

# Unsupervised Contrastive Cross-modal Hashing

Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng

**Abstract**—In this paper, we study how to make unsupervised cross-modal hashing (CMH) benefit from contrastive learning (CL) by overcoming two challenges. To be exact, i) to address the performance degradation issue caused by binary optimization for hashing, we propose a novel momentum optimizer that performs hashing operation learnable in CL, thus making on-the-shelf deep cross-modal hashing possible. In other words, our method does not involve binary-continuous relaxation like most existing methods, thus enjoying better retrieval performance; ii) to alleviate the influence brought by false-negative pairs (FNPs), we propose a Cross-modal Ranking Learning loss (CRL) which utilizes the discrimination from all instead of only the hard negative pairs, where FNP refers to the within-class pairs that were wrongly treated as negative pairs. Thanks to such a global strategy, CRL endows our method with better performance because CRL will not overuse the FNPs while ignoring the true-negative pairs. To the best of our knowledge, the proposed method could be one of the first successful contrastive hashing methods. To demonstrate the effectiveness of the proposed method, we carry out experiments on five widely-used datasets compared with 13 state-of-the-art methods. The code is available at <https://github.com/penghu-cs/UCCH>.

**Index Terms**—Cross-modal retrieval, unsupervised cross-modal hashing, common Hamming space, contrastive hashing network.

## 1 INTRODUCTION

CROSS-MODAL retrieval aims to retrieve semantically relevant samples from one modality (e.g., image) by utilizing a query from another modality (e.g., text). The key challenge of cross-modal retrieval is to bridge the gap between different modalities. To narrow such a so-called heterogeneity gap, a variety of methods have been proposed and achieved promising performance [1], [2], [3], [4]. However, these approaches suffer from high costs in storage and computation because the representation learned by these methods is with continuous values that are less attractive to large-scale cross-modal retrieval. Therefore, it is still an open issue to efficiently bridge the heterogeneity gap for large-scale cross-modal retrieval.

To efficiently narrow the heterogeneity gap and boost retrieval performance, cross-modal hashing has been of considerable interest in the community [5], [6], [7], [8], [9], [10]. The basic idea of cross-modal hashing is projecting the high-dimensional multimodal data into compact binary bits [11], [12], [13]. Thanks to the bit-wise similarity measurement (i.e., XOR), the hashing process will be

more efficient than continuous-value methods [7], [8], [14] in terms of storage and computation. Most existing cross-modal hashing approaches could be roughly classified into supervised and unsupervised categories. More specifically, the supervised approaches [8], [9], [15], [16] often learn the hash codes from the multimodal data by using the labeled semantic information and have achieved promising performance. However, they need a large amount of labeled data and the data annotation is labor-intensive [13], [17]. Different from supervised methods, unsupervised cross-modal hashing methods [13], [18], [19] could avoid intensive data annotation, which is more attractive in practice. In this paper, we mainly focus on the unsupervised learning paradigm.

All existing unsupervised cross-modal hashing approaches are based on either shallow or deep models. In brief, shallow methods learn one-layer linear or nonlinear transformations to project different modalities into a common Hamming space [13], [18], [20]. However, these shallow models cannot capture the highly-level nonlinear information well [21], and thus they would achieve sub-optimal performance. To address this issue, Deep Neural Networks (DNNs) [13], [19], [22], [23] are used to learn the hashing functions given their advantages in modeling nonlinearity. Motivated by the success of recent contrastive learning, it is highly expected to investigate how to conduct unsupervised contrastive learning for cross-modal hashing. Although such an idea seems straightforward, it is nontrivial due to the following two challenges. First, contrastive learning is always treated as a pre-train step, and there exists a gap with the downstream cross-modal hashing retrieval. In fact, contrastive learning usually takes a continuous value optimization strategy which is inconsistent with the binary output of cross-modal hashing (see Fig. 1(a)), thus probably resulting in performance degradation. Second, to bridge hashing learning and cross-modal retrieval, most existing cross-modal hashing methods employ a max-margin ranking loss whose performance heavily depends on the

*This work is supported by the National Natural Science Foundation of China (Grants No. 62102274, U19A2078, U21B2040, 61971296, and 62176171), Sichuan Science and Technology Planning Project (Grants No. 2021YFS0389, 2021YFG0317, and 2021YFG0301), China Postdoctoral Science Foundation (No. 2021M692270), AME Programmatic Funding Scheme (Project No. A18A2b0046), Career Development Fund (Project C210812033) and RobotHTPO Seed Fund (Project C211518008), Open Research Projects of Zhejiang Lab under Grant 2021KH0AB02, EDB OSTIN Space Technology and Development Programme (Project S22-19016-STD), A\*STAR under its AME Programmatic Funds (Projects No. A18A2b0046, and A1892b0026).*

*Peng Hu and Xi Peng are with the College of Computer Science, Sichuan University, Chengdu 610065, China.*

*Hongyuan Zhu and Jie Lin are with the Institute for Infocomm Research, Agency for Science, Technology and Research (A\*STAR), Singapore.*

*Dezhong Peng is with the College of Computer Science, Sichuan University, Chengdu 610065, China, with Chengdu Ruibei Yingte Information Technology Ltd. Company, Chengdu 610094, China, and with Sichuan Zhiqian Technology Ltd. Company, Chengdu 610094, China.*

*Yin-Ping Zhao is with the School of Software, Northwestern Polytechnical University, Xi'an 710072, China.*

*Corresponding author: Xi Peng (email: pengx.gm@gmail.com).*

established positive and negative pairs (see Fig. 1(b)). In an unsupervised setting, however, it is difficult to well establish the positive and negative samples due to the unavailability of labels. Hence, unsupervised cross-modal hashing generally treats the co-occurred samples as positives and the other samples as negative. Clearly, such an approach will lead to a novel noise, *i.e.*, a number of within-class samples are wrongly treated as negative. To the best of our knowledge, such a false-negative pair (FNP) problem is less touched so far and no existing solution is available for cross-modal hashing.

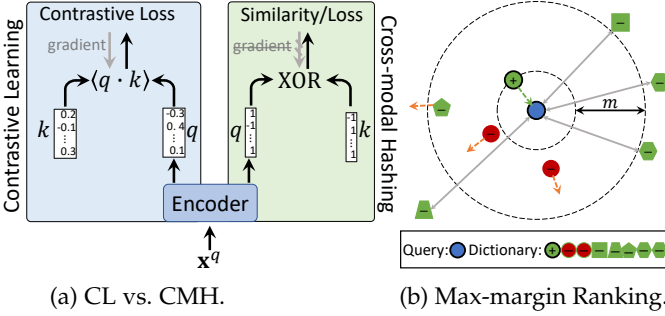


Fig. 1: The limitations of existing methods. In this figure, we take a query as an example. (a) shows the difference/gap between contrastive learning (CL) and cross-modal hashing (CMH). More specifically, CL is optimized with continuous values in a differentiable manner. On the contrary, CMH is performed with binary codes and could not compute the gradients. In the figure,  $q$  and  $k$  are the query and key samples, respectively. (b) shows the traditional max-margin ranking will ignore the cross-modal samples outside the margin, leading to more attention paid to false-negative ones (red circles). In the figure, green lines represent relevant/positive cross-modal correlations; orange lines represent irrelevant/negative cross-modal correlations; gray lines denote ignored correlations; blue items represent query samples; green items denote true-negative samples, and red items represent false-negative points.

To tackle the aforementioned two problems, we propose a deep unsupervised cross-modal hashing method, termed Unsupervised Contrastive Cross-modal Hashing (UCCH). To be specific, UCCH employs a novel momentum-based binarization optimizer to endow hashing operation learnable, thus making on-the-shelf deep cross-modal hashing possible. Second, to overcome the FNP challenge, we propose a Cross-modal Ranking Learning loss (CRL) which utilizes the discrimination from all instead of the hardest negative pairs (see Fig. 4). Such a remedy is proposed to avoid the performance degradation caused by a property of the max-margin loss, *i.e.*, the traditional triplet loss with max-margin [19], [24], [25] is apt to overfit on the FNPs while ignoring the true-negative pairs (TNPs) because FNPs are usually more attractive and harder than TNPs to DNN optimization (see Section 4.3.7).

Different from the well-studied contrastive learning models, our UCCH is a task-specified contrastive learning method. More specifically, almost all existing contrastive learning methods [26], [27], [28] aim to learn a model in a self-supervised manner, and then fine-tune the model to fit

the downstream tasks. Limited by such a two-stage strategy, there exists a performance gap between the contrastive learning model and the downstream tasks. To bridge the performance gap, our UCCH is specifically designed for achieving cross-modal hashing in a one-stage fashion. In addition, different from the max-margin loss, our CRL could exploit more discrimination from all negative pairs than hard ones, because the former contains more TNPs (see Section 4.3.5 and Section 4.3.7), thus embracing better performance. To the best of our knowledge, this could be one of the first studies on FNPs in cross-modal hashing.

The main contribution and novelty of this work could be summarized as follows:

- To the best of our knowledge, the proposed UCCH could be the first method that endows contrastive learning with unsupervised cross-modal hashing.
- A novel momentum optimizer is proposed to make the binary memory bank learnable, thus narrowing the gap between contrastive learning and hashing.
- A Cross-modal Ranking Learning loss (CRL) is proposed to overcome the FNP challenge by utilizing the discrimination from all instead of hard negative pairs. Thanks to CRL, our method embraces better performance and robustness to FNPs.
- Extensive experiments verify the efficacy of our method on five widely-used benchmark multimodal datasets compared with 13 state-of-the-art methods.

## 2 RELATED WORK

Over the past decades, a variety of methods have been proposed to learn a common Hamming space to bridge the heterogeneity gap existing into cross-modal data. In this section, we briefly review some related works from the aspects of supervised cross-modal hashing methods, unsupervised cross-modal hashing methods, and contrastive learning.

