Multimodal Mutual Information Maximization: A Novel Approach for Unsupervised Deep Cross-Modal Hashing

Tuan Hoang, Thanh-Toan Do, Tam V. Nguyen, Senior Member, IEEE, and Ngai-Man Cheung, Senior Member, IEEE

Abstract—In this article, we adopt the maximizing mutual information (MI) approach to tackle the problem of unsupervised learning of binary hash codes for efficient cross-modal retrieval. We proposed a novel method, dubbed cross-modal info-max hashing (CMIMH). First, to learn informative representations that can preserve both intramodal and intermodal similarities, we leverage the recent advances in estimating variational lower bounds of MI to maximizing the MI between the binary representations and input features and between binary representations of different modalities. By jointly maximizing these MIs under the assumption that the binary representations are modeled by multivariate Bernoulli distributions, we can learn binary representations, which can preserve both intramodal and intermodal similarities, effectively in a mini-batch manner with gradient descent. Furthermore, we find out that trying to minimize the modality gap by learning similar binary representations for the same instance from different modalities could result in less informative representations. Hence, balancing between reducing the modality gap and losing modality-private information is important for the cross-modal retrieval tasks. Quantitative evaluations on standard benchmark datasets demonstrate that the proposed method consistently outperforms other state-of-the-art cross-modal retrieval methods.

Index Terms—Cross-modal retrieval, multi-modal, mutual information (MI), representation learning, unsupervised hashing, variational information maximization.

I. INTRODUCTION

The last few years have witnessed an exponential surge in the amount of information available online in heterogeneous modalities, e.g., images, tags, text documents, videos, and subtitles. Thus, it is desirable to have a single efficient system that can facilitate large-scale multimedia searches.

Manuscript received 7 September 2020; revised 11 April 2021 and 6 September 2021; accepted 5 December 2021. Date of publication 4 January 2022; date of current version 1 September 2023. This work was supported in part by the Singapore University of Technology and Design (SUTD) under Project PIE-SGP-AI-2018-01 and in part by the National Research Foundation Singapore through its AI Singapore Program under Award A1SG-100E2018-005. (Corresponding author: Tuan Hoang.)

Tuan Hoang and Ngai-Man Cheung are with the Information Systems Technology and Design (ISTD), Singapore University of Technology and Design (SUTD), Singapore 487372 (e-mail: hnanhtuan.91@gmail.com; ngaiman_cheung@sutd.edu.sg).

Thanh-Toan Do is with the Department of Data Science and AI, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia (e-mail: toan.do@monash.edu).

Tam V. Nguyen is with the Department of Computer Science, University of Dayton, Dayton, OH 45469 USA (e-mail: tamnguyen@udayton.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3135420.

Digital Object Identifier 10.1109/TNNLS.2021.3135420

In general, this system should support both single and cross-modality searches, i.e., the system returns a set of semantically relevant results of all modalities given a query in any modality. In addition, to be used in large-scale applications, the system should have efficient storage and fast searching. Several cross-modality hashing approaches have been proposed to handle the above challenges, in both supervised [1]–[22] and unsupervised [23]–[37] manners. Furthermore, as the unsupervised hashing does not require any label information, it is suitable for large-scale retrieval problem in which the label information is mostly unavailable. Thus, in this work, we focus on the unsupervised setting of the cross-modality hashing problem for retrieval tasks.

When learning binary representations for the cross-modal retrieval task, it is essential to preserve both intramodal and intermodal similarities in a common Hamming space. Equivalently, the binary representations should satisfy several requirements: 1) the representations necessarily capture information from the input features, i.e., preserve intramodal similarity; 2) for the representations of a modality to effectively retrieve samples of other modalities (i.e., the intermodal similarity is preserved), the representations of this modality should capture as much information about other modalities as possible; and 3) the modality gap (i.e., heterogeneous gap) between the representations of different modalities should be minimized, i.e., binary codes of all modalities should be in the same common space and binary codes from different modalities of the same instance (which contain the same information) should be as similar as possible [29], [38], [39]. Minimizing the modality gap is necessary for the similarity between different modalities to be measured directly.

To preserve both intramodal and intermodal similarities in the unsupervised setting, many existing cross-modality hashing methods, both convolutional neural network (CNN)-based and non-CNN-based, relied on similarity matrices/graphs (one for each modality [23], [25], [28], [39], [40] or to a joint similarity matrix for all modalities [29]). Then, they learn hash codes via the eigenvalue decomposition of the similarity matrices. However, constructing the similarity matrix could be challenging and computationally expensive for large datasets. Furthermore, eigenvalue decomposition decreases the mapping quality substantially when increasing the hash code length [41]. Matrix factorization (MF)-based methods could avoid the large-scale graph constructing and eigen-decomposition process by finding a shared latent semantic
space [26], [41] that can reconstruct input data well for all modalities. However, only simple and capability-limited linear projection is used in MF-based methods. In addition, scaling up MF-based methods for much larger datasets is nontrivial. Recently, Li et al. [42] proposed unsupervised coupled cycle generative adversarial hashing networks [unsupervised coupled hashing (UCH)], which used pair-coupled generative adversarial networks (GAN) to learn representations for individual modality and generate compact hash codes. Even though this approach can achieve very competitive performance, training the minimax loss of GAN can be challenging.

Taking a different approach, inspired by recent advances in unsupervised representation learning [43]–[46], in this article, we propose to learn informative binary representations for unsupervised cross-modal hashing via maximizing mutual information (MI). We learn to preserve the intramodal and intermodal similarities via maximizing the MI between representations and input and the MI between representations of different modalities. More specifically, by adopting the variational information maximization method [47], we can use the binary representations to be modeled by multivariate Bernoulli distributions. As a result, the binary representations can be learned easily and effectively by maximizing the MI between themselves and input via maximizing the estimated variational MI lower bounds [47]–[51] using gradient descent optimization in the mini-batch manner.

Furthermore, we find out that trying to minimize the modality gap by learning similar binary representations for the same instance from different modalities could result in undesirable side effect. Specifically, the modality-private information (i.e., the information of one modality that does not share with any other modality) is discarded. Consequently, the representations may become less representative for the input. Hence, balancing between reducing modality gap and losing modality-private information is important for the cross-modal retrieval tasks.

In addition to the above requirements for cross-modal retrieval tasks, independence and balance are well known to be important properties of informative hash codes [52]–[57]. The independence property, i.e., different bits in the binary codes are independent of each other, is to ensure that hash codes do not capture redundant information. The balance property, i.e., each bit has a 50% chance of being 0 or 1, is to ensure that hash codes contain a maximum amount of information [54]. By assuming the binary representations to be modeled by multivariate Bernoulli distributions, we propose to leverage the total correlation (TC) [58] as a regularizer (i.e., minimizing TC) to enhance the independence between hash bits. Furthermore, the balanced property can also be achieved by regularizing the Bernoulli distributions such that the averaged probabilities over a training set for a bit to be 0 or 1 are equal and equal to 50%.

In summary, by adopting the maximizing MI approach, we propose a novel framework, dubbed cross-modal info-max hashing (CMIMH), whose main contributions are as follows.

1) We propose to adopt the maximizing MI approach to learn binary representations for the cross-modal retrieval tasks. Besides maximizing MI between the representations and inputs, we explicitly maximize the MI between representations of different modalities, which is important to learn informative representations for cross-modal retrieval tasks.

2) We find out that minimizing modality gap by learning similar binary representations for the same instance from different modalities could result in less informative representations. Since both informative representations and modality gap are important for cross-modal retrieval tasks, properly balancing these two factors is important to achieve good performance, as shown in our experiments. To the best of our knowledge, our work is the first work that provides in-depth analyses about the tradeoff between these two factors.

3) We propose to leverage the TC as a regularizer to enhance the independence between hash bits. The experimental results confirm that minimizing TC results in more independence hash bits and higher performance.

4) We compare our proposed method against various state-of-the-art unsupervised cross-modal hashing methods on three standard cross-modal benchmark datasets, i.e., MIR-Flickr25K, NUS-WIDE, and MS-COCO. Quantitative results justify our contributions and demonstrate that CMIMH outperforms the compared methods on various evaluation metrics and settings.

II. RELATED WORKS

In this section, we briefly discuss the noticeable methods proposed cross-modal hashing.

