



# Semantic-guided autoencoder adversarial hashing for large-scale cross-modal retrieval

Mingyong Li<sup>1</sup> · Qiqi Li<sup>1</sup> · Yan Ma<sup>1</sup> · Degang Yang<sup>1</sup>

Received: 26 May 2021 / Accepted: 26 November 2021 / Published online: 4 January 2022  
© The Author(s) 2021

## Abstract

With the vigorous development of mobile Internet technology and the popularization of smart devices, while the amount of multimedia data has exploded, its forms have become more and more diversified. People's demand for information is no longer satisfied with single-modal data retrieval, and cross-modal retrieval has become a research hotspot in recent years. Due to the strong feature learning ability of deep learning, cross-modal deep hashing has been extensively studied. However, the similarity of different modalities is difficult to measure directly because of the different distribution and representation of cross-modal. Therefore, it is urgent to eliminate the modal gap and improve retrieval accuracy. Some previous research work has introduced GANs in cross-modal hashing to reduce semantic differences between different modalities. However, most of the existing GAN-based cross-modal hashing methods have some issues such as network training is unstable and gradient disappears, which affect the elimination of modal differences. To solve this issue, this paper proposed a novel Semantic-guided Autoencoder Adversarial Hashing method for cross-modal retrieval (SAAH). First of all, two kinds of adversarial autoencoder networks, under the guidance of semantic multi-labels, maximize the semantic relevance of instances and maintain the immutability of cross-modal. Secondly, under the supervision of semantics, the adversarial module guides the feature learning process and maintains the modality relations. In addition, to maintain the inter-modal correlation of all similar pairs, this paper use two types of loss functions to maintain the similarity. To verify the effectiveness of our proposed method, sufficient experiments were conducted on three widely used cross-modal datasets (MIRFLICKR, NUS-WIDE and MS COCO), and compared with several representatives advanced cross-modal retrieval methods, SAAH achieved leading retrieval performance.

**Keywords** Deep hashing · Cross-modal retrieval · Adversarial autoencoder · Adversarial learning

## Introduction

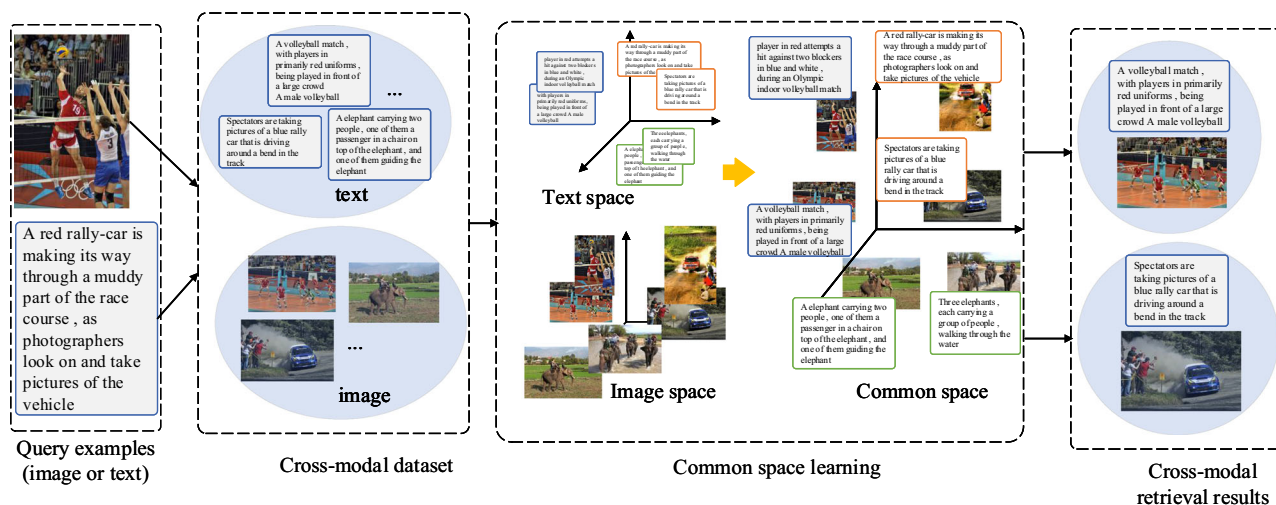
In recent years, with the widespread popularity of the Internet and mobile devices, the scale of multimodal data (text, image, video, audio, etc.) has increased dramatically. While the amount of multimedia data has exploded, its forms have become more and more diversified. People's demand for information is no longer satisfied with single-modal data retrieval, and cross-modal retrieval has become a research hotspot in recent years. For example, given a query image, it may be necessary to retrieve a set of text that best describes the image, or match the given text to a set of visually

related images. Cross-modal retrieval tasks can efficiently analyze multi-modal data semantic relevance, to achieve mutual matching between different modalities. To reduce the cost of finding the nearest neighbor, Approximate Nearest Neighbor (ANN) [1] has become the most commonly used retrieval method in cross-modal retrieval tasks. In recent years, the hash feature representation of data has the advantages of small storage space and fast retrieval speed, so it has received extensive attention in the field of large-scale information retrieval [1–3,6,27,28].

As the data of different modalities are heterogeneous and their distribution and presentation are inconsistent, the key to cross-modal retrieval is “modality gap”, that is, how to measure the similarity between different modal representations [8,29]. The current mainstream method is the common space learning method, the purpose of this method is to learn the features of different modalities in an intermediate common

✉ Yan Ma  
cqnu\_mayan@163.com

<sup>1</sup> College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China



**Fig. 1** A brief illustration of common space learning method for cross-modal retrieval, which can present retrieval results with different modalities by a query of any modality

space and measure their similarity [29]. A brief description of the common space learning method is shown in Fig. 1.

cross-modal hashing (CMH) method is considered as one of the best methods to solve the cross-media retrieval problem [9–11,30]. It encodes samples of different modalities into short binary codes, the search of hash code can efficiently carry out cross-media retrieval. Among the existing cross-media hashing methods, deep cross-media hashing has achieved great success [12–14,21,31,32].

Although some breakthroughs have been made in this field, there are still some problems in deep cross-modal hashing. First of all, in cross-media adversarial learning, the researcher often use GAN as the adversarial module. However, most of the existing GAN-based cross-modal retrieval methods mainly use the original GAN loss function and training strategy, which leads to the problems of unstable network training and gradients disappear, which affect the elimination of modal differences to a certain extent. At present, there is also a small amount of work that uses autoencoders for cross-modal retrieval. The existing cross-modal retrieval methods based on autoencoders mainly adopt the reconstruction strategy based on mean square error. Compared with the original input, the decoded output has a certain information loss, and the original features cannot be better preserved.

To solve the above problems, in this paper, a novel Semantic-guided Autoencoder Adversarial Hashing method (SAAH) is proposed. As shown in Fig. 2, this is an end-to-end neural network structure that can perform both feature representation and hash coding. To facilitate feature learning and make up for the modal gap, this paper designed two kinds of adversarial autoencoder modules (inter-modal and intra-modal) based on semantic multi-labels. The intra-modal adversarial network improves the intra-modal reconstruction process of the autoencoder, and uses the idea of adversarial

learning to make it difficult to distinguish the input features and reconstruction features. The inter-modal adversarial network is used to reduce the difference between the modals, so that the samples with the same semantics from different modalities can generate uniform semantic features and binary codes. Through the minimax training strategy, the learned features are optimized during the adversarial learning process to achieve the consistency of the distribution of different representation modalities. Finally, image modal data and text modal data are difficult to distinguish.

The main contributions are summarized as follows:

1. This paper proposed a novel semantic-guided adversarial autoencoder hashing method (SAAH). we designed two kinds of adversarial autoencoder networks (Inter-modal adversarial network and Intra-modal adversarial network). Under semantic supervision, the adversarial networks guide the feature learning process and maintain the modal relationship between common feature space and common hamming space. The joint optimization of the two types of adversarial autoencoder networks can effectively eliminate the distribution differences between modalities and improve retrieval accuracy.
2. The proposed adversarial cross-modal hashing method integrates three loss functions, including the inter-modal triplet loss, classification prediction loss and inter-modal pairwise loss. Therefore, a more discriminative hash code can be generated.
3. To verify the effectiveness of our proposed method, sufficient experiments were conducted on three widely used cross-modal datasets (MIRFLICKR, NUS-WIDE and MS COCO), and compared with several representative advanced cross-modal retrieval methods, SAAH achieved leading retrieval performance.

