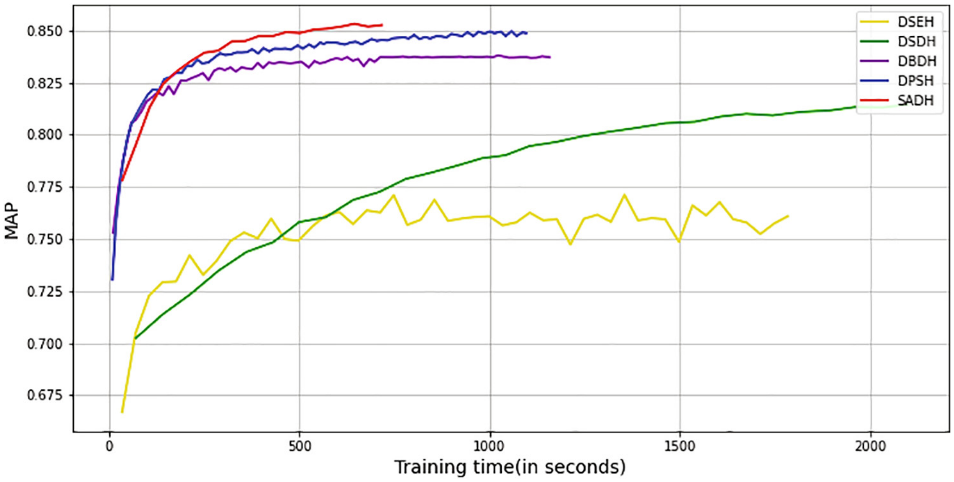


Fig. 11. Training time of SADH compared to 4 methods on NUS-WIDE with code length of 64.

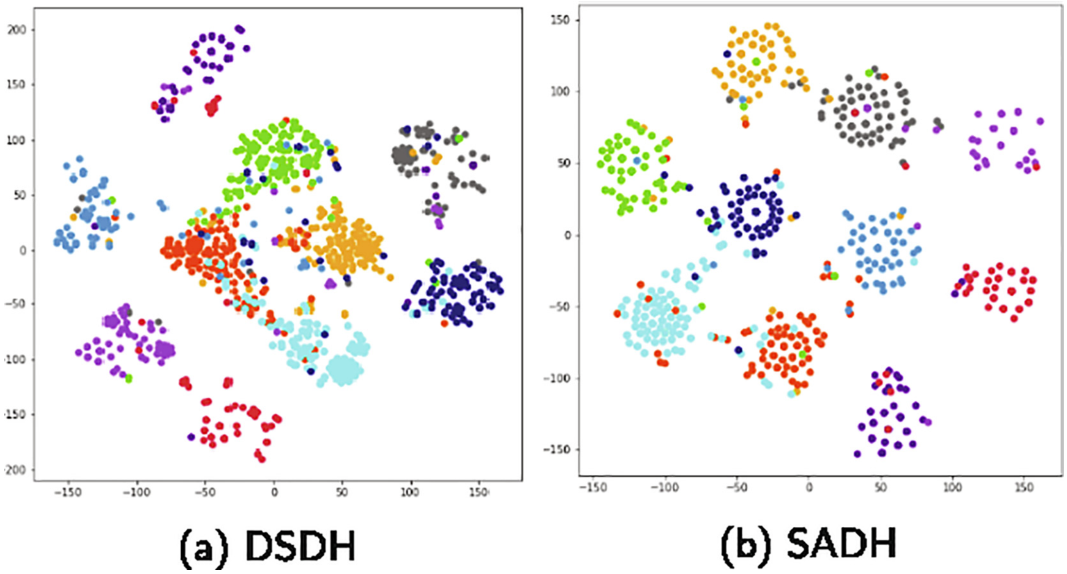


Fig. 12. The t-SNE visualization of hash codes learned by DSDH and SADH.

