# Self-auxiliary Hashing for Unsupervised Cross Modal Retrieval

Jingnan Xu, Tieying Li, Chong Xi, and Xiaochun Yang(✉)

Northeastern University, Shenyang 110169, China
{1901816,1910624}@stumail.neu.edu.cn, yangxc@mail.neu.edu.cn

**Abstract.** Recently, cross modality hashing has attracted significant attention for large scale cross-modal retrieval owing to its low storage overhead and fast retrieval speed. However, heterogeneous gap still exist between different modalities. Supervised methods always need additional information, such as labels, to supervise the learning of hash codes, while it is laborious to obtain these information in daily life. In this paper, we propose a novel self-auxiliary hashing for unsupervised cross modal retrieval (SAH), which makes sufficient use of image and text data. SAH uses multi-scale features of pairwise image-text data and fuses them with the uniform feature to facilitate the preservation of intra-modal semantic, which is generated from Alexnet and MLP. Multi-scale feature similarity matrices of intra-modality preserve semantic information better. For inter-modality, the accuracy of the generated hash codes is guaranteed by the collaboration of multiple inter-modal similarity matrices, which are calculated by uniform features of both modalities. Extensive experiments carried out on two benchmark datasets show the competitive performance of our SAH than the baselines.

**Keywords:** Cross-modal retrieval · Multi-scale fusion · Cross-modal hashing

## 1 Introduction

With the development of science and technology, more and more multimedia data, such as images and texts, appear on the Internet. Owing to the explosive increase of these data, the requirement of cross-modal retrieval increases sharply. Cross-modal retrieval aims to search semantically related images (texts) with text (image) query and vice versa. Image retrieval hashing is a long-established research task to retrieve images with similar contents [17], it is common for us to process images with VGG [19] or some other neural networks. For text, Word2Vec technology is widely used, which also try to exploit latent semantic [23]. One of the biggest challenges of cross-modal retrieval is how to bridge the heterogeneous gap between two different modalities. The cause of the heterogeneous gap is the difference distribution between the feature from different modalities. Data from intra-modality also have heterogeneous information, which can be tackled from multiple views [5]. To tackle the problem of the heterogeneous gap between modalities, many cross-modal hashing methods have been

Y. Sun et al. (Eds.): ChineseCSCW 2021, CCIS 1492, pp. 431–443, 2022.
https://doi.org/10.1007/978-981-19-4549-6_33

proposed because of the advantages of low storage cost and high query speed by mapping data into binary codes.

The development of cross-modal retrieval can be divided into two phases: shallow cross-modal hashing and deep learning-based cross-modal hashing. Shallow cross-modal hashing is based on hand-crafted features and learns the hash codes by linear functions. The advantage of these methods is easily implemented, while they cannot fully explore the semantic information of two modalities. Recently, with the development of deep learning, the deep neural network(DNN) has been deployed to cross-modal hashing. DNN-based cross-modal hashing can be divided into two categories: supervised hashing and unsupervised hashing. Supervised methods, with label information, such as tags, always perform remarkably. While in the real-life, it is a waste of time for us to obtain labels of the image-text pairs. Unsupervised methods that do not use label information in the training phase have also shown remarkable performance in recent years. Unsupervised methods focus more on the information of raw features. As a result, the quality of the hash codes that used in retrieval task is dramatically concerned with the feature learning stage.

However, there are still some issues that should be tackled. Firstly, at the feature extraction phase, these methods only focus on the single source feature, neglecting the rich semantic information gained from multiple views. Secondly, the general similarity matrix of features can not bridge the heterogeneous gap well, because the distribution information or similarities of different scales are not considered. In this paper, we propose a novel self-auxiliary hashing (SAH) method for unsupervised cross-modal retrieval. SAH provides a two-branch network for each modality, including the uniform branch and the auxiliary branch. Each branch will generate specific features and hash codes. Moreover, based on the features and hash codes of two branches, we construct multiple similarity matrices for inter-modality and intra-modality. These similarity matrices will be calculated to preserve more semantic and similarity information. Extensive experiments demonstrate the superior performance of our method.