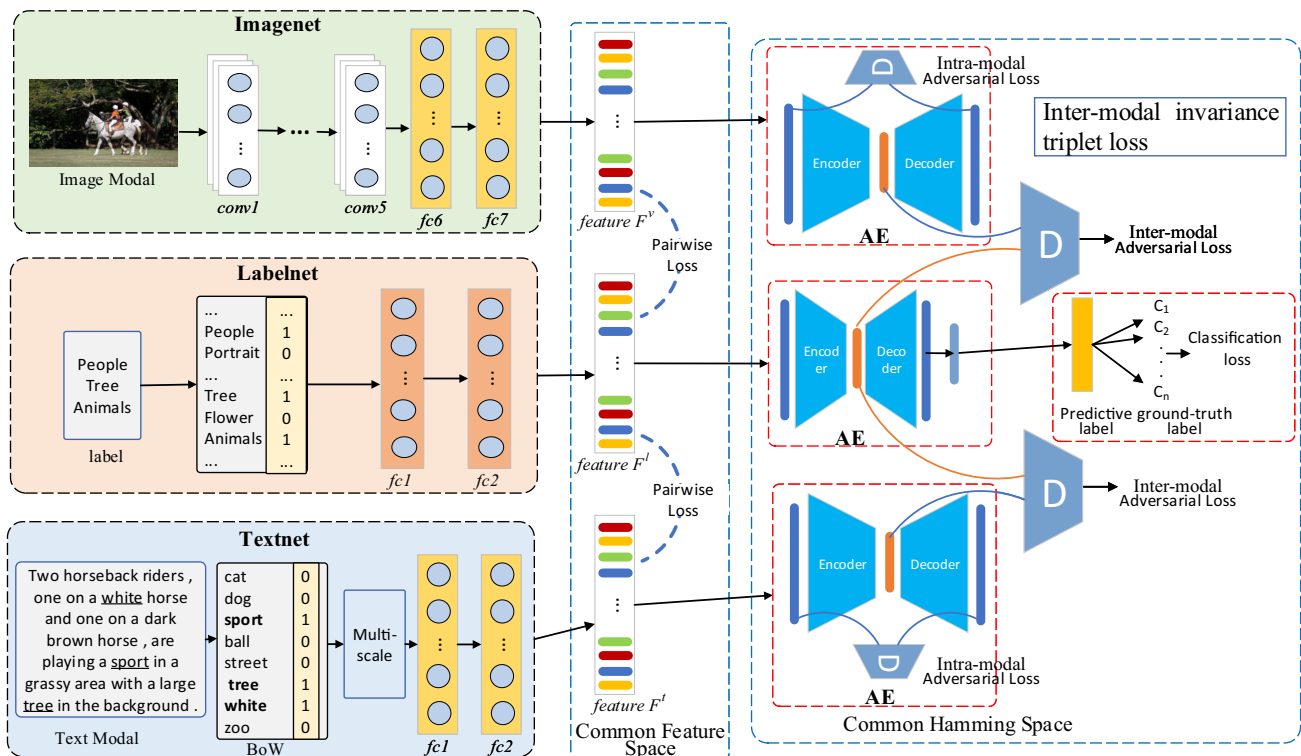


Fig. 2 The framework of the proposed SAAH

## Related works

Cross-modal retrieval is achieved by modeling the relationship between different modes. Therefore, the first problem to be solved by cross-modal retrieval is the heterogeneous problem caused by the different distribution and structure among different modality [8,12,18,29,33,34,38].

## Non-GAN-based cross-modal retrieval methods

According to whether the supervised information is used in the training, the cross-modal hashing methods can be divided into unsupervised method [4,5,16,17] and supervised method [18–20]. CVH [16] proposed by Kumar et al. enables data representing different views of the same objects to have the same hash codes. LSSH [5] uses sparse coding and matrix decomposition to capture the latent semantic features of images and texts. CMFH [4] proposed by Ding et al. uses collaborative Matrix decomposition to decompose data of different modalities into basis Matrix and coefficient Matrix, and uses Matrix decomposition to learn the hidden factors of different modalities and generate unified hash codes. STMH [17] models image as potential semantic concepts, models text as multiple semantic topics, and then learns the relationship between text and image in the potential semantic space.

The supervised methods always perform better because of the usage of label information. SCM [19] proposed by

Zhang et al. uses non-negative matrix decomposition and a nearest-neighbor preserving algorithm to maintain semantic consistency both inter-modal and intra-modal. CMSSH [18] proposed by Zhu et al. used supervised similarity learning to map input data from two arbitrary spaces to hamming space. SePH [20] proposed by Lin et al. approximates the probability distribution of training data into the hash codes in hamming space by minimizing the KL entropy, and then uses the kernel logistic regression to learn the nonlinear hash function from each view.

Shallow cross-modal hashing methods are mostly unable to describe complex cross-modal associations. Recent cross-modal deep hashing methods [21,22,31,35,36] have shown better performance in preserving similarities between different modalities. DCMH [21] proposed by Jiang et al. introduced deep learning into the cross-modal hash retrieval algorithm, this method integrates the feature learning process and hash code learning process in the same end-to-end deep neural network to learn more effective hash codes. As an improvement, PRDH [22] proposed by Yang et al. explored Pairwise constraints from inter-modal and intra-modal to find heterogeneous associations between different modalities and maintain the semantic similarity of hash codes learned. DMFH [40] introduced a multi-scale fusion network to enable more effective feature extraction when learning hash codes. SCAHN [47] introduces the attention mechanism for common representation enhancement while increasing

the weight of each hash code to characterize each bit's coding ability. DSCA [48] proposed a correlation-aligned multi-semantic image-text hashing framework. The similarity between the modalities is produced by the semantic label and the original characteristics of the data. The constructed covariance matrix achieves more effective cross-modal correlation modeling through alignment. MS2GAN [49] divides common representation into modal independent representation and modal consistent representation, and uses interval loss based on given semantic annotations to maintain the structure based on common representation to improve representation learning ability.

### GAN-based adversarial cross-modal retrieval

At this stage, the cross-modal retrieval method based on GAN [7,14,24,34,37] has become a new research hotspot. GAN consists of two parts: Generator ( $G$ ) and Discriminator ( $D$ ). The purpose of  $G$  is to learn the distribution close to the real sample to confuse  $D$ , and the function of  $D$  is to distinguish whether the data comes from the real sample or the sample generated by  $G$ .  $G$  and  $D$  perform a mini-max game. In an ideal state,  $G$  can generate  $G(z)$  that is enough to "make the fake and the real", but for  $D$ , it is difficult to determine whether the output of  $G$  is real or not.

ACMR [24] and SSAH [14] are two typical early research work, and they first introduce GAN so that they can align the code distribution in different modalities, reduce the heterogeneous gap and improve the performance of cross-modal retrieval. The GANs improved the performance of cross-modal retrieval by reducing the heterogeneity gap of different modalities. ACMR [24] is the first method to apply GANs framework into cross-modal retrieval. It uses a minimax strategy to train the network, the feature mapper and the modality classifier interact with each other, and the process of minimization and maximization between the two parts eliminates the difference of feature representation of different modalities. The same idea of the adversarial network is adopted, SSAH [14] further combines adversarial learning with hash technique, which utilizes two adversarial networks to maximize the semantic relevance and consistency of the representations of different modality, and designed a self-supervised semantic network, supervised the training of the other two networks. CM-GANs [7] realizes the consistency of the modal through the intra-modal and inter-modal discriminator, at the same time, the data is reconstructed by generating adversary, so as to learn more discriminative common representation features. UCH [34] and CYC-DGH [37] are GAN-based hashing method for unsupervised cross-modal retrieval. UCH [34] consists of a two-cycle generated adversarial hashing network. The outer-cycle GAN is used to learn common representations, while the inner-cycle GAN is used to generate reliable hash codes.

In the last 2 years, some of the latest GAN-based work has appeared. DAML [41] non-linearly project the data of different modalities to the latent feature space, the purpose is to learn the representation of the invariance between the modalities. MHTN [44] realizes the transfer of knowledge from the single-modal source domain to the target source domain, and learns cross-modal public representation. UGACH [45] proposed a graph-based generative adversarial hash learning framework. Given the data in one modal, the generative model selects data pairs from other modalities based on the shared learning to challenge the discriminant model. The discriminant model distinguishes the generated data pair from the actual data pair collected in the relationship graph. The framework is further extended to handle five-modal data and perform cross-modal retrieval in a more general sense [46]. CPAH [39] proposed a multi-task consistency-maintaining adversarial network for image-text hashing. Two modules were developed, namely the consistency refinement module (CR) and the multi-task adversarial learning module (MA) to learn semantic consistency information.

However, the cost of continuously training the discriminator to approach the optimal discriminator is that the loss function of the discriminator converges quickly, and the gradient of the generator cannot be updated continuously, causing the gradient of the generator to disappear. At the same time, most of the existing GAN-based cross-modal retrieval methods mainly use the original GAN loss function and training strategy, which leads to the instability of network training and the disappearance of gradients, which affect the elimination of modal differences to a certain extent.

### Autoencoder-based cross-modal retrieval

Since autoencoders naturally have the ability to generate compact binary codes, in recent years, researchers have proposed a cross-modal retrieval method based on deep autoencoders. Autoencoder consists of two parts: Encoder and Decoder. The output data after training are the hidden feature (encoding feature) of the autoencoder. The encoding operation is the process of projecting data from the input layer to the hidden layer, and the corresponding decoding operation is the reconstruction process of projecting the encoded features obtained from the hidden layer as the input to the output layer. The main feature of the autoencoder is to encode high-dimensional data to reduce the dimensionality, and then reconstruct the input data through decoding. The cross-modal retrieval method based on deep autoencoders is mainly to conduct correlation learning on the hidden features of two single-modal autoencoders.