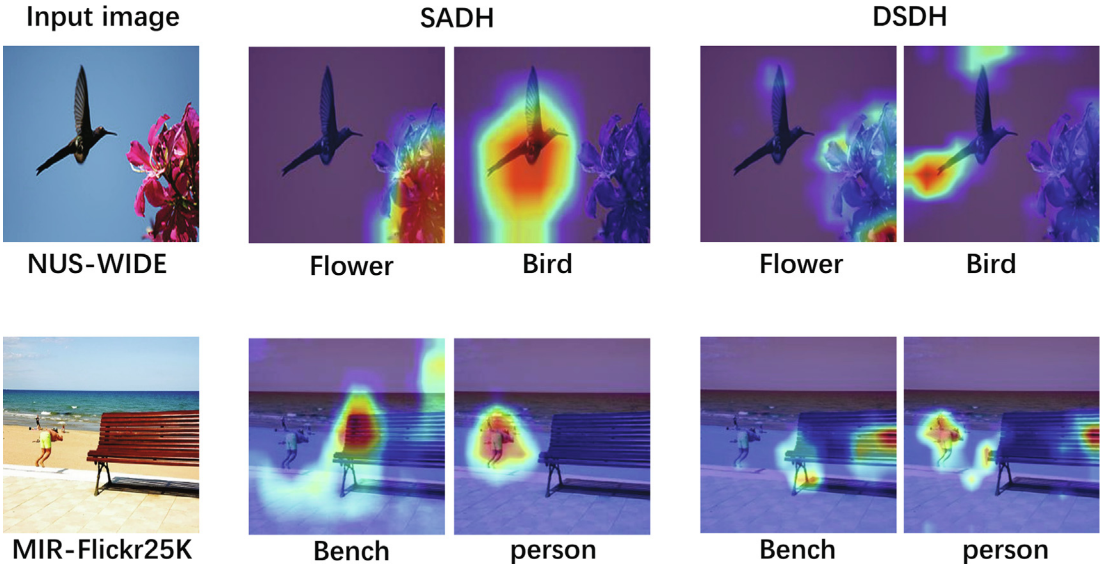


Fig. 13. Grad-CAM visualization of SADH and DSDH for images sampled from multi-label benchmarks with respect to different ground-truth categories.

Table 7
MAP@ALL on NUS-WIDE for cross-modal retrieval.

Task	Method	NUS-WIDE(MAP@ALL)		
		16bits	48bits	64bits
Image to Text	SSAH [68]	0.6163	0.6278	0.6140
	PRDH [77]	0.5919	0.6059	0.6116
	DCMH [76]	0.5445	0.5597	0.5803
	SADH-c	**0.6536**	**0.6614**	**0.6663**
Text to Image	SSAH [68]	0.6204	0.6251	0.6349
	PRDH [77]	0.6155	0.6286	0.6349
	DCMH [76]	0.5793	0.5922	0.6014
	SADH-c	**0.6748**	**0.6821**	**0.6857**

Table 8
MAP@ALL on MS-COCO for cross-modal retrieval.

Task	Method	MS-COCO(MAP@ALL)		
		16bits	48bits	64bits
Image to Text	SSAH [68]	0.5204	0.5187	0.5272
	PRDH [77]	0.5538	0.5672	0.5572
	DCMH [76]	0.5228	0.5438	0.5419
	SADH-c	**0.6362**	**0.6679**	**0.6929**
Text to Image	SSAH [68]	0.4789	0.4753	0.4888
	PRDH [77]	0.5122	0.5190	0.5404
	DCMH [76]	0.4883	0.4942	0.5145
	SADH-c	**0.6347**	**0.6673**	**0.6834**

4.4.4. Heatmap visualization of focused regions

The Grad-CAM visualization of our SADH and DSDH following [94] for sampled images on NUS-WIDE and MIR-Flickr25K is illustrated in Fig. 13. For each selected class of interest, Grad-CAM highlights the focused regions of convolutional feature maps. We observe that, comparing to DSDH, our SADH can correlates selected semantics with corresponding regions more accurately, which is a strong proof for robust semantic feature preserving capacity of our SADH especially for multi-label scenarios.