### 2.1 Supervised Cross-modal Hashing Methods

By utilizing the discriminative information rooted in the label, almost all of supervised cross-modal methods learn different modality-specific hashing functions to project the multimodal data into a common Hamming space [15], [29], [30], [31]. One typical approach utilizes max-margin ranking loss to apply the ranking information for cross-modal hashing learning [15], [32], [33]. In [32], Ding *et al.* propose a novel ranking-based hashing framework to map different modalities into a common Hamming space by explicitly employing the ranking information with a max-margin ranking loss. To utilize the semantic ranking information for cross-modal hashing learning, Liu *et al.* present a ranking-based deep cross-modal hashing approach to learn unified Hamming representations from different modalities with a max-margin loss [33]. Furthermore, Jiang *et al.* propose a Discrete Latent Factor model that employs cross-modal Hashing (DLFH) to directly learn the binary hash codes by using the semantic information [23]. Xu *et al.* [34] develop a Discrete Cross-modal Hashing (DCH) method which directly learns unified discriminative binary codes by retaining the discrete constraints for multimodal data. These supervised cross-modal hashing methods have achieved

promising performance for cross-modal retrieval thanks to the semantic information of annotated data. However, to achieve desirable performance, these supervised methods require huge amount of labeled data which is cost- and time-prohibitive.

## 2.2 Unsupervised Cross-modal Hashing Methods

Unsupervised cross-modal hashing methods learn the unified binary codes by minimizing the correlation of the cross-modal pairs (such as image-text pairs) that are cheaper than the data annotation. Hence, unsupervised cross-modal hashing methods have attracted much attention from academic and industrial researchers [13], [18], [20]. In [20], Kumar *et al.* propose a Cross-view Hashing (CVH) method to learn the common hash codes by using the intra- and inter-modal similarities of the multimodal data. In [35], a Collective Matrix Factorization Hashing (CMFH) method is proposed to learn a common Hamming space by using collective matrix factorization with a latent factor model. Liu *et al.* [18] propose Fusion Similarity Hashing (FSH) method which explicitly embeds the graph-based fusion similarity between distinct modalities into the common hash representations. As the aforementioned methods are shallow methods, they cannot capture the highly nonlinear semantics in the multimodal data. To address this problem, some DNN-based methods have been recently proposed. In [19], an Unsupervised Generative Adversarial Cross-modal Hashing (UGACH) approach is proposed to employ GAN's ability to exploit the underlying manifold structure of cross-modal data with max-margin ranking loss. In [13], an unsupervised method, termed Unsupervised coupled Cycle generative adversarial Hashing networks (UCH), is proposed to learn the unified binary representations by employing outer- and inner-cycle network.

## 2.3 Unsupervised Contrastive Hashing

In recent, contrastive learning [26], [27], [28] has attracted considerable attention from the community. For example, Wu *et al.* [27] observed that the apparent similarity can be learned from data themselves without explicit guidance. Thus, instead of learning from label-level discrimination, contrastive learning is proposed to learn the discrimination at the instance level [26], [27]. Inspired by the huge success of contrastive learning, some contrastive hashing methods [36], [37], [38] were proposed to learn binary representations from unimodal data, and achieved promising performance. In brief, Li *et al.* [36] proposed a self-supervised hashing approach based on dual pseudo agreement by using an end-to-end differentiable network, a contrastive loss, a hashing loss, and a balance loss. Qiu *et al.* [38] designed a general probabilistic hashing method to learn binary hashing codes by minimizing the contrastive loss while reducing the mutual information between the codes and original input data. In addition, Jang *et al.* [37] proposed an unsupervised deep quantization-based image retrieval method, dubbed Self-supervised Product Quantization (SPQ) network. In short, SPQ jointly learns the feature extractor and the codewords by combining contrastive learning with Product Quantization (PQ) codewords. However, these methods are developed to handle unimodal

data and less effort has been devoted to exploring how to incorporate contrastive learning into cross-modal hashing as far as we know. To enable cross-modal hashing to benefit from contrastive learning, there are two challenges at least, *i.e.*, the discrete optimization and the FNP issue as discussed in Introduction.

## 3 THE PROPOSED METHOD

In numerous real-world applications, one instance can be described by different modalities, such as image, text, audio, etc. Without loss of generality, we focus on bimodal (*i.e.*, image and text) hashing problem in this paper. As shown in Fig. 2, our UCCH consists of feature extraction and hashing learning modules. Specifically, the feature extraction module aims to extract the features using the given extractors from the original multimedia inputs to represent the corresponding image/text samples. The detail implementation of this module will be explained in Section 4. Our hashing learning module attempts to project different modalities into a latent common Hamming space, where the correlated samples are compacted and uncorrelated ones are scattered. In this section, we present the details of the proposed approach including the problem formulation and the hashing learning algorithm.

### 3.1 Problem Formulation

For ease of presentation, we first give some definitions for cross-modal Hashing problem. Boldface uppercase letters (*e.g.*,  $\mathbf{X}$ ) and boldface lowercase letters (*e.g.*,  $\mathbf{x}$ ) denote matrices and vectors, respectively. Let  $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$  denote a cross-modal dataset with  $n$  image-text pairs/instances, where  $\mathbf{x}_i \in \mathbb{R}^{d_x \times 1}$  is the  $i$ -th sample of the image modality,  $\mathbf{y}_i \in \mathbb{R}^{d_y \times 1}$  is the text modality related to  $\mathbf{x}_i$ , and  $d_x$  and  $d_y$  are the dimension of the image and text features.

The goal of cross-modal hashing is to project different modalities into a common Hamming space. In the space, the unified codes of image and text are denoted as:  $\mathcal{B}^x = \{\mathbf{b}_i^x\}_{i=1}^n$  for the image modality and  $\mathcal{B}^y = \{\mathbf{b}_i^y\}_{i=1}^n$  for the text modality, where  $\mathbf{b}_i^* \in \{-1, +1\}^L$ ,  $*$   $\in \{x, y\}$  and  $L$  is the length of hash codes. The Hamming distance is used to evaluate the similarity between image and text samples. More specifically, if the  $i$ -th image and the  $j$ -th text are similar, the Hamming distance between  $\mathbf{b}^x$  and  $\mathbf{b}^y$  should be small. Otherwise, the Hamming distance between dissimilar samples should be large. To facilitate the computation of Hamming distance, we can use the inner product  $\langle \mathbf{b}^x, \mathbf{b}^y \rangle$  to compute the Hamming distance  $d(\mathbf{b}^x, \mathbf{b}^y)$  as  $d(\mathbf{b}^x, \mathbf{b}^y) = \frac{1}{2}(L - \langle \mathbf{b}^x, \mathbf{b}^y \rangle)$ . Thus, the similarity between the  $i$ -th image and the  $j$ -th text can be quantified by the inner product  $\langle \mathbf{b}^x, \mathbf{b}^y \rangle$  in the Hamming space.

To transform different modalities to the unified binary codes, we learn two modality-specific hash functions for the cross-modal inputs. Here, we design a couple of modality-specific networks to this end. In details, the two hash functions are formulated as  $f^x(\mathbf{x}, \Theta^x)$  and  $f^y(\mathbf{y}, \Theta^y)$  for image and text, where  $\Theta^x$  and  $\Theta^y$  are the corresponding modality-specific network parameters to be learned. In our UCCH, the outputs of the hash functions are defined as

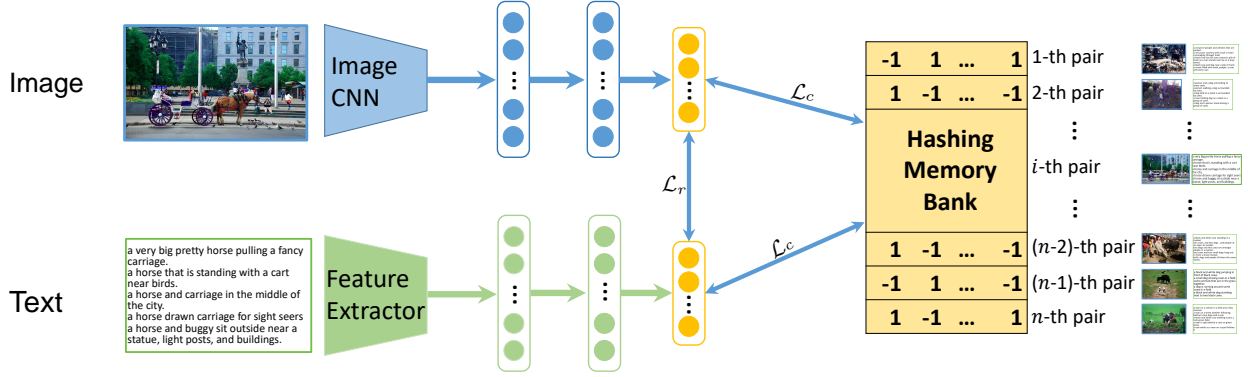


Fig. 2: The pipeline of the proposed method and we take a bimodal case as an example. In the example, two modality-specific networks learn unified binary representations for different modalities. The outputs of networks directly interact with the hash codes to learn the latent discrimination by using instance-level contrast without continuous relaxation, *i.e.*, contrastive hashing learning ( $\mathcal{L}_c$ ). The cross-modal ranking loss  $\mathcal{L}_r$  is utilized to bridge cross-modal hashing learning to cross-modal retrieval.

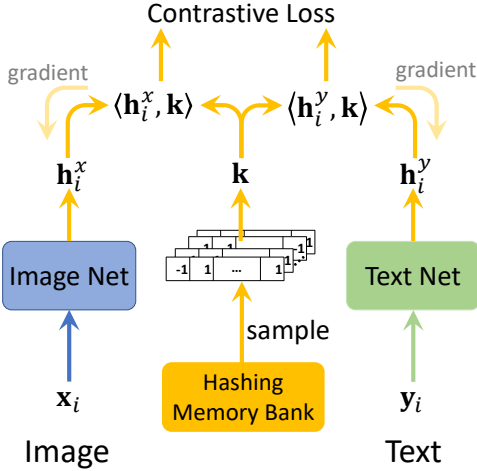


Fig. 3: Contrastive hashing learning adopts a contrastive loss to train the modality-specific networks by matching an image-text query  $(\mathbf{x}_i, \mathbf{y}_i)$  to a dictionary that is sampled from the hashing memory bank. By contrasting with the positive and negative keys, relevant cross-modal pairs directly approximate the corresponding unified binary codes and separate from their irrelevant pairs without continuous relaxation. The memory bank is driven by a momentum update with the corresponding pairs.