Correspondence Autoencoder (Corr-AE) [42] correlates the implicit representations of two single-modal autoencoders, constructs an optimal goal, and minimizes the correlation learning error between the two modal implicit



representations of the autoencoder. Multi-modal Semantic Autoencoder (MMSAE) [43] learns multi-modal mapping in two stages, projects multi-modal data to obtain low-dimensional embedding, and uses autoencoder to achieve cross-modal reconstruction. Existing cross-modal retrieval methods based on autoencoders mainly adopt a reconstruction strategy based on mean square error. Compared with the original input, the decoded output has a certain information loss, and the original features cannot be better preserved.

Aiming at the above shortcomings, this paper proposes a novel cross-modal retrieval method based on the Adversarial Autoencoder (AAE), as shown in Fig. 1. Two types of adversarial autoencoder networks are designed (intra-modal adversarial network and inter-modal adversarial network). The intra-modal adversarial network improves the intra-modal reconstruction process of the autoencoder. The discriminator module tries to distinguish between input features and reconstruction features. Finally, it is difficult to distinguish the input features and reconstruction features. The inter-modal adversarial network is used to reduce the differences between the modalities, so that samples with the same semantics from different modalities generate unified semantic features and binary codes in the common semantic space and the Hamming space. Through the minimax training strategy, in the end, image modal data and text modal data are difficult to distinguish. The method in this paper combines two types of adversarial autoencoder models, which can effectively eliminate the distribution differences between modalities and improve retrieval accuracy.

## Proposed method

### Problem definition

Let us start with some of the notations used in this paper. Given a cross-modal dataset  $O = \{o_i\}_{i=1}^n$ ,  $o_i \in (v_i, t_i, l_i)$ , where  $v_i \in \mathbb{R}^{d_v}$  is the original image feature representation of the  $i$ -th sample, and  $t_i \in \mathbb{R}^{d_t}$  is text feature representation. A semantic label vector  $l_i = [l_{i1}, \dots, l_{ik}] \in \mathbb{R}^k$  is assigned for  $o_i$ , where  $k$  is the total class number.  $o_i$  and  $o_j$  are associated with similarity label  $s_{ij}$ , where  $s_{ij} = 1$  implies  $o_i$  and  $o_j$  are similar, or otherwise  $s_{ij} = 0$ . Considering the samples are multi-label, define  $s_{ij} = 1$  if  $o_i$  and  $o_j$  share at least one label, and  $s_{ij} = 0$  if  $o_i$  and  $o_j$  have no common label. Our goal is to learn the unified hash code for image and text modalities:  $b^{v,t} \in \{-1, 1\}^K$ . The detailed symbols definition is shown in Table 1.

Like Euclidean distance, Hamming distance is a measure of distance, which is used to measure the similarity of binary code. The Hamming distance can be calculated as the inner product of two hash codes. For two binary codes  $b_i$  and  $b_j$ , their hamming distance  $\text{dis}_H(b_i, b_j)$  and

**Table 1** The notation of the proposed method

Notation	Illustration
$O$	The cross-modal dataset
$o_i \in (v_i, t_i, l_i)$	The $i$ -th sample
$v_i; t_i; l_i$	The original image feature; the original text feature; the semantic label vector
$S$	The label similar matrix ( $s_{ij} \in S$ )
$s_{ij}$	$s_{ij} = 1$ implies $o_i$ and $o_j$ are similar, or otherwise $s_{ij} = 0$
$h$	Hash code
$H$	The hash code space
$B$	The binary code space
$K$	The code length
$\theta^{v,t,l}$	The neural network parameter
$f_i^l$	The semantic feature from Labelnet
$f_j^{v,t}$	The semantic feature generated by Imagenet and Textnet
$F$	The feature representation space
$D^{v,l}; D^{t,l}$	The image discriminator; the text discriminator

inner product  $\langle b_i, b_j \rangle$  can be formulated as:  $\text{dis}_H(b_i, b_j) = \frac{1}{2}(K - \langle b_i, b_j \rangle)$ , where  $K$  is the length of the binary code, so the similarity between two binary codes can be quantized using the inner product. Given  $S$ , the probability of  $S$  under condition  $b_i$  and  $b_j$  is defined as a likelihood function:

$$p(s_{ij}|b_i, b_j) = \begin{cases} \sigma(\varphi_{ij}), & s_{ij} = 1 \\ 1 - \sigma(\varphi_{ij}), & s_{ij} = 0, \end{cases} \quad (1)$$

where  $\sigma(\varphi_{ij}) = \frac{1}{1+e^{-\varphi_{ij}}}$  is the sigmoid function, and  $\varphi_{ij} = \frac{1}{2}b_i^T b_j$ . We can see that the smaller hamming distance  $\text{dis}_H(b_i, b_j)$  is, the larger their inner product  $\langle b_i, b_j \rangle$ . A larger condition probability  $p(1|b_i, b_j)$  implies  $b_i$  and  $b_j$  should be similar; otherwise, a larger condition probability  $p(0|b_i, b_j)$  means  $b_i$  and  $b_j$  should be dissimilar.

### Framework overview

The SAAH framework proposed in this paper is shown in Fig. 2. The framework consists of two parts: feature generation part (left) and adversarial learning part (right).

*The feature generation part* In this part, three neural networks are adopted, namely Imagenet, Labelnet, and Textnet, which are used to extract the features of the original samples and map them to a common feature space. Imagenet is used for image modality. It adopts the classic convolutional neural network CNN-F, and its output is generated into

image feature representation in the common feature space. The semantic features of Imagenet are finally input into the autoencoder to generate hash codes, this autoencoder belongs to the adversarial learning part. Similarly, Textnet is used for text modal, which contains two fully connected layers and a multi-scale module. The semantic features extracted by Textnet are also used as input to the corresponding autoencoder to generate corresponding hash codes. However, the role of Labelnet is different from Imagenet and Textnet. The labelnet learns semantic features from multi-label information, then its most important role is to supervise the features learning of image and text modal.

The adversarial learning part comprises 3 autoencoders and 2 types of discriminators (4 in total). A kind of discriminators (2 in total) is used in inter-modal adversary to progressively reduce the distribution differences of image and text features by the adversarial learning way. Another discriminators (2 in total) is used in intra-modal, the aim is to reduce the feature representation error after the reconstruction of the autoencoder.

### Supervised semantic generated by Labelnet

As shown in Fig. 2, this paper selected a sample in the MIRFLICKR-25K dataset, this example is annotated with multi-labels, such as ‘tree’, ‘people’ and ‘animals’. Therefore, we can use multi-label annotation as a kind of supervised information to establish the semantic relation between image and text modalities. The established Labelnet adopts the end-to-end fully connected model, which can be used to model the semantic association between image and text. Labelnet extracts semantic features of multi-labels vectors to monitor the learning process of Imagenet and Textnet. A triplet  $(v_i, t_i, l_i)$  is used to describe the same  $i$ -th sample, we regard  $l_i$  as semantic information for  $v_i$  and  $t_i$ .

In common feature space, Labelnet is used to extract rich semantic associations in label information. The logarithmic maximum estimated by hash code mapping can be expressed as:

$$\begin{aligned} \log p(H^l|S) &\propto \log p(S|H^l)p(H^l) \\ &= \sum_{s_{ij} \in S} \log p(s_{ij}|h_i^l, h_j^l)p(h_i^l, h_j^l), \end{aligned} \quad (2)$$

where  $\log p(S|H^l)$  is the likelihood function, and  $p(H^l)$  is the prior distribution.  $H^l$  denote the hash codes in a common hamming space for labels.  $s_{ij}$  indicates whether sample  $i$  and  $j$  contain at least one same label, and if so,  $S_{ij} = 1$ , indicating that sample  $i$  and  $j$  are semantically similar. If not included,  $S_{ij} = 0$ , indicating that the sample  $i$  and  $j$  are not semantically similar. To represent the similarity of features generated by the labels of sample  $i$  and  $j$ , the loss

function can be defined as follows according to the negative log likelihood of pairs of labels:

$$\begin{aligned} \min_{F^l, \theta^l} L_{\text{pairwise}}^l &= -\log p(S|F^l) \\ &= -\sum_{i,j=1}^n (s_{ij} \langle f_i^l, f_j^l \rangle - \log(1 + e^{\langle f_i^l, f_j^l \rangle})), \end{aligned} \quad (3)$$

where  $\langle f_i^l, f_j^l \rangle$  represents the cosine similarity of the semantic features generated by the label of sample  $i$  and  $j$ . When  $S_{ij} = 1$ ,  $\min L_{\text{pairwise}}^l = -\sum_{i,j=1}^n (\log(\frac{e^{\langle f_i^l, f_j^l \rangle}}{1+e^{\langle f_i^l, f_j^l \rangle}})) = \max \langle f_i^l, f_j^l \rangle$ , the loss function maximizes the cosine similarity of the semantic feature generated by the label of sample  $i$  and  $j$ ; When  $S_{ij} = 0$ ,  $\min L_{\text{pairwise}}^l = -\sum_{i,j=1}^n (\log(\frac{1}{1+e^{\langle f_i^l, f_j^l \rangle}})) = \min \langle f_i^l, f_j^l \rangle$ , the loss function minimizes their cosine similarity. This is entirely consistent with the goal of maintaining similarity between semantic features.