## 2 Related Work

Cross modality hashing can be roughly divided into supervised cross modality hashing and unsupervised cross modality hashing. The task of cross modality retrieval is to retrieve images (or texts) with similar semantics to the input text (or image). Shallow cross-modal hashing methods [12,13,15,16] and deep cross-modal hashing methods [1,2,11,22] are two stages of cross-modal hashing methods development. Shallow Cross-Modal Hashing uses hand-crafted features to learn the binary vector projection which is mapped from instances. However, most shallow cross-modal hashing retrieval methods just deal the feature with only a single layer and map data in a linear or nonlinear way. In recent years, the deep learning algorithm proposed in machine learning has been applied to cross modality retrieval. Deep cross-modal retrieval [18] also can be divided into unsupervised methods and supervised methods.

Supervised hashing methods [7,10,18,22] explore relative information, such as semantic information, or some other relative information by labels or tags, to enhance the ability of cross modality retrieval. Deep cross-modal hashing (DCMH) [7] is an end-to-end hashing method with deep neural networks, which can jointly learn hash codes and feature. In deep cross-modal hashing methods, generative adversarial network (GAN) is used to make adversarial learning. Self-Supervised Adversarial Hashing Networks (SSAH) [10] and Wang *et al.* [22] use image and text adversarial networks to generate hashing codes of both modalities, the learned features are used to keep the semantic relevance and preserve the semantic of different modalities.

Although some supervised methods perform well in practical applications, supervised information, such as label, is hard for us to collect, which is not suitable in reality.

Unsupervised hashing methods aim to learn hashing functions without supervised information, such as labeled data. For example, inter-media hashing (IMH) [20] considers the inter-media consistency and intra-media consistency with linear hash functions, and learns the hash function of image modality and text modality jointly. CVH [9] proposes a principled method to learn a hash function of different modality instances. Collective Matrix Factorization Hashing (CMFH) [4] learns the hash codes of an instance from two modalities and proposes the upper and lower bound. Latent Semantic Sparse Hashing (LSSH) [26] copes the instances of image and text with different methods and performs search by Sparse Coding and Matrix Factorization. Unsupervised Deep Cross-Modal Hashing (UDCMH) [24] makes a combination of deep learning and matrix factorization, considering the neighbour information and the weight assignment of optimization stage. Deep joint semantics reconstructing hashing (DJSRH) [21] considers the neighborhood information of different modalities.

Although the performance of these methods are remarkable, the features they focus on are not comprehensive. Moreover, they neglect the deep similarity information of two modalities and have bad performance at bridging the "heterogeneity gap".

## 3 Proposed Method

### 3.1 Problem Fomulation

Assume the training dataset of our methods is a collection of the pairwise image-text instances, written as $O = (X, Y)$. $X$ is the instance of image modality and $Y$ is the text modal instance. The number of instances of each modality is $n$. The goal of our method is to learn the modality-specific hash function for image modality and text modality which can generate hash codes with rich semantic information. For each modality, two branches (uniform branch and modality-specific auxiliary branch) are used to generate different features for each modality. $MF_{*_i}$ are the $i$-th multi-scale features of image or text modality, which generate from the auxiliary branch with different dimensions. $MH_{*_i}$ denotes the $i$-th multi-scale hash code of image or text which is generated from

$MF_{*_i}$. $F_*$ and $H_*$ are the feature and hash code gained from the uniform branch which is same as other unsupervised methods. The notations used in SAH are summarized in Table 1.

**Table 1.** Notations and their descriptions.

| Notations | Descriptions | Modality |
|---|---|---|
| $S^F_{x,y}$ | Similarity of uniform feature | $(x, y) \in \{(I, I), (T, T), (I, T)\}$ |
| $S^H_{x,y}$ | Similarity of uniform hash code | $(x, y) \in \{(I, I), (T, T), (I, T)\}$ |
| $S^{CH}_{x,y}$ | Similarity of complex hash code | $(x, y) \in \{(I, T)\}$ |
| $S^{MH}_{x,y}$ | Similarity of multi-scale hash code | $(x, y) \in \{(I, I), (T, T)\}$ |
| $F_*$ | Uniform feature of image or text | $* \in \{I, T\}$ |
| $MF_{*_i}$ | The multi-scale feature of image or text | $* \in \{I, T\}$ |
| $H_*$ | Uniform hash code of image or text | $* \in \{I, T\}$ |
| $H_{*\_mix}$ | The mix hash code of image or text | $* \in \{img, txt\}$ |
| $MH_{*_i}$ | The i-th multi-scale hash code of image or text | $* \in \{I, T\}$ |
| $MH_{*\_com}$ | The comprehensive hash code of image or text | $* \in \{I, T\}$ |
| $CH_*$ | The complex hash code of image or text | $* \in \{I, T\}$ |

