Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval

Chao Li[1], Cheng Deng*[1], Ning Li[1], Wei Liu*[2], Xinbo Gao[1], and Dacheng Tao[3]

[1]School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2]Tencent AI Lab, Shenzhen, China
[3]UBTECH Sydney AI Centre, SIT, FEIT, University of Sydney, Australia,
li_chao@stu.xidian.edu.cn, {chdeng.xd, ningli2017}@gmail.com, wliu@ee.columbia.edu,
xbgao@mail.xidian.edu.cn, dacheng.tao@sydney.edu.au

Abstract

*Thanks to the success of deep learning, cross-modal retrieval has made significant progress recently. However, there still remains a crucial bottleneck: how to bridge the modality gap to further enhance the retrieval accuracy. In this paper, we propose a self-supervised adversarial hashing (**SSAH**) approach, which lies among the early attempts to incorporate adversarial learning into cross-modal hashing in a self-supervised fashion. The primary contribution of this work is that two adversarial networks are leveraged to maximize the semantic correlation and consistency of the representations between different modalities. In addition, we harness a self-supervised semantic network to discover high-level semantic information in the form of multi-label annotations. Such information guides the feature learning process and preserves the modality relationships in both the common semantic space and the Hamming space. Extensive experiments carried out on three benchmark datasets validate that the proposed SSAH surpasses the state-of-the-art methods.*

1. Introduction

Owing to the explosive increase in multimedia data from a great variety of search engines and social media, cross-modal retrieval has become a compelling topic in recent years [20, 21, 22, 23, 24, 25, 29, 35, 36, 41, 42, 45]. Cross-modal retrieval aims to search semantically similar instances in one modality (*e.g.*, image) by using a query from another modality (*e.g.*, text). In order to satisfy the requirements of low storage cost and high query speed in real-world applications, hashing has been of considerable interest in the field of cross-modal retrieval, which maps high-dimensional multi-modal data into a common hash code space in such a way that gives similar cross-modal items similar hash codes. Since the instances from different modalities are hetrogeneous in terms of their feature representation and distribution, *i.e.*, their modality gap, it is necessary to explore their semantic relevance in sufficient detail to bridge this modality gap. Most existing shallow cross-modal hashing methods (in both unsupervised [2, 10, 14, 18] and supervised settings [7, 17, 19, 26, 30, 40, 33]), always capture the semantic relevance in a common Hamming space. Compared with their unsupervised counterparts, supervised cross-modal hashing methods can achieve superior performance by exploiting semantic labels or information concerning relevance, thereby distilling a cross-modal correlation. However, almost all these existing shallow cross-modal hashing methods are based on hand-crafted features, which may limit the discriminative representation of instances and thus degrade the accuracy of the learned binary hash codes.

In recent years, deep learning has become very successful at learning highly discriminative features for various applications [1][13]. However, only a few works have performed deep learning for cross-modal hashing [3, 9, 12, 31, 43], which can capture nonlinear correlations among cross-modal instances more effectively. It is worth noting that there are still some common disadvantages hindering the current deep cross-modal hashing methods. First, these methods simply and directly adopt single-class labels to measure the semantic relevance across modalities [9][12]. In fact, in standard cross-modal benchmark datasets such as NUS-WIDE [6] and Microsoft COCO [15], an image instance can be assigned to multiple category labels [27], which is beneficial as it permits semantic relevance to be described more accurately across different modalities. Second, these methods enforce a narrowing of the modality gap by constraining the corresponding hash codes with certain pre-defined loss functions [4]. The code length is usually

*Corresponding authors

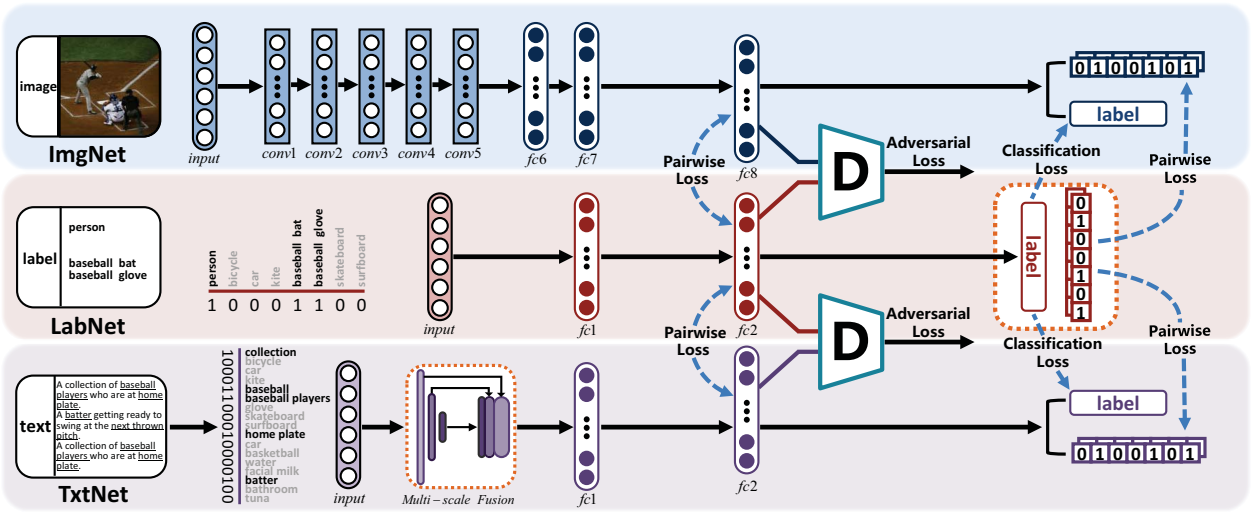


Figure 1: The framework of our proposed SSAH.

less than 128 bits. This means that most of the useful information is neutralized, making the hash codes incapable of capturing the inherent modality consistency. In comparison, high-dimensional modality-specific features contain more abundant information that helps to bridge the modality gap. Therefore, how to encourage more abundant semantic relevance and build more accurate modality relationships becomes crucial to achieve satisfactory performance in real-world retrieval applications.