In addition, this paper use binary regularization to reduce the error of hash value discretization. The regularization term is defined as follows:

$$\min_{B^l, \theta^l} L_{\text{regular}}^l = \|H^l - B^l\|_F^2, \quad (4)$$

where  $B^l$  is the binary code obtained by the symbol operation of  $H^l$ .  $L_{\text{regular}}^l$  is the approximate loss of the binarization of the hash code, which makes  $H^l$  and  $B^l$  as close as possible, so that the elements in the hash vector are as close as possible  $\{-1, 1\}$ , and the loss is reduced  $H \rightarrow B$ .

Finally, to maintain accurate classification information when training Labelnet, this paper remapped the hash codes obtained from the common Hamming space to the original label space.  $\hat{L}^l$  is the prediction labels recovered by the feature. Therefore, the predicted label can be written as:  $\hat{L}^l = W^T H^l + b$ ,  $W$  is the mapping weight. Define the following loss to minimize the distance between the predicted value  $\hat{L}$  and the ground truth value  $L$ :

$$\min_{\hat{L}^l, \theta^l} L_{\text{predict}}^l = \|\hat{L}^l - L\|_F^2, \quad (5)$$

$L_{\text{predict}}^l$  represents the classification loss of the feature between the original label and the predicted label, so that the recovered label is as same as the original label feature as far as possible. Therefore, the total generation objective function of Labelnet is as follows:

$$\min_{B^l, \theta^l, \hat{L}^l} L_{\text{lab}} = L_{\text{pairwise}}^l + \alpha L_{\text{regular}}^l + \beta L_{\text{predict}}^l$$

$$= - \sum_{i,j=1}^n (s_{ij} \langle f_i^l, f_j^l \rangle - \log(1 + e^{\langle f_i^l, f_j^l \rangle})) + \alpha \|H^l - B^l\|_F^2 + \beta \|\hat{L}^l - L\|_F^2, \tag{6}$$

where  $\alpha, \beta$  are hyper-parameters that balance the weight of  $L_{\text{pairwise}}^l, L_{\text{regular}}^l$  and  $L_{\text{predict}}^l$ .

### Feature learning for image and text modality

In this paper, the feature learning of image and text modalities is supervised, the semantic information generated by Labelnet supervises the learning process of these two modalities.

For image modality, the image feature learning network (Imagenet) established by us adopts CNN-F [23] structure, which projects images into the common feature space. Image feature learning is carried out under the supervision of Labelnet, so that Imagenet and Labelnet keep the same semantic correlation. Similarly, for text modality, this paper relies on label features generated by Labelnet to monitor the learning process of Textnet features. This paper uses a multi-scale model to extract text features.

We hope to define such an objective function to retain semantic information generated by Labelnet in the Textnet and Imagenet during the training process, therefore, we hope the predicted label is similar to the real label. The features and hash codes of text and image extracted by Imagenet and Textnet are as same as the features and hash codes generated by Labelnet. Therefore, when learning the image and text feature, the supervised information also constrains the similarity of feature extraction and feature generation.

In the common feature space of Labelnet and Imagenet, if the sample pair  $v_i$  and  $v_j$  are similar, their corresponding feature representations  $f_i^v$  and  $f_j^v$  should also be similar. Similarly, for text modality, if the sample pairs  $t_i$  and  $t_j$  are similar, their corresponding feature representations  $f_i^t$  and  $f_j^t$  should also be similar. Under the supervision of the semantic features of Labelnet, the semantic features  $F^v$  of Imagenet and Labelnet can be described as follows:

$$\begin{aligned} \min_{F^v, \theta^v} L_{\text{pairwise}}^v &= - \log p(S|F^v) \\ &= - \sum_{i,j=1}^n (s_{ij} \langle f_i^v, f_j^v \rangle - \log(1 + e^{\langle f_i^v, f_j^v \rangle})), \end{aligned} \tag{7}$$

where  $f_i^l$  is the semantic features of Labelnet, and  $f_j^v$  is the semantic features generated by Imagenet.  $\langle f_i^l, f_j^v \rangle$  represents the cosine similarity between the semantic features generated by the label of sample  $i$  and the semantic features extracted from the sample input (image).

The goal is to get  $h^v$  (the extracted hash) as close as possible to  $h^l$  (the hash generated by the label). The approximate loss of learning hash code binarization is defined as follows:

$$\min_{B^v, \theta^v} L_{\text{regular}}^v = \|H^v - B^v\|_F^2. \tag{8}$$

Accordingly, the overall objective function of Imagenet is defined as follows:

$$\begin{aligned} \min_{B^v, \theta^v} L_{\text{img}} &= L_{\text{pairwise}}^v + \alpha L_{\text{regular}}^v + \gamma L_{\text{adv\_inter}}^v \\ &\quad + \eta L_{\text{adv\_intra}}^v + \delta L_{\text{triplet}}^v, \end{aligned} \tag{9}$$

where  $\alpha, \gamma, \eta$  and  $\delta$  is the weight parameter of each loss of  $L_{\text{pairwise}}^v, L_{\text{regular}}^v, L_{\text{adv\_inter}}^v, L_{\text{adv\_intra}}^v$  and  $L_{\text{triplet}}^v$ .  $L_{\text{adv\_inter}}^v$  and  $L_{\text{adv\_intra}}^v$  are adversarial loss for inter-modal and intra-modal, respectively.  $L_{\text{triplet}}^v$  is inter-modal invariance triplet loss, the details are in ‘‘Inter-modal triplet loss’’. Similarly, the total generating objective function of text modality is as follows:

$$\begin{aligned} \min_{B^t, \theta^t} L_{\text{txt}} &= L_{\text{pairwise}}^t + \alpha L_{\text{regular}}^t + \gamma L_{\text{adv\_inter}}^t \\ &\quad + \eta L_{\text{adv\_intra}}^t + \delta L_{\text{triplet}}^t. \end{aligned} \tag{10}$$

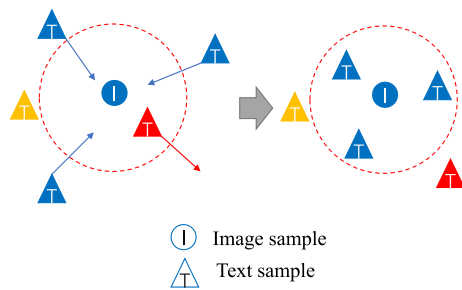
### Inter-modal triplet loss

Modality similarity is maintained by minimizing the distance between all semantic similar instances representations from different modalities, meanwhile, maximizing the distance between dissimilar instance representations. Like the pairwise loss, the triplet loss is a commonly used objective function. Inspired by ACMR [24], to reduce the computational overhead of triplet sampling, samples are taken from the marked instances in each small batch, rather than from the entire instance space. The triplet form of image modal is constructed as follows:  $(v_i, t_j^+, t_k^-)$ , text instance  $t_k^-$  is semantically unrelated to image  $v_i$ , while  $t_j^+$  is the opposite. Similarly, the text modality triplet form  $(t_i, v_j^+, v_k^-)$ . Inter-modal triplet loss across image and text modalities are as follows:

$$L_{\text{triplet}}^v = \sum_{i,j,k} \max \left( \lambda \cdot \|f_{v_i} - f_{t_j^+}\|_2^2 - \|f_{v_i} - f_{t_k^-}\|_2^2, 0 \right), \tag{11}$$

$$L_{\text{triplet}}^t = \sum_{i,j,k} \max \left( \lambda \cdot \|f_{t_i} - f_{v_j^+}\|_2^2 - \|f_{t_i} - f_{v_k^-}\|_2^2, 0 \right). \tag{12}$$

The optimization objective of the loss function (Eq. (11)) is shown in the Fig. 3. Where  $\lambda$  is the margin parameter,



**Fig. 3** A simple demonstration of the optimization goal about the loss function. Taking the image sample as an example, it pulls the text samples with the same semantics (represented by the same color) closer, while pushing the text samples with different semantics farther

$(f_{v_i}, f_{t_j^+}, f_{t_k^-})$  and  $(f_{t_i}, f_{v_j^+}, f_{v_k^-})$  represent their feature representations, respectively. Combining Eqs. (11) and (12), the total inter-modal triplet loss is:

$$J_{\text{triplet}} = L_{\text{triplet}}^v + L_{\text{triplet}}^t. \quad (13)$$

### Adversarial learning and optimization

Semantic associations can be maintained in different ways under the guidance of Labelnet monitoring information. However, the goal of generating a uniform hash code faces some difficulties because the distribution of features extracted from different modalities is quite different. We want the feature representations of instances with the same semantics to be as close as possible.

The common strategy is to take some methods to eliminate the gap between modalities and finally improve the retrieval accuracy. Inspired by ACMR [24], we learn the common Hamming subspace of different modalities in an adversarial way. In the common Hamming space, this paper added two different types of discriminators for image and text modalities, two of which are used to distinguish modal features from image (text) features or label semantic features. The other two discriminators are used to minimize the loss between the input and output of the image (text) autoencoder.