### 3.2 Network Architecture

Figure 1 is a flowchart of our SAH. Our method is composed of two networks, image network and text network, both of them can be divided into the uniform branch and the modality-specific auxiliary branch. For image network, the uniform branch is composed of AlexNet [8]. Image auxiliary branch, shown by Fig. 2, deals the input image with a fully connected layer and gains the auxiliary data. For text network, the uniform branch consists of MLP. Text auxiliary branch, drawn in Fig. 3, tackles the input text data with a pooling layer first and gets the auxiliary data which is convenient for the later procession.

**Feature Extraction.** For image modality, we adopt the pre-trained AlexNet as the uniform feature extractor which is widely used in unsupervised methods. However, features gained from AlexNet are not comprehensive enough, which is the common disadvantage of previous works. Features obtained from a single scale often comes from the same measurement perspective, ignoring the details that may be obtained from other perspectives. Benefit from feature learning at multiple scales, multi-scale features can better represent the semantics of instances.

To gain multi-scale features, we process the input of image modality by fully connected layer and three pooling layers respectively. Then we get three sizes of image feature which we called the auxiliary data. And we resize them into the same size. These three auxiliary data are single-channel, we make them expand to three channels. To tackle with the auxiliary data of image modality, we deal

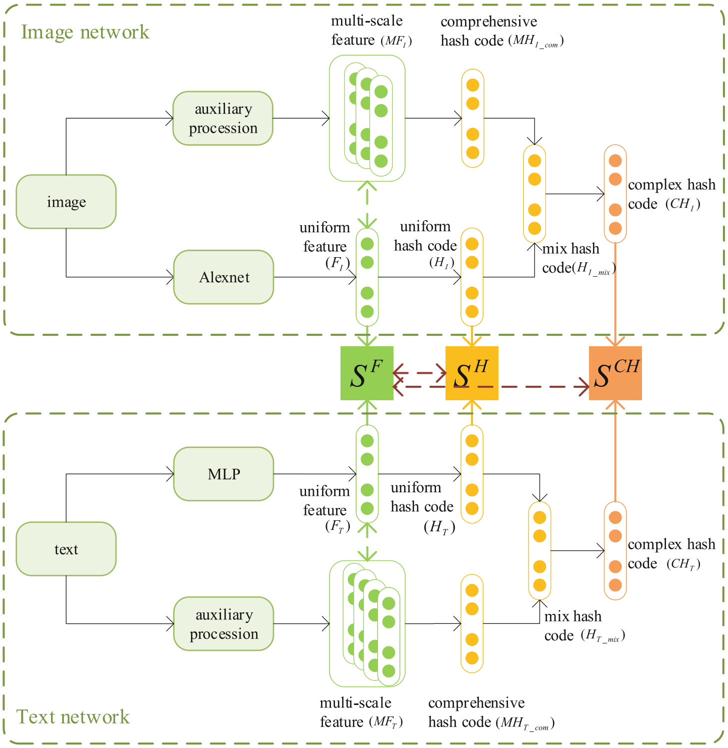


Image network
multi-scale feature ($MF_I$)
comprehensive hash code ($MH_{I\_com}$)
auxiliary procession
image
complex hash code ($CH_I$)
uniform feature ($F_I$)
uniform hash code ($H_I$)
mix hash code($H_{I\_mix}$)
Alexnet
$S^F$
$S^H$
$S^{CH}$
MLP
uniform feature ($F_T$)
uniform hash code ($H_T$)
text
complex hash code ($CH_T$)
mix hash code ($H_{T\_mix}$)
auxiliary procession
Text network
multi-scale feature ($MF_T$)
comprehensive hash code ($MH_{T\_com}$)