In this paper, we propose a novel self-supervised adversarial hashing (SSAH) method to aid in cross-modal retrieval. Specifically, we employ two adversarial networks to jointly learn the high-dimensional features and their corresponding hash codes for different modalities. At the same time, a self-supervised semantic network is seamlessly integrated to discover semantic information in the form of multi-label annotations, with which the adversarial learning is supervised to maximize the semantic relevance and the feature distribution consistency between modalities. The highlights of our work can be outlined as follows:

- We propose a novel self-supervised adversarial hashing method for cross-modal retrieval. As far as we know, this is one of the first attempts to utilize adversarial learning in an aim to tackle the cross-modal hashing problem.
- We integrate self-supervised semantic learning with adversarial learning in order to preserve the semantic relevance and the representation consistency across modalities as much as possible. In this way, we can effectively bridge the modality gap.
- Extensive experiments conducted on three benchmark datasets demonstrate that our proposed SSAH significantly outperforms the current state-of-the-art cross-modal hashing methods, including both traditional and deep-learning-based methods.

The rest of this paper is organized as follows. Related work in cross-modal hashing is introduced in Section 2. Our proposed SSAH model and the learning algorithm are presented in Section 3. Experiments are shown in Section 4 and Section 5 concludes this work.

2. Related Work

Cross-modal hashing methods can be roughly categorized into unsupervised and supervised settings. Unsupervised hashing methods [8, 34, 38, 46] learn hashing functions by discovering the inter-modality and intra-modality information belonging to the unlabeled training data. Ding *et al.* [8] learned a unified binary code by performing a matrix factorization with a latent factor model. The work of Song *et al.* [34] learns functions that can map features from different modalities into the common Hamming space.

Supervised hashing methods [2, 4, 14, 16, 39, 40, 44] aim to exploit available supervised information (such as labels or the semantic affinities of training data) to improve performance. Brostein *et al.* [2] present a cross-modal hashing approach by preserving the intra-class similarity via eigen-decomposition and boosting. Semantic correlation maximization (SCM) [44] utilizes label information to learn a modality-specific transformation, which preserves the maximal correlation between modalities. Semantics-preserving hashing (SePH) [16] generates a unified binary code by modeling an affinity matrix in a probability distribution while at the same time minimizing the Kullback-Leibler divergence. Most of these methods depend on hand-crafted features that have to be extracted by shallow architectures; as such, these methods make it difficult to effectively exploit the heterogeneous relationships across modalities.

Recently, some works have reported on deep cross-modal hashing retrieval [3, 9, 12, 37]. Deep cross-modal hashing (DCMH) [12] performs an end-to-end learning framework, using a negative log-likelihood loss to preserve

the cross-modal similarities. Adversarial cross-modal retrieval (ACMR) [37] directly borrows from the adversarial-learning approach and tries to discriminate between different modalities using a classification manner that is the one most closely related to ours. In comparison to [37], however, our SSAH utilizes two adversarial networks to jointly model different modalities and thereby further capture their semantic relevance and representation consistence under the supervision of the learned semantic feature.

3. Proposed SSAH

Without loss of generality, we focus on cross-modal retrieval for bimodal data (*i.e.*, image and text). Fig. 1 is a flowchart showing the general principles of the proposed SSAH method. This method mainly consists of three parts, including a self-supervised semantic generation network called *LabNet*, and two adversarial networks called *ImgNet* and *TexNet* for image and text modalities, respectively.

Specifically, the target of *LabNet* is framed in a way that allows it to learn semantic features from multi-label annotations. It can then be regarded as a common semantic space in which to supervise modality-feature learning over two phases. In the first phase, modality-specific features from separate generator networks are associated with each other in a common semantic space. Since each output layer in a deep neural network contains semantic information, associating modality-specific features in a common semantic space can help to promote the semantic relevance between modalities. In the second phase, semantic features and modality-specific features are simultaneously fed into two discriminator networks. As a result, the feature distributions of the two modalities tend to become consistent under the supervision of the same semantic feature. In this section, we present the details about our SSAH method, including the methods behind the model formulation and the learning algorithm.

3.1. Problem Formulation

Let $O = \{o_i\}_{i=1}^{n}$ denote a cross-modal dataset with n instances, $o_i = (v_i, t_i, l_i)$, where $v_i \in \mathbb{R}^{1\times d_v}$ and $t_i \in \mathbb{R}^{1\times d_t}$ are the original image and text features for the i-th instance, and $l_i = [l_{i1}, \ldots, l_{ic}]$ is the multi-label annotation assigned to o_i, where c is the class number. If o_i belongs to the j-th class $l_{ij} = 1$, otherwise $l_{ij} = 0$. The image-feature matrix is defined as V, the text-feature matrix as T, and the label matrix as L for all instances. The pairwise multi-label similarity matrix S is used to describe semantic similarities between each of the two instances, where $S_{ij} = 1$ means that o_i is semantically similar to o_j, otherwise $S_{ij} = 0$. In a multi-label setting, two instances (o_i and o_j) are annotated by multiple labels. Thus, we define $S_{ij} = 1$ if o_i and o_j share as least one label, otherwise $S_{ij} = 0$.

The goal of cross-modal hashing is to learn a unified hash code for the two modalities: $B^{v,t} \in \{-1, 1\}^K$, where K is the length of the binary code. The similarity between two binary codes is evaluated using the Hamming distance. The relationship between their Hamming distance $dis_H(b_i, b_j)$ and their inner product $\langle b_i, b_j\rangle$ can be formulated using $dis_H(b_i, b_j) = \frac{1}{2}(K - \langle b_i, b_j\rangle)$. So, we can use the inner product to quantify the similarity of two binary codes. Given S, the probability of S under the condition B can be expressed as:

$$p(S_{ij}|B) = \begin{cases} \delta(\Psi_{ij}), & S_{ij} = 1 \\ 1 - \delta(\Psi_{ij}), & S_{ij} = 0 \end{cases} \tag{1}$$

where $\delta(\Psi_{ij}) = \frac{1}{1+e^{-\Psi_{ij}}}$, and $\Psi_{ij} = \frac{1}{2}\langle b_i, b_j\rangle$. Therefore, two instances with a larger inner product should be similar with a high probability. The problem of quantifying the similarity between binary codes in the Hamming space can thereby be transformed into a calculation of the inner product of the codes' original features.