$\mathbf{h}_i^x = f^x(\mathbf{x}_i)$  and  $\mathbf{h}_i^y = f^y(\mathbf{y}_i)$  for the  $i$ -th image and the  $i$ -th text points, respectively. With the learned hash functions, the binary representation of a sample is calculated by applying the sign function to  $\mathbf{h}_i^*$ :

$$\mathbf{b}_i^* = \text{sgn}(\mathbf{h}_i^*), \quad * \in \{x, y\}, \quad (1)$$

where  $\text{sgn}(x)$  is the sign function, whose value is 1 if  $x \geq 0$ , and -1 otherwise. To learn the hash functions, we propose a novel unsupervised objective function to enforce the networks to eliminate the cross-modal discrepancy. Unlike the supervised methods, our UCCN adopts the contrastive learning to excavate apparent similarity among image-text pairs instead of labels. The overall objective function of our

UCCN is formulated as follows:

$$\arg \min_{\Theta^x, \Theta^y} (\beta \mathcal{L}_c + (1 - \beta) \mathcal{L}_r), \quad (2)$$

where  $\beta$  ( $0 < \beta < 1$ ) is a trade-off hyper-parameter to balance the contrastive hashing loss  $\mathcal{L}_c$  and the cross-modal ranking loss  $\mathcal{L}_r$ , which are detailedly introduced in the following sections.

### 3.2 Contrastive Cross-Modal Hashing Learning

Contrastive learning [26] aims at learning discriminative representations by using the similar and dissimilar relationship of the query-key pairs. This also can be thought of as a dictionary look-up problem [28], [39]. Different from exiting contrastive learning methods, we present a task-specific Contrastive Cross-modal Hashing learning method (CCH) by leveraging a unified binary dictionary for all modalities. Such a problem has been less touch so far to the best of our knowledge. Without continuous-value relaxation, for a given query  $\mathbf{h}_i^*$  ( $* \in \{x, y\}$ ), it aims to directly retrieve the correlated/positive keys from the hashed points  $\{\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_n\}$  of the dictionary. Moreover, the  $i$ -th key  $\mathbf{k}_i$  of the dictionary corresponds to the  $i$ -th image-text pair. In the unsupervised cross-modal case, there is a single positive key (denoted as  $\mathbf{k}_i^+$ ) in the dictionary, which is matched to the query  $\mathbf{h}_i^*$  ( $* \in \{x, y\}$ ). The contrastive loss [26], [28] evaluates the similarity between a query  $\mathbf{h}_i^*$  ( $* \in \{x, y\}$ ) and its retrieved results  $\{\mathbf{k}_i\}_{i=1}^n$ , whose value is low when  $\mathbf{h}_i^*$  ( $* \in \{x, y\}$ ) is similar to its positive key  $\mathbf{k}_i^+$  and dissimilar to all other keys (considered as negative keys for the query).

In practice, it is infeasible to retrieve on a large dictionary for a large-scale dataset. To address this large scale learning issue, we randomly sample a part of the whole dictionary  $\{\mathbf{k}_i\}_{i=1}^n$  (considered as hashing memory bank) as a new small dictionary for the retrieval (Fig. 3). Specifically, in contrast to the positive key, we randomly sample  $K$  points from the hashing memory bank to construct a negative key set  $\{\mathbf{k}_j^-\}_{j=1}^K$ , where  $\mathbf{k}_j^- = \text{sgn}(\mathbf{v}_r)$ ,  $\mathbf{v}_r$  is the corresponding continuous-valued source key of  $\mathbf{k}_j^-$  (the detail definition is given in the following part) and  $r$  is the

corresponding random index. Following [27], we enforce  $\|\mathbf{h}^*\| = 1$  ( $* \in \{x, y\}$ ) and  $\|\mathbf{k}^*\| = 1$  ( $* \in \{+, -\}$ ) via the  $\ell_2$ -normalization. As the aforementioned discussion, the similarity between different hash points is measured by dot product. With the dot-product similarity, an effective contrastive loss function, called InfoNCE [40], is adopted to maximize the instance-level discrimination and minimize the cross-modal discrepancy:

$$\mathcal{L}_c = -\sum_{i=1}^n \log P(i|\mathbf{h}_i^x) - \sum_{i=1}^n \log P(i|\mathbf{h}_i^y), \quad (3)$$

where  $P(i|\mathbf{h}_i^x)$  and  $P(i|\mathbf{h}_i^y)$  are the probability of  $\mathbf{h}_i^x$  and  $\mathbf{h}_i^y$  being recognized as the  $i$ -th point. The formulations are defined as:

$$P(i|\mathbf{h}_i^*) = \frac{\exp(\langle \mathbf{h}_i^*, \mathbf{k}_i^+ \rangle / \tau)}{\exp(\langle \mathbf{h}_i^*, \mathbf{k}_i^+ \rangle / \tau) + \sum_{j=1}^K \exp(\langle \mathbf{h}_i^*, \mathbf{k}_j^- \rangle / \tau)}, \quad (4)$$

where  $* \in \{x, y\}$ , and  $\tau$  is a temperature hyper-parameter [27]. Intuitively, this loss function could also be seen as the negative log-likelihood of a  $(K+1)$ -way non-parameter softmax-based classifier. Different from the traditional softmax based classifier, the above formulation aims at classifying the  $i$ -th image-text pair (i.e.,  $\mathbf{h}_i^x$  and  $\mathbf{h}_i^y$ ) as the corresponding positive key (i.e., the  $i$ -th hash point  $\mathbf{k}_i^+$ ) from the memory bank.

Another challenge of the dictionary is binary-value optimization. Thus, the hashing contrastive loss in Equation (3) enforces the samples to approximate the hash codes of their positive keys and distinguish from the discrete representations of their negative keys. Different from the existing cross-modal hashing methods, our UCCN directly learns the discrete representations without continuous relaxation. However, directly optimizing the discrete memory bank is an NP-hard problem [20], [41], [42]. To make the hashing memory bank be learnable, we define its signed magnitude as  $\{\mathbf{v}_i\}_{i=1}^n$ . The hash keys could be obtained by  $\mathbf{k}_i = \text{sgn}(\mathbf{v}_i)$  correspondingly. Then, a momentum mechanism is used to update the memory bank  $\{\mathbf{v}_i\}_{i=1}^n$  as follows:

$$\mathbf{v}_{i'} = \delta \mathbf{v}_{i'} + (1 - \delta) \frac{\mathbf{h}_i^x + \mathbf{h}_i^y}{2}, \quad (5)$$

where  $\delta \in [0, 1)$  is a momentum coefficient, and  $\mathbf{v}_{i'}$  is the value of  $i'$ -th location of the memory bank for the positive key  $\mathbf{k}_i^+$  which is derived from a sampled multimodal pair  $\{\mathbf{x}_i, \mathbf{y}_i\}$  in the batch  $\{\mathbf{x}_i, \mathbf{y}_i\}_{j=1}^{N_b}$ , where  $N_b$  is the batch size. Note that  $i'$  is the corresponding location in the memory bank for the  $i$ -th pair of a mini-batch.

### 3.3 Cross-Modal Ranking Learning

Besides retrieving from the unified hash dictionary, we also need to bridge the model training to the performance of the downstream task (i.e., cross-modal retrieval). To achieve this goal, the similarity of relevant pairs is enforced to be larger than that of irrelevant cross-modal samples. Specifically, an image query  $\mathbf{h}_i^x$  is first used to retrieve the co-occurred sample  $\mathbf{h}_i^y$  from the text dictionary  $\{\mathbf{h}_i^y\}_{i=1}^n$ . Intuitively, the similarity between the query  $\mathbf{h}_i^x$  and the relevant point  $\mathbf{h}_i^y$  should be larger than the similarities between  $\mathbf{h}_i^x$  and the irrelevant samples  $\{\mathbf{h}_j^y\}_{j \neq i}^n$ . This similarly applies to the text

query  $\mathbf{h}_i^y$  and the relevant image point  $\mathbf{h}_i^x$ . To achieve that, a bidirectional max-margin ranking loss is widely adopted to enforce this constraint in multimodal learning [43], [44]. The max-margin ranking loss is formulated as follows:

$$\mathcal{L}'_r = \mathcal{L}_r^{xy} + \mathcal{L}_r^{yx} \quad (6)$$

where

$$\mathcal{L}_r^{*} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j \neq i}^n \max(0, m + M_{ij}^* - M_{ii}^*), \quad (7)$$

$* \in \{xy, yx\}$ ,  $M_{ij}^{xy} = \langle \mathbf{h}_i^x, \mathbf{h}_j^y \rangle$ ,  $M_{ij}^{yx} = \langle \mathbf{h}_i^y, \mathbf{h}_j^x \rangle$ , and  $m$  is a positive margin value.

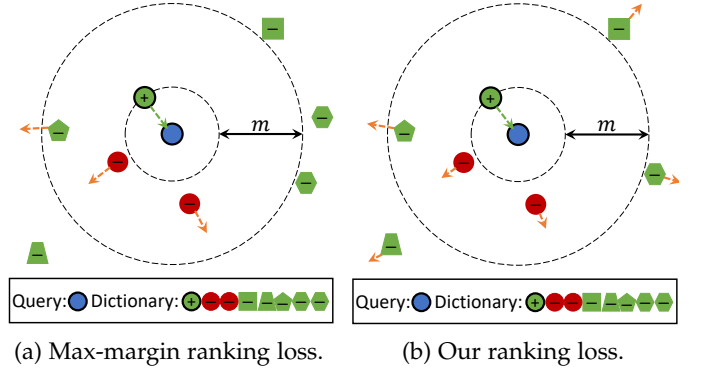


Fig. 4: The major difference between the max-margin ranking loss and our loss is that the former does not utilize the samples outside the margin whereas the latter does. To be specific, for a given query, one aims to retrieve the most relative sample from a given dictionary, where the query and dictionary lie into two modalities and different shapes denote different classes. Due to the existence of FNP, the vanilla max-margin ranking loss probably lead to wrong optimization result, as shown in (a). Instead of only pushing within-marginal between-pairs samples away, our loss simultaneously maximizes the intra-pair similarity while minimizing all inter-pair similarities. Thus, our ranking loss could fully utilize all negative samples so that the influence of FNP is alleviated.

Equation (6) shows that the max-margin loss focuses on optimizing the hard negative pairs whose similarity is not  $m$  smaller than positive pairs (i.e.,  $M_{ii}^* - M_{ij}^* < m$ ), resulting in ignoring the easier ones. In other words, the vanilla triplet loss mainly focuses on harder negative pairs, and its performance heavily depends on the well-established negative samples. In unsupervised setting, however, it is difficult to guarantee the correctness of the negative pairs because the pairwise/co-occurred samples are always used as positive pairs and the other samples are treated as negative. Clearly, such a pair construction strategy will wrongly treat a number of within-class samples as negative, and these FNPs will lead to wrong optimization direction of the vanilla triplet loss. More specifically, the max-margin loss will emphasize separating the FNPs while ignoring the true negative pairs (TNPs) because the former is harder to separate than the latter due to the semantic correlation among FNPs. As a result, it is difficult to obtain encouraging results as verified in Sections 4.3.5 and 4.3.7.