**Fig. 1.** The overview of our proposed SAH.

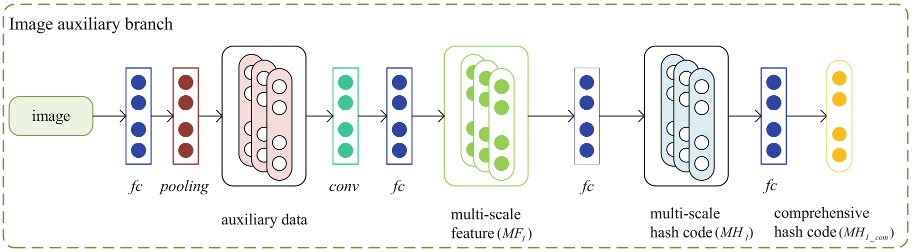


Image auxiliary branch
image
fc
pooling
auxiliary data
conv
fc
multi-scale feature ($MF_I$)
fc
multi-scale hash code ($MH_I$)
fc
comprehensive hash code ($MH_{I\_com}$)

**Fig. 2.** Image auxiliary branch.

them with five convolution layer and three fully connected layer networks and obtain three multi-scale features $MF_{I_1}$, $MF_{I_2}$ and $MF_{I_3}$. For text modality, we set four pooling layers to tackle with the input text data respectively and get four different size auxiliary data. Due to the character of text data is sparse, we deal these four data just with a fully connected layer and get four multi-scale features $MF_{T_1}$, $MF_{T_2}$, $MF_{T_3}$ and $MF_{T_4}$ of text modality.

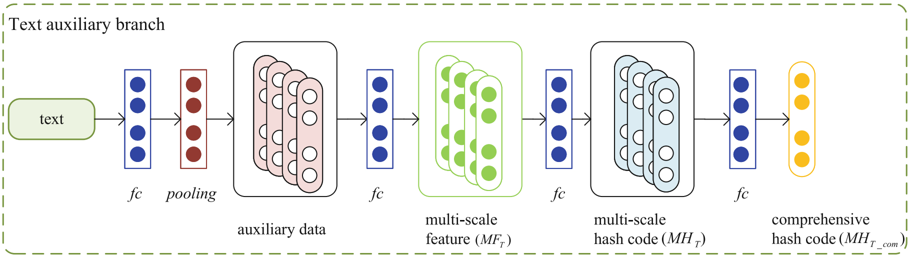


**Fig. 3.** Text auxiliary branch.

The reason for the difference in the amount of auxiliary data between two modalities is that, image always contains more comprehensive information than text. Therefore, we should explore the semantic information of text more comprehensively. To this end, we can obtain rich semantic features of each modality, which can be utilized to construct similarity matrices and guide hash codes learning. We calculate the similarity matrix of uniform feature based on cosine similarity. $S_{IT}^F$ is intra-modal similarity matrix, $S_{II}^F$ and $S_{TT}^F$ are inter-modal similarity matrices, defined as follows:

$$\begin{aligned} S_{x,y}^F &= cos(F_x, F_y), \\ s.t.(x,y) &\in (I,I),(T,T),(I,T). \end{aligned} \tag{1}$$

**Hash Code Generation.** We will generate two kinds of hash code, uniform hash code and comprehensive hash code for two modalities respectively. The uniform hash codes ($H_I$ and $H_T$) is obtained by the uniform feature in uniform branch with a simple hash layer for each modalities.

For image modality, we process three auxiliary image features and get three same size hash codes $MH_{I_1}$, $MH_{I_2}$, and $MH_{I_3}$ with auxiliary branch of image modality. We concatenate these three hash codes together through hash layer $HILayer$ to obtain the comprehensive hash code $H_{I\_com}$ which contains multi-scale semantics. The concatenation will not change the semantic of each bit dramatically, it can be seen as a way of data enhancement.

$$H_{I\_com} = HILayer(ConCat(MH_{I_1}, MH_{I_2}, MH_{I_3})), \tag{2}$$

where $ConCat()$ denotes the concatenation of vectors.

For text modality, we have four auxiliary features, and we process them with four different hash layer and get four same size auxiliary hash codes, $MH_{T_1}$, $MH_{T_2}$, $MH_{T_3}$, $MH_{T_4}$. We also concatenate them four and make this hash code into a hash layer $HTLayer$ and get the comprehensive hash code $H_{T\_com}$.

$$H_{T\_com} = HTLayer(ConCat(MH_{T_1}, MH_{T_2}, MH_{T_3}, MH_{T_4})). \tag{3}$$

Concatenation is a compression of semantic information which can preserve different scales of semantic information and similarity. Furthermore, we fuse the uniform hash code and the comprehensive hash code according to a certain proportion $\mu$ $(0 < \mu < 1)$ and get the mixed hash code. The mixed hash code can maintain more semantic information than the uniform hash code of each modality.