Here, we frame a couple of adversarial networks (*ImgNet* and *TxtNet*) to learn separate hash functions for image and text modalities (*i.e.,* $H^{v,t} = f^{v,t}(v, t; \theta^{v,t})$). At the same time, we construct an end-to-end self-supervised semantic network (*LabNet*) in order to model the semantic relevance between image and text modality in a common semantic space while learning the hash function for the semantic feature (*i.e.,* $H^{l} = f^{l}(l; \theta^{l})$). Here, $f^{v,t,l}$ are hash functions, and $\theta^{v,t,l}$ are the network parameters to be learned. With the learned $H^{v,t,l}$, binary codes $B^{v,t,l}$ can be generated by applying a sign function to $H^{v,t,l}$:

$$B^{v,t,l} = sign(H^{v,t,l}) \in \{-1, 1\}^K \tag{2}$$

To make this easier to understand, we additionally use $F^{v,t,l} \in \mathbb{R}^{s\times n}$ to denote the semantic features in a common semantic space for images, text and labels, s is the dimension of the semantic space. In practice, $F^{v,t,l}$ correspond to certain output layers of deep neural networks (*ImgNet*, *TxtNet* and *LabNet*, respectively).

3.2. Self-supervised Semantic Generation

Taking the Microsoft COCO dataset as an example, there is an instance that is annotated with multiple labels, such as "person", "baseball bat" and "baseball glove". In this scenario, the most natural thought is that it is possible to take the multi-label annotation as a conduciveness with which to bridge the semantic relevance between modalities at a more fine-grained level. We have designed an end-to-end full-connected deep neural network, named *LabNet*, to model semantic relevance between different modalities. Given a multi-label vector for an instance, *LabNet* extracts abstract semantic features layer by layer; with these we can supervise the feature-learning process in both *ImgNet* and *TxtNet*. Since a triplet (v_i, t_i, l_i) is used to describe the same i-th instance, we regard l_i as self-supervised semantic information for v_i and t_i. In *LabNet*, semantic features are projected into their corresponding hash codes through nonlinear transformation. Our intention is that the similarity relationships between semantic features and their corresponding hash codes

is well preserved; this is the fundamental premise behind the efficient association between different modalities. Accordingly, for *LabNet*, the final objective can be formulated as follows:

$$\begin{aligned}\min_{B^l,\theta^l,\hat{L}^l} \mathcal{L}^l =&\alpha\mathcal{J}_1+\gamma\mathcal{J}_2+\eta\mathcal{J}_3+\beta\mathcal{J}_4\\ =&-\alpha\sum_{i,j=1}^{n}\left(S_{ij}\Delta_{ij}^l-\log\left(1+e^{\Delta_{ij}^l}\right)\right)\\ &-\gamma\sum_{i,j=1}^{n}\left(S_{ij}\Gamma_{ij}^l-\log\left(1+e^{\Gamma_{ij}^l}\right)\right)\\ &+\eta\left\|H^l-B^l\right\|_F^2+\beta\left\|\hat{L}^l-L\right\|_F^2\\ &s.t.\quad B^l\in\{-1,1\}^K\end{aligned} \tag{3}$$

where $\Delta_{ij}^l=\frac{1}{2}(F_{*i}^l)^\top(F_{*j}^l)$, $\Gamma_{ij}^l=\frac{1}{2}(H_{*i}^l)^\top(H_{*j}^l)$, H^l are predicted hash codes and $\hat{L}^l$ are predicted labels. α, γ, η and β are hyper-parameters. In (3), $\mathcal{J}_1$ and $\mathcal{J}_2$ are two negative-log likelihood functions. $\mathcal{J}_1$ is used to preserve the similarity between semantic features, and $\mathcal{J}_2$ is used to preserve the instances where the similar label information has similar hash codes. $\mathcal{J}_3$ is the approximation loss for the binarization of the learned hash codes, and $\mathcal{J}_4$ is the classification loss of the original label and the predicted label.

3.3. Feature Learning

As described above, the different modalities belonging to a multi-modal instance are semantically relevant. In order to preserve this semantic relevance, we supervise the feature-learning process for two modalities under *LabNet*'s guidance, including the supervision of the semantic features and the learned binary codes. To address image modality, we have designed an end-to-end feature-learning network, named *ImgNet*, which can project images into hash codes. By supervising the image-feature learning using the semantic network, we can keep the same semantic relevance between *ImgNet* and the semantic network. This is the self-supervised role of the semantic network when used in *ImgNet*. Similarly, when considering text modality, we use the semantic network to supervise the feature-learning process of *TxtNet* in the same way. Thus, the objective function of self-supervised feature learning for different modalities in v and t can be written as:

$$\begin{aligned}\min_{B^{v,t},\theta^{v,t}} \mathcal{L}^{v,t} =&\alpha\mathcal{J}_1+\gamma\mathcal{J}_2+\eta\mathcal{J}_3+\beta\mathcal{J}_4\\ =&-\alpha\sum_{i,j=1}^{n}\left(S_{ij}\Delta_{ij}^{v,t}-\log\left(1+e^{\Delta_{ij}^{v,t}}\right)\right)\\ &-\gamma\sum_{i,j=1}^{n}\left(S_{ij}\Gamma_{ij}^{v,t}-\log\left(1+e^{\Gamma_{ij}^{v,t}}\right)\right)\\ &+\eta\|H^{v,t}-B^{v,t}\|_F^2+\beta\left\|\hat{L}^{v,t}-L\right\|_F^2\\ &s.t.\quad B^{v,t}\in\{-1,1\}^K\end{aligned} \tag{4}$$

where $\Delta_{ij}^{v,t}=\frac{1}{2}(F_{*i}^l)^\top(F_{*j}^{v,t})$, and $\Gamma_{ij}^{v,t}=\frac{1}{2}(H_{*i}^l)^\top(H_{*j}^{v,t})$. $H^{v,t}$ are predicted hash codes and $\hat{L}^{v,t}$ are predicted labels for images and text, respectively. α, γ, η and β are hyper-parameters. $\mathcal{J}_1$ and $\mathcal{J}_2$ are two negative-log likelihood functions. $\mathcal{J}_3$ and $\mathcal{J}_4$ are approximation loss and classification loss defined in a way that is similar to that used in *LabNet*. It should be noted that although (3) and (4) are similar in structure they have different meanings. As such, we use the supervised information F_{*i}^l and H_{*i}^l (learned from the semantic network) to guide the process of learning in *ImgNet* and *TxtNet*. The relevance can be established using the semantic network. As a result, the modality gap can then be alleviated.