4.4.5. Extention: experiments on cross-modal hashing

As discussed earlier in Section 3.5, our SADH algorithm can be seamlessly extended to cross-modal hashing. We devise a image-text cross-modal hashing framework namely SADH-c by maintaining the network architecture of Image-Network and Semantic-Network and add a 3-layer MLP network with a multi-scale fusion module to extract textual features and learn hash codes, which is the same as the TxtNet used in SSAH. Table 7 and Table 8 show the MAP result of our method and three other state-of-the-art deep supervised cross-modal hashing methods: DCMH [76], PRDH [77], SSAH [68] on MS-COCO and NUS-WIDE for cross-modal retrieval between image data and text data, the according precision-recall curves are shown in Fig. 14. Our approach substantially outperforms all comparison methods with particularly superior performance in MS-COCO which has 80 semantics in total, this is a strong evidence of the robustness of our method in multi-label datasets. Comparing to SSAH, which utilizes point-to-point symmetric association and logarithm Maximum a Posterior (MAP) estimation (16), the remarkable performance of our proposed method is capacitated by the margin scalable pairwise constraint and asymmetric guidance mechanism.

5. Conclusion

In this paper, we present a novel Deep Hashing with Self-Supervised Asymmetric Semantic Excavation and Margin-Scalable Constraint. To improve the reliability of retrieval performance in multi-labeled scenarios, the proposed SADH preserve and refine abundant semantic information from semantic labels in two semantic dictionaries to supervise the 2nd framework Image-Network with asymmetric guidance mechanism. A margin-scalable constraint is designed to precisely search similarity information in fine-grained level. Additionally, the proposed method is seamlessly extended to cross-modal scenarios. Comprehensive empirical evidence shows that SADH outperforms several state-of-the-art methods including traditional methods and deep hashing methods on FOUR widely used benchmarks. In the future, we

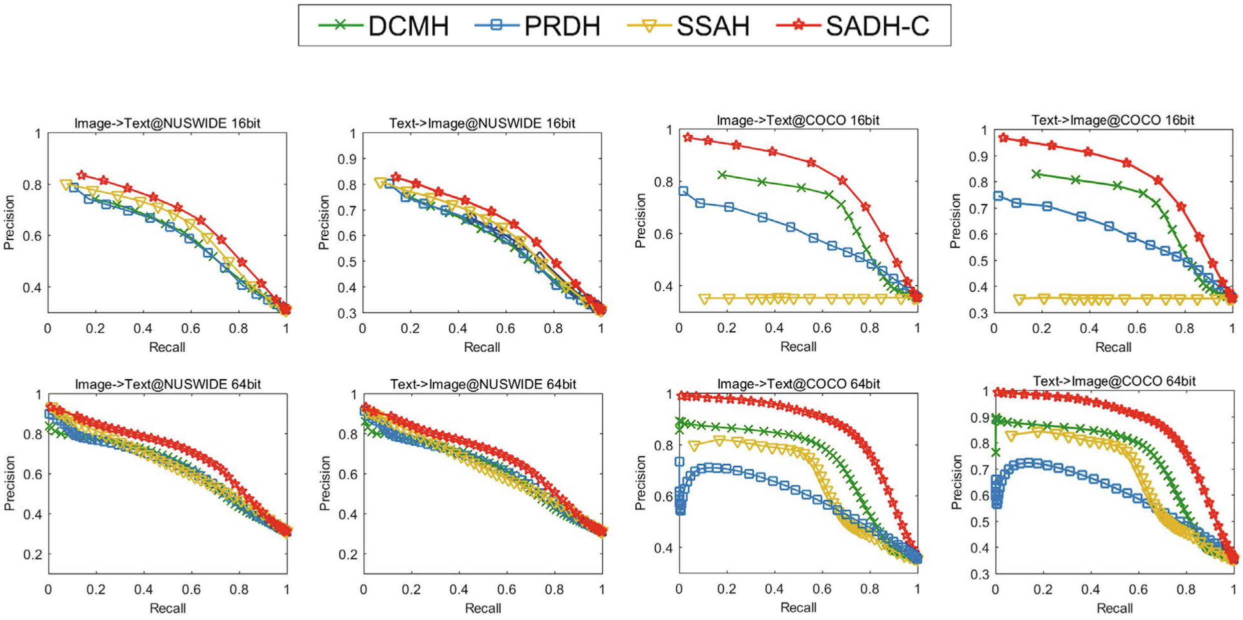


Fig. 14. Precision-recall curve on NUS-WIDE and MS-COCO for cross-modal hashing.

will explore to more detailedly investigate the proposed SADH method in deep hashing for multi-modal data retrieval.