$$H_{img\_mix} = \mu MH_{I\_com} + (1-\mu)H_I. \tag{4}$$

$$H_{txt\_mix} = \mu MH_{t\_com} + (1-\mu)H_T. \tag{5}$$

The inter-similarity matrix of uniform hash codes can be calculated by similarity function:

$$\begin{aligned} S^H_{x,y} &= cos(H_x, H_y), \\ s.t.(x,y) &\in (I,I),(T,T),(I,T). \end{aligned} \tag{6}$$

**Similarity Matrices Learning.** Since dimension reduction during the procession from features to hash codes will cause some semantic lose, we aim to keep the semantic consistency of instance pairs. To this end, we introduce a loss function that can measure the semantic consistency between hash codes and features of intra-modality and inter-modality. The loss function $L_1$ can be written as follows:

$$L_1 = \sum_{i=1}^{n}\sum_{j=1}^{n} \parallel S^H_{x,y}(i,j) - S^F_{x,y}(i,j) \parallel . \tag{7}$$

For intra-modality, the multi-scale hash codes and the uniform hash codes are generated from features of different scale, they preserve richer semantic information of different view. Uniform feature similarity matrix $S^F$ offers us the degree of similarity among different instances in a single modality. Loss function $L_1$ makes the multi-scale hash codes retains the semantic consistency, too. To ensure the accuracy of hash codes, the similarity matrix of hash codes should approximate to the feature similarity matrix. Therefore, we can minimize the distance between the similarity of the multi-scale hash codes of each modality and its intra-modality feature similarity. The loss function $L_2$ can be written as follows:

$$L_2 = \sum_{ni=1}^{3} \parallel S^{MH}_{I_{ni}} - S^F_{I,I} \parallel + \sum_{mi=1}^{4} \parallel S^{MH}_{T_{mi}} - S^F_{T,T} \parallel . \tag{8}$$

For inter-modality, the similarity matrix of features should also contains the inherent pair-wise information. The feature similarity of pair-wise instance of different modalities can be seen as converging to the maximum value in the cosine similarity. Apart from that, complex hash code will generate from the mix hash code to make sure the mixed hash codes still retain similarity consistency. To this end, loss function $L_3$ and $L_4$ can be written as:

$$L_3 = \sum_{i=1}^{n} \mid S^{CH}_{I,T} - E \mid + \sum_{i=1}^{n} \mid S^{CH}_{I,T} - S^F_{I,T} \mid . \tag{9}$$

$$L_4 = \sum_{i=1}^{n} \mid S_{I,T}^{F} - E \mid . \tag{10}$$

where $E$ is an identity matrix.

### 3.3 Optimization

As mentioned above, the final loss function can be written as follows:

$$min\ \alpha L_1 + \beta L_2 + \gamma L_3 + \delta L_4.$$

The goal of our method is to generate hash codes, a kind of discrete data. The optimization of our objective function should satisfy the discrete condition. The sign function can map the input into $-1$ or 1. The gradient of this function is zero for all non-zero inputs, and may cause gradient explosion in backpropagation:

$$\lim_{\delta \to \infty} tanh(\eta x) = sgn(x). \tag{11}$$

where $\eta$ is a hyper-parameter and will rise during network training.

With sign function as the activation function, the network will finally converge to our hash layer by changing the problem into a sequence of smoothed optimization problems.

## 4 Experiment

### 4.1 Datasets

**MIRFlickr25k** [6] contains $25,000$ image-text pairs collected from the image website Flickr. The image-text pairs are labeled from 24 categories. All the images are denoted as SIFT feature. We use BoW vector to form the text tags with 1386 dimensions.

**NUS-WIDE** [3] consists of $269,648$ pairs of images and texts. There are 81 label categories in the dataset, but we only used the top 10 most frequent categories, resulting in a total of 186,577 image-text pairs that can be used. The setup for this dataset is the same as the other methods. We use BoW vector to form the text tags with 500 dimensions.

### 4.2 Baselines and Evaluaton

We compare our SAH with 6 baseline methods, including CVH [9], IMF [20], CMFH [4], LSSH [25], UDCMH [24], and DJSRH [21].

**Evaluation Criterion.** Mean Average Precision (mAP) [14] and the top-K precision curves are used to evaluate the performance of the proposed SAH and baselines. Two instances of different modalities are considered semantically similar if they have the same label.