In comparison to image modality, an instance in text modality, generally represented by a bag-of-words (BoW) vector, easily results in sparsity. Therefore, BoW is unsuitable when aiming to discover valuable information needed for learning hash codes. To solve the problem, we have designed a multi-scale fusion model, which consists of multiple average pooling layers and a 1×1 convolutional layer. Multiple average pooling layers are used to extract multiple scale features for text data, following which the 1×1 convolutional layer is used to fuse multiple features. Through this process, the correlation between different words can also be captured, which is useful when building semantic relevance for text modality. More detailed parameter information is given in Section 3.6.

3.4. Adversarial Learning

Under the supervision of *LabNet*, the semantic relevance can be preserved across different modalities. However, different modalities usually are inconsistently distributed, which is not beneficial if we want to generate unified hash codes. In order to bridge this modality gap and enable more accurate retrieval, we have studied the distribution agreement for different modalities in an adversarial learning manner. We have built two discriminators for image and text modalities to discover their distribution differences. For the image (text) discriminator, the inputs are image (text) modality features and semantic features generated through *LabNet*, and the output is one single value, either "0" or "1". Specifically, we define the modality label for the semantic feature that has been generated from a label as "1" and define the modality label for image (text) semantic modality features generated from *ImgNet* (*TxtNet*) as "0". We feed F^v and F^l into the discriminator that has been designed for images and feed F^t and F^l into another discriminator that has been designed for text. To formulate this structure, let $Y=\{y_i\}_{i=1}^{3\times n}$, $y_i\in\{0,1\}$ denote the modality label assigned to the semantic feature in the shared common space. Let $Y^l=\{y_i^l\}_{i=1}^n$, $y_i^l=1$ denote the modality labels for the label. Let $Y^{v,t}=\{y_i^{v,t}\}_{i=1}^n$ and $y_i^{v,t}=0$ denote the modality labels for image and text, respectively. When training our model, these two discriminators act as the two adversaries. As such, the objective function can be written as follows:

Algorithm 1 Pseudopod showing the optimization of our SSAH

Require: Image set V ; Text set T ; Label set L ;
Ensure: Optimal code matrix B
Initialization
Initialize parameters: $\theta^{v,t,l}$, $\theta^{v,t}_{adv}$, $\alpha, \gamma, \eta, \beta$
learnrate: μ, mini-batch size: $N^{v,t,l} = 128$, maximum iteration number: T_{max}.
repeat
for t iteration **do**
Update θ^l by BP algorithm:
$\theta^l \leftarrow \theta^l - \mu \cdot \nabla_{\theta^l} \frac{1}{n} \left(\mathcal{L}_{gen} - \mathcal{L}_{adv}\right)$
Update the parameter $\theta^{v,t}$ by BP algorithm:
$\theta^\star \leftarrow \theta^\star - \mu \cdot \nabla_{\theta^\star} \frac{1}{n} \left(\mathcal{L}_{gen} - \mathcal{L}_{adv}\right), \star = v, t$
Update $\theta^\star_{adv}$ by BP algorithm:
$\theta^\star_{adv} \leftarrow \theta^\star_{adv} - \mu \cdot \nabla_{\theta^\star_{adv}} \frac{1}{n} \left(\mathcal{L}_{(gen)} - \mathcal{L}_{adv}\right), \star = v, t$
end for
Update the parameter B by
$B = sign(H + F + G)$
until convergence

$$\min_{\theta^{\star,l}_{adv}} \mathcal{L}^{\star,l}_{adv} = \sum_{i=1}^{2\times n} \left\| D^{\star,l}(x_i^{\star l}) - {y_i}^{\star,l} \right\|_2^2, \ \star = v, t \tag{5}$$

where $x_i^{v,t,l}$ is the semantic feature in the common semantic space, while the modality label is $y_i^{v,t,l}$, $2 \times n$, denoting the number of instances that are fed into each discriminator. The result of (5) is that the discriminators act as two binary classifiers, classifying the input semantic feature into class "1" and class "0".

3.5. Optimization

It is noted that three kinds of hash codes can be generated using our SSAH: $B^{v,t,l} = sign(H^{v,t,l})$. During the training process, we make $B = sign(H^v + H^t + H^l)$ to train our model to generate similar binary codes for semantically similar instances. As mentioned above, the overall objective function can be written as follows:

$$\begin{aligned} \mathcal{L}_{gen} &= \mathcal{L}^v + \mathcal{L}^t + \mathcal{L}^l \\ \mathcal{L}_{adv} &= \mathcal{L}^v_{adv} + \mathcal{L}^t_{adv} \end{aligned} \tag{6}$$

If we put them together, we can obtain:

$$\begin{aligned} (B, \theta^{v,t,l}) &= \underset{B,\theta^{v,t,l}}{argmin}\ \mathcal{L}_{gen}(B, \theta^{v,t,l}) - \mathcal{L}_{adv}(\hat{\theta}_{adv}) \\ \theta_{adv} &= \underset{\theta_{adv}}{argmax}\ \mathcal{L}_{gen}(\hat{B}, \hat{\theta}^{v,t,l}) - \mathcal{L}_{adv}(\theta_{adv}) \\ &s.t. \quad B \in \{-1, 1\}^K \end{aligned} \tag{7}$$

Due to the discreteness of parameter B and the vanishing-gradient problem caused by the minimax loss, the optimization of (7) is intractable. Hence, we optimize the objective (7) through iterative optimization. Firstly, we optimize the $\mathcal{L}^l$ over θ^l, B^l, and $\hat{L}^l$ by exploring label information. Then, we optimize $\mathcal{L}^v$ over θ^v and B^v by fixing θ^l and B^l. Similarly, we leave θ^l and B^l fixed to learn θ^t and B^t, allowing the optimization of $\mathcal{L}^t$. During this process, two kinds of modality features are learned in a self-supervised learning manner. Finally, we optimize $\mathcal{L}^{v,t}_{adv}$ over $\theta^{v,t}$ by fixing $\theta^{v,t,l}$. It is noted that all network parameters are learned by utilizing the stochastic gradient descent (SGD) with the back-propagation (BP) algorithm, which is widely adopted in existing deep-learning methods. Algorithm 1 outlines the whole learning algorithm in detail.