CRediT authorship contribution statement

Zhengyang Yu: Conceptualization, Methodology, Writing – original draft. **Song Wu:** Methodology, Validation, Formal analysis, Supervision, Writing – review & editing. **Erwin M. Bakker:** Writing – review & editing.

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.

Acknowledgements

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

References

[1] A. Gionis, P. Indyk, R. Motwani, et al., Similarity search in high dimensions via hashing, Vldb 99 (1999) 518–529.
[2] Y. Gong, S. Lazebnik, A. Gordo, F. Perronnin, Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2012) 2916–2929.
[3] R. Salakhutdinov, G. Hinton, Semantic hashing, Int. J. Approximate Reasoning 50 (2009) 969–978.
[4] J. Wang, H.T. Shen, J. Song, J. Ji, Hashing for similarity search: A survey, CoRR abs/1408.2927 (2014)..
[5] K. He, F. Wen, J. Sun, K-means hashing: An affinity-preserving quantization method for learning binary compact codes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2938–2945.
[6] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, S.-F. Chang, Supervised hashing with kernels, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2074–2081.
[7] B. Kulis, T. Darrell, Learning to hash with binary reconstructive embeddings, in: Advances in neural information processing systems, 2009, pp. 1042–1050..
[8] Y. Weiss, A. Torralba, R. Fergus, Spectral hashing, Advances in neural information processing systems 21 (2008) 1753–1760.
[9] M. Norouzi, D.J. Fleet, Minimal loss hashing for compact binary codes, in: ICML, 2011..
[10] M. Norouzi, D.J. Fleet, R.R. Salakhutdinov, Hamming distance metric learning, in: Advances in neural information processing systems, 2012, pp. 1061–1069..
[11] W.-J. Li, S. Wang, W.-C. Kang, Feature learning based deep supervised hashing with pairwise labels, IJCAI (2016) 1711–1717.
[12] H. Lai, Y. Pan, Y. Liu, S. Yan, Simultaneous feature learning and hash coding with deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3270–3278.
[13] H. Zhu, M. Long, J. Wang, Y. Cao, Deep hashing network for efficient similarity retrieval, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016..
[14] Z. Cao, M. Long, J. Wang, P.S. Yu, Hashnet: Deep learning to hash by continuation, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 5608–5617.
[15] Y. Cao, M. Long, B. Liu, J. Wang, Deep cauchy hashing for hamming space retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1229–1237.
[16] B. Liu, Y. Cao, M. Long, J. Wang, J. Wang, Deep triplet quantization, in: Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 755–763.
[17] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: Proceedings of the ACM international conference on image and video retrieval, 2009, pp. 1–9.
[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755.
[19] M.J. Huiskes, M.S. Lew, The mir flickr retrieval evaluation, in: Proceedings of the 1st ACM international conference on Multimedia information retrieval, 2008, pp. 39–43.
[20] Q.-Y. Jiang, X. Cui, W.-J. Li, Deep discrete supervised hashing, IEEE Trans. Image Process. 27 (2018) 5996–6009.
[21] H. Liu, R. Wang, S. Shan, X. Chen, Deep supervised hashing for fast image retrieval, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2064–2072.
[22] H. Jun, B. Ko, Y. Kim, I. Kim, J. Kim, Combination of multiple global descriptors for image retrieval, arXiv preprint arXiv:1903.10663 (2019)..
[23] R. Xia, Y. Pan, H. Lai, C. Liu, S. Yan, Supervised hashing for image retrieval via image representation learning., in: AAAI, volume 1, 2014, p. 2..