A. Supervised Cross-Modal Hashing

Supervised hashing methods can explore the semantic information to enhance the data correlation from different modalities (i.e., reduce modality gap) and reduce the semantic gap. Many supervised cross-modal hashing methods with shallow architectures have been proposed, for instance, coregularized hashing (CRH) [1], heterogeneous translated hashing (HTH) [2], supervised multimodal hashing (SMH) [3], quantized correlation hashing (QCH) [4], semantics-preserving hashing (SePH) [5], discrete cross-modal hashing (DCH) [6], and supervised matrix factorization hashing (SMFH) [8]. All of these methods are based on handcrafted features, which cannot effectively capture the heterogeneous correlation between different modalities and may therefore result in unsatisfactory performance. Unsurprisingly, recent deep learning-based works [10]–[16], [59] can capture heterogeneous cross-modal correlations more effectively. Deep cross-modal hashing (DCMH) [12] simultaneously conducts feature learning and hash code learning in a unified framework. Pairwise relationship-guided deep hashing (PRDH) [15], in addition, takes intramodal and intermodal constraints into consideration. Deep visual-semantic hashing (DVSH) [16] uses CNNs, long short-term memory (LSTM), and a deep visual-semantic fusion network (unifying CNN and LSTM) for learning isomorphic hash codes in a joint embedding space. However, the text modality in DVSH is only limited to sequence texts (e.g., sentences). In cross-modal deep variational hashing, Liong et al. [14], [60] first proposed to learn...
shared binary codes from a fusion network and then learn genera-
tive modality-specific networks for encoding out-of-sample
inputs. In cross-modal hamming hashing, Cao et al. [61] pro-
posed an exponential focal loss that puts higher losses on
pairs of similar samples with Hamming distance much larger
than 2 (in comparison with the sigmoid function with the
inner product of binary codes). Mandal et al. [9] proposed
generalized semantic preserving hashing (GSPH), which can
work for unpaired inputs (i.e., given a sample in one modality,
there is no paired sample in other modality.). Song et al. [18]
took advantage of the memory mechanism to design a memory
network that can learn to store supporting information and
retrieve the necessary information in reference. Xie et al. [19]
proposed multitask consistency-preserving adversarial hashing
(CPAH), which consists of two modules: consistency refined
module to learn modality-common and modality-private rep-
resentations and multitask adversarial learning module to pre-
serve the semantic consistency information between different
modalities. Ji et al. [62] proposed an attribute-guided net-
work (AgNet) framework to narrow the semantic gap brought
due to modality heterogeneity and category migration for the zero-
shot cross-modal retrieval.

Although supervised hashing typically achieves very high
performance, it requires a labor-intensive process to obtain
large-scale labels, especially for multimodalities, in many real-
world applications. In contrast, unsupervised hashing does
not require any label information. Hence, it is suitable for
large-scale image search in which the label information is
usually unavailable.

B. Unsupervised Cross-Modal Hashing

Cross-view hashing (CVH) [23] and intermedia hashing
(IMH) [25] adopt spectral hashing [52] for the cross-modality
hashing problem. These two methods, however, produce dif-
ferent sets of binary codes for different modalities, which
may result in limited performance. Linear cross-modal hash-
ing (LCM) [40] reduces the training complexity of IMH
by representing training data with some cluster centers to
avoid the large-scale graph construction process. In predictable
dual-view hashing (PDH), Rastegari et al. [24] introduced the
predictability to explain the idea of learning linear hyperplanes
that each one divides a particular space into two subspaces re-
presented by −1 or 1. The hyperplanes, in addition, are learned
in a self-taught manner, i.e., to learn a certain hash bit of a
sample by looking at the corresponding bit of its nearest neigh-
bors. Collective matrix factorization hashing (CMFH) [26]
aims to find consistent hash codes from different views by
collective matrix factorization. Latent semantic sparse hashing
(LSSH) [27] was proposed to learn hash codes in two steps.
First, latent features from images and texts are jointly learned
with sparse coding, and then, hash codes are achieved by using
matrix factorization. Subsequently, Wang et al. [41] proposed
robust and flexible discrete hashing (RFDH), which directly
optimizes and generates the unified binary codes for various
views in the unsupervised manner via matrix factorization
such that large quantization errors caused by relaxation can
be relieved to some extent. Inspired by canonical correlation
analysis-iterative quantization (CCA-ITQ) [53], Irie et al. [58]
proposed alternating coquantization (ACQ) to alternately min-
imize the binary quantization error for each of modalities.
Liu et al. [29] applied nearest neighbor similarity [63] to
construct a fusion anchor graph (FSH) from text and image
modalities for learning binary codes. Recently, the proposed
collective affinity learning method (CALM) [35] collectively
and adaptively learns hashing functions in an unsupervised
manner with an anchor graph constructed on partial multi-
modal data. We would like to refer readers to [64] for a more
comprehensive survey on non-deep neural network (DNN)-
based cross-modal retrieval methods.

In addition to the aforementioned shallow methods, several
works [31]–[33] utilized (stacked) autoencoders for learning
binary codes. These methods try to minimize the distance
between hidden spaces of modalities to preserve intermodal
semantic to a certain extent. Deep binary reconstruction
(DBRC) [65] proposed to minimize the reconstruction error
based on the shared binary representation. DBRC addition-
ally proposed a scalable tanh activation with a learnable
parameter, which can mitigate the gradient problem of the
discrete domain of {−1, 1} during training. Recently, Zhang
and Peng [66] proposed multipathway generative adversarial
hashing (MGAH), which consists of a generative model and a
discriminative model, in which the generative model fits the
distribution over the manifold structure and selects informative
data of other modalities. While the discriminative model learns
to discriminate data generated from the generative model
and data sampled a correlation graph (a graph that captures
the underlying manifold structure across different modalities).
Wu et al. [39] proposed unsupervised deep cross-modal hash-
ing (UDCMH) that enables the feature learning to be jointly
optimized with the binarization. Su et al. [38] proposed a
joint-semantics affinity matrix, which integrates the neighbor-
hood information of two modals, for mini-batch samples to
train deep network in an end-to-end manner. Li et al. [42]
proposed UCH, which used pair-coupled generative adversarial
networks to learn representations for individual modality and
generate compact hash codes. Given a multimodal unpaired
data, Wu et al. [67] adopted the cycle-consistent loss [68] to
learn hashing functions. Although these methods make great
progress, the performance of these systems still has room for
improvement.

C. Representation Learning With MI

MIHash Cakir et al. [69] proposed to use MI to learn
hash codes for online hashing. In specific, given two
Hamming distance distributions of a sample with its neigh-
bors and its nonneighbors, the MI is used to measure
the separability of these two distributions, which gives a
good quality indicator for online hashing. However, this
method is only proposed for the single modality case in
a supervised manner, while our proposed method aims
to learn binary representations for cross-modal retrieval in
an unsupervised manner. Guo et al. [70] only adopted MI
to learn to preserve intramodal similarity, while our pro-
posed method utilized MI to preserve both intramodal and
intermodal similarities. In addition, in contrast with [70]
which learns real-valued representations, our method learns
binary representations for large-scale retrieval. Besides, several
works [43], [45], [46], [71]–[73] rely on MI to learn representations in unsupervised/self-supervised manner for single-view and/or multiview settings. Specifically, Hjelm et al. [43] proposed to learn a global image representation such that the MI between this global representation and local features is maximized. Bachman et al. [45] further improved [43] by maximizing the MI between a global representation and local features of the different views (i.e., different images generated by data augmentation). Different from [43], [45], [74], the work adopted the information bottleneck (IB) objective [75] with a variational approximation to minimize the MI between the input and the representation while still ensuring that the global representations can fulfill the target task (e.g., classification). This learning method could result in more robust representations. In [46], with the assumption that any single view can fully contain information of labels of a downstream task (e.g., classification), the authors aimed to learn robust representations by capturing the shared information between views and discarding the private information (i.e., information that is exclusively contained in a particular view). Information competing process (ICP) [71] is another intriguing MI-based representation learning method. ICP aims to learn diversified representations by first separating a representation into two parts with different MI constraints and then forcing separated parts to accomplish the downstream task independently without any knowledge of what the other part has learned. However, these works [43], [45], [46], [71]–[73] mainly focus on learning a single real-valued representation from single-/multiview inputs for classification. In contrast, our work aims to learn binary representations for multimodal retrieval.