As for out-of-sample extensions: the proposed framework can be applied to cross-modalities. Indeed, it is not limited to two modalities; rather, it can easily be adapted to solve the problems in situations with more than two modalities. Hash codes for the unseen data-point, which may come from different modalities, images or text, can be directly obtained by feeding the original feature into our model:

$$b_q^{v,t,l} = sign(f^{v,t,l}(b_q; \theta^{v,t,l})) \tag{8}$$

Moreover, by feeding the label information into $LabNet$ we can obtain hash codes for the label information, which can then be used to retrieve the related results from both images and text at same time.

3.6. Implementation Details

Self-Supervised Semantic Network: We built *LabNet* with four-layer feed-forward neural networks, which are used to project a label into hash codes ($L \rightarrow 4096 \rightarrow 512 \rightarrow N$). The nodes of output layer N are related to the length of the hash code K and the total class labels c for different datasets, $N = K + c$.

Generative Network for Images: We built *ImgNet* based on CNN-F [5] neural networks. In order to apply CNN to our SSAH model, we reserve the first seven layers (which were the same as those in CNN-F). Following this, a middle layer fc8 (with 512 nodes) and final output layer (with N nodes) are framed. In addition, we also evaluated our method using the vgg19 [32] network; here, we replaced the CNN-F network with the vgg19 network and left the rest remain unchanged.

Generative Network for Text: We built *TxtNet* using a three-layer feed-forward neural network and a multi-scale (MS) fusion model (T→MS→4096→512→N). MS consists of a five-level pooling layer (1×1, 2×2, 3×3, 5×5, and 10×10).

Adversarial Networks: We built the discriminator networks using a three-layer feed-forward neural network ($F^{v,t,l} \rightarrow 4096 \rightarrow 4096 \rightarrow 1$).

Regarding the activate function used in SSAH: $sigmoid$ activation is used to output the predicted label; $tanh$ activation is used to output the hash codes; and the rest of the layers are all uniformly activated by the $relu$ function. In addition, SSAH is implemented via TensorFlow and is run on a server with two NVIDIA TITAN X GPUs.

4. Experiment

4.1. Datasets

The ***MIRFLICKR-25K*** dataset [11] contains 25,000 instances collected from Flickr. Each image is labeled with its associated textual tags. Here, we follow the experimental protocols given in DCMH [12]. In total, 20,015 data

Table 1: Statistics of the datasets used in our experiments.

Dataset	Total	Train / Test	Labels
MIRFLICKR-25K	20,015	10,000 / 2,000	24
NUS-WIDE	190,421	10,500 / 2,100	21
MS COCO	85,000	10,000 / 5,000	80

points have been selected for our experiment. The text for each point is represented as a 1,386-dimensional BoW vector, and each point is manually annotated with at least one of the 24 unique labels.

The ***NUS-WIDE*** dataset [6] is a public web image dataset containing 269,648 web images. There are 81 ground-truth concepts that have been manually annotated for search evaluation. After pruning the data that is without any label or tag information, a subset of 190,421 image-text pairs that belong to some of the 21 most-frequent concepts are selected to serve as our dataset.

The ***MS COCO*** dataset [15] contains about 80,000 training images and 40,000 validation images. Five thousand images from the validation set are selected randomly. In total, there are 85,000 data items have been used in our experiment. Each data item consists of one image-text pair for two different modalities, and each text is represented as a 2,000-dimension BoW vector. Table 1 summarizes the statistics of the three datasets.

4.2. Evaluation and Baselines

Evaluation: The Hamming ranking and hash lookup are two classical retrieval protocols used to evaluate the performance of a cross-modal retrieval task. In our experiments, we use three evaluation criteria: mean average precision (MAP), which is used to measure the accuracy of the Hamming distances; the precision-recall (PR) curve, which is used to measure the accuracy of the hash lookup protocol; and the precision at n (P@n) curve used to evaluate precision by considering only the number of top returned points.

Baselines: We compare our SSAH using six state-of-the-art methods, including several shallow-structure-based methods (CVH [14], STMH [38], CMSSH [2], SCM [44], SePH [16]), and a deep-structure-based method (DCMH [12]). In order to conduct a fair comparison, we utilize both CNN-F [5] and vgg19 [32], which have both been pre-trained on the ImageNet datasets [28] in order to extract deep features for all shallow-structure-based baselines.

In order to determine the hyper-parameters α, γ, η, and β, we randomly select some data points (2,000 for each dataset) from the retrieval database to serve as our validation set. A sensitivity analysis of these hyper-parameters is provided in Fig. 2. It can be seen that high performance can always be achieved when α=γ=1 and η=β=10^{-4}. For image modality, we initialize the first seven layers of *ImgNet* with the CNN-F network pre-trained on the ImageNet dataset. For text modality, *TxtNet* randomly is initialized. The learning rate is chosen from 10^{-4} to 10^{-8}. Following this, we show the average results of the 10 runs.

4.3. Performance

Hamming Ranking: Table 2 reports the MAP results for both our SSAH and the other compared methods with CNN-F features on three popular datasets (MIRFLICKR-25K, NUS-WIDE and MS COCO) in cross-modal retrieval. "I→T" denotes that the query is image and the database is text-based, and "T→I" denotes that the query is text and the database is image-based. Compared with the shallow baselines of CVH, STMH, CMSSH, SCM and SePH, our SSAH achieves absolute more than a 10% increase on MAP for I→T/T→I on the MIRFLICKR-25K dataset. While when comparing our SSAH with the deep-learning-based method (DCMH), we run the source code provided by the author. Here, it can be seen that SSAH can achieve more than a 5% increase on MAP. For another two datasets NUS-WIDE and MS COCO with more instances and complex content, which are more challenging, SSAH always provides superior performance than other comparison methods, as presented in Table 2. This may be because, during the learning process, the proposed self-supervised adversarial network more effectively facilitate the learning of semantic relevance between different modalities, which means that more discriminative representations can be learned using our SSAH. As a result, SSAH can more accurately capture correlations between modalities.