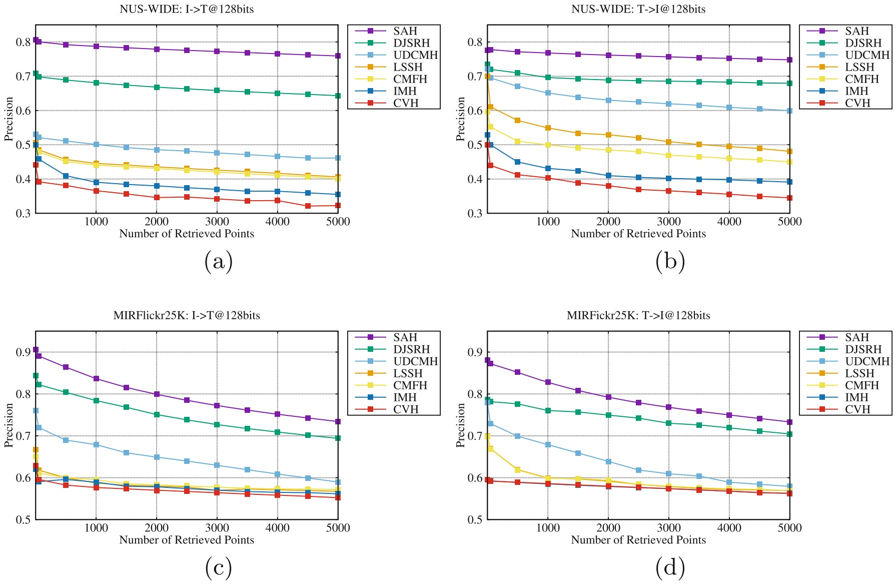
### 4.3 Implementation Details

The network of each modality is composed of the uniform branch and the auxiliary branch. For image modality, our uniform branch is composed of AlexNet which is same with UDCMH [24] for the sake of fairness. The auxiliary branch deal with the input image data and get three scale auxiliary data of image, 1024 × 1024, 512 × 512, 256 × 256, respectively. For text modality, MLP is uniform branch. In auxiliary branch, the lengths of text auxiliary data in four scale are 1024, 512, 256 and 128, respectively. For hyper-parameter, we set $\alpha = 1$, $\beta = 0.1$, $\gamma = 1$, $\delta = 1$, and $\mu = 0.1$ to achieve best performance. We implement our method by PyTorch on the NVIDIA RTX 1660Ti. We fix batch size as 32 and the learning rate for image network and text network is 0.005. During the optimization phase, we employ mini-batch optimizer to optimize our networks of two modalities.

### 4.4 Comparison with Existing Methods

**Results on MIRFlickr25k.** Table 2 shows the MAP@50 on MIRFlickr25K dataset of our proposed SAH and other previous methods. As can be seen, the proposed SAH significantly outperforms the baselines. We show the curve of 128 bits length hash code and can easily find that our SAH has the best performance. For the I→T retrieval, we get more than 50% improvement in MAP in 128 bits compared with CVH. Compare with the latest method DJSRH, we get 3.1% enhancement in 128 bits. For the T→I retrieval, we also achieve the superior performance compare with methods. The difference value of I→T and T→I has a shrink than any other works, which means that the auxiliary data bridges the heterogeneous gap (Fig. 4).

**Results on NUS-WIDE.** Table 3 also shows the MAP@50 on NUS-WIDE dataset of six methods, which shows that our SAH performs better than other methods. It can be seen that we get the best performance on four kinds of code length for two datasets, which means that our method is effective for cross modality retrieval. The results indicate that the auxiliary data of both modalities could mine more latent information in both modalities and remain the similarity consistency.