To overcome the FNP challenge, we propose a new learning paradigm termed Cross-modal Ranking Learning (CRL) which uses all negative pairs for optimization. Fig. 4 visually illustrates the difference between the max-margin loss and CRL. As shown, one could see that max-margin ranking loss only focuses on inter-pair samples inside the margin while ignoring the inter-pair samples outside the margin. Without semantic labels, it may focus on the FNPs, and ignore the TNPs as shown in Fig. 4(a). Different from the max-margin loss [19], [24], [33], our CRL could fully utilize all negative samples, and alleviate the influence of FNPs as shown in Fig. 4(b).

To utilize all negative pairs including the ignored ones, we construct the following upper bound of max-margin by simultaneously considering negative samples both inside and outside the margin. First, let

$$S_{ij}^* = \begin{cases} M_{ij}^*, & M_{ii}^* - M_{ij}^* \leq m \\ M_{ij}^* - \xi, & \text{otherwise,} \end{cases} \quad (8)$$

where  $\xi > 0$ . Thus,  $S_{ij}^*$  is never larger than  $M_{ij}^*$ .

**Lemma 1.** For any  $\kappa \in \mathbb{R}^+$ , it holds that

$$\max_{j=1, \dots, n} (S_{ij}^*) \leq \kappa \log \left( \sum_{j=1}^n \exp(S_{ij}^*/\kappa) \right) \quad (9)$$

where  $*$   $\in \{xy, yx\}$ .

*Proof.* To proof the above lemma, we have that

$$\begin{aligned} \max_{j=1, \dots, n} (S_{ij}^*) &= \max_{j=1, \dots, n} \kappa \log \left( (\exp(S_{ij}^*))^{1/\kappa} \right) \\ &= \kappa \log \left( \max_{j=1, \dots, n} (\exp(S_{ij}^*))^{1/\kappa} \right) \\ &\leq \kappa \log \left( \sum_{j=1}^n \exp(S_{ij}^*/\kappa) \right) \end{aligned} \quad (10)$$

**Theorem 1.** For any  $\kappa \in \mathbb{R}^+$ , it holds that

$$\begin{aligned} &\sum_{j=1}^n \max(0, m + S_{ij}^* - S_{ii}^*) \\ &\leq n \left( m + \kappa \log \left( \sum_{j=1}^n \exp(S_{ij}^*/\kappa) \right) - S_{ii}^* \right) \end{aligned} \quad (11)$$

where  $*$   $\in \{xy, yx\}$ .

*Proof.* By Lemma 1, we could easily have that

$$\begin{aligned} &\sum_{j=1}^n \max(0, m + S_{ij}^* - S_{ii}^*) \\ &= |\mathcal{A}_i| m + \sum_{S_{ij} \in \mathcal{A}_i} S_{ij}^* - |\mathcal{A}_i| S_{ii}^* \\ &\leq |\mathcal{A}_i| \left( m + \max_{j=1, \dots, n} (S_{ij}^*) - S_{ii}^* \right) \\ &\leq |\mathcal{A}_i| \left( m + \kappa \log \left( \sum_{j=1}^n \exp(S_{ij}^*/\kappa) \right) - S_{ii}^* \right) \\ &\leq n \left( m + \kappa \log \left( \sum_{j=1}^n \exp(S_{ij}^*/\kappa) \right) - S_{ii}^* \right), \end{aligned} \quad (12)$$

where  $\mathcal{A}_i = \{S_{ij}^* | S_{ii}^* - S_{ij}^* \leq m; i \neq j, j = 1, 2, \dots, n\}$ , and  $|\mathcal{A}_i|$  is the size of  $\mathcal{A}_i$ .  $\square$

Therefore, we could easily obtain the following inequality:

$$\mathcal{L}_r^* \leq \frac{1}{n} \sum_{i=1}^n \left( m + \kappa \log \left( \sum_{j=1}^n \exp(S_{ij}^*/\kappa) \right) - S_{ii}^* \right). \quad (13)$$

Then, we can transform the minimizing optimization of Equation (6) into minimizing its upper bound as follows:

$$\begin{aligned} \mathcal{L}_r &= \frac{1}{n} \sum_{i=1}^n \left( m + \kappa \log \left( \sum_{j=1}^n \exp(S_{ij}^{xy}/\kappa) \right) - S_{ii}^{xy} \right) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left( m + \kappa \log \left( \sum_{j=1}^n \exp(S_{ij}^{yx}/\kappa) \right) - S_{ii}^{yx} \right) \end{aligned} \quad (14)$$

Here,  $m$  is a margin constraint as shown in Fig. 4. By minimizing Equation (14), the cross-modal networks are trained to scatter all inter-pair samples. Different from the traditional max-margin ranking loss  $\mathcal{L}_r'$ , our cross-modal ranking loss  $\mathcal{L}_r$  can simultaneously scatter the top to bottom ranking irrelevant samples instead of only focusing on the top ones. Furthermore, our method pays more attention to the top irrelevant samples (considered as more difficult points) than the bottom ones, thus the top ones can be sufficiently apart from the queries. Simultaneously, the loss also compacts the positive samples (*i.e.*, the relevant image-text pairs) to eliminate the cross-modal discrepancy.

### 3.4 Optimization

The process of learning the optimal hashing functions is conducted by jointly minimizing the contrastive hashing loss  $\mathcal{L}_c$  and cross-modal ranking loss  $\mathcal{L}_r$  as Equation (2). The joint loss is as follows:

$$\mathcal{L} = \beta \mathcal{L}_c + (1 - \beta) \mathcal{L}_r. \quad (15)$$

Our UCCH (Equation (2)) could be iteratively optimize in a batch-by-batch manner. By minimizing  $\mathcal{L}_c$ , our UCCH learns to capture the apparent similarity through instance-level discrimination learning [27], and encodes the multimodal data to binary codes without continuous relaxation. Furthermore, the cross-modal retrieval metric is directly injected into the learning process to bridge the cross-modal gap. The whole model of our UCCH could be optimized by using any one stochastic gradient descent optimization algorithm, like Adam [45]. The optimization process of our UCCH is summarized in Algorithm 1.

## 4 EXPERIMENT STUDY

To verify the effectiveness of our UCCH, we carry out experiments on five widely-used multimodal datasets, *i.e.*, MIRFLICKR-25K [46], IAPR TC-12 [47], NUS-WIDE [48], MS-COCO [49], and Flickr30K [50]. Our method is implemented with PyTorch [51] on a single NVIDIA GEFORCE RTX 2080 Ti GPU.

**Algorithm 1** Optimization process of our UCCH

**Input:** The training image-text pairs  $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ , the length of the hash codes  $L$ , batch size  $N_b$ , balance parameter  $\beta$ , momentum coefficient  $\delta$ , margin constraint parameter  $m$ , number of negative samples  $K$ , and learning rate  $\alpha$ .

- 1: Randomly initialize  $\Theta_x, \Theta_y$ .
- 2: **while** not converge **do**
- 3: Randomly sample  $N_b$  image-text pairs from  $\mathcal{D}$  to construct an image-text mini-batch  $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N_b}$ .
- 4: Randomly sample  $K$  negative keys  $\{\mathbf{k}_j^-\}_{j=1}^K$  and select the corresponding positive key  $\mathbf{k}_i^+ = \text{sgn}(\mathbf{v}_{i'})$  for each pair  $\{\mathbf{x}_i, \mathbf{y}_i\}$  from the hashing memory bank.
- 5: Calculate the representation of each point in the mini-batch by using the corresponding hash function.
- 6: Compute the hashing contrastive loss and cross-modal ranking loss according to Equation (3) and Equation (14) on the mini-batch, respectively.
- 7: Update the parameters of the view-specific hash networks by minimizing  $\mathcal{L}$  in Equation (15) with descending their stochastic gradient:  

$$\Theta_* = \Theta_* - \alpha \left( \beta \frac{\partial \mathcal{L}_c}{\partial \Theta_*} + (1 - \beta) \frac{\partial \mathcal{L}_r}{\partial \Theta_*} \right) \quad (* \in \{x, y\})$$
- 8: Update the corresponding positive keys of the sampled mini-batch  $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N_b}$  in the hashing memory bank through Equation (5):  

$$\mathbf{v}_{i'}^+ = \delta \mathbf{v}_{i'}^+ + (1 - \delta) \frac{\mathbf{h}_i^x + \mathbf{h}_i^y}{2} \quad (i = 1, \dots, N_b)$$
- 9: **end while**

**Output:** Optimized UCCH model.

## 4.1 Datasets

### 4.1.1 MIRFLICKR-25K [46]

It is a widely-used cross-modal dataset for cross-modal hashing retrieval. This dataset consists of 25,000 image-text pairs where each pair contains an image and its corresponding multiple textual tags manually annotated with a multi-label from 24 unique semantic classes. After pruning pairs without class information, 20,015 pairs are totally left for our experiments. For a fair comparison, we exactly follow the data partition strategy of [13] to randomly select 2,000 image-text pairs as query set and the remaining ones are used as retrieval database. For the supervised baselines, we randomly select 5,000 pairs from the retrieval database as their training set. Image and text samples are respectively represented as a 4,096-dimensional vector extracted by the pretrained 19-layer VGGNet [52] and 1,386-dimensional bag-of-words (BoW) vector.

### 4.1.2 IAPR TC-12 [47]

This dataset totally contains 20,000 image-text pairs which are annotated with multi-labels from 255 unique semantic categories. Unlike other datasets, the entire IAPR TC-12 is used for our experiments. The image of each pair is represented as a 4,096-dimensional vector extracted by the pretrained CNN-F [53], each text is represented by a 2,912-dimensional BoW vector. Like MIRFLICKR-25K, we randomly select 2,000 image-text pairs as the query set and the remaining ones are used as retrieval database. Furthermore, we also randomly select 5,000 image-text pairs from the

retrieval database as the training set for the supervised baselines.