We further verify our SSAH using vgg19 features [32] that have been pre-trained on the ImageNet dataset. Table 3 shows the MAP results on three different datasets. As shown in Table 3, we can see that almost all methods that are based on vgg19 outperform those based on CNN-F. Not only that, but it becomes evident that our SSAH consistently achieves the best performance. Compared with the shallow baselines (CVH, STMH, CMSSH, SCM and SePH), SSAH achieves absolute more than 5% increase on an average MAP for I→T/T→I on the MIRFLICKR-25K dataset. This means that the proposed SSAH can be applied to other networks and can achieve more accurate retrieval when equipped with an effective deep-network structure.

Hash Lookup: When considering the lookup protocol, we compute the PR curve for the returned points given any Hamming radius. The PR curve can be obtained by varying the Hamming radius from 0 to 16 with a step-size of 1. Fig. 4 shows the PR curves of all the current state-of-the-art methods with 16-bit hash codes on three benchmark datasets. In this way, it can be seen that our SSAH significantly outperforms all of its state-of-the-art competitors.

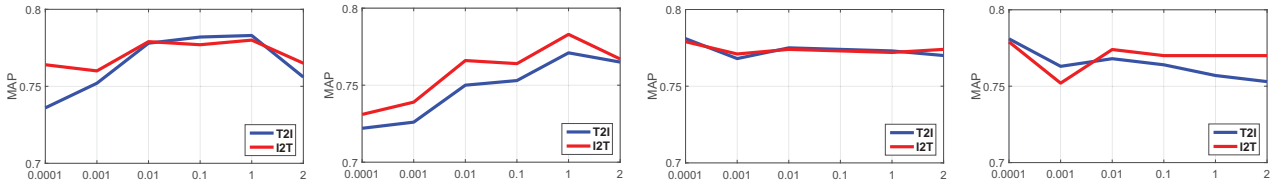
Ablation study of SSAH: We also verify the impact of different network modules on our SSAH's performance. Three variants are designed as baselines of our SSAH networks: (a) SSAH-1 is built by removing the self-supervised semantic network; (b) SSAH-2 is built by replacing *TxtNet* with three full-connected layers; (c) SSAH-3 is built by removing the adversarial learning module. Fig. 3 shows the

Table 2: MAP. The best accuracy is shown in boldface. The baselines are based on CNN-F features.

TASK	Method	Flickr-25K			NUS-WIDE			MS COCO		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
I→ T	CVH [14]	0.557	0.554	0.554	0.400	0.392	0.386	0.412	0.401	0.400
	STMH [38]	0.602	0.608	0.605	0.522	0.529	0.537	0.422	0.459	0.475
	CMSSH [2]	0.585	0.584	0.572	0.511	0.506	0.493	0.512	0.495	0.482
	SCM [44]	0.671	0.682	0.685	0.533	0.548	0.557	0.483	0.528	0.550
	SePH [16]	0.657	0.660	0.661	0.478	0.487	0.489	0.463	0.487	0.501
	DCMH [12]	0.735	0.737	0.750	0.478	0.486	0.488	0.511	0.513	0.527
	OURS	**0.782**	**0.790**	**0.800**	**0.642**	**0.636**	**0.639**	**0.550**	**0.558**	**0.557**
T→ I	CVH [14]	0.557	0.554	0.554	0.372	0.366	0.363	0.367	0.359	0.357
	STMH [38]	0.600	0.606	0.608	0.496	0.529	0.532	0.431	0.461	0.476
	CMSSH [2]	0.567	0.569	0.561	0.449	0.389	0.380	0.429	0.408	0.398
	SCM [44]	0.697	0.707	0.713	0.463	0.462	0.471	0.465	0.521	0.548
	SePH [16]	0.648	0.652	0.654	0.449	0.454	0.458	0.449	0.474	0.499
	DCMH [12]	0.763	0.764	0.775	0.638	0.651	0.657	0.501	0.503	0.505
	OURS	**0.791**	**0.795**	**0.803**	**0.669**	**0.662**	**0.666**	**0.537**	**0.538**	**0.529**

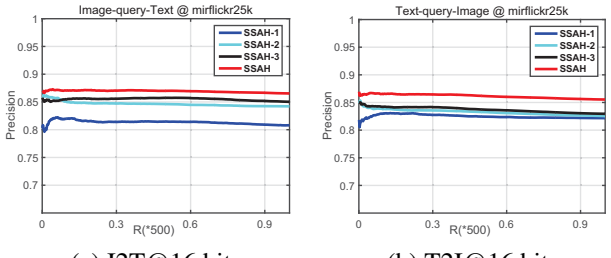
Table 3: MAP. The best accuracy is shown in boldface. The baselines are based on vgg19 features.

TASK	Method	Flickr-25K			NUS-WIDE			MS COCO		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
I→ T	CVH [14]	0.557	0.554	0.554	0.405	0.397	0.391	0.441	0.428	0.402
	STMH [38]	0.591	0.606	0.613	0.471	0.516	0.549	0.445	0.482	0.502
	CMSSH [2]	0.593	0.592	0.585	0.508	0.506	0.495	0.504	0.495	0.492
	SCM [44]	0.685	0.693	0.697	0.497	0.502	0.499	0.498	0.556	0.565
	SePH [16]	0.709	0.711	0.716	0.479	0.501	0.492	0.489	0.502	0.499
	DCMH [12]	0.677	0.703	0.725	0.590	0.603	0.609	0.497	0.506	0.511
	OURS	**0.797**	**0.809**	**0.810**	**0.636**	**0.636**	**0.637**	**0.550**	**0.577**	**0.576**
T→ I	CVH [14]	0.557	0.554	0.554	0.385	0.379	0.373	0.413	0.402	0.388
	STMH [38]	0.600	0.613	0.616	0.472	0.526	0.550	0.446	0.478	0.506
	CMSSH [2]	0.585	0.570	0.569	0.377	0.389	0.388	0.417	0.420	0.416
	SCM [44]	0.707	0.714	0.719	0.567	0.583	0.597	0.492	0.556	0.568
	SePH [16]	0.722	0.723	0.727	0.487	0.493	0.488	0.485	0.495	0.485
	DCMH [12]	0.705	0.717	0.724	0.620	0.634	0.643	0.507	0.520	0.527
	OURS	**0.782**	**0.797**	**0.799**	**0.653**	**0.676**	**0.683**	**0.552**	**0.578**	**0.578**



(a) α (b) γ (c) η (d) β

Figure 2: A sensitivity analysis of the hyper-parameters.