III. PROPOSED METHOD

Given a multimodality dataset of $N$ instances, denoted as $O = \{o_i\}_{i=1}^N$, in which each instance is described by an image–text pair $o_i = (x'_i, t'_i)$, where $x'_i \in \mathbb{R}^{D'_i}$ and $t'_i \in \mathbb{R}^{D_t}$ are the $j$th $D'_i$ and $D_t$ dimensional features of image and text modalities, respectively. We aim to learn the corresponding $L$-bit binary representations $h'_i \in \{0, 1\}^L$ for each image and text pair $(x'_i, t'_i)$.

For representations being suitable for the cross-modal retrieval task, the representations should satisfy several requirements.

1) The representations should well represent the input data, i.e., they necessarily capture information from the input features.

2) For the image representations to effectively retrieve text samples, the image representations should capture as much information about the text modality as possible. Analogously, the text representation should contain as much information about the image modality as possible to retrieve image samples effectively.

3) The representations of different modalities should be well aligned with each other (i.e., the modal gap is minimized).

A. MI Maximization

MI has been proven to be an important quantity in data science to measure the dependence of two random variables since it can capture nonlinear statistical dependencies between variables [50]. Recent representation learning methods [43], [45] showed that MI maximization between inputs and encoder outputs can help to learn informative representations. Hence, to achieve the first requirement, we aim to maximize the MI between the binary representations and the input data. Noticeably, in the ideal case, when the representations fully capture all input information; the MI between image and text representations would be maximized and be equal to the MI between the image and text input data (which is a constant).

Equivalently, the second requirement would be satisfied. However, in practice, the representations may not fully capture all input information. Hence, we propose to further enforce the second requirement by explicitly maximizing the MI between the representations of image and text modalities. Our initial objective now can be written as follows:

$$\max I(x'; h') + I(x'; h') + I(h'; h').$$ (1)

However, MI is well known to be notoriously difficult to compute. To handle this trouble, we propose to assume the image and text representations to be random variables so that we can leverage recent advances in estimating variational lower bounds of MI [47]–[51] to maximize the objective function (1).

B. Variational Lower Bounds of MI

1) Variational Information Maximization: Directly optimizing $I(x'; h')$ and $I(x'; h')$ in the objective (1) is infeasible as the true posterior distributions (i.e., $P(x|h'), P(x|h')$) requiring for computing the MI is still unknown. Fortunately, we can use the variational information maximization [47], [48] to compute the MI lower bound, in which $Q_S(x|h')$ can be used to approximate the true posterior distribution, as follows:

$$I(x'; h') = H(x') - H(x'|h') = H(x') + \mathbb{E}_{P(x)}[\mathbb{E}_{P_h(x|h')}[\log P(x|h')]]$$

$$= H(x') + \mathbb{E}_{P(x)}[\mathbb{E}_{P_h(x|h')}[\log Q_S(x|h')]]$$

$$\geq H(x') + \mathbb{E}_{P(x)}[\mathbb{E}_{P_h(x|h')}[\log Q_S(x|h')]]$$

(2)

where $H(\cdot)$ is the entropy function of a random variable, $\mathbb{E}$ is expectation, and $\theta_i$ and $\phi_i$ represent the model parameters of the encoder and decoder distributions, respectively. Similarly, we have

$$I(x'; h') \geq H(x') + \mathbb{E}_{P(x)}[\mathbb{E}_{P_h(x|h')}[\log Q_S(x|h')]].$$ (3)

Note that $H(x')$ and $H(x')$ are constant for the given input data. To be concise, from now on, we skip the subscript about model parameters in the encoder and decoder distributions whenever the context is clear.

As we aim to obtain binary representations for the cross-modal hashing, we adopt the multivariate Bernoulli distributions to model the encoder distributions $P(h'|x')$ and $P(h'|x')$, i.e., $P(h'|x') := \text{Bern}(\mu')$ and $P(h'|x') := \text{Bern}(\mu')$. In addition, we assume that the decoder distributions $Q(x|h')$ and $Q(x|h')$ are Gaussian. Therefore, the log likelihoods in (2) and (3) can be maximized by minimizing the $L2$ reconstruction loss.
2) Reparameterization Trick: Following [76], [77], we can reparameterize $h \sim \text{Bern}(\mu)$ as $h = \text{sign}(z)$ (i.e., $\text{sign}(z) = 1$ if $z \geq 0$ and $\text{sign}(z) = 0$ if $z < 0$), where $z$ is a vector of independent logistic random variables defined as follows:

$$z = g(u, \mu) = \log \frac{\mu}{1 - \mu} + \log \frac{\mu}{1 - \mu}$$

where $u \sim \text{Uniform}(0, 1)$. Even though this reparameterization trick can help to avoid sampling from the Bernoulli distribution, this trick still requires a discrete threshold function, which hinders the gradient descent optimization. To handle this difficulty, we resort to the straight-through estimator (STE) [78], i.e., $(\partial \mu / \partial z)_z = 1$, to approximate the gradients propagating through the sign function.

3) Sample-Based Differentiable MI Lower Bound: Different from $I(x'; h')$ and $I(x'; h')$, in $I(h'; h')$, we can access samples from two random variables independently.

This allows us to maximize the MI between the two representations $I(h'; h')$ using a sample-based differentiable MI lower bound, which could have a tighter bound than (2) and (3) in practice [79]. Furthermore, as our primary interest is to maximize the MI and not to find its precise value; we can rely on non-Kullback–Leibler (KL) divergences estimator, i.e., a Jensen–Shannon MI estimator ($I_{JS}$), which is observed to work better in practice (e.g., more stable) than the KL divergences MI estimator (e.g., Donsker–Varadhan (DV) representation [80] or $f$-divergences [49]) [43]. The sample-based differentiable Jensen–Shannon MI estimator ($I_{JS}$) could be defined as follows:

$$I(h'; h') \leq I_{JS}(h'; h') = D_{JS}[p(h'; h') || p(h) p(h')]$$

$$\geq \sup_{T \in F} \mathbb{E}_{p(h', h')}[T] + \mathbb{E}_{p(h') p(h')}[\log(2 - \exp(-T))]$$

where $F$ is a family of functions $T(h'; h') \in \mathbb{R}$, parameterized by neuron networks, which is jointly optimized during the training procedure to classify whether a pair of samples are from the joint distribution $p(h, h')$ or the product of marginal distributions $p(h') p(h)$, i.e., pairs of $(h'; h')$ are produced from the pairs of inputs sampled from joint distributions $p(x', x')$ or sampled from the product of marginal distributions $p(x') p(x')$ [49], [50]. In addition, $\bar{T} = \log(2) - \log(1 + \exp(-T))$ [49].

From (5), we can see that if the function $T$ can correctly classify between samples from the joint and product of marginal distributions with high confident, the MI lower bound will be maximized. However, we found that the process of jointly training the function $T$ (with binary inputs sampled from $h \sim \text{Bern}(\mu)$ and $h' \sim \text{Bern}(\mu')$) and the encoders may result in an undesirable side effect. In particular, besides encouraging the encoders to learn the hidden variables for different modalities such that the MI of these hidden variables is maximized, the function $T$ also promotes the encoders to reduce the stochasticity in $p(h')$ and $p(h')$ (i.e., $\mu \rightarrow 0$ or $\mu \rightarrow 0$). Intuitively, reducing stochasticity in $p(h')$ and $p(h')$ (i.e., less noise) allows the function $T$ to correctly classify samples easier. Consequently, this side effect may impact on the variational information maximization in (2) and (3) as $P(h' | x')$ and $P(h' | x')$ become more deterministic. Arguably, using multiple pairs of $(h', h')$ could help to mitigate this problem. However, this requires a higher computational cost. To effectively address the problem, we propose to directly classify whether pairs of multivariate Bernoulli distributions (Bern($\mu$), Bern($\mu'$)) (from which a pair of $(h', h')$ is sampled) are produced from the pairs of inputs sampled from joint distributions $p(x', x')$ or sampled from the product of marginal distributions $p(x') p(x')$, i.e., $T(\mu, \mu')$ instead of $T(h', h')$. This would help to eliminate noise in the inputs of the function $T$ while still being able to reflect the relationship between hidden variables of different modalities. In addition, using the Bernoulli variables as the inputs is also helpful in a gradient descent optimization process. The gradients from the function $T$ do not flow through the STE, which is a biased gradient estimator [78].