### 4.1.3 NUS-WIDE [48]

This dataset consists of 269,498 web images with their textual tags categorized into one or multiple labels from 81 concept categories. In the dataset, 186,557 image-text pairs, which belong to the 10 most frequent classes, are selected for our experiments. We follow the data partition strategy of [22] to randomly select 2,100 image-text pairs as the query set and the left ones are used as retrieval set. Each text point is represented as a 1,000-dimensional BoW vector. The feature of each image sample is a 4,096-dimensional vector extracted by the pretrained 19-layer VGGNet. Moreover, 5,000 image-text pairs are selected from the retrieval set to construct the training set for the supervised baselines.

### 4.1.4 MS-COCO [49]

This dataset totally contains 123,287 images. Each image is described with five annotated sentences with their annotations classified into 80 categories. After removing the pairs without any label information, 122,218 image-text pairs are totally remained for our experiments. Different from other datasets, the text of each pair is represented by a 300-dimensional vector extracted by the pretrained Doc2Vec [54]. Each image is represented as a 4,096-dimensional vector extracted by the pretrained 19-layer VGGNet. We randomly select 5,000 image-text pairs as query set and the remaining ones are used as the retrieval set. Like other datasets, 5,000 image-text pairs are randomly selected as the training set for the supervised approaches.

### 4.1.5 Flickr30K [50]

This dataset consists of 31,000 images with five text annotations for each image. We use the default splits of [55], i.e., the training set includes 29,000 images and 145,000 texts, the validation set contains 1,000 images and 5,000 texts, and the testing set consists of 1,000 images and 5,000 texts. Following [55], each image is represented as a 4,096-dimensional vector extracted from FC7 of the pretrained 19-layer VGGNet. Different four previous datasets, Flickr30K is an unlabeled dataset. Thus, we could only conduct image-text matching on Flickr30K instead of cross-modal retrieval based on semantics, which is one kind of cross-modal retrieval based on instances.

## 4.2 Evaluation Protocol and Baseline

### 4.2.1 Evaluation Protocol

For each dataset, some samples are randomly selected from the total set as the query set and the left ones are used as the retrieval database following [22], [56], [57]. To evaluate the performance of the cross-modal hashing methods, we perform two different cross-modal retrieval tasks: retrieving relevant text points using an image query (Image  $\rightarrow$  Text), and retrieving relevant image points using a text query (Text  $\rightarrow$  Image). The ground-truth relevant neighbors are defined as the cross-modal points which share at least one same semantic category. To evaluate the accuracy of the retrieved results, the widely-used Hamming ranking and hash lookup are used as retrieval protocols in the experiments. The

evaluation metric utilizes the widely-used Mean Average Precision (MAP), which is the mean value of Average Precision (AP) scores for each query, to measure the accuracy scores of the Hamming ranking results. MAP is extensively used to evaluate the performance of cross-modal retrieval since it simultaneously considers both retrieval precision and the ranking of returned results. In addition to MAP, we adopt the precision-recall curves as the hash lookup protocol to visually evaluate the performance of cross-modal retrieval. Note that all MAP scores are computed on all returned retrieval results in the experiments (*i.e.*, MAP@ALL). Besides the comparison under the above two category-level metrics, we adopt Recall@K (R@K, higher is better) for different values of K to measure the performance for instance-level image-text matching following [55]. In brief, R@K is the percentage of tested queries for which at least one correct item is among the top K ranking results [55].

#### 4.2.2 Baselines

In our experiments, 13 state-of-the-art cross-modal hashing methods are used as baselines, including four supervised cross-modal hashing methods (DLFH [23], MTFH [16], FOMH [58], DCH [34]), and nine unsupervised approaches (CVH [20], LSSH [59], CMFH [60], FSH [18], UGACH [61], DJSRH [62], JDSH [63], UCH [13], and DGCPN [64]). All of these methods are shallow cross-modal hashing models except for UGACH, DJSRH, JDSH, UCH, and DGCPN which are seven recently proposed deep hashing methods. For a fair comparison, all methods use the same features to learn the hash codes and the extractors (or backbones) are not fine-tuned for deep methods during training. We randomly sample 2000 instances from the retrieval databases as the validation set. The hyper-parameters of other methods are adopted as the default parameters given by the authors. For our UCCH, we utilize validation sets to choose the hyper-parameter  $\beta$ . Other hyper-parameters are empirically set as fixed values for all experiments, *i.e.*,  $\delta = 0.4$ ,  $\tau = 0.9$ ,  $K = 4096$ ,  $m = 0.2$ , and  $\alpha = 0.0001$ . For Flickr30K, we transplant DJSRH [62], JDSH [63], and our UCCH on the framework of VSE++ [55] for a fair comparison.

### 4.3 Experimental Analysis

#### 4.3.1 Hamming Ranking

Two cross-modal retrieval tasks (*i.e.*, Image  $\rightarrow$  Text, and Text  $\rightarrow$  Image) are conducted on five widely-used benchmark multimodal datasets to evaluate the performance of our UCCH and other baselines. The MAP@ALL/Recall@K scores of these tasks are reported in Tables 1 and 2, *i.e.*, Table 1 is the result on the MIRFLICKR-25K and IAPR TC-12 datasets, Table 2 is on the NUS-WIDE and MS-COCO datasets, and Table 4 is on the Flickr30K dataset. From the experimental results shown in the tables, one could observe that our UCCH outperforms all the other baselines for the different code lengths (*i.e.*, 16, 32, 64, and 128). From the experimental results, we can draw the following observations:

- 1) DNN-based cross-modal hashing methods (UGACH, UCH, and our UCCH) outperform

most other shallow baselines, indicating that highly-level nonlinearity of DNN can improve the performance of cross-modal retrieval.

- 2) Although the supervised methods can achieve promising performance with sufficient labeled data, they cannot outperform most of unsupervised methods with insufficient labeled data, which indicates that unsupervised approaches have a lot of potential for amounts of unlabeled data. The supervised methods are too dependent on costly labeled data, but the unsupervised ones can handle this problem. Thus, the unsupervised cross-modal hashing methods have great advantages for large-scaled multimodal data.
- 3) The cross-modal hashing methods (DLFH and our UCCH), which directly learns the discrete representations without any continuous relaxation, outperforms their relaxation-based continuous counterparts (*i.e.*, supervised methods for DLFH, and unsupervised approaches for our UCCH). Thus, hashing learning without continuous relaxation can improve the retrieval performance.
- 4) Instance-level image-text matching is much more sensitive to hashing than category-level cross-modal retrieval, probably because instance-level retrieval is much more complicated than category-level retrieval. Even so, our method could still achieve a competitive hashing performance, which indicates that our method could capture the instance-level discrimination for cross-modal hashing well.

#### 4.3.2 Hash Lookup

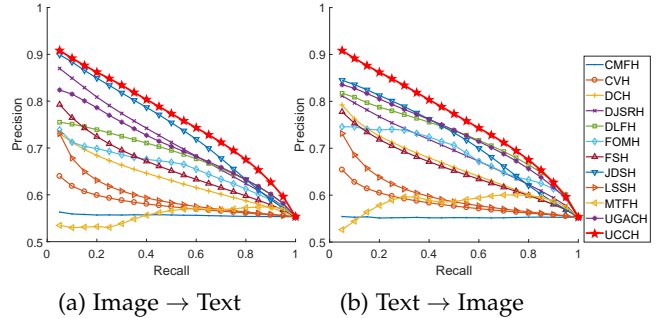


Fig. 5: The precision-recall curves on the MIRFLICKR-25K dataset. The code length is 128.

Besides the Hamming ranking, the precision and recall are calculated by the returned results with the Hamming distance following [9], [13], [61]. The precision-recall curves with code length 128 are drawn to evaluate the performance of the cross-modal hashing methods on the MIRFLICKR-25K, IAPR TC-12, NUS-WIDE, and MS-COCO datasets as shown in Figs. 5–8. One could see that the precision-recall evaluations in these figures are consistent with the MAP scores of Hamming ranking, where our UCCH is superior to all compared cross-modal hashing methods. In addition, the proposed UCCH also outperforms other approaches for other cases with other code length (*i.e.*, 16, 32, 64), whose curves are omitted due to space limitation. In summary, our UCCH could achieve the best performance for cross-modal



TABLE 1: Performance comparison in terms of MAP scores on the MIRFLICKR-25K and IAPR TC-12 datasets. The highest score is shown in **boldface**.

Method	MIRFLICKR-25K								IAPR TC-12							
	Image $\rightarrow$ Text				Text $\rightarrow$ Image				Image $\rightarrow$ Text				Text $\rightarrow$ Image			
	16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
CVH [20]	0.620	0.608	0.594	0.583	0.629	0.615	0.599	0.587	0.392	0.378	0.366	0.353	0.398	0.384	0.372	0.360
LSSH [59]	0.597	0.609	0.606	0.605	0.602	0.598	0.598	0.597	0.372	0.386	0.396	0.404	0.367	0.380	0.392	0.401
CMFH [60]	0.557	0.557	0.556	0.557	0.553	0.553	0.553	0.553	0.312	0.314	0.314	0.315	0.306	0.306	0.306	0.306
FSH [18]	0.581	0.612	0.635	0.662	0.576	0.607	0.635	0.660	0.377	0.392	0.417	0.445	0.383	0.399	0.425	0.451
DLFH [23]	0.638	0.658	0.677	0.684	0.675	0.700	0.718	0.725	0.342	0.358	0.374	0.395	0.358	0.380	0.403	0.434
MTFH [16]	0.507	0.512	0.558	0.554	0.514	0.524	0.518	0.581	0.277	0.324	0.303	0.311	0.294	0.337	0.269	0.297
FOMH [58]	0.575	0.640	0.691	0.659	0.585	0.648	0.719	0.688	0.312	0.316	0.317	0.350	0.311	0.315	0.322	0.373
DCH [34]	0.596	0.602	0.626	0.636	0.612	0.623	0.653	0.665	0.336	0.336	0.344	0.352	0.350	0.358	0.374	0.391
UGACH [61]	0.685	0.693	0.704	0.702	0.673	0.676	0.686	0.690	0.462	0.467	0.469	0.480	0.447	0.463	0.468	0.463
DJSRH [62]	0.652	0.697	0.700	0.716	0.662	0.691	0.683	0.695	0.409	0.412	0.470	0.480	0.418	0.436	0.467	0.478
JDSH [63]	0.724	0.734	0.741	0.745	0.710	0.720	0.733	0.720	0.449	0.472	0.478	0.484	0.447	0.477	0.473	0.486
DGCPN [64]	0.711	0.723	0.737	0.748	0.695	0.707	0.725	0.731	0.465	0.485	0.486	0.495	0.467	0.488	0.491	0.497
UCH [13]	0.654	0.669	0.679	/	0.661	0.667	0.668	/	0.447	0.471	0.485	/	0.446	0.469	0.488	/
UCCH	<b>0.739</b>	<b>0.744</b>	<b>0.754</b>	<b>0.760</b>	<b>0.725</b>	<b>0.725</b>	<b>0.743</b>	<b>0.747</b>	<b>0.478</b>	<b>0.491</b>	<b>0.503</b>	<b>0.508</b>	<b>0.474</b>	<b>0.488</b>	<b>0.503</b>	<b>0.508</b>

TABLE 2: Performance comparison in terms of MAP scores on the NUS-WIDE and MS-COCO datasets. The highest score is shown in **boldface**.