(a)

(b)

(c)

(d)

**Fig. 4.** Precision@top-K curves on two datasets at 128 bits.

**Table 2.** Mean average precision (MAP@50) comparison results.

| Task | Method | MIRFlickr25K | | | | NUS-WIDE | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 16 bit | 32 bit | 64 bit | 128 bit | 16 bit | 32 bit | 64 bit | 128 bit |
| I→T | CVH | 0.606 | 0.599 | 0.596 | 0.598 | 0.372 | 0.362 | 0.406 | 0.390 |
| | IMH | 0.612 | 0.601 | 0.592 | 0.579 | 0.470 | 0.473 | 0.476 | 0.459 |
| | CMFH | 0.642 | 0.662 | 0.676 | 0.685 | 0.529 | 0.577 | 0.614 | 0.645 |
| | LSSH | 0.584 | 0.599 | 0.602 | 0.614 | 0.481 | 0.489 | 0.507 | 0.507 |
| | UDCMH | 0.689 | 0.698 | 0.714 | 0.717 | 0.511 | 0.519 | 0.524 | 0.558 |
| | DJRSH | 0.810 | 0.843 | 0.862 | 0.876 | 0.724 | 0.773 | 0.798 | 0.817 |
| | **OURS** | **0.852** | **0.879** | **0.889** | **0.903** | **0.753** | **0.779** | **0.804** | **0.818** |
| Task | Method | MIRFlickr25K | | | | NUS-WIDE | | | |
| | | 16 bit | 32 bit | 64 bit | 128 bit | 16 bit | 32 bit | 64 bit | 128 bit |
| T→I | CVH | 0.591 | 0.583 | 0.576 | 0.576 | 0.401 | 0.384 | 0.442 | 0.432 |
| | IMH | 0.603 | 0.595 | 0.589 | 0.580 | 0.478 | 0.483 | 0.472 | 0.462 |
| | CMFH | 0.642 | 0.662 | 0.676 | 0.685 | 0.529 | 0.577 | 0.614 | 0.645 |
| | LSSH | 0.584 | 0.599 | 0.602 | 0.614 | 0.455 | 0.459 | 0.468 | 0.473 |
| | UDCMH | 0.692 | 0.704 | 0. 718 | 0.733 | 0.637 | 0.653 | 0.695 | 0.716 |
| | DJRSH | 0.786 | 0.822 | 0.835 | 0.847 | 0.712 | 0.744 | 0.771 | 0.789 |
| | **OURS** | **0.852** | **0.864** | **0.878** | **0.885** | **0.765** | **0.772** | **0.786** | **0.791** |

### 4.5 Ablation Study

We verify our method with 3 variants as diverse baselines of SAH:

(a) SAH-1 is built by removing the intra-modality multi-scale hash codes semantic enhancement;
(b) SAH-2 is built by removing the similarity matrices difference between uniform hash codes and uniform features;
(c) SAH-3 is built by removing the consistency between complex hash codes similarity and uniform features similarity.

Table 3 shows the results on MIRFlickr25K dataset with 64 bits and 128 bits. From the results, we can observe that each part is important to our method. Especially the part of similarity consistency between features and complex hash codes, which ensures the semantic consistency.

**Table 3.** The mAP@50 results for ablation analysis on MIRFlickr25k.

| Method | 64 bits | | 128 bits | |
|---|---|---|---|---|
| | I→T | T→I | I→T | T→I |
| SAH | 0.889 | 0.878 | 0.903 | 0.885 |
| SAH-1 | 0.881 | 0.869 | 0.885 | 0.883 |
| SAH-2 | 0.877 | 0.863 | 0.898 | 0.880 |
| SAH-3 | 0.855 | 0.849 | 0.863 | 0.834 |

## 5 Conclusion

In this paper, we propose a novel unsupervised deep hashing model named self-auxiliary hashing. We propose a two-branch network for each modality, mixing the uniform hash codes and the comprehensive hash codes, which can preserve richer semantic information and bridge the gap of different modalities. Moreover, we make a full use of inter-modality similarity matrices and the multi-scale intra-modality similarity matrices to learn the similarity information. Extensive experiments conducted on two datasets show that our SAH outperforms several baseline methods for cross modality retrieval.

## References

1. Cao, Y., Long, M., Wang, J., Yang, Q., Yu, P.S.: Deep visual-semantic hashing for cross-modal retrieval. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016, pp. 1445–1454. ACM (2016). https://doi.org/10.1145/2939672.2939812