[24] Q. Li, Z. Sun, R. He, T. Tan, Deep supervised discrete hashing, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates Inc, 2017. https://proceedings.neurips.cc/paper/2017/file/e94f63f579e05cb49c05c2d050ead9c0-Paper.pdf..
[25] F. Shen, C. Shen, W. Liu, H. Tao Shen, Supervised discrete hashing, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 37–45.
[26] C. Da, G. Meng, S. Xiang, K. Ding, S. Xu, Q. Yang, C. Pan, Nonlinear asymmetric multi-valued hashing, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2018) 2660–2676.
[27] A. Shrivastava, P. Li, Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips), arXiv preprint arXiv:1405.5869 (2014)..
[28] Z. Zhang, Z. Lai, Z. Huang, W.K. Wong, G.-S. Xie, L. Liu, L. Shao, Scalable supervised asymmetric hashing with semantic and latent factor embedding, IEEE Trans. Image Process. 28 (2019) 4803–4818.
[29] B. Neyshabur, P. Yadollahpour, Y. Makarychev, R. Salakhutdinov, N. Srebro, The power of asymmetry in binary hashing, arXiv preprint arXiv:1311.7662 (2013)..
[30] Q.-Y. Jiang, W.-J. Li, Asymmetric deep supervised hashing, AAAI (2018) 3342–3349.
[31] P. Indyk, R. Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality, in: Proceedings of the thirtieth annual ACM symposium on Theory of computing, 1998, pp. 604–613.
[32] M.S. Charikar, Similarity estimation techniques from rounding algorithms, in: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing, 2002, pp. 380–388..
[33] B. Wang, X. Liu, K. Xia, K. Ramamohanarao, D. Tao, Random angular projection for fast nearest subspace search, Pacific Rim Conference on Multimedia, Springer (2018) 15–26.
[34] J. Ji, J. Li, Q. Tian, S. Yan, B. Zhang, Angular-similarity-preserving binary signatures for linear subspaces, IEEE Trans. Image Process. 24 (2015) 4372–4380.
[35] Y. Xu, X. Liu, B. Wang, R. Tao, K. Xia, X. Cao, Fast nearest subspace search via random angular hashing, IEEE Trans. Multimedia 23 (2021) 342–352.
[36] Y. Gong, S. Lazebnik, A. Gordo, F. Perronnin, Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2012) 2916–2929.
[37] X. Lu, X. Zheng, X. Li, Latent semantic minimal hashing for image retrieval, IEEE Trans. Image Process. 26 (2016) 355–368.
[38] Z. Lai, Y. Chen, J. Wu, W.K. Wong, F. Shen, Jointly sparse hashing for image retrieval, IEEE Trans. Image Process. 27 (2018) 6147–6158.
[39] T. Yuan, W. Deng, J. Hu, Distortion minimization hashing, IEEE Access 5 (2017) 23425–23435.
[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[41] A. Joly, O. Buisson, Random maximum margin hashing, in: CVPR 2011, IEEE, 2011, pp. 873–880..
[42] H. Yang, X. Bai, Y. Liu, Y. Wang, L. Bai, J. Zhou, W. Tang, Maximum margin hashing with supervised information, Multimedia Tools Appl. 75 (2016) 3955–3971.
[43] Z. Li, J. Tang, L. Zhang, J. Yang, Weakly-supervised semantic guided hashing for social image retrieval, Int. J. Comput. Vis. 128 (2020).
[44] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (2017) 84–90.
[45] J. Tang, Z. Li, X. Zhu, Supervised deep hashing for scalable face image retrieval, Pattern Recogn. 75 (2018) 25–32.
[46] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M.S. Lew, Deep learning for visual understanding: A review, Neurocomputing 187 (2016) 27–48.
[47] H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, A convolutional neural network cascade for face detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5325–5334.
[48] Z. Liu, X. Li, P. Luo, C.-C. Loy, X. Tang, Semantic image segmentation via deep parsing network, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1377–1385.
[49] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 1904–1916.
[50] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448..
[51] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99..