Method	NUS-WIDE								MS-COCO							
	Image $\rightarrow$ Text				Text $\rightarrow$ Image				Image $\rightarrow$ Text				Text $\rightarrow$ Image			
	16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
CVH [20]	0.487	0.495	0.456	0.419	0.470	0.475	0.444	0.412	0.503	0.504	0.471	0.425	0.506	0.508	0.476	0.429
LSSH [59]	0.442	0.457	0.450	0.451	0.473	0.482	0.471	0.457	0.484	0.525	0.542	0.551	0.490	0.522	0.547	0.560
CMFH [60]	0.339	0.338	0.343	0.339	0.306	0.306	0.306	0.306	0.366	0.369	0.370	0.365	0.346	0.346	0.346	0.346
FSH [18]	0.557	0.565	0.598	0.635	0.569	0.604	0.651	0.666	0.539	0.549	0.576	0.587	0.537	0.524	0.564	0.573
DLFH [23]	0.385	0.399	0.443	0.445	0.421	0.421	0.462	0.474	0.522	0.580	0.614	0.631	0.444	0.489	0.513	0.534
MTFH [16]	0.297	0.297	0.272	0.328	0.353	0.314	0.399	0.410	0.399	0.293	0.295	0.395	0.335	0.374	0.300	0.334
FOMH [58]	0.305	0.305	0.306	0.314	0.302	0.304	0.300	0.306	0.378	0.514	0.571	0.601	0.368	0.484	0.559	0.595
DCH [34]	0.392	0.422	0.430	0.436	0.379	0.432	0.444	0.459	0.422	0.420	0.446	0.468	0.421	0.428	0.454	0.471
UGACH [61]	0.613	0.623	0.628	0.631	0.603	0.614	0.640	0.641	0.553	0.599	0.598	0.615	0.581	0.605	0.629	0.635
DJSRH [62]	0.502	0.538	0.527	0.556	0.465	0.532	0.538	0.545	0.501	0.563	0.595	0.615	0.494	0.569	0.604	0.622
JDSH [63]	0.647	0.656	0.679	0.680	0.649	0.669	0.689	0.699	0.579	0.628	0.647	0.662	0.578	0.634	0.659	0.672
DGCPN [64]	0.610	0.614	0.635	0.641	0.617	0.621	0.642	0.647	0.552	0.590	0.602	0.596	0.564	0.590	0.597	0.597
UCH [13]	/	/	/	/	/	/	/	/	0.521	0.534	0.547	/	0.499	0.519	0.545	/
UCCH	<b>0.698</b>	<b>0.708</b>	<b>0.737</b>	<b>0.742</b>	<b>0.701</b>	<b>0.724</b>	<b>0.745</b>	<b>0.750</b>	<b>0.605</b>	<b>0.645</b>	<b>0.655</b>	<b>0.665</b>	<b>0.610</b>	<b>0.655</b>	<b>0.666</b>	<b>0.677</b>

TABLE 3: Ablation study on different datasets. The highest score is shown in **boldface**.

Dataset	Method	Image $\rightarrow$ Text				Text $\rightarrow$ Image			
		16	32	64	128	16	32	64	128
IAPR TC-12	UCCH (with $\mathcal{L}_c$ only)	0.457	0.469	0.478	0.482	0.447	0.469	0.483	0.486
	UCCH (with $\mathcal{L}'_{r,m=0.1}$ only)	0.410	0.426	0.432	0.438	0.421	0.434	0.461	0.460
	UCCH (with $\mathcal{L}'_{r,m=0.5}$ only)	0.423	0.446	0.463	0.470	0.434	0.450	0.471	0.479
	UCCH (with $\mathcal{L}'_{r,m=0.9}$ only)	0.444	0.460	0.472	0.480	0.450	0.472	0.469	0.476
	UCCH (with $\mathcal{L}_r$ only)	0.461	0.482	0.496	0.495	0.457	0.476	0.492	0.488
	Full UCCH	<b>0.478</b>	<b>0.491</b>	<b>0.503</b>	<b>0.508</b>	<b>0.474</b>	<b>0.488</b>	<b>0.503</b>	<b>0.508</b>
MS-COCO	UCCH (with $\mathcal{L}_c$ only)	0.577	0.605	0.621	0.624	0.579	0.610	0.626	0.627
	UCCH (with $\mathcal{L}'_{r,m=0.1}$ only)	0.495	0.512	0.548	0.555	0.483	0.503	0.534	0.549
	UCCH (with $\mathcal{L}'_{r,m=0.5}$ only)	0.499	0.525	0.554	0.579	0.498	0.527	0.546	0.566
	UCCH (with $\mathcal{L}'_{r,m=0.9}$ only)	0.529	0.535	0.554	0.558	0.525	0.545	0.546	0.560
	UCCH (with $\mathcal{L}_r$ only)	0.563	0.574	0.599	0.602	0.563	0.576	0.606	0.609
	Full UCCH	<b>0.605</b>	<b>0.645</b>	<b>0.655</b>	<b>0.665</b>	<b>0.610</b>	<b>0.655</b>	<b>0.666</b>	<b>0.677</b>

hashing retrieval comparing with these cross-modal hashing approaches.

#### 4.3.3 Sensitivity to Parameters

To investigate the impact of the hyper-parameter  $\beta$ , Fig. 9 plots the MAP scores of cross-modal retrieval versus different  $\beta$  on the IAPR TC-12 and MS-COCO datasets. From the figure, one could see that both contrastive hashing loss ( $\mathcal{L}_c$ ) and cross-modal ranking loss ( $\mathcal{L}_r$ ) contribute to exploiting the discrimination from the multimodal data. More ablation

evaluations could be found in Section 4.3.5. From the experimental results, one could find that the proposed approach is robust to the choice of the parameters. Note that, the values are determined by the achieved retrieval performance on the corresponding validation sets in the other evaluation parts.

#### 4.3.4 Convergence Analysis

Fig. 10 plot the convergence curves of our UCCH on the IAPR TC-12 and MS-COCO datasets, where x-axis indicates the values of the loss function (*i.e.*,  $\mathcal{L}$ ) and y-axis indicates

TABLE 4: Performance comparison in terms of Recall@K scores on Flickr30K. The highest score is shown in **boldface**.

Bit	Method	Image $\rightarrow$ Text			Text $\rightarrow$ Image		
		R@1	R@5	R@10	R@1	R@5	R@10
64	VSE++ [55]	10.7	28.0	39.2	8.3	25.4	37.1
	DJSRH [62]	3.6	14.4	22.1	3.4	11.6	18.5
	JDSH [63]	10.0	28.6	39.3	8.0	23.6	34.5
	UCCH	<b>14.5</b>	<b>37.6</b>	<b>50.8</b>	<b>10.9</b>	<b>32.3</b>	<b>44.0</b>
128	VSE++ [55]	11.3	31.1	42.6	9.2	27.7	40.4
	DJSRH [62]	7.7	27.2	37.8	5.9	19.9	30.3
	JDSH [63]	10.7	30.0	42.5	8.2	25.6	37.3
	UCCH	<b>17.9</b>	<b>44.9</b>	<b>55.4</b>	<b>14.0</b>	<b>37.0</b>	<b>50.1</b>
512	VSE++ [55]	13.5	34.7	48.2	10.8	31.1	43.6
	DJSRH [62]	17.9	43.5	56.3	13.3	36.3	48.9
	JDSH [63]	13.6	35.6	49.4	9.8	29.1	42.6
	UCCH	<b>22.8</b>	<b>48.1</b>	<b>61.0</b>	<b>16.9</b>	<b>41.8</b>	<b>54.9</b>

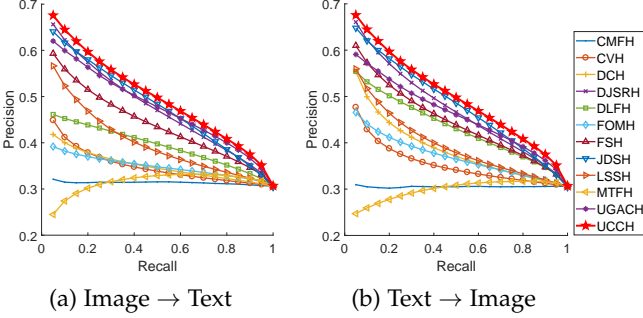


Fig. 6: The precision-recall curves on the IAPR TC-12 dataset. The code length is 128.

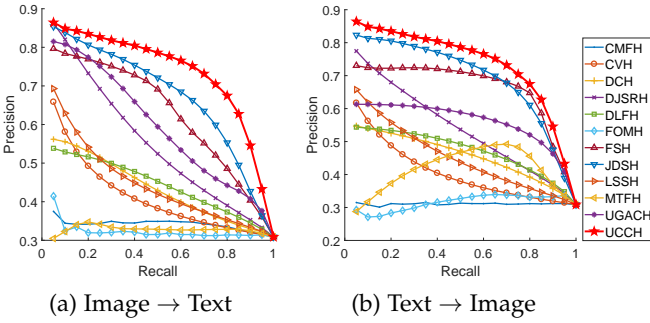


Fig. 7: The precision-recall curves on the NUS-WIDE dataset. The code length is 128.