### Adversarial learning for inter-modal

For the image (text) discriminator  $D^{v,l}, D^{t,l}$  with parameters  $\theta_D$ , the input is the image (text) modality hash code and the hash code generated through Labelnet. The input of text discriminator is  $H^v$  and  $H^l$ , and the input of image discriminator is  $H^t$  and  $H^l$ . These two discriminators act as opponents because they are trained in an adversarial manner, inter-modal adversarial loss is as follows:

$$\min_{\theta_D^{*,l}} L_{\text{adv}}^{*,l} = -\frac{1}{n} \sum_{i=1}^n (\log D^{*,l}(h_i^l; \theta_D^{*,l}))$$

$$+ \log(1 - D^{*,l}(h_i^*; \theta_D^{*,l})), * \in \{v, t\}, \quad (14)$$

where  $h_i^v, h_i^t, h_i^l$  is the hash code of image modality, text modality and label, respectively,  $L_{\text{adv}}^{v,l}$  is the cross-entropy loss of image and label modal classification of all instances  $o_i, i = 1, \dots, n$  used in each iteration training, and  $L_{\text{adv}}^{t,l}$  is the cross-entropy loss of text and label modal.  $D^{v,l}(h_i^v; \theta_D^{v,l})$  is the image modal probability generated by each item in the instance  $o_i$ , and  $D^{t,l}(h_i^t; \theta_D^{t,l})$  is the generated text modality probability.

### Adversarial learning for intra-modal

Although the structure of intra-modal adversarial loss is similar to formula (14), some details are different and the optimization objectives are also different. For the image (text) discriminator  $D^{vae}, D^{tae}$ , the input is the image (text) modal feature representation and the feature representation after the reconstruction of the autoencoder. The intra-modal adversarial loss is as follows:

$$\min_{\theta_D^{*ae}} L_{\text{adv}}^{*ae} = -\frac{1}{n} \sum_{i=1}^n (\log D^{*ae}(f_i^{*ae-\text{in}}; \theta_D^{*ae}) + \log(1 - D^{*ae}(f_i^{*ae-\text{out}}; \theta_D^{*ae}))), * \in \{v, t\}. \quad (15)$$

The objective function of the whole feature generation part is as follows:

$$L_{\text{gen}} = L_{\text{img}} + L_{\text{txt}} + L_{\text{lab}}, \quad (16)$$

$L_{\text{img}}, L_{\text{txt}}, L_{\text{lab}}$  represent the loss of image feature extraction, text feature extraction and label generation, respectively. The objective function of the whole adversarial loss part is as follows:

$$L_{\text{adv}} = L_{\text{adv}}^{v,l} + L_{\text{adv}}^{t,l} + L_{\text{adv}}^{vae} + L_{\text{adv}}^{tae}. \quad (17)$$

This paper train the multi-modal feature extraction network (Imagenet, Textnet, Labelnet) by a way of adversarial learning. The process of learning the optimal semantic features is a joint optimization process, which is carried out by jointly minimizing the generated losses and maximizing the adversarial losses. The feature generation loss and the adversarial loss are shown by Eqs. (16) and (17), respectively. Since the optimization objectives of the two objective functions are opposite, this is a minimum–maximum game:

$$\begin{aligned} (\hat{\theta}^v, \hat{\theta}^t, \hat{\theta}^l) &= \arg \min_{\theta^v, \theta^t, \theta^l} (L_{\text{gen}}(\theta^v, \theta^t, \theta^l) - L_{\text{adv}}(\hat{\theta}_D)) \\ \hat{\theta}_D &= \arg \max_{\theta_D} (L_{\text{gen}}(\hat{\theta}^v, \hat{\theta}^t, \hat{\theta}^l) - L_{\text{adv}}(\theta_D)). \end{aligned} \quad (18)$$



**Algorithm 1** The algorithm for SAAH.

**Input:** Image set  $V = \{v_1, v_2, \dots, v_n\}$ ; Text set  $T = \{t_1, t_2, \dots, t_n\}$ ;  
Label set  $L = \{l_1, l_2, \dots, l_n\}$ ;  
**Initialize:** Initialize the hyperparameters:  $\alpha, \beta, \gamma, \delta, \eta, \lambda$ ;  
**For**  $i = 0; i \leq t; i++$  **do**  
  Update Labelnet parameters  $\theta^l$  by descending their gradients:  
  
$$\theta^l = \theta^l - \pi \cdot \nabla_{\theta^l} \frac{1}{n} (L_{gen} - L_{adv})$$
  
  Update Imagenet and Textnet parameters  $\theta^{v,t}$  by descending their gradients:  
  
$$\theta^* = \theta^* - \pi \cdot \nabla_{\theta^*} \frac{1}{n} (L_{gen} - L_{adv}), * \in \{v, t\}$$
  
  Update  $\theta_D^{v,l}, \theta_D^{t,l}, \theta_D^{vae}, \theta_D^{tae}$ , by BP algorithm:  
  
$$\theta_D^{*,l} = \theta_D^{*,l} + \pi \cdot \nabla_{\theta_D^{*,l}} \frac{1}{n} (L_{gen} - L_{adv}), * \in \{v, t\}$$
  
  
$$\theta_D^{*,vae} = \theta_D^{*,vae} + \pi \cdot \nabla_{\theta_D^{*,vae}} \frac{1}{n} (L_{gen} - L_{adv}), * \in \{v, t\}$$
  
**Endfor**  
**Output:** optimal binary hash codes  $B$ ,  
parameters of three generate network and four adversarial network.

## Experiment

This paper conducted adequate experiments on three popular benchmark datasets MIRFLICKR-25K [26], NUS-WIDE [25] and MS COCO [50] to prove its performance. The deep learning framework used in the experiment was TensorFlow V1.15.4, and the deep learning acceleration card was NVIDIA GTX 1080TI GPU.

## Datasets

MIRFLICKR-25K [26] contains 25015 images, each of which has a corresponding text description, so each instance sample is an image-text pair. There are 24 categories in this dataset, and each instance sample is marked by at least one tag. In our experiment, we only kept the tags with more than 20 times of marking, and removed the remaining tags to obtain 20,015 samples. For each instance sample, each text sample is represented as a 1386-dimensional BoW vector.

NUS-WIDE [25] contains 269,648 images and a total of 81 labels. Each image corresponds to some text description. This dataset is a multi-label dataset, that is, each instance sample is tagged by one or more tags. This paper selected 21 categories with the highest frequency and left 195,834 image-text pairs for the experiment. For each instance sample, each text sample is represented as a 1000-dimensional BoW vector.

MS COCO dataset [50], its training set size is 80,000, and the verification set size is 40,000. This paper randomly select 5000 image-text pairs as the validation set of our experiment, so a total of 85000 image-text pairs are selected as the training set of the experiment. Each data item is composed of two image-text pairs with different modalities, and the text adopts 2000-dimensional BoW vector features. In our experiment, the specific implementation details of the two cross-modal datasets are shown in Table 2.

## Evaluation metric

In the experiments, this paper uses two kinds of retrieval tasks for cross-modal retrieval: retrieving text by image query (image  $\rightarrow$  text) and retrieving image by text query (text  $\rightarrow$  image). By the way, this paper also compare the effect of single-modal query in our proposed cross-modal method: retrieving image by image query (image  $\rightarrow$  image) and retrieving text by text query (text  $\rightarrow$  text). Three widely used evaluation metrics are used to evaluate the quality of retrieval: Mean Average Precision (MAP), precision–recall curve (PR-curve) and the objective function loss curve (Loss-curve).

## Experiment results

Some representative methods were selected for comparison to verify the effectiveness of the proposed SAAH method. The shallow hashing methods including: CVH [16], STMH [17], CMSSH [18], SCM [19] and SePH [20], and deep cross-modal hashing methods including: DCMH [21], PRDH [22], SSAH [14], AGAH [15] and CPAH [39]. For fairness, the comparison method applies the same Settings as in the original work.

### Results on MIRFLICKR-25K

Table 3 presents the MAP results of all baselines and our method on FLICKR-25K, with both *Image*  $\rightarrow$  *Text* task and *Text*  $\rightarrow$  *Image* task. The best accuracy is indicated in boldface. From the results we can know that deep cross-modal methods achieve better performance than all the shallow hashing methods, our proposed SAAH is obviously superior to all of the comparative method. As the length of the code increases, more information is retained, so the length of the code affects the result. In our experiment, performance was best when the code length was 64 bits in the MIRFLICKR-25K dataset. By comparing the best shallow hashing method and deep hashing method, the method this paper proposed achieved the best results. In particular, compared to SePH, our proposed approach achieved a more than 13% lead in both retrieval tasks. Compared with the latest representative deep methods (CPAH, CPAH\* and AGAH), our MAP results are still the best in the two tasks of image query text and text query image. Our work is based on CNN-F features. However, CPAH is not only based on CNN-F features, but also uses features based on VGG16. For a comprehensive comparison, this paper also compared CPAH with VGG16 features, denoted by CPAH\*.