Noticeably, a more direct way to enforce the intermodal similarity is to maximize $I(x'; h') + I(x'; h')$. However, we find out that maximizing $I(x'; h') + I(x'; h')$ results in similar performances compared with maximizing $I(h'; h')$ while requiring a higher computational cost.

C. Minimizing Modality Gap

In the cross-modal retrieval task, besides having the representations that well capture input information and having MI between representations of different modalities maximized (the first and second requirements), it is also desirable for the gap between different modalities to be minimized (the third requirement). In other words, binary codes from different modalities of the same pair should be as similar as possible [29], [38], [39]. To achieve this requirement, we propose to minimize the symmetrized KL divergence between the two multivariate Bernoulli distributions [i.e., $P(h' | x')$ and $P(h' | x')$] of the same pairs as follows:

$$D_{skl}[P(h' | x') || P(h' | x') ] + D_{skl}[P(h' | x') || P(h' | x') ]$$

with the KL divergence between two multivariate Bernoulli distributions as

$$D_{kl}[P(h' | x') || P(h' | x') ]$$

$$= \sum_{i=1}^{l} \left( \mu_{i}' \log \frac{\mu_{i}'}{\mu_{i}} + (1 - \mu_{i}' \log \frac{1 - \mu_{i}'}{1 - \mu_{i}}) \right)$$

where $\mu_{i}'$ and $\mu_{i}'$ are the $i$th element of $\mu'$ and $\mu'$, respectively.

However, we found that strictly enforcing this property could result in an undesirable outcome, specifically, discarding modality-private information [46]. In particular, considering that the amount of information $h'$ contains which is unique to $x'$ and not shared by $x'$ (i.e., $I(x'; h' | x')$), $I(x'; h' | x')$ can

1Note that the MI lower bound in (2) is derived for random variables [47], [48] and may not be applicable for deterministic variables.

2The MI of $x'$ and $h'$ given $x'$. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
be expressed as

\[ I(x'; h'|x') = \mathbb{E}_{x':x' \sim p(x',x')} \mathcal{L}h \mathcal{L}_{h(h'|x')} \cdot \frac{P_{h}(h'|x' = x')}{P_{h}(h'|x' = x')} \]

\[ = \mathbb{E}_{x:x \sim p(x',x')} \mathcal{L}h \mathcal{L}_{h(h'|x')} \cdot \frac{P_{h}(h'|x' = x')}{P_{h}(h'|x' = x')} \]

\[ + \mathbb{E}_{x:x' \sim p(x',x')} \mathcal{L}h \mathcal{L}_{h(h'|x')} \cdot \frac{P_{h}(h'|x' = x')}{P_{h}(h'|x' = x')} \]

\[ = D_{KL}[P_{h}(h'|x')||P_{h}(h'|x')] - D_{KL}[P_{h}(h'|x')||P_{h}(h'|x')] \]

\[ \leq D_{KL}[P_{h}(h'|x')||P_{h}(h'|x')] \]

(8)

Analogously

\[ I(x'; h'|x') \leq D_{KL}[P_{h}(h'|x')||P_{h}(h'|x')] \]

(9)

Note that the upper bounds in (8) and (9) are tight as the case)

\[ TC(z) = \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{q(\tilde{z})} \right] \approx \mathbb{E}_{q(z)} \left[ \log \frac{D(z)}{1 - D(z)} \right] \]

(11)

in which the classifier \( D \in [0, 1] \) is jointly trained to classify between samples from \( q(z) \) and \( q(\tilde{z}) \) and the classifier outputs the probability \( D(z) \) that the input is a sample from \( q(z) \) rather than from \( q(\tilde{z}) \). Similar to the function \( T \) for estimating MI lower bound discussed in Section III-B3, we also use the Bernoulli variables as the input for the classifier \( D(\cdot) \)

\[ \mathcal{L}_{ind} = \mathbb{E}_{q(z)} \left[ \log \frac{D(z)}{1 - D(z)} \right] + \mathbb{E}_{q(\tilde{z})} \left[ \log \frac{D(\tilde{z})}{1 - D(\tilde{z})} \right] \]

(12)

2) Balance: To obtain balanced hash codes, we regularize the encoders such that the averaged probabilities (over the training set) for a bit to be 0 or 1 are equal and equal to 50%. Equivalently, we have

\[ L_{bal} = \sum_{l=1}^{L} \frac{1}{N} \sum_{j=1}^{N} (|z_j'| - 0.5) \]

(13)

D. Additional Properties of Good Hash Code: Independence and Balance

The encoded binaries in hashing algorithms are in general short in length. To maximize hash code representational capability, we additionally include the independent and balancing regularizers on the binary codes, i.e., different bits in the binary codes are independent to each other and each bit has 50% chance of being 0 or 1, respectively [52], [54]. The independence property is to minimize redundant information captured in hash codes, and the balance property is to ensure that hash codes contain a maximum amount of information [54].

1) Independence: To enhance the independence between hash bits, we aim to minimize the TC [58], which is a popular measure of dependence for multiple random variables (i.e., multiple Bernoulli variables of multiple hash bits in our case) \( TC(z) := D_{KL}[q(z)||q(\tilde{z})] \).

\[ q(z) := \prod_{j=1}^{L} q(z_j) \]

However, the TC is intractable since both \( q(z) \) and \( q(\tilde{z}) \) involve mixtures with an exponential number of components. Fortunately, being able to access to samples from both \( q(z) \)

A more detail derivation is provided in Appendix A.

and \( q(\tilde{z}) \) distributions allows us to minimize their KL divergence using the density-ratio trick [51] as illustrated in [81] as follows:

\[ TC(z) = \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{q(\tilde{z})} \right] \approx \mathbb{E}_{q(z)} \left[ \log \frac{D(z)}{1 - D(z)} \right] \]

(11)

E. Final Objective Function and Reference Stage

In summary, the final objective function of our proposed method is defined as follows:

\[ \max \ I(x'; h') + I(x'; h') + \lambda_1 I(h'; h') \]

\[ - \lambda_2 D_{KL}(h'; h') - \lambda_3 L_{ind} - \lambda_4 L_{bal} \]

(14)

in which \( \lambda_1, \lambda_2, \lambda_3, \) and \( \lambda_4 \) are hyperparameters.

For reference, it is undesirable to have different binary codes for a query sample under different retrieval runs; hence, we obtain the deterministic binary codes by simply applying a threshold function on the Bernoulli variables, i.e., \( h = \text{sign}(\mu - 0.5) \).

IV. EXPERIMENT

In this section, we conduct a wide range of experiments to validate our proposed method on three standard benchmark datasets for the cross-modal retrieval task, i.e., MIR-Flickr25k [82], NUS-WIDE [83], and MS-COCO [84].

A. Experiment Setting

1) Datasets: The MIR-Flickr25K dataset [82] is collected from Flickr website, which contains 25,000 image–text pairs together with 24 provided labels. The texts are represented as 1386-dimensional tagging vectors. In addition, we remove the pairs whose texts do not contain any tag in the 1386 common tags results. As a result, 20,015 pairs are preserved. Following [38], [39], we randomly sample 2000 instances for the query set, while the remaining instances are used as the database. In addition, 5000 instances are randomly sampled from the database to form the training set.
The **NUS-WIDE** dataset [83] is a multilabel image dataset crawled from Flickr, which contains 296,648 images with associated tags. Each image–tag pair is annotated with one or more labels from 81 concepts. In this dataset, each text is represented by a 1000-dimensional preprocessed bag of words (BOW) feature. Following the common practice [13], [29], [38], we select image–tag pairs that have at least one label belonging to the top ten most frequent concepts and the corresponding 186,577 annotated instances are preserved. We randomly sample 2000 instances as queries. The remaining instances are used as the database, and 5000 instances are randomly sampled from the database to form the training set.

The **MS-COCO-2017** consists of 118,287 training images and 5000 validation images. Each image includes at least five sentences annotations (captions). We randomly select one sentence and use the pretrained bidirectional encoder representations from transformers (BERT) model [85] to extract the sentence embedding as the text representations. Following [67], we use the provided 80 image segmentation categories as ground-truth labels for the image–sentence pairs. We use the validation set as the query set. By removing image–sentence pairs that have no category information, we obtain 117,266 database samples and 4952 query samples. Similar to the MIR-Flickr25K and NUS-WIDE datasets, we randomly sample 5000 instances from the database for training.