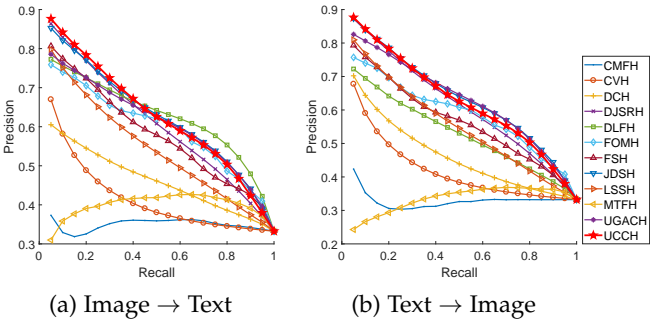


Fig. 8: The precision-recall curves on the MS-COCO dataset. The code length is 128.

different number of epochs. From the figures, one could see that our UCCH quickly converges from the 50-th to 100-th epochs, and the loss remarkably decreases in the first 20 epochs. Thus, the maximum epoch is set as 20 for our UCCH

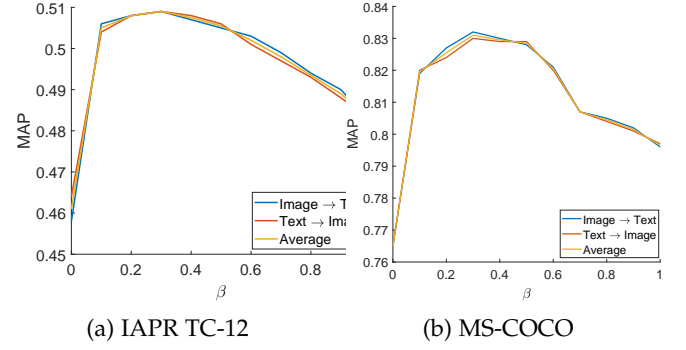


Fig. 9: Cross-modal retrieval performance of our UCCH in terms of MAP scores versus different values of  $\beta$  on the validation sets of the IAPR TC-12 and MS-COCO datasets, respectively. The code length is 128.

on all datasets.

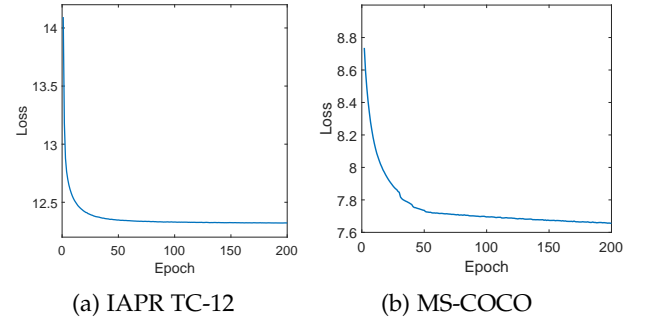


Fig. 10: Convergence curves of our UCCH on the validation sets of IAPR TC-12 and MS-COCO. The code length is 128.

#### 4.3.5 Ablation Study

In this section, we investigate the contributions of different components (*i.e.*,  $\mathcal{L}_c$  and  $\mathcal{L}_r$ ) to the cross-modal hashing retrieval. In order to completely evaluate the performance of each component, we compare our UCCH with its three variations on IAPR TC-12 and MS-COCO datasets, *i.e.*, UCCH with  $\mathcal{L}_c$  only, UCCH with  $\mathcal{L}'_r$  only, and UCCH with  $\mathcal{L}_r$  only. All variants are trained with the same setting as our UCCH for a fair comparison. The experimental results are shown in Table 3. From the table, one could see that the performance of UCCH without  $\mathcal{L}_c$  or  $\mathcal{L}_r$  are worse than our UCCH on the two databases. Thus, both of these two components contribute to the retrieval performance, and the mutual cooperation of  $\mathcal{L}_c$  and  $\mathcal{L}_r$  could improve the retrieval performance. In the table, we also could see that our proposed cross-modal ranking loss  $\mathcal{L}_r$  (*i.e.*, Equation (14)) is efficient to improve the traditional max-margin ranking loss  $\mathcal{L}'_r$  (*i.e.*, Equation (6)), which demonstrates the effectiveness of considering all samples. Moreover, from the comparisons among  $\mathcal{L}'_{r,m=0.1}$ ,  $\mathcal{L}'_{r,m=0.5}$ , and  $\mathcal{L}'_{r,m=0.9}$ , one could see that more negative pairs are used, the performance becomes better. This verifies our claim on the max-margin loss, *i.e.*, it will ignore the TNPs. Besides, we also illustrate the MAP curves to show the performance of different variations in Fig. 11. From the results, one could see that CRL is more stable than the max-margin loss thanks to the use of all negative pairs.

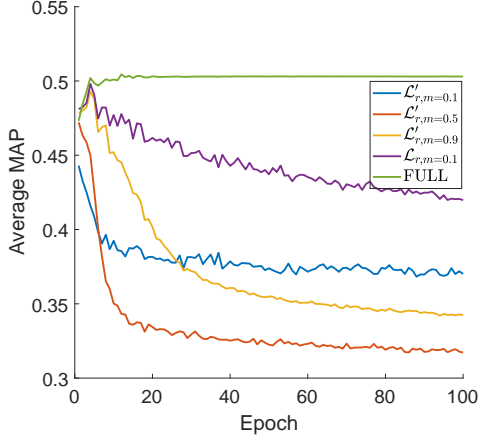


Fig. 11: Different epochs vs. average MAP scores of cross-modal retrieval with different variants on the validation set of IAPR TC-12. The code length is 128.

#### 4.3.6 Effectiveness of Momentum-based Binarization

In this section, we compare the proposed CCH with contrastive learning without hashing (CL w/o Hashing) [65] to investigate the effectiveness of our binarization mechanism. Fig. 12 shows their average MAP scores of cross-modal retrieval with increasing epochs on the validation set of IAPR TC-12. From the results, one could see that CL performs unstable for cross-modal hashing retrieval. With the help of our binarization strategy, CCH achieves better and more stable performance. Hence, one could conclude that it is nontrivial to develop cross-modal hashing based on CL.

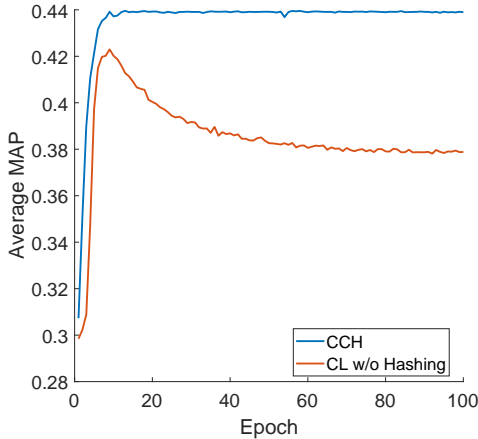


Fig. 12: Performance evolution of CCH and CL w/o Hashing in terms of average MAP scores for cross-modal retrieval on the validation set of IAPR TC-12.

#### 4.3.7 Analysis on False-Negative Pairs

To further investigate the influence of FNPs during training, we take the max-margin loss to train the DNN model with different margin values (i.e.,  $\mathcal{L}'_{r,m=0.1}$ ,  $\mathcal{L}'_{r,m=0.5}$ ) on IAPR TC-12. Accordingly, Fig. 13 demonstrates the evolution curves of FNPs and Fig. 13(a) shows the average number of valid negative pairs for all batches of each epoch. From the figures, one could see that a number of true- and false-negative samples will be pushed outside the margin. For

the lower margins, almost all negatives samples would be pushed out of the margin, e.g.,  $m = 0.1$  and  $0.5$ . In other words, there are almost no negative pairs participating in further training. Furthermore, Fig. 13(b) illustrates that the FNP rate increases in the first few epochs and quickly descends with further training for lower margins (i.e.,  $m = 0.1$  and  $0.5$ ). This could attribute to that TNPs are easier to separate than FNPs. For a larger margin (i.e.,  $m = 0.9$ ), more negative samples would be remained inside the margin as show in Fig. 13(a). Fig. 13(b) shows that in the first several epochs the model will use a number of negative samples. As a result, TNPs will be rapidly pushed out the margin, while the FNPs will dominate the negative pairs and thus resulting in performance degradation as shown in Fig. 12. In contrast, our method will not encounter this problem and thus enjoy better performance as shown in Table 3.

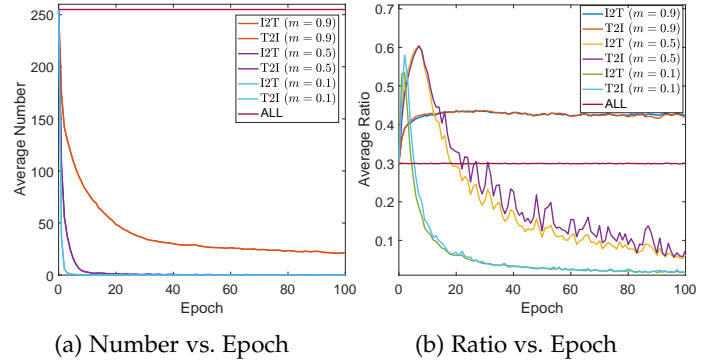


Fig. 13: False-negative samples analysis on IAPR TC-12. With different margin values ( $m = 0.1, 0.5, 0.9$ ), (a) shows the number of valid negative pairs vs. different epochs. (b) shows the ratio of FNPs vs. different epochs. The negative pairs could be obtained from image query text (I2T) and text query image (T2I), respectively. “ALL” denotes all negative pairs.

#### 4.3.8 Efficiency Comparison with State-of-the-art Methods

In this section, we evaluate the efficiency of the proposed method comparing with some state-of-the-art unsupervised cross-modal hashing approaches on the MIRFLICKR-25K dataset. The testing platform is with an Intel i9-10900X CPU@3.70GHz and a GeForce RTX 2080Ti GPU. From Table 5, we could see that our DNN-based UCCH could need more training time comparing with the shallow approaches (e.g., CVH, FSH, CMFH, etc), which is a common phenomenon in deep methods since the multi-layer neural network iteration optimization. However, the inference time is much less than the training time and comparable to the shallow methods. Furthermore, comparing with a DNN-based cross-modal hashing method (i.e., UGACH), we need much less time in both training and inference stages. Furthermore, UCCH with contrastive learning ( $\mathcal{L}_c$ ) will cost more time and memory than UCCH without  $\mathcal{L}_c$  in the training stage, but it is in the acceptable range (about 0.67s per epoch and 9.64% more memory cost). In addition, they have the same inference efficiency since there is no contrastive learning in the inference stage. For a fair comparison, the shallow methods use the default hyper-parameters provided by the authors. For the DNN-based methods, the

maximal training epochs are set as 20. Our UCCH could approach convergence within 20 epochs as shown in Fig. 10, but UGACH needs more epochs. Compared to the graph-based method UGACH [61], our method could remarkably reduce the training time and memory cost, i.e., decreasing training time from >12h to 78.18s and 69.62% less memory consumption.