2. Cao, Y., Long, M., Wang, J., Zhu, H.: Correlation autoencoder hashing for supervised cross-modal search. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR 2016, New York, New York, USA, 6–9 June 2016, pp. 197–204. ACM (2016). https://doi.org/10.1145/2911996.2912000
3. Chua, T., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: a real-world web image database from national university of Singapore. In: Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009, Santorini Island, Greece, 8–10 July 2009. ACM (2009)
4. Ding, G., Guo, Y., Zhou, J.: Collective matrix factorization hashing for multimodal data. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, 23–28 June 2014, pp. 2083–2090. IEEE Computer Society (2014)
5. Du, G., Zhou, L., Yang, Y., Lü, K., Wang, L.: Deep multiple auto-encoder-based multi-view clustering. Data Sci. Eng. **6**(3), 323–338 (2021)
6. Huiskes, M.J., Lew, M.S.: The MIR flickr retrieval evaluation. In: Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2008, Vancouver, British Columbia, Canada, 30–31 October 2008, pp. 39–43. ACM (2008)
7. Jiang, Q., Li, W.: Deep cross-modal hashing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 3270–3278. IEEE Computer Society (2017)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
9. Kumar, S., Udupa, R.: Learning hash functions for cross-view similarity search. In: Walsh, T. (ed.) IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, 16–22 July 2011, pp. 1360–1365. IJCAI/AAAI (2011)
10. Li, C., Deng, C., Li, N., Liu, W., Gao, X., Tao, D.: Self-supervised adversarial hashing networks for cross-modal retrieval. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 4242–4251. IEEE Computer Society (2018)
11. Liong, V.E., Lu, J., Tan, Y., Zhou, J.: Cross-modal deep variational hashing. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017, pp. 4097–4105. IEEE Computer Society (2017)
12. Liu, H., Ji, R., Wu, Y., Huang, F., Zhang, B.: Cross-modality binary code learning via fusion similarity hashing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 6345–6353. IEEE Computer Society (2017)
13. Liu, W., Mu, C., Kumar, S., Chang, S.: Discrete graph hashing. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, 8–13 December 2014, Montreal, Quebec, Canada, pp. 3419–3427 (2014). https://proceedings.neurips.cc/paper/2014/hash/f63f65b503e22cb970527f23c9ad7db1-Abstract.html
14. Liu, W., Mu, C., Kumar, S., Chang, S.F.: Discrete graph hashing (2014)
15. Liu, W., Wang, J., Ji, R., Jiang, Y., Chang, S.: Supervised hashing with Kernels. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012, pp. 2074–2081. IEEE Computer Society (2012)

16. Liu, X., Nie, X., Zeng, W., Cui, C., Zhu, L., Yin, Y.: Fast discrete cross-modal hashing with regressing from semantic labels. In: 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, 22–26 October 2018, pp. 1662–1669. ACM (2018)
17. Lu, J., Chen, M., Sun, Y., Wang, W., Wang, Y., Yang, X.: A smart adversarial attack on deep hashing based image retrieval. In: Proceedings of the 2021 International Conference on Multimedia Retrieval, ICMR 2021, pp. 227–235. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3460426.3463640
18. Nie, X., Wang, B., Li, J., Hao, F., Jian, M., Yin, Y.: Deep multiscale fusion hashing for cross-modal retrieval. IEEE Trans. Circ. Syst. Video Technol. **31**(1), 401–410 (2021). https://doi.org/10.1109/TCSVT.2020.2974877
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015)
20. Song, J., Yang, Y., Yang, Y., Huang, Z., Shen, H.T.: Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, 22–27 June 2013, pp. 785–796. ACM (2013). https://doi.org/10.1145/2463676.2465274
21. Su, S., Zhong, Z., Zhang, C.: Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), 27 October–2 November 2019, pp. 3027–3035. IEEE (2019)
22. Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, 23–27 October 2017, pp. 154–162. ACM (2017)
23. Wawrzinek, J., Pinto, J., Wiehr, O., Balke, W.T.: Exploiting latent semantic subspaces to derive associations for specific pharmaceutical semantics. Data Sci. Eng. **5**, 333–345 (2020)
24. Wu, G., et al.: Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval. In: Lang, J. (ed.) Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, 13–19 July 2018, Stockholm, Sweden, pp. 2854–2860. ijcai.org (2018)
25. Zhou, J., Ding, G., Guo, Y.: Latent semantic sparse hashing for cross-modal similarity search. In: The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2014, Gold Coast, QLD, Australia, 6–11 July 2014, pp. 415–424. ACM (2014)
26. Zhou, J., Ding, G., Guo, Y., Liu, Q., Dong, X.: Kernel-based supervised hashing for cross-view similarity search. In: IEEE International Conference on Multimedia and Expo, ICME 2014, Chengdu, China, 14–18 July 2014, pp. 1–6. IEEE Computer Society (2014)