For images of all datasets, we extract FC7 features from the PyTorch pretrained AlexNet network [86] and then apply principal component analysis (PCA) to compress to 1024-dimension.

2) **Evaluation Metrics:** The evaluations are presented in both cross-modal retrieval tasks (i.e., Img → Txt and Txt → Img) and single-modal retrieval tasks (i.e., Img → Img and Txt → Txt), in which images (Img)/texts (Txt) are used as queries to retrieve image/text database samples accordingly. The quantitative performance is evaluated by the standard evaluation metrics: 1) mean average precision of top 1000 returned samples (mAP@1k) and 2) precision curve at top-K retrieved images (Prec@K). The image–text pairs are considered to be similar if they share at least one common label. Otherwise, they are considered to be dissimilar.

3) **Implementation Details:** Both encoder and decoder consist of multilayer perceptrons (MLPs) of two hidden rectified linear units (ReLUs) of size 1024. The critic $T$ for $I_\text{S}$ estimator (5) is a separable function $f(x; y) = \phi(x)\phi(y)$, where $\phi(\cdot)$ and $\phi(\cdot)$ are MLPs with two hidden layers of size 512 and Leaky-ReLU activations. The classifier $D$ to estimate $T_C$ also consists of an MLP of two hidden Leaky-ReLU units of size 512.

In addition, we employ the stochastic gradient descent (SGD) optimizer with a mini-batch size of 128, a momentum of 0.9, and a weight decay of 0.0001. The learning rate is set as 0.01 for the encoders, the critic $T$, and the classifier $D$, and set as 0.001 for the decoders. The hyperparameters $\lambda_1, \lambda_2, \lambda_3$, and $\lambda_4$ are empirically set by cross validation as 1, 5, 1, 0.25, and 0.01, respectively, for MIR-Flickr25k and NUS-WIDE datasets and set as 4, 1.5, 0.25, and 0.01, respectively, for MS-COCO dataset.

B. **Ablation Study and Parameter Analysis**

1) **Necessity of Explicitly Maximizing the MI Between Hash Codes of Different Modalities:** In this section, we conduct experiments on the MIR-Flickr25k and NUS-WIDE datasets using 32-bit hash codes with various values of $I(h^i, h^j)$ weight (i.e., $\lambda_i$). The experimental results in terms of mAP@1k are presented in Fig. 1. As can be seen, when using a very small weight for $I(h^i, h^j)$ (i.e., $\lambda_i \leq 0.1$), the retrieval performance is significantly lower for all four retrieval tasks in comparison with larger $I(h^i, h^j)$ weights (i.e., $\lambda_i \geq 1$). With a reasonable large weight for $I(h^i, h^j)$, the model is enforced to retain the information that is shared across modalities. The information that is shared among different modalities information is generally more useful for both the cross- and single-modal retrieval tasks, as, intuitively, this type of information is more likely to contain the ground-truth information. The experimental results confirm the importance and necessity of explicitly maximizing the MI between hash codes of different modalities. Besides, at a too large $I(h^i, h^j)$ weight (i.e., $\lambda_i \geq 5$), we also observe small decreases in retrieval performance. This fact is also understandable as the model pays less attention on maximizing $I(x^i, h^j)$ and $I(x^i, h^j)$, which results in less informative hash codes.

2) **The Benefit of Using Bernoulli Variables as the Input of Function $T$ in Estimating $I(h^i, h^j)$ Lower Bound:** In this section, we conduct experiments on the MIR-Flickr25 dataset with $L = 32$ bits to validate the benefit of using Bernoulli variables as the input of function $T$ in estimating $I(h^i, h^j)$ lower bound in comparison with using multiple pairs of binary samples as the input. The retrieval performance is presented in Table I. We also present the histograms of $\mu^i$ when using $T(\mu^i, \mu^j)$ and $T(h^i, h^j)$ (with single sample) for the MI lower bound estimator in Fig. 2. When using single pair of binary samples as the input for the function $T(h^i, h^j)$, we can observe that the majority values of $\mu^i$ become very small or very large. Approximately 67% of $\mu^i$ are in [0, 0.01] or [0.99, 1] (i.e., to be 0 or 1, respectively, with 99% confidence), in comparison with about 22% when using $T(\mu^i, \mu^j)$. This effect significantly affects the retrieval performance. When using multiple binary samples, the performance improves. However, the best performance is achieved when using Bernoulli variables, as this not only helps to eliminate input noise but it also helps the gradients to not propagate through the biased gradient estimator STE, which introduces noise in gradients. Furthermore, we note that using $n$ binary
we present the mAP@1K cost. We can observe that too large weights for \( D_{\text{SKL}} \) varies on the MIR-Flickr25k and NUS-WIDE datasets. First, \( n \) samples would require approximately 6296 IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 34, NO. 9, SEPTEMBER 2023 304 curves as the weight of \( h_t \) for the MI lower bound estimator. The experiment is conducted on MIR-Flickr25k with 32-bit hash codes. We observe similar histograms for \( \mu \). (a) \( T(\mu^I, \mu^T) \). (b) \( T(h^I, h^T) \).

**Fig. 3.** mAP@1K curves as the weight of \( D_{\text{SKL}} \) varies on MIR-Flickr25k and NUS-WIDE using 32-bit hash codes. (a) MIR-Flickr25k. (b) NUS-WIDE.

**TABLE I**

| Task       | \( T(\mu^I, \mu^T) \) | \# of pairs of bin. samples for \( T(h^I, h^T) \) |
|------------|-------------------------|-----------------------------------------------|
| Img \( \rightarrow \) Txt | 81.93 | 80.24, 81.27, 81.75, 81.74 |
| Txt \( \rightarrow \) Img | 81.43 | 79.41, 80.86, 81.22, 81.29 |

**Fig. 2.** Histogram of \( \mu^I \) when using \( T(\mu^I, \mu^T) \) and \( T(h^I, h^T) \) (single pair of binary samples) for the MI lower bound estimator. The experiment is conducted on MIR-Flickr25k with 32-bit hash codes. We observe similar histograms for \( \mu \). (a) \( T(\mu^I, \mu^T) \). (b) \( T(h^I, h^T) \).

| Configuration | 32 bits | 48 bits |
|---------------|---------|---------|
| \( I(h^I, h^T) \) | \( D_{\text{SKL}} \) | \( \mathcal{L}_{\text{ind}} \) | \( \mathcal{L}_{\text{bal}} \) |
| ✓ | ✓ | ✓ | ✓ | 75.13 | 73.14 | 77.57 | 75.49 |
| ✓ | ✓ | ✓ | ✓ | 77.61 | 75.53 | 78.89 | 78.39 |
| ✓ | ✓ | ✓ | ✓ | 75.52 | 73.69 | 77.84 | 75.83 |
| ✓ | ✓ | ✓ | ✓ | 77.87 | 75.95 | 79.34 | 78.69 |
| ✓ | ✓ | ✓ | ✓ | 81.28 | 80.73 | 82.07 | 81.51 |
| ✓ | ✓ | ✓ | ✓ | 81.35 | 80.93 | 82.11 | 81.82 |
| ✓ | ✓ | ✓ | ✓ | 81.43 | 81.14 | 82.16 | 81.78 |
| ✓ | ✓ | ✓ | ✓ | 81.93 | 81.43 | 82.92 | 82.18 |

**TABLE II**

**3) Effect the Symmetrized KL Divergence:** In Fig. 3, we present the mAP@1K curves as the weight of \( D_{\text{SKL}} \) varies on the MIR-Flickr25k and NUS-WIDE datasets. First, we can observe that too large weights for \( D_{\text{SKL}} \) (i.e., \( \lambda_2 \geq 5 \)) have significant impacts on the retrieval performance for all four retrieval tasks. This observation is consistent with our discussion in Section III-C that too large \( D_{\text{SKL}} \) weights will force the model to discard a large amount of modality-private information in the representations. As a result, the binary representations do not well represent the input data. For too small \( D_{\text{SKL}} \) weights (i.e., \( \lambda_2 < 0.5 \)), the retrieval performance on Img \( \rightarrow \) Txt andTxt \( \rightarrow \) Img retrieval tasks is unsurprisingly low as the binary representations of image and text modalities are poorly aligned with each other and not suitable for the cross-modal retrieval tasks. However, different from the case of too large \( D_{\text{SKL}} \) weights, too small \( D_{\text{SKL}} \) weights only result in minor performance drops for the Img \( \rightarrow \) Img andTxt \( \rightarrow \) Txt retrieval tasks, which means that the learned binary representations still well capture information of the input data. The small performance drops for the Img \( \rightarrow \) Img andTxt \( \rightarrow \) Txt retrieval tasks potentially indicate that the binary representations also capture information from the input data that do not share with the ground truth (e.g., noise).