[52] S. Wu, A. Oerlemans, E.M. Bakker, M.S. Lew, Deep binary codes for large scale image retrieval, Neurocomputing 257 (2017) 5–15. Machine Learning and Signal Processing for Big Multimedia Analysis..

[53] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M.S. Lew, Deep learning for visual understanding: A review, Neurocomputing 187 (2016) 27–48. Recent Developments on Deep Big Vision..

[54] X. Wang, X. Zou, E.M. Bakker, S. Wu, Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval, Neurocomputing 400 (2020) 255–271.

[55] Y. Chen, Z. Lai, Y. Ding, K. Lin, W.K. Wong, Deep supervised hashing with anchor graph, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9796–9804.

[56] H. Lai, P. Yan, X. Shu, Y. Wei, S. Yan, Instance-aware hashing for multi-label image retrieval, IEEE Trans. Image Process. 25 (2016) 2469–2479.

[57] K. Lin, H.-F. Yang, J.-H. Hsiao, C.-S. Chen, Deep learning of binary hash codes for fast image retrieval, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2015, pp. 27–35.

[58] H.-F. Yang, K. Lin, C.-S. Chen, Supervised learning of semantics-preserving hash via deep convolutional neural networks, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2017) 437–451.

[59] T. Yao, F. Long, T. Mei, Y. Rui, Deep semantic-preserving and ranking-based hashing for image retrieval., in: IJCAI, volume 1, 2016, p. 4..

[60] F. Shen, C. Shen, W. Liu, H. Tao Shen, Supervised discrete hashing, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 37–45.

[61] D. Xie, C. Deng, C. Li, X. Liu, D. Tao, Multi-task consistency-preserving adversarial hashing for cross-modal retrieval, IEEE Trans. Image Process. 29 (2020) 3626–3637.

[62] C. Da, G. Meng, S. Xiang, K. Ding, S. Xu, Q. Yang, C. Pan, Nonlinear asymmetric multi-valued hashing, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2018) 2660–2676.

[63] N. Li, C. Li, C. Deng, X. Liu, X. Gao, Deep joint semantic-embedding hashing, IJCAI (2018) 2397–2403.

[64] F. Zhao, Y. Huang, L. Wang, T. Tan, Deep semantic ranking based hashing for multi-label image retrieval, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1556–1564.

[65] Z. Zhang, Q. Zou, Y. Lin, L. Chen, S. Wang, Improved deep hashing with soft pairwise similarity for multi-label image retrieval, IEEE Trans. Multimedia 22 (2019) 540–553.

[66] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, IEEE, 2006, pp. 1735–1742.

[67] R. Kang, Y. Cao, M. Long, J. Wang, P.S. Yu, Maximum-margin hamming hashing, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8252–8261.

[68] C. Li, C. Deng, N. Li, W. Liu, X. Gao, D. 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.

[69] W. Gu, X. Gu, J. Gu, B. Li, Z. Xiong, W. Wang, Adversary guided asymmetric hashing for cross-modal retrieval, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, 2019, pp. 159–167.

[70] J. Song, Y. Yang, Y. Yang, Z. Huang, H.T. Shen, Inter-media hashing for large-scale retrieval from heterogeneous data sources, in: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, 2013, pp. 785–796.

[71] J. Zhou, G. Ding, Y. Guo, Latent semantic sparse hashing for cross-modal similarity search, in: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, 2014, pp. 415–424.

[72] G. Ding, Y. Guo, J. Zhou, Y. Gao, Large-scale cross-modality search via collective matrix factorization hashing, IEEE Trans. Image Process. 25 (2016) 5427–5440.

[73] S. Kumar, R. Udupa, Learning hash functions for cross-view similarity search, in: Twenty-second international joint conference on artificial intelligence, 2011.

[74] D. Zhang, W.-J. Li, Large-scale supervised multimodal hashing with semantic correlation maximization, in: Proceedings of the AAAI conference on artificial intelligence, volume 28, 2014..