**Table 2** Setup of the two cross-modal datasets

Dataset	Total	Train	Test	Labels	Image feature	Text feature
MIRFLICKR-25K	20,015	10,000	2000	24	4096d CNN-F	1386d BoW
NUS-WIDE	195,834	10,500	2100	21	4096d CNN-F	1000d BoW
MS COCO	85,000	10,000	5000	80	4096d CNN-F	2000d BoW

**Table 3** The MAP of two retrieval tasks on MIRFlickr-25K dataset, the baselines are based on CNN-F features

Methods	Image-query-text			Text-query-image		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
CVH [16]	0.557	0.554	0.554	0.557	0.554	0.554
STMH [17]	0.602	0.608	0.605	0.600	0.606	0.608
CMSSH [18]	0.585	0.584	0.572	0.567	0.569	0.561
SCM [19]	0.671	0.682	0.685	0.697	0.707	0.713
SePH [20]	0.657	0.66	0.661	0.648	0.652	0.654
DCMH [21]	0.735	0.737	0.75	0.763	0.764	0.775
PRDH [22]	0.722	0.740	0.755	0.755	0.764	0.777
SSAH [14]	0.771	0.784	0.792	0.768	0.774	0.784
AGAH [15]	<b>0.792</b>	0.795	0.807	0.788	0.79	0.805
CPAH [39]	0.775	0.791	0.787	0.777	0.787	0.789
CPAH* [39]	0.789	<b>0.796</b>	0.795	0.778	0.786	0.785
<b>OURS</b>	<b>0.792</b>	<b>0.796</b>	<b>0.815</b>	<b>0.795</b>	<b>0.803</b>	<b>0.806</b>

CPAH\* means based on VGG16 features

**Table 4** The MAP of two retrieval tasks on NUS-WIDE dataset, the baselines are based on CNN-F features

Methods	Image-query-text				Text-query-image			
	16 bits	32 bits	64 bits	80 bit	16 bits	32 bits	64 bits	80 bit
CVH [16]	0.4	0.392	0.386	–	0.372	0.366	0.363	–
STMH [17]	0.522	0.529	0.537	–	0.496	0.529	0.532	–
CMSSH [18]	0.511	0.506	0.493	–	0.449	0.389	0.38	–
SCM [19]	0.533	0.548	0.557	–	0.463	0.462	0.471	–
SePH [20]	0.478	0.487	0.489	–	0.449	0.454	0.458	–
DCMH [21]	0.566	0.601	0.600	–	0.534	0.587	0.592	–
PRDH [22]	0.593	0.633	0.624	–	0.594	0.610	0.601	–
SSAH [14]	0.602	0.622	0.646	–	0.612	0.637	0.640	–
AGAH [15]	<b>0.646</b>	<b>0.66</b>	0.651	–	0.631	0.642	0.634	–
CPAH [39]	0.607	0.627	0.634	–	0.642	0.662	0.665	–
CPAH* [39]	0.613	0.629	0.630	–	0.649	<b>0.669</b>	<b>0.668</b>	–
<b>OURS</b>	0.628	0.646	<b>0.656</b>	<b>0.67</b>	<b>0.651</b>	0.663	0.659	<b>0.672</b>

CPAH\* means based on VGG16 features

**Table 5** The MAP of two retrieval tasks on MS-COCO dataset, the baselines are based on CNN-F features

Methods	Image-query-text			Text-query-image		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
CVH [16]	0.412	0.401	0.400	0.367	0.359	0.357
STMH [17]	0.422	0.459	0.475	0.431	0.461	0.476
CMSSH [18]	0.512	0.495	0.482	0.429	0.408	0.398
SePH [20]	0.463	0.487	0.501	0.449	0.474	0.499
DCMH [21]	0.511	0.513	0.527	0.501	0.503	0.505
SSAH [14]	0.550	0.558	0.557	0.537	0.538	0.529
<b>OURS</b>	<b>0.573</b>	<b>0.576</b>	<b>0.571</b>	<b>0.558</b>	<b>0.551</b>	<b>0.537</b>

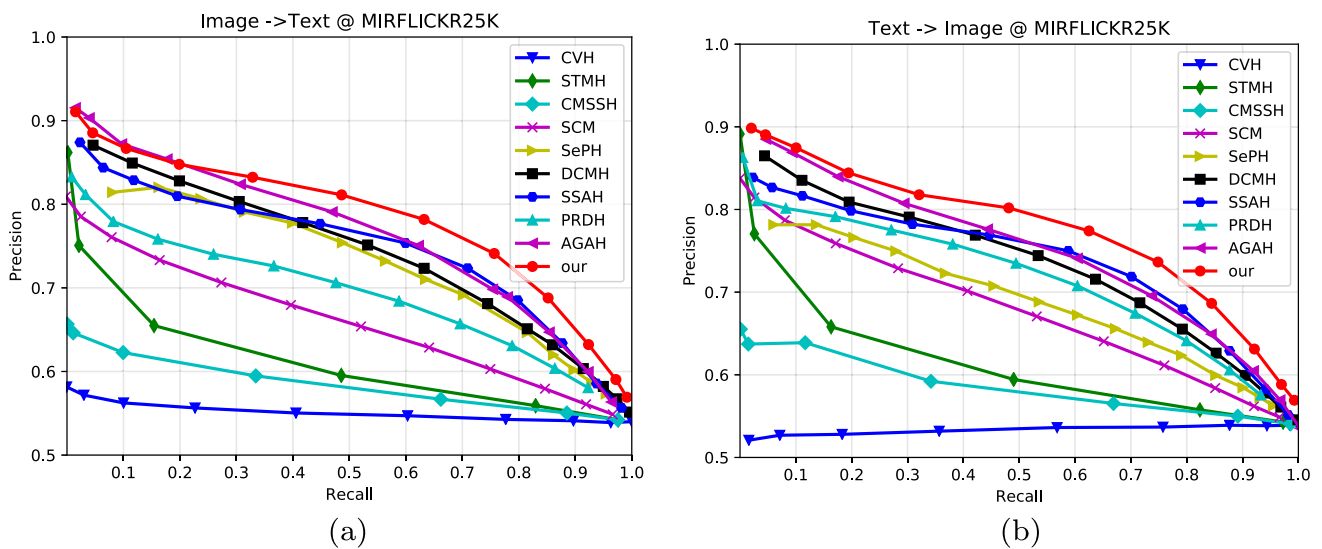


Fig. 4 The precision–recall curves on MIRFLICKR25K with 16bit hash codes

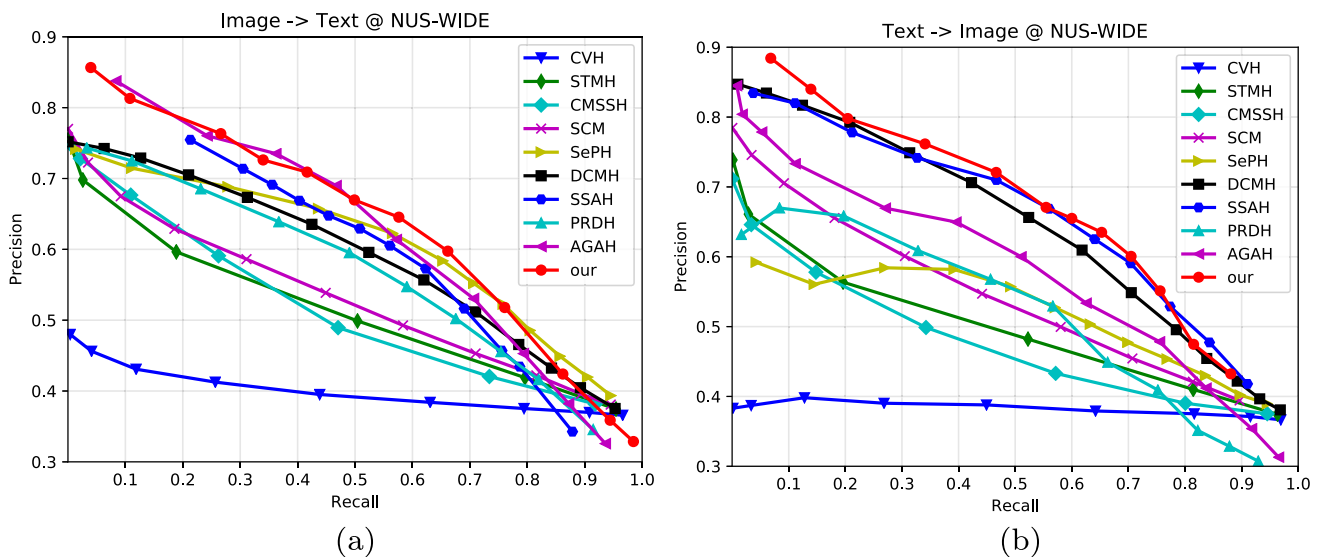


Fig. 5 The precision–recall curves on NUS-WIDE with 16-bit hash codes

**Results on NUS-WIDE**

Table 4 lists the MAP results of all methods on NUS-WIDE. Compared with MIRFLICKR-25K, which has more samples and more complex contents. Our approach still leads, but by a smaller margin than MIRFLICKR-25K. compared with CPAH, CPAH\* and AGAH, in the image query text task, our MAP achieved the best results when the hash code length is 64 bits, 80 bits. AGAH has the best result when the hash code length is 16, 32. In the text query image task, our MAP achieved the best results when the hash code length was 16 bits, 32 bits, and 80 bits. CPAH\* has the best result when the hash code length is 32 bits and 64 bits. As can be seen from the P–R curve, the precision of our method shows an