TABLE 5: Efficiency Comparison on MIRFLICKR-25K with 128 code length. The inference time is the total time cost to hash the whole dataset excluding the retrieval process. The training time/memory is the time/memory cost to complete the training of the corresponding method on the training set.

Method	Inference Time	Training Time	Training Memory
CVH [20]	0.15s	12.61s	3.11G
FSH [18]	0.28s	78.03s	3.80G
CMFH [60]	0.23s	5.76s	5.05G
LSSH [59]	7.78s	180.99s	5.68G
UGACH [61]	26.59s	>12h	14.22G
UCCH w/o $\mathcal{L}_c$	0.41s	64.75s	3.94G
UCCH	0.41s	78.18s	4.32G

## 5 CONCLUSION

In this paper, we propose a novel cross-modal hashing approach, termed Unsupervised Contrastive Cross-modal Hashing (UCCH), which projects different modalities into a common Hamming space. UCCH consists of two task-specific learning parts, namely, Contrastive Cross-modal Hashing (CCH) and Cross-modal Ranking Learning (CRL). On the one hand, CH enforces different modalities to fit the unified binary representations with a novel momentum-based binarization optimizer. Thanks to the optimizer, CCH could endow contrastive learning with unsupervised cross-modal hashing. On the other hand, CRL exploits the discrimination from all instead of the hardest negative pairs, which will alleviate the influence of FNPs to facilitate cross-modal retrieval. Extensive experimental results on five widely-used benchmark datasets and the comprehensive analysis have demonstrated the effectiveness and efficiency of the proposed method comparing with 13 state-of-the-art methods. In the future, we plan to investigate how to further improve the performance of our method by utilizing a few labeled data.

## REFERENCES

- [1] Zongsheng Yue, Hongwei Yong, Deyu Meng, Qian Zhao, Yee Leung, and Lei Zhang. Robust multiview subspace learning with nonindependently and nonidentically distributed complex noise. *IEEE transactions on neural networks and learning systems*, 31(4):1070–1083, 2019.
- [2] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. COMIC: Multi-view clustering without parameter selection. In *International Conference on Machine Learning*, pages 5092–5101, 2019.
- [3] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR’19*, pages 635–644, New York, NY, USA, 2019. ACM.
- [4] Fan Ma, Deyu Meng, Xuanyi Dong, and Yi Yang. Self-paced multi-view co-training. *Journal of Machine Learning Research*, 2020.
- [5] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 785–796, 2013.
- [6] Yuming Shen, Li Liu, Ling Shao, and Jingkuan Song. Deep binaries: Encoding semantic-rich cues for efficient textual-visual cross retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4097–4106, 2017.
- [7] Erkun Yang, Tongliang Liu, Cheng Deng, and Dacheng Tao. Adversarial examples for hamming space search. *IEEE transactions on cybernetics*, 2018.
- [8] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing*, 27(8):3893–3903, 2018.
- [9] Peng Hu, Xu Wang, Liangli Zhen, and Dezhong Peng. Separated variational hashing networks for cross-modal retrieval. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1721–1729, 2019.
- [10] Jingkuan Song, Xiaosu Zhu, Lianli Gao, Xin-Shun Xu, Wu Liu, and Heng Tao Shen. Deep recurrent quantization for generating sequential binary codes. In *28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.
- [11] Fumin Shen, Yadong Mu, Yang Yang, Wei Liu, Li Liu, Jingkuan Song, and Heng Tao Shen. Classification by retrieval: Binarizing data and classifiers. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 595–604, 2017.
- [12] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4242–4251, 2018.
- [13] Chao Li, Cheng Deng, Lei Wang, De Xie, and Xianglong Liu. Coupled cyclegan: Unsupervised hashing network for cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 176–183, 2019.
- [14] Xiang Zhou, Fumin Shen, Li Liu, Wei Liu, Liqiang Nie, Yang Yang, and Heng Tao Shen. Graph convolutional network hashing. *IEEE Transactions on Cybernetics*, 50(4):1460–1472, April 2020.
- [15] Kai Li, Guo-Jun Qi, Jun Ye, and Kien A Hua. Linear subspace ranking hashing for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (9):1825–1838, 2017.
- [16] Xin Liu, Zhikai Hu, Haibin Ling, and Yiu-ming Cheung. Mtfh: A matrix tri-factorization hashing framework for efficient cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [17] Peng Hu, Hongyuan Zhu, Xi Peng, and Jie Lin. Semi-supervised multi-modal learning with balanced spectral decomposition. In *34th AAAI Conference on Artificial Intelligence (AAAI)*, New York, NY, USA, Feb., 2020.
- [18] Hong Liu, Rongrong Ji, Yongjian Wu, Feiyue Huang, and Baoshang Zhang. Cross-modality binary code learning via fusion similarity hashing. In *Computer Vision and Pattern Recognition*, pages 6345–6353, 2017.
- [19] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. In *AAAI Conference on Artificial Intelligence*, 2018.
- [20] Shaishav Kumar and Raghavendra Udapa. Learning hash functions for cross-view similarity search. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.
- [21] Xi Peng, Shijie Xiao, Jiashi Feng, Wei-Yun Yau, and Zhang Yi. Deep subspace clustering with sparsity prior. In *International Joint Conference on Artificial Intelligence*, pages 1925–1931, New York, NY, USA, 9-15 July 2016.
- [22] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3232–3240, 2017.
- [23] Qing-Yuan Jiang and Wu-Jun Li. Discrete latent factor model for cross-modal hashing. *IEEE Transactions on Image Processing*, 28(7):3490–3501, 2019.
- [24] Shuli Cheng, Liejun Wang, and Anyu Du. Deep semantic-preserving reconstruction hashing for unsupervised cross-modal retrieval. *Entropy*, 22(11):1266, 2020.
- [25] Mingyong Li and Hongya Wang. Unsupervised deep cross-modal hashing by knowledge distillation for large-scale cross-modal retrieval. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 183–191, 2021.
- [26] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, volume 2, pages 1735–1742. IEEE, 2006.

- [27] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.
- [29] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Semantics-preserving hashing for cross-view retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3864–3872, 2015.
- [30] Zijia Lin, Guiguang Ding, Jungong Han, and Jianmin Wang. Cross-view retrieval via probability-based semantics-preserving hashing. *IEEE transactions on cybernetics*, 47(12):4342–4355, 2017.
- [31] Devraj Mandal, Kunal N Chaudhury, and Soma Biswas. Generalized semantic preserving hashing for n-label cross-modal retrieval. In *Computer Vision and Pattern Recognition*, pages 2633–2641. IEEE, 2017.
- [32] Kun Ding, Bin Fan, Chunlei Huo, Shiming Xiang, and Chunhong Pan. Cross-modal hashing via rank-order preserving. *IEEE Transactions on Multimedia*, 19(3):571–585, 2017.
- [33] Xuanwu Liu, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Yazhou Ren, and Maozu Guo. Ranking-based deep cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4400–4407, 2019.
- [34] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing*, 26(5):2494–2507, 2017.
- [35] Guiguang Ding, Yuchen Guo, and Jile Zhou. Collective matrix factorization hashing for multimodal data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2075–2082, 2014.
- [36] Yang Li, Yapeng Wang, Zhuang Miao, Jiabao Wang, and Rui Zhang. Contrastive self-supervised hashing with dual pseudo agreement. *IEEE Access*, 8:165034–165043, 2020.
- [37] Young Kyun Jang and Nam Ik Cho. Self-supervised product quantization for deep unsupervised image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12085–12094, 2021.
- [38] Zexuan Qiu, Qinliang Su, Zijong Ou, Jianxing Yu, and Changyou Chen. Unsupervised hashing with contrastive information bottleneck. *arXiv preprint arXiv:2105.06138*, 2021.
- [39] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.
- [40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [41] Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen. Supervised discrete hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 37–45, 2015.
- [42] Jiwen Lu, Venice Erin Liong, Xiuzhuang Zhou, and Jie Zhou. Learning compact binary face descriptor for face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):2041–2056, 2015.
- [43] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.
- [44] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017.
- [45] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–13, San Diego, USA, 2015.
- [46] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *ACM International Conference on Multimedia Information Retrieval*, pages 39–43. ACM, 2008.
- [47] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villaseñor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer vision and image understanding*, 114(4):419–428, 2010.
- [48] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *International Conference on Multimedia*, pages 251–260. ACM, 2010.
- [49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [50] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.
- [51] Adam Paszke, Sam Gross, Francisco Massa, and et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 32, pages 8024–8035. Curran Associates, Inc., 2019.
- [52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.
- [53] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.
- [54] Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Workshop on Representation Learning for NLP*, pages 78–86. Association for Computational Linguistics, 2016.
- [55] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. 2018.
- [56] Rong-Cheng Tu, Xian-Ling Mao, Bing Ma, Yong Hu, Tan Yan, Wei Wei, and Heyan Huang. Deep cross-modal hashing with hashing functions and unified hash codes jointly learning. *IEEE Transactions on Knowledge and Data Engineering*, 2020.
- [57] Xinzhi Wang, Xitao Zou, Erwin M Bakker, and Song Wu. Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval. *Neurocomputing*, 400:255–271, 2020.
- [58] Xu Lu, Lei Zhu, Zhiyong Cheng, Jingjing Li, Xiushan Nie, and Huaxiang Zhang. Flexible online multi-modal hashing for large-scale multimedia retrieval. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1129–1137, 2019.
- [59] Jile Zhou, Guiguang Ding, and Yuchen Guo. Latent semantic sparse hashing for cross-modal similarity search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 415–424, 2014.
- [60] Guiguang Ding, Yuchen Guo, Jile Zhou, and Yue Gao. Large-scale cross-modality search via collective matrix factorization hashing. *IEEE Transactions on Image Processing*, 25(11):5427–5440, 2016.
- [61] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [62] Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3027–3035, 2019.
- [63] Song Liu, Shengsheng Qian, Yang Guan, Jiawei Zhan, and Long Ying. Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1379–1388, 2020.
- [64] Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4626–4634, 2021.
- [65] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI* 16, pages 776–794. Springer, 2020.