(a) I2T@16 bit (b) T2I@16 bit

Figure 3: Precision@top1000 curves on MIRFLICKR-25K.

comparison results with 16 bits on the MIRFLICKR-25K dataset. From the results, we can see that our method can achieve a more accurate performance when using the designed modules and that the self-supervised semantic network significantly improves the performance.

Training efficiency: Fig. 5 shows the variation between MAP and training time for SSAH and DCMH. We can see that our approach reduces training time by a factor of 10 over DCMH. In comparison to DCMH, SSAH exploits *LabNet* to learn more sufficient supervised information from

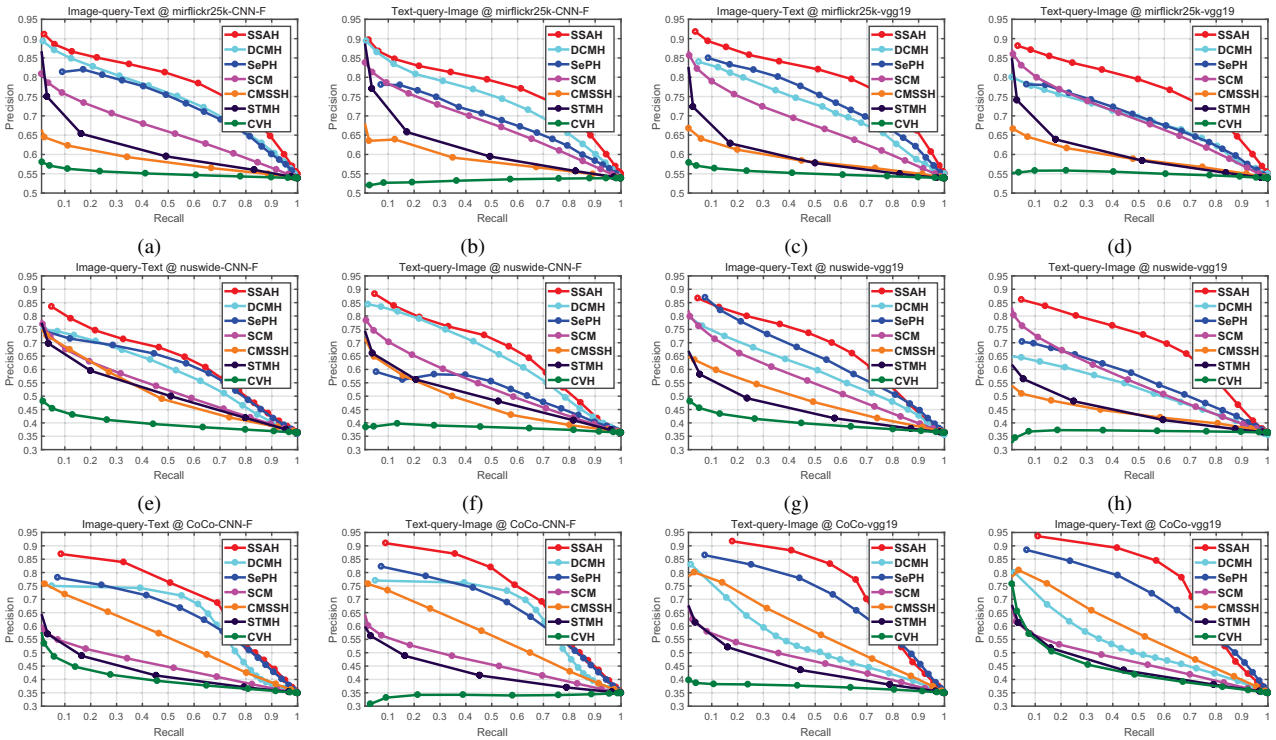
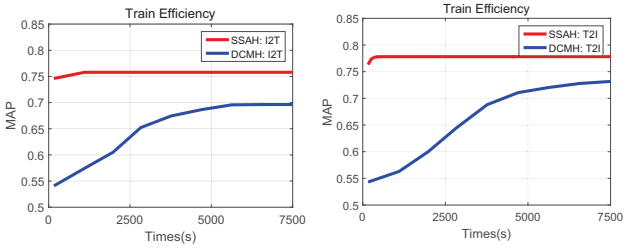


Figure 4: Precision-recall curves. The baselines are based on CNN-F features. The code length is 16.



(a) I2T@16 bit (b) T2I@16 bit

Figure 5: Training Efficiency of SSAH and DCMH.

high-dimensional semantic features and hash codes, using it to train *ImgNet* and *TxtNet* efficiently. Thus, more accurate correlations between different modalities can be captured and the modality gap can be bridged more effectively.

Comparison with ACMR: According to our current best knowledge, ACMR [37] is the first work that borrows from adversarial learning approach for cross-modal retrieval. However, ACMR is not a hashing-based method. So as to be fairly compared with ACMR, we follow the experiment settings used in ACMR. SSAH is conducted on an NUS-WIDE-10k dataset, which is constructed by randomly selecting 10,000 image/text pairs from the 10 largest categories within the NUS-WIDE dataset. Table 4 shows the experiment results. The underlined results are reported in ACMR. It can be seen that our method outperforms ACMR significantly. This may be because two adversarial networks are used in our framework, with which SSAH can more accurately learn the distribution of different modalities and can thus capture the correlation more effectively.

Table 4: MAP with CNN-F features on NUS-WIDE.