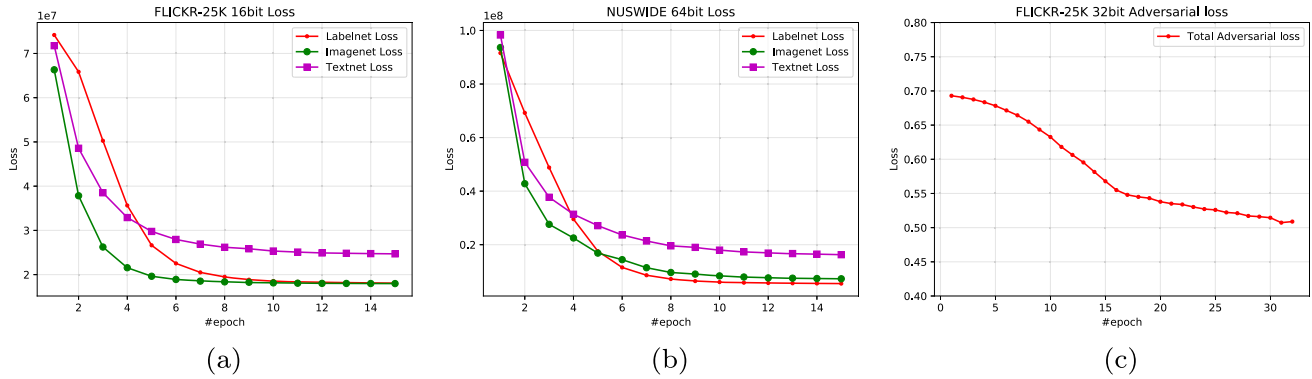
upward trend as the code length increases. We found that both image-query-text and text-query-image tasks produced the best results at 80 bits. This shows that our proposed method is better when the length of the hash code is longer.

**Results on MS COCO**

The MAP of two retrieval tasks on MS-COCO dataset as shown in Table 5. In the two tasks of image query text and text query image, the method proposed in this paper achieves the best MAP value. This paper did not test the MAP value of the comparison methods on the MS-COCO dataset, the MAP value of the comparison method is directly obtained from the original paper. Since some methods (PRDH, AGAH

**Table 6** The MAP scores of our method on two datasets with single-modality retrieval tasks

Task	Flickr-25K			NUS-WIDE			
	16 bits	32 bits	64 bits	16 bits	32 bits	64bits	80bits
Image–image	0.828	0.829	0.844	0.698	0.706	0.712	0.723
Text–text	0.764	0.767	0.773	0.595	0.613	0.606	0.614



**Fig. 6** Convergence of four kinds of loss

and CPAH) were not tested on the MS-COCO dataset in the original paper, this paper did not include them in the comparison.

**PR curves analysis**

The precision–recall (PR) curves are used to measure the accuracy of the results returned within a certain Hamming radius. We plotted the P-R curves for all the methods in Figs. 4 and 5 at 16 bit code length. The *x*-coordinate represents the recall rate, and the *y*-coordinate represents the precision value. The left figure is the PR curve of searching text by the image query, and the right figure is the PR curve of searching image by the text query. The results of each method are represented by lines with different nodes and colors. We can also see from the curve that the performance of the deep hashing method is significantly better than that of the shallow hashing method in both types of retrieval tasks. The SAAH proposed by us achieves the optimal performance, this is further proof of the superiority of our method.

The proposed cross-modal method is also suitable for single-modality retrieval, and has better retrieval precision than the existing single-modality retrieval methods. The single-modality retrieval MAP of our method is shown in Table 6.

**Convergence analysis**

Figure 6 shows the training loss changes with the epoch, the convergence curve is drawn according to Eq. (16) and (17). We can see that in the training process, the loss of each epoch is monotonously decreasing, and with the training, the

**Table 7** The MAP results at 64 bits for ablation analysis on MIRFLICKR-25K

Method	Configuration	I2T	T2I
SAAH	$L_{gen} + L_{adv}$	0.815	0.806
SAAH-1	$SAAH - (L_{adv}^{v,l} + L_{adv}^{t,l})$	0.778	0.769
SAAH-2	$SAAH - (L_{adv}^{vae} + L_{adv}^{tae})$	0.801	0.797
SAAH-3	$SAAH - L_{triplet}$	0.807	0.796

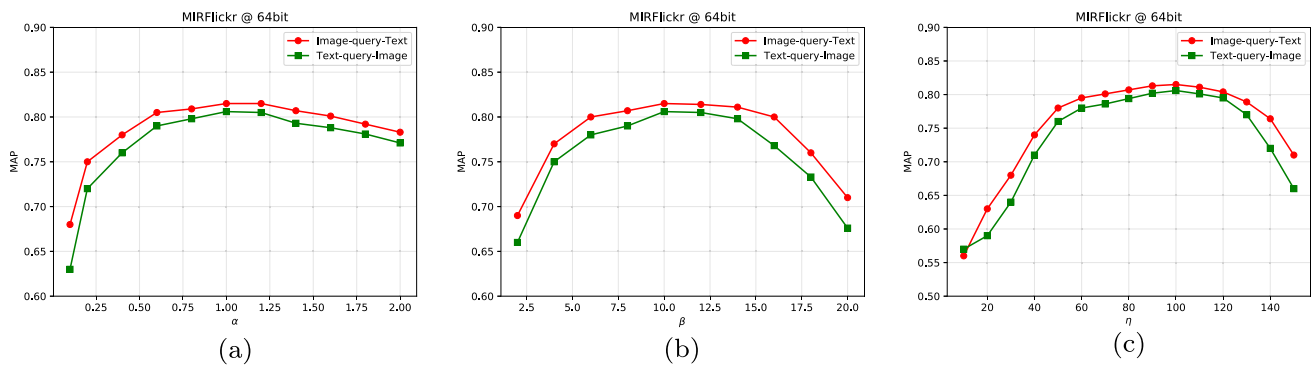
loss becomes small and stable. Figure 6a, b shows that the total losses of the feature generation module become small and stable with fewer epoch numbers. Therefore, this shows that the feature generation module is effective and accurately maintains the cross-modal correlation. Figure 6c shows the total adversarial losses change with the epoch, and the total adversarial losses rapidly converge and stabilize. This shows that our proposed cross-modal adversarial method is effective and can accurately maintain cross-media correlation.

**Ablation study**

To further demonstrate the effectiveness of each part in SAAH, We design several variants to evaluate the impacts of different modules and demonstrate the superiority of SAAH. The three variants are listed as follows:

- (1) SAAH-1 is the variant without inter-modal adversarial loss;
- (2) SAAH-2 is the variant without intra-modal adversarial loss;
- (3) SAAH-3 is the variant without inter-modal triplet loss.





**Fig. 7** Parameter sensitivity analysis of  $\alpha$ ,  $\beta$  and  $\eta$  on MIRFlickr25K

Table 7 shows the results on MIRFlickr25K datasets with 64 bits. As can be observed, each module plays a certain role in SAAH. Specifically, the results of SAAH-1 indicate that the inter-modal adversarial module is a crucial component, which can eliminate the difference in feature distribution between different modalities, thus further improving the MAP results on different datasets. The performance of SAAH-2 shows that the intra-modal adversarial loss can reduce the feature representation error after the reconstruction of the autoencoder. Besides the performance of SAAH-3 shows that the inter-modal triplet loss will improve the MAP results, so the inter-modal triplet loss is also an important component. However, it is less important than SAAH-1.

### Parameter sensitivity

Finally, this paper further analyzed the impact of the trade-off parameters  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\eta$  and  $\delta$ , and discussed the sensitivity of our method to different hyper-parameter values. Figure 7 shows the effect of these three hyper-parameters on MIRFlickr25K dataset with hash code lengths of 64, the MAP scores include the value of image-query-text and text-query-image results. When one hyper-parameter is evaluated, the others are fixed. From the results in Fig. 7a, our approach is not sensitive to the choice of  $\alpha$  in the range [1, 1.2], in our experiments, we set  $\alpha$  as 1. Similarly, in Fig. 7b, c,  $\beta$  is not sensitive in the range [10, 14], and  $\eta$  in the range [90, 110]. In addition, from the figure, the best results can be achieved when  $\alpha = 1$ ,  $\beta = 10$ , and  $\eta = 100$ . Similarly, after cross-validation, we set  $\gamma = \delta = 1$ . For simplicity, we used the same parameter settings in both datasets (MIRFlickr25K and NUS-WIDE).