**4) Effectiveness of Using TC as a Regularizer to Enhance Hash Bit Independence:** We conduct experiments on the MIR-Flickr25k dataset with and without the independence regularizer \( \mathcal{L}_{\text{ind}} \). In Table II, we show the mean square error (MSE) between the correlation matrix of the binary code of the database and the identity matrix (i.e., \( \text{Corr MSE} = \| \frac{1}{N} \hat{H}^I \hat{H}^I - I_L \|_F \), where \( \hat{H} = [\hat{h}_j]_{j=1}^L \in \{-1, 1\}^{N \times L} \) is the set of \( L \)-bit hash codes of a dataset and \( \hat{h}_j = [2h_j - 1] \) together with the retrieval performance at different code lengths. As can be seen, \( \mathcal{L}_{\text{ind}} \) consistently helps to reduce the Corr MSEs for both image and text modalities at various code lengths. A smaller Corr MSE indicates that the hash bits are more independence and consequently leads to higher performance. Interestingly, we also notice that, at high code lengths, even though the classifiers can easily predict if a sample is from \( q(z) \) with very high confident (i.e., high TC\(^5\)), \( \mathcal{L}_{\text{ind}} \) is still helpful in reducing correlation between hash bits and improving performance.

\(^5\)The probability that the classifier predicts a sample from \( q(z) \) can be computed as \( 1/(1 + \exp(-TC)) \) (e.g., \( 1/(1 + \exp(-6.0)) \approx 99.75\% \)).
5) A Summary of Effectiveness of Different Components: We additionally present in Table III the cross-modal retrieval performance \([mAP@1K \text{ (\%)}]\) for MIR-Flickr25k dataset with 32- and 48-bit hash codes with different combinations of components in the loss function. We can observe that the two terms \(I(h_i; h_t)\) and \(D_{SKL}(h_i; h_t)\) play very important roles in our proposed cross-modal hashing method. Without either \(I(h_i; h_t)\) or \(D_{SKL}(h_i; h_t)\), the cross-modal retrieval performance is significantly degraded. The ablation study also shows that the independence and balance terms are beneficial for hashing methods. However, even without these two terms, our proposed method CMIMH still achieves very good performance.

C. Comparison With the States of the Art

In this section, we compare our proposed method against recent state-of-the-art cross-modal hashing methods, i.e., CVH [23], PDH [24], CMFH [26], ACQ [28], fusion similarity hashing (FSH) [29], and deep joint-semantics reconstructing hashing (DJSRH) [38]. From the experimental results in terms of \(mAP@1K\) shown in Table IV, we can observe that our proposed method outperforms the state-of-the-art methods at majority of encoding lengths, datasets, and retrieval tasks. For the MS-COCO dataset with at \(L = 48\), CMIMH achieves lower performance than FSH and DJSRH on the Txt→Txt retrieval tasks while still outperforming DJSRH by clear margins in other settings.

In addition, Figs. 4 and 5, respectively, show the \(Pre@K\) curves and precision–recall (PR) curves for Img→Txt and Txt→Img tasks with 32-bit hash codes. Compared with other methods, ours still significantly outperforms the state-of-the-art baselines over the three benchmark datasets for both metrics (i.e., \(Pre@K\) curves and PR curve). These results confirm the advantages of our proposed method in unsupervised cross-modal retrieval.

1) Comparison With UDCMH [39]: Following UDCMH, we report the \(mAP\) of top-50 retrieved results (\(mAP@50\)). The experiment results are shown in Table V. We observe that our proposed method can outperform UDCMH by large margins for both MIR-Flickr25k and NUS-WIDE datasets.

2) Comparison With MGAH [66] and Unsupervised Knowledge Distillation for Cross-Modal Hashing (UKD) [87]: We conduct additional experiments on the MIR-Flickr25k and NUS-WIDE datasets. For a fair comparison, we follow the experiment settings from [66] and [87]. Specifically, the FC7 features of the pretrained 19-layer VGGNet are used for images. The 1000-dimensional BOW features are used for texts in both datasets; 1% sample of the NUS-WIDE dataset and 5% samples of the MIR-Flickr25k dataset are used as the

### Table IV

| Task   | Method   | MIR-Flickr25k | NUS-WIDE | MS-COCO |
|--------|----------|--------------|----------|---------|
|        |          | 16 | 32 | 48 | 16 | 32 | 48 | 16 | 32 | 48 |
| Img→Txt | CVH [23] | 68.18 | 66.95 | 66.32 | 56.43 | 57.16 | 57.47 | 61.49 | 62.15 | 60.06 |
|        | PDH [24] | 78.16 | 79.62 | 81.10 | 70.98 | 74.21 | 75.13 | 61.66 | 65.60 | 67.27 |
|        | CMFH [26] | 77.97 | 78.69 | 78.30 | 69.13 | 70.96 | 71.18 | 59.24 | 64.62 | 66.35 |
|        | ACQ [28] | 76.16 | 76.50 | 76.93 | 67.35 | 70.05 | 70.88 | 60.66 | 63.24 | 65.66 |
|        | FSH [29] | 77.55 | 79.36 | 80.32 | 69.45 | 70.48 | 72.72 | 62.75 | 66.24 | 69.04 |
|        | DJSRH [38] | 79.05 | 79.53 | 81.66 | 71.23 | 74.86 | 76.32 | 59.39 | 67.32 | 68.30 |
|        | CMIMH | 80.68 | 81.93 | 82.92 | 73.92 | 76.37 | 77.21 | 65.32 | 69.21 | 70.20 |

| Tx→Img | CVH | 68.08 | 66.89 | 66.40 | 57.40 | 58.30 | 58.51 | 62.45 | 63.46 | 61.22 |
|        | PDH | 76.79 | 78.64 | 79.22 | 69.61 | 72.24 | 73.88 | 63.24 | 67.69 | 69.66 |
|        | CMFH | 76.81 | 76.83 | 77.36 | 66.98 | 69.14 | 70.32 | 60.20 | 66.06 | 68.39 |
|        | ACQ | 74.46 | 75.22 | 75.39 | 65.53 | 68.22 | 69.54 | 61.83 | 64.44 | 66.96 |
|        | FSH | 75.10 | 77.10 | 78.47 | 67.39 | 69.03 | 70.25 | 65.05 | 69.07 | 71.48 |
|        | DJSRH | 77.44 | 78.65 | 80.10 | 68.18 | 73.29 | 74.72 | 56.19 | 67.95 | 71.11 |
|        | CMIMH | 79.77 | 81.43 | 82.18 | 72.75 | 75.02 | 75.68 | 66.08 | 70.35 | 72.21 |

| Txt→Txt | CVH | 69.49 | 68.36 | 67.76 | 59.64 | 60.41 | 61.25 | 60.66 | 61.97 | 60.94 |
|        | PDH | 79.65 | 81.46 | 82.86 | 74.23 | 77.06 | 78.08 | 61.10 | 65.28 | 67.06 |
|        | CMFH | 81.45 | 82.54 | 83.13 | 75.33 | 77.72 | 78.35 | 60.31 | 65.53 | 67.78 |
|        | ACQ | 78.91 | 78.94 | 79.58 | 71.13 | 76.63 | 74.88 | 60.07 | 62.46 | 64.76 |
|        | FSH | 80.22 | 82.39 | 83.68 | 73.28 | 75.57 | 76.59 | 61.70 | 65.64 | 68.47 |
|        | DJSRH | 82.39 | 83.17 | 84.07 | 77.29 | 79.29 | 80.27 | 59.87 | 66.77 | 68.29 |
|        | CMIMH | 83.79 | 85.74 | 86.76 | 78.74 | 81.35 | 81.95 | 65.38 | 69.24 | 70.16 |