Method	ACMR		SSAH	
Task	I→ T	T→ I	I→ T	T→ I
MAP	<u>0.544</u>	<u>0.538</u>	**0.617**	**0.641**

5. Conclusion

In this work, we proposed a novel deep hashing approach, dubbed self-supervised adversarial hashing (SSAH), in order to address the problem of cross-modal retrieval more effectively. The proposed SSAH incorporates a self-supervised semantic network coupled with multi-label information, and carries out adversarial learning to maximize the semantic relevance and feature distribution consistency between different modalities. The extensive experiments show that SSAH achieves state-of-the-art retrieval performance on three benchmark datasets.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 61572388 and Grant 61703327, the Key R&D Program/The Key Industry Innovation Chain of Shaanxi under Grant 2017ZDCXL-GY-05-04-02 and Grant 2017ZDCXL-GY-05-04-02, and ARC FL-170100117, DP-180103424, DP-140102164, LP-150100671.

References

[1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013.

[2] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, pages 3594–3601, 2010.

[3] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu. Deep visual-semantic hashing for cross-modal retrieval. In *ACM SIGKDD*, pages 1445–1454, 2016.

[4] Y. Cao, M. Long, J. Wang, and H. Zhu. Correlation autoencoder hashing for supervised cross-modal search. In *ACM ICMR*, pages 197–204, 2016.

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[6] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM CIVR*, page 48, 2009.

[7] C. Deng, X. Tang, J. Yan, W. Liu, and X. Gao. Discriminative dictionary learning with common label alignment for cross-modal retrieval. *IEEE Trans. Multimed.*, 18(2):208–218, 2016.

[8] G. Ding, Y. Guo, and J. Zhou. Collective matrix factorization hashing for multimodal data. In *CVPR*, pages 2083–2090, 2014.

[9] V. Erin Liong, J. Lu, Y.-P. Tan, and J. Zhou. Cross-modal deep variational hashing. In *ICCV*, 2017.

[10] F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. pages 7–16, 2014.

[11] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *ACM CIVR*, pages 39–43, 2008.

[12] Q.-Y. Jiang and W.-J. Li. Deep cross-modal hashing. In *CVPR*, 2017.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[14] S. Kumar and R. Udupa. Learning hash functions for cross-view similarity search. In *IJCAI*, volume 22, page 1360, 2011.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[16] Z. Lin, G. Ding, M. Hu, and J. Wang. Semantics-preserving hashing for cross-view retrieval. In *CVPR*, 2015.

[17] H. Liu, R. Ji, Y. Wu, F. Huang, and B. Zhang. Cross-modality binary code learning via fusion similarity hashing. In *CVPR*, 2017.

[18] W. Liu, C. Mu, S. Kumar, and S.-F. Chang. Discrete graph hashing. In *NIPS*, pages 3419–3427, 2014.

[19] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang. Supervised hashing with kernels. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2074–2081. IEEE, 2012.

[20] W. Liu, J. Wang, S. Kumar, and S.-F. Chang. Hashing with graphs. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1–8. Citeseer, 2011.

[21] X. Liu, C. Deng, B. Lang, D. Tao, and X. Li. Query-adaptive reciprocal hash tables for nearest neighbor search. *IEEE Transactions on Image Processing*, 25(2):907–919, 2016.

[22] X. Liu, B. Du, C. Deng, M. Liu, and B. Lang. Structure sensitive hashing with adaptive product quantization. *IEEE transactions on cybernetics*, 46(10):2252–2264, 2016.

[23] X. Liu, L. Huang, C. Deng, B. Lang, and D. Tao. Query-adaptive hash code ranking for large-scale multi-view visual search. *IEEE Transactions on Image Processing*, 25(10):4514–4524, 2016.

[24] X. Liu, Z. Li, C. Deng, and D. Tao. Distributed adaptive binary quantization for fast nearest neighbor search. *IEEE Transactions on Image Processing*, 26(11):5324–5336, 2017.

[25] D. Mandal, K. N. Chaudhury, and S. Biswas. Generalized semantic preserving hashing for n-label cross-modal retrieval. In *CVPR*, 2017.

[26] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber. Multimodal similarity-preserving hashing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(4):824–830, 2014.

[27] V. Ranjan, N. Rasiwasia, and C. V. Jawahar. Multi-label cross-modal retrieval. In *ICCV*, 2015.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

[29] F. Shen, W. Liu, S. Zhang, Y. Yang, and H. T. Shen. Learning binary codes for maximum inner product search. In *2015 IEEE International Conference on Computer Vision (ICCV)*, volume 11, pages 4148–4156. IEEE, 2015.

[30] F. Shen, C. Shen, W. Liu, and H. T. Shen. Supervised discrete hashing. In *CVPR*, volume 2, page 5, 2015.

[31] Y. Shen, L. Liu, L. Shao, and J. Song. Deep binaries: Encoding semantic-rich cues for efficient textual-visual cross retrieval. In *ICCV*, 2017.

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[33] D. Song, W. Liu, R. Ji, D. A. Meyer, and J. R. Smith. Top rank supervised binary coding for visual search. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1922–1930, 2015.

[34] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *ACM SIGMOD*, pages 785–796, 2013.

[35] S. Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013.

[36] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. Deep learning for content-based image retrieval: A comprehensive study. In *ACM MM*, pages 157–166, 2014.

[37] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen. Adversarial cross-modal retrieval. In *ACM MM*, pages 154–162, 2017.

[38] D. Wang, X. Gao, X. Wang, and L. He. Semantic topic multimodal hashing for cross-media retrieval. In *IJCAI*, pages 3890–3896, 2015.

[39] J. Wang, W. Liu, A. X. Sun, and Y.-G. Jiang. Learning hash codes with listwise supervision. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3032–3039. IEEE, 2013.

[40] F. Wu, Y. Zhou, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang. Sparse multi-modal hashing. *IEEE Trans. Multimed.*, 16(2):427–439, 2014.

[41] Y. Wu, S. Wang, and Q. Huang. Online asymmetric similarity learning for cross-modal retrieval. In *CVPR*, 2017.

[42] C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[43] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao. Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI*, pages 1618–1625, 2017.

[44] D. Zhang and W. J. Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI*, pages 2177–2183, 2014.

[45] T. Zhang and J. Wang. Collaborative quantization for cross-modal similarity search. In *CVPR*, 2016.

[46] J. Zhou, G. Ding, and Y. Guo. Latent semantic sparse hashing for cross-modal similarity search. In *ACM SIGIR*, pages 415–424, 2014.