### Conclusion

This paper proposed a semantic-guided adversarial hashing method, the adversarial learning based on semantic infor-

mation supervision not only eliminates the modal gap, but also keeps the invariance among the modalities. Two kinds of adversarial autoencoder networks are designed to maximize the semantic correlation of similar instances, and the adversarial learning process of adversarial modules is conducted under the supervision of semantic information, and modal relations can be maintained. In addition, to maintain the inter-modal correlation of all similar pairs, we use two types of loss functions to maintain the similarity. To verify the effectiveness of our proposed method, sufficient experiments were conducted on three widely used cross-modal datasets (NUS-WIDE, MIRFLICKR and MS COCO), and compared with several representative advanced cross-media retrieval methods, SAAH achieved leading retrieval performance.

**Acknowledgements** This work was supported by National Natural Science Foundation of China (no. 61370205), the Fundamental Research Funds for the Central Universities and Graduate Student Innovation Fund of Donghua University(No: CUSF-DH-D-2020092) and science and technology project of Chongqing Education Commission of China (KJQN201900520).

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

### References

- Gionis A, Indyk P, Motwani R et al (1999) Similarity search in high dimensions via hashing. In: VLDB, vol 99, pp 518–529

2. Datar M, Immorlica N, Indyk P, Mirrokni VS (2004) Locality sensitive hashing scheme based on p-stable distributions. In: Proceedings of the twentieth annual symposium on Computational geometry. ACM, pp 253–262
3. Ji J, Li J, Yan S, Tian Q, Zhang B (2013) Min-max hash for Jaccard similarity. In: 2013 IEEE 13th international conference on data mining. IEEE, pp 301–309
4. Ding G, Guo Y, Zhou J et al (2016) Large-scale cross-modality search via collective matrix factorization hashing. *IEEE Trans Image Process* 25(11):5427–5440
5. Zhou J, Ding G, Guo Y (2014) Latent semantic sparse hashing for cross-modal similarity search. In: Proceedings of the 37th ACM SIGIR international conference on research and development in information retrieval, Gold Coast, QLD, Australia. ACM, pp 415–424
6. Xia R, Pan Y, Lai H, Liu C, Yan S (2014) Supervised hashing for image retrieval via image representation learning. In: Proceedings of the AAAI conference on artificial intelligence, pp 2156–2162
7. Peng Y, Qi J (2019) CM-GANs: cross-modal generative adversarial networks for common representation learning. *ACM Trans Multimed Comput Commun Appl (TOMM)* 15(1):22
8. Wang K, Yin Q, Wang W, Wu S, Wang L (2016) A comprehensive survey on cross-modal retrieval. arXiv preprint [arXiv:1607.06215](https://arxiv.org/abs/1607.06215)
9. Ding G, Guo Y, Zhou J (2014) Collective matrix factorization hashing for multimodal data. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2075–2082
10. Kumar S, Udupa R (2011) Learning hash functions for cross-view similarity search. In: IJCAI proceedings-international joint conference on artificial intelligence, vol 22, p 1360
11. Zhang D, Li W-J (2014) Large-scale supervised multimodal hashing with semantic correlation maximization. In: AAAI, vol 1, p 7
12. Cao Y, Long M, Wang J, Yang Q, Yu PS (2016) Deep visual-semantic hashing for cross-modal retrieval. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1445–1454
13. Liu L, Shen F, Shen Y, Liu X, Shao L (2017) Deep sketch hashing: fast free-hand sketch-based image retrieval. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2298–2307
14. Li C, Deng C, Li N, Liu W, Gao X, Tao D (2018) Self-supervised adversarial hashing networks for cross-modal retrieval. In: CVPR, pp 4242–4251
15. Gu W, Gu X, Gu J, Li B, Xiong Z, Wang W (2019) Adversary guided asymmetric hashing for cross-modal retrieval. *ICMR* 159–167
16. Kumar S, Udupa R (2011) Learning hash functions for cross-view similarity search. In: IJCAI, vol 22, p 1360
17. Wang D, Gao X, Wang X, He L (2015) Semantic topic multimodal hashing for cross-media retrieval. In: IJCAI, pp 3890–3896
18. Bronstein MM, Bronstein AM, Michel F, Paragios N (2010) Data fusion through cross-modality metric learning using similarity-sensitive hashing. In: CVPR, pp 3594–3601
19. Zhang D, Li WJ (2014) Large-scale supervised multimodal hashing with semantic correlation maximization. In: AAAI, pp 2177–2183
20. Lin Z, Ding G, Hu M, Wang J (2015) Semantics-preserving hashing for cross-view retrieval. In: CVPR
21. Jiang Q-Y, Li W-J (2017) Deep cross-modal hashing. In: CVPR
22. Yang E, Deng C, Liu W, Liu X, Tao D, Gao X (2017) Pairwise relationship guided deep hashing for cross-modal retrieval. In: AAAI, pp 1618–1625
23. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: delving deep into convolutional nets. arXiv preprint [arXiv:1405.3531](https://arxiv.org/abs/1405.3531)
24. Wang B, Yang Y, Xu X, Hanjalic A, Shen HT (2017) Adversarial cross-modal retrieval. In: ACMMM, pp 154–162
25. Chua T-S, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) NUS-WIDE: a real-world web image database from National University of Singapore. In: ACM CIVR, p 48
26. Huiskes MJ, Lew MS (2008) The MIR Flickr retrieval evaluation. In ACM: CIVR, pp 39–43
27. Liu H, Wang R, Shan S, Chen X (2016) Deep supervised hashing for fast image retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2064–2072
28. Liu W, Wang J, Ji R, Jiang Y-G, Chang S-F (2012) Supervised hashing with kernels. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2074–2081
29. Peng Y, Huang X, Zhao YZ (2017) An overview of cross-media retrieval: concepts, methodologies, benchmarks and challenges. *IEEE Trans Circuits Syst Video Technol* (2017)
30. Lin Z, Ding G, Hu M, Wang J (2015) Semantics-preserving hashing for cross-view retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3864–3872
31. Cao Y, Long M, Wang J (2016) Correlation hashing network for efficient cross-modal retrieval (2016). CoRR abs/1602.06697. [arXiv:1602.06697](https://arxiv.org/abs/1602.06697)
32. Liong VE, Lu J, Tan Y-P, Zhou J (2017) Cross-modal deep variational hashing. In: 2017 IEEE international conference on computer vision (ICCV). IEEE, pp 4097–4105
33. Su S, Zhong Z, Zhang C (2019) Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. *ICCV*
34. Li C, Deng C, Wang L et al (2019) Coupled CycleGAN: unsupervised hashing network for cross-modal retrieval. AAAI
35. Chen Z-D, Yu W-J, Li C-X, Nie L, Xu X-S (2018) Dual deep neural networks cross-modal hashing. In: AAAI, pp 274–281
36. Zhang J, Peng Y (2018) Query-adaptive image retrieval by deep-weighted hashing. *IEEE Trans Multimed* 20(9):2400–2414
37. Wu L, Wang Y, Shao L (2019) Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Trans Image Process* 28(4):1602–1612
38. Li K et al (2019) Visual semantic reasoning for image-text matching. *ICCV*
39. Xie D, Deng C, Li C, Liu X, Tao D (2020) Multi-task consistency-preserving adversarial hashing for cross-modal retrieval. *IEEE Trans Image Process* 29:3626–3637. <https://doi.org/10.1109/TIP.2020.2963957>
40. Nie X, Wang B, Li J, Hao F, Jian M, Yin YL (2021) Deep multiscale fusion hashing for cross-modal retrieval. *IEEE Trans Circuits Syst Video Technol* 31(1):401–410
41. Xu X, He L, Lu HM et al (2019) Deep adversarial metric learning for cross-modal retrieval. *World Wide Web* 22(2):657–672
42. Feng F, Wang X, Li R (2014) Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 22nd ACM international conference on multimodal, New York, USA. ACM, pp 7–16
43. Wu Y, Wang S, Huang Q (2019) Multi-modal semantic autoencoder for cross-modal retrieval. *Neurocomputing* 331(28):165–175
44. Huang X, Peng Y, Yuan M (2020) MHTN: modal-adversarial hybrid transfer network for cross-modal retrieval. *IEEE Trans Cybern* 50(3):1047–1059
45. Zhang J, Peng YX, Yuan MK (2018) Unsupervised generative adversarial cross-modal hashing. In: Proceedings of the 32nd AAAI conference on artificial intelligence. AAAI
46. Zhang J, Peng Y (2020) Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval. *IEEE Trans Multimed* 22(1):174–187

47. Wang X, Zou X, Bakker EM, Wu S (2020) Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval. *Neurocomputing* 400:255–271
48. Zhang M, Li J, Zhang HX, Liu L (2020) Deep semantic cross modal hashing with correlation alignment. *Neurocomputing* 381:240–251
49. Wu F, Jing X, Wu Z, Ji Y, Dong X, Luo X, Huang QH, Wang R (2020) Modality-specific and shared generative adversarial network for cross-modal retrieval. *Pattern Recognit* 104:107335
50. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollr P, Zitnick CL (2014) Microsoft coco: common objects in context. In: *ECCV*, pp 740–755

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.