| Img→Txt | CVH | 67.47 | 66.79 | 66.72 | 57.45 | 60.13 | 61.50 | 64.17 | 66.84 | 64.96 |
|        | PDH | 75.72 | 76.66 | 78.04 | 67.61 | 71.02 | 71.63 | 64.27 | 70.16 | 72.36 |
|        | CMFH | 74.48 | 74.92 | 75.38 | 65.92 | 68.00 | 69.64 | 62.16 | 69.09 | 71.17 |
|        | ACQ | 72.77 | 73.91 | 74.18 | 63.68 | 67.13 | 68.72 | 63.28 | 66.38 | 69.45 |
|        | FSH | 73.35 | 75.03 | 76.28 | 65.89 | 67.55 | 69.10 | 67.05 | 72.78 | 75.67 |
|        | DJSRH | 75.34 | 76.48 | 77.51 | 67.22 | 71.01 | 72.27 | 64.55 | 72.44 | 74.94 |
|        | CMIMH | 77.36 | 78.42 | 79.01 | 70.18 | 72.33 | 72.72 | 68.15 | 73.28 | 74.89 |
query sets, and the rest as training set and also the retrieval database. We present that the retrieval performance in terms of mAP (of all retrieved samples) is shown in Table VI. We can observe that our proposed CMIMH consistently outperforms MGAH for both MIR-Flickr25k and NUS-WIDE datasets, especially for MIR-Flickr25k where the improvement gaps are greater than 5%. In comparison with UKD [87], our proposed method achieves lower performance on the NUS-WIDE dataset—Txt→Img task while still outperforming UKD on the NUS-WIDE dataset—Img→Txt task and MIR-Flickr25k for both tasks. These results show that our proposed method is still more favorable than UKD.

3) Comparison With Learning Disentangled Representation for Cross-Modal Retrieval With Deep MI Estimation (LDR) [70]: To have a fair comparison, we follow the experiment setting mentioned in [70]. Specifically, we extract
TABLE V
Comparison With UDCMH [39] Using mAP@50 on the MIR-Flickr25k and NUS-WIDE Datasets. The Results of UDCMH Are Cited From [39].

| Task   | Method | MIR-Flickr25k | NUS-WIDE |
|--------|--------|---------------|----------|
| Img → Ttxt | UDCMH | 68.9 | 69.8 | 71.4 | 51.1 | 51.9 | 52.4 |
| Txt → Img | CMIMH | 83.2 | 86.5 | 88.2 | 74.52 | 78.23 | 79.75 |

TABLE VI
Comparison With MGAH [66] and UKD [87] Using mAP (%) on the MIR-Flickr25k and NUS-WIDE Datasets. The Results of MGAH and UKD Are Cited From Original Articles [66], [87].

| Task   | Method | MIR-Flickr25k | NUS-WIDE |
|--------|--------|---------------|----------|
| Img → Ttxt | MGAH | 68.5 | 69.3 | 70.4 | 61.3 | 62.3 | 62.8 |
| Txt → Img | UKD-SS | 71.4 | 71.8 | 72.3 | 61.4 | 63.7 | 63.8 |
| Img → Ttxt | CMIMH | 74.15 | 74.65 | 75.98 | 62.72 | 63.83 | 64.75 |
| Txt → Img | UKD-SS | 67.3 | 67.8 | 68.6 | 60.3 | 61.4 | 64.0 |
| Img → Ttxt | CMIMH | 72.27 | 72.44 | 75.41 | 62.41 | 63.70 | 64.23 |

TABLE VII
Comparison With LDR [70] Using RecallOne@K (%) on MS-COCO. The Results of LDR Are Cited From [70].

| Method       | K = 1 | K = 10 |
|--------------|-------|--------|
| LDR (1024-D) | 53.4  | 91.3   |
| CMIMH (128 bits) | 65.53 | 94.35  | 69.82 | 93.32 |

VGG19 FC7 feature for images. We randomly select 1000 images as query, 1000 images as validation, and the remaining images as database as well as training set. We report in Table VII the retrieval performance in terms of RecallOne@K, the percent of queries for which the ground truth is one of the first K retrieved. Guo et al. [70] reported the RecallOne@K for 1024-dimensional real-valued features (32 768 bits), and we find out that even with 128 bit hash codes, our proposed method can outperform LDR by large margins for both RecallOne@1 and RecallOne@10.

a) Comparison with state of the art using handcrafted image feature: Following the experiment setting of collective reconstructive embedding (CRE) [30] and FSH [29], we conduct experiments with handcrafted features on NUS-WIDE datasets. For the NUS-WIDE dataset, each image is represented by 500-dimensional BOW scale-invariant feature transform (SIFT) features and each text is represented by a 1000-dimensional preprocessed BOW feature. We randomly select 2000 pairs as the query set; the remaining are used as the database. We also sample 20,000 pairs from the database as the training set. We present the experiment results in terms of mAP (of all returned samples) and Pre@100 (as used in [29] and [30]) in Table VIII. We can observe that our proposed method can also work well with handcrafted features and outperforms all compared methods.

b) Effect of training size: Different from most of the state-of-the-art methods [23], [24], [26], [28], [29], [39], our proposed method can be fully optimized using gradient descent in a mini-batch manner. Therefore, our proposed method can be easily trained with much larger training sets. We further analyze the effects on retrieval performance when varying the training size on the NUS-WIDE dataset. We also compare our retrieval performance with the retrieval performance of DJSRH [38], which is one of our most competitive methods and can also be trained in a mini-batch manner. The retrieval performance in terms of mAP@1k when using different training set sizes is shown in Table IX. We can observe that our proposed method can achieve higher mAP@1k when utilizing more training data. Furthermore, our proposed method also consistently outperforms DJSRH [38] with different training sizes.

V. Conclusion
In this article, inspired by recent advances in learning representation by maximizing MI, we proposed a novel framework,
dubbed CMIMH. By assuming the binary representations to be modeled by multivariate Bernoulli distributions, we can maximize the MI effectively using gradient descent optimization in a mini-batch manner via maximizing their estimated variational lower bounds. We additionally find out that trying to minimize the modality gap by learning similar binary representations for the same instance from different modalities could result in modality-private information loss. Properly balancing the modality gap and modality-private information loss is important to achieve better performance. Experiment results confirm the effectiveness of our proposed method for both cross- and single-modal retrieval tasks. In addition, the ablation studies clearly justify the advantages of different components in our proposed method.

VI. MORE DETAIL DERIVATION OF (8)

$I(\mathbf{x}^i; \mathbf{h}^i|\mathbf{x}^j)$ as shown at the top of the page.

Note that since $\mathbf{h}^i$ is completely determined by $\mathbf{x}^i$, $H(\mathbf{h}^i|\mathbf{x}^j)=0$. Consequently, $\mathbf{h}^i$ and $\mathbf{x}^i$ are conditionally independent given $\mathbf{x}^j$, i.e.,

$$P_0(\mathbf{h}^i = h|\mathbf{x}^j = x^j, \mathbf{x}^j = x^j) = P_0(\mathbf{h}^i = h|\mathbf{x}^j = x^j).$$  \hfill (16)