[75] Z. Lin, G. Ding, M. Hu, J. Wang, Semantics-preserving hashing for cross-view retrieval, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3864–3872.

[76] Q.-Y. Jiang, W.-J. Li, Deep cross-modal hashing, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3232–3240.

[77] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, X. Gao, Pairwise relationship guided deep hashing for cross-modal retrieval, in: proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017..

[78] D. Xie, C. Deng, C. Li, X. Liu, D. Tao, Multi-task consistency-preserving adversarial hashing for cross-modal retrieval, IEEE Trans. Image Process. 29 (2020) 3626–3637.

[79] Z. Li, J. Tang, T. Mei, Deep collaborative embedding for social image understanding, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2018) 2070–2083.

[80] L. Jin, Z. Li, J. Tang, Deep semantic multimodal hashing network for scalable image-text and video-text retrievals, IEEE Trans. Neural Networks Learn. Syst. (2020).

[81] X. Lu, Y. Chen, X. Li, Hierarchical recurrent neural hashing for image retrieval with hierarchical convolutional features, IEEE Trans. Image Processing 27 (2017) 106–120.

[82] H.-F. Yang, K. Lin, C.-S. Chen, Supervised learning of semantics-preserving hash via deep convolutional neural networks, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2017) 437–451.

[83] T. Yao, F. Long, T. Mei, Y. Rui, Deep semantic-preserving and ranking-based hashing for image retrieval., in: IJCAI, volume 1, 2016, p. 4..

[84] Y. Chen, X. Lu, Deep discrete hashing with pairwise correlation learning, Neurocomputing 385 (2020) 111–121.

[85] X. Wang, Y. Shi, K.M. Kitani, Deep supervised hashing with triplet labels, in: Asian conference on computer vision, Springer, 2016, pp. 70–84..

[86] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009)..

[87] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755..

[88] R. Zhang, L. Lin, R. Zhang, W. Zuo, L. Zhang, Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification, IEEE Trans. Image Process. 24 (2015) 4766–4779.

[89] P. Zhang, W. Zhang, W.-J. Li, M. Guo, Supervised hashing with latent factor models, in: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, 2014, pp. 173–182.

[90] X. Zheng, Y. Zhang, X. Lu, Deep balanced discrete hashing for image retrieval, Neurocomputing (2020).

[91] L. Yuan, T. Wang, X. Zhang, F.E. Tay, Z. Jie, W. Liu, J. Feng, Central similarity quantization for efficient image and video retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3083–3092.

[92] D.P. Kingma, J. Ba, in: ICLR (Poster) (Ed.), Adam: A method for stochastic optimization, 2015.

[93] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition (2013)..

[94] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vision 128 (2020) 336–359.

Zhengyang Yu received his B.S. degree in the College of Computer Information Science, Southwest University. His current research interests include computer vision and pattern recognition, large-scale instance retrieval and person reIdentification.

Song Wu received his B.S. degree and M.S. degree in Computer Science from the Southwest University, Chongqing, China, in 2009 and 2012, respectively. He received his Ph.D from the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Netherlands. He is a member of the Overseas High-level Talent Program in Chongqing and currently working at the College of Computer and Information Science of Southwest University. His current research interests include large-scale image retrieval and classification, big data technology and deep learning based computer vision (the co-author of the most cited paper of journal Neurocomputing: Deep learning for visual understanding: A review).

Zhihao Dou received his B.S degree in Automation in the College of Computer Information Science, Southwest University in 2020. He works as an intern in State Grid Corporation of China. His research interests are computer version and data mining.

Erwin M. Bakker is co-director of the LIACS Media Lab at Leiden University. He has published widely in the fields of image retrieval, audio analysis and retrieval and bioinformatics. He was closely involved with the start of the International Conference on Image and Video Retrieval (CIVR) serving on the organizing committee in 2003 and 2005. Moreover, he regularly serves as a program committee member or organizing committee member for scientific multimedia and human-computer interaction conferences and workshops.