REFERENCES

[1] Y. Zhen and D.-Y. Yeung, “Co-regularized hashing for multimodal data,” in Proc. NIPS, 2012, pp. 1376–1384.
[2] Y. Wei, Y. Song, Y. Zhen, B. Liu, and Q. Yang, “Heterogeneous translated hashing: A scalable solution towards multi-modal similarity search,” ACM Trans. Knowl. Discovery Data, vol. 10, no. 4, pp. 1–28, Jul. 2016.
[3] D. Zhang and W.-J. Li, “Large-scale supervised multimodal hashing with semantic correlation maximization,” in Proc. AAAI, 2014, pp. 2177–2183.
[4] B. Wu, Q. Yang, W.-S. Zheng, Y. Wang, and J. Wang, “Quantized correlation hashing for fast cross-modal search,” in Proc. IJCAI, 2015, pp. 3946–3952.
[5] Z. Lin, G. Ding, M. Hu, and J. Wang, “Semantics-preserving hashing for cross-view retrieval,” in Proc. CVPR, 2015, pp. 3864–3872.
[6] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, “Learning discriminative binary codes for large-scale cross-modal retrieval,” IEEE Trans. Image Process., vol. 26, no. 5, pp. 2494–2507, May 2017.
[7] Q.-Y. Jiang and W.-J. Li, “Discrete latent factor model for cross-modal hashing,” IEEE Trans. Image Process., vol. 28, no. 7, pp. 3490–3501, Jul. 2019.
[8] J. Tang, K. Wang, and L. Shao, “Supervised matrix factorization hashing for cross-modal retrieval,” IEEE Trans. Image Process., vol. 25, no. 7, pp. 3157–3166, Jul. 2016.
[9] D. Mandal, K. N. Chaudhury, and S. Biswas, “Generalized semantic preserving hashing for cross-modal retrieval,” IEEE Trans. Image Process., vol. 28, no. 1, pp. 102–112, Jan. 2019.
[10] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, “Triplet-based deep hashing network for cross-modal retrieval,” IEEE Trans. Image Process., vol. 27, no. 8, pp. 3893–3903, Aug. 2018.
[11] Z. Cao, M. Long, J. Wang, and Q. Yang, “Transitive hashing network for heterogeneous multimedia retrieval,” in Proc. AAAI, 2017, pp. 81–87.
[12] Q.-Y. Jiang and W.-J. Li, “Deep cross-modal hashing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3270–3278.
[13] Z.-D. Chen, W.-J. Yu, C.-X. Li, L. Nie, and X.-S. Xu, “Dual deep neural networks cross-modal hashing,” in Proc. AAAI, 2018, pp. 274–281.
[14] V. E. Liong, J. Lu, Y. Tan, and J. Zhou, “Cross-modal deep variational hashing,” in Proc. ICCV, 2017, pp. 4077–4085.
[15] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, “Pairwise relationship guided deep hashing for cross-modal retrieval,” in Proc. AAAI, 2017, pp. 1618–1625.
[16] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, “Deep visual-semantic hashing for cross-modal retrieval,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1445–1454.
[17] C.-X. Li, T.-K. Yan, X. Luo, L. Nie, and X.-S. Xu, “Supervised robust discrete multimodal hashing for cross-media retrieval,” IEEE Trans. Multimedia, vol. 21, no. 11, pp. 2863–2877, Nov. 2019.
[18] G. Song, D. Wang, and X. Tan, “Deep memory network for cross-modal retrieval,” IEEE Trans. Multimedia, vol. 21, no. 5, pp. 1261–1275, May 2019.
[19] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, “Multi-task consistency-preserving adversarial hashing for cross-modal retrieval,” IEEE Trans. Image Process., vol. 29, pp. 3626–3637, 2020.
[20] L. Jin, K. Li, Z. Li, F. Xiao, G.-J. Qi, and J. Tang, “Deep semantic-preserving ordinal hashing for cross-modal similarity search,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 5, pp. 1429–1440, May 2019.
[21] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, “Adversarial cross-modal retrieval,” in Proc. 25th ACM Int. Conf. Multimedia, Oct. 2017, pp. 154–162.
[22] F. Wu et al., “Modality-specific and shared generative adversarial network for cross-modal retrieval,” Pattern Recognit., vol. 104, Aug. 2020, Art. no. 107335.
[23] S. Kumar and R. Udupa, “Learning hash functions for cross-view similarity search,” in Proc. IJCAI, 2011, pp. 1360–1365.
[24] M. Rastegari, J. Choi, S. Fakhrzai, D. Hal, and L. Davis, “Predictable dual-view hashing,” in Proc. ICML, 2013, pp. 1328–1336.
[25] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, “Inter-media hashing for large-scale retrieval from heterogeneous data sources,” in Proc. ACM SIGMOD, 2013, pp. 783–796.
[26] G. Ding, Y. Guo, J. Zhou, and Y. Gao, “Large-scale cross-modality search via collective matrix factorization hashing,” IEEE Trans. Image Process., vol. 25, no. 11, pp. 5427–5440, Nov. 2016.
D. Barber and F. Agakov, “The IM algorithm: A variational approach for unsupervised DCMH,” in Proc. ACM SIGIR, 2014, pp. 415–424.

G. Irie, H. Arai, and Y. Taniguchi, “Alternating co-quantization for cross-modal hashing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4217–4226.

J. Zhou, G. Ding, and Y. Guo, “Latent semantic sparse hashing for cross-modal similarity search,” in Proc. IEEE Trans. Inf. Theory, vol. 56, no. 10, pp. 5394–5406, Oct. 2010.

S. Watanabe, “Information theoretical analysis of multivariate correlation,” IBM J. Res. Develop., vol. 4, no. 1, pp. 66–82, Jan. 1960.

T. Hoang, T.-T. Do, H. H. Le-Tan, and N.-M. Cheung, “Simultaneous compression and quantization: A joint approach for efficient unsupervised hashing,” Comput. Vis. Image Understand., vol. 191, Feb. 2020, Art. no. 102852.

J. Guo and W. Zhu, “Collective affinity learning for partial cross-modal hashing,” IEEE Trans. Circuits Syst. Video Technol., vol. 27, no. 7, pp. 1140–1153, Jul. 2017.

T. Hoang, T.-T. Do, H. H. Le-Tan, and N.-M. Cheung, “Simultaneous compression and quantization: A joint approach for efficient unsupervised hashing,” Comput. Vis. Image Understand., vol. 191, Feb. 2020, Art. no. 102852.

J. Wang, S. Kumar, and S. Chang, “Semi-supervised hashing for scalable image retrieval,” in Proc. CVPR, 2010, pp. 3424–3431.

J. Zhou, G. Ding, and Y. Guo, “Latent semantic sparse hashing for cross-modal similarity search,” in Proc. ACM SIGIR, 2014, pp. 415–424.
[82] M. J. Huiskes and M. S. Lew, “The MIR Flickr retrieval evaluation,” in ACM Int. Conf. Multimedia Inf. Retr., 2008, pp. 39–43.

[83] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, “NUSWIDE: A real-world web image database from National University of Singapore,” in Proc. CIVR, 2009, pp. 1–8.

[84] T. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. ECCV, 2014, pp. 740–755.

[85] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805.

[86] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. NeurIPS, 2012, pp. 2–9.

[87] H. Hu, L. Xie, R. Hong, and Q. Tian, “Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 3120–3129.

Tuan Hoang received the bachelor’s degree in electrical engineering from Portland State University, Portland, OR, USA, in 2014. He is currently pursuing the Ph.D. degree with the Singapore University of Technology and Design (SUTD), Singapore. His research interests are image retrieval and image hashing.

Thinh-Toan Do received the Ph.D. degree in computer science from the French National Institute for Research in Computer Science and Control (INRIA), Rennes, France, in 2012. From 2013 to 2016, he was a Research Fellow with the Singapore University of Technology and Design, Singapore. From 2016 to 2018, he was a Research Fellow with the Australian Centre for Robotic Vision and The University of Adelaide, Adelaide, SA, Australia. From 2018 to 2020, he was a Lecturer with the University of Liverpool, Liverpool, U.K. He is currently a Senior Lecturer with the Faculty of Information Technology, Monash University, Clayton, VIC, Australia. His research interests include computer vision and machine learning.

Tam V. Nguyen (Senior Member, IEEE) received the Ph.D. degree from the National University of Singapore (NUS), Singapore, in 2013. He is currently an Assistant Professor and the Director of the Vision and Mixed Reality Laboratory, Department of Computer Science, University of Dayton, Dayton, OH, USA. His research interests include computer vision, machine learning, mixed reality, and multimedia analysis.

Ngai-Man Cheung (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, CA, USA. He is currently an Associate Professor with the Singapore University of Technology and Design (SUTD), Singapore. His research interests include image and signal processing, computer vision, and AI. His research has resulted in 14 U.S. patents granted with several pending. Two of his inventions have been licensed to companies. One of his research results has led to a SUTD spinoff on AI for wound care. His research has also been featured in the National Artificial Intelligence Strategy.

Dr. Cheung has received several research recognitions, including recently the Best Paper Finalist at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2019, the Finalist of Super AI Leader (SAIL) Award at the World AI Conference (WAIC) 2019, and the Excellence in Research Award from SUTD.