Cycle-Consistent Deep Generative Hashing for Cross-Modal Retrieval

Lin Wu, Yang Wang, and Ling Shao, *Senior Member, IEEE*

***Abstract*— In this paper, we propose a novel deep generative approach to cross-modal retrieval to learn hash functions in the absence of paired training samples through the cycle consistency loss. Our proposed approach employs adversarial training scheme to learn a couple of hash functions enabling translation between modalities while assuming the underlying semantic relationship. To induce the hash codes with semantics to the input-output pair, cycle consistency loss is further delved into the adversarial training to strengthen the correlation between the inputs and corresponding outputs. Our approach is generative to learn hash functions, such that the learned hash codes can maximally correlate each input–output correspondence and also regenerate the inputs so as to minimize the information loss. The learning to hash embedding is thus performed to jointly optimize the parameters of the hash functions across modalities as well as the associated generative models. Extensive experiments on a variety of large-scale cross-modal data sets demonstrate that our proposed method outperforms the state of the arts.**

***Index Terms*— Cross-modal retrieval, generative hash, cycle-consistency.**

I. Introduction

THE sheer volumes of big multimedia data with different modalities, including images, videos, and texts, are now mixed together and represent comprehensive knowledge to perceive the real world. It thus attracts increasing attention to approximate nearest neighbor search across different media modalities that brings both computation efficiency and search quality. Naturally, entities in correspondence from heterogenous modalities may endow semantic correlations, and it tends to entail cross-modal retrieval that returns relevant search results from one modality as response to query of another modality, e.g., retrieval of texts/images by using a query image/text, as shown in Fig.2.

A viable solution to large-volume cross-modal retrieval is to develop hash methods that learns compact binary codes as similar/dissimilar as possible if they have the same/different semantics. However, effective and efficient cross-modal hashing remains a challenge due to the heterogeneity across divergent modalities [1], and the semantic gap between low-level features and high-level semantics. A large body of cross-modal hashing methods are proposed to learn projections from different modalities into an independent semantic embedding space with respect to characterize the model-specific relationship [2], [3]. However, these shallow methods essentially learn a single pair of linear or non-linear projections to map each example into its binary codes. Optimizing single projections towards each modality is suboptimal, and also the low-level descriptions on images are limited in expressing their high-level semantics [4]–[7]. Recent models based on deep learning are developed for cross-modal hashing [8]–[14]. These supervised deep models utilize semantic labels to enhance the correlation of cross-modal data wherein the feature transformations and hash functions can be jointly learned in an end-to-end manner.

It is admitted that cross-modal retrieval has been made towards substantial progression by the promotion on deep learning models. We remark that there are two major challenges remained open to be addressed: First, the cross-modal hashing is performed to learn the mapping between an input image/text and an output text/image using a training set of labeled *aligned* pairs. The supervision of paired correspondence is to enhance the correlation of cross-modal data such that the hashing can be guided by preserving the semantics. For instance, Zhang and Li [15] performed semantic correlation maximization using label information to learn modality-specific transformations. However, for many

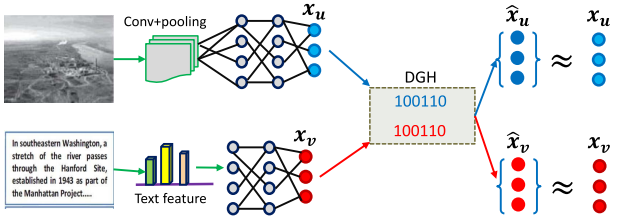


Fig. 1. The architecture overview of our approach. Given two modalities of images and texts, we learn coupled generative hash functions in the absence of paired correspondence through the cycle-consistent adversarial learning. This is achieved by training two mappings: $\boldsymbol{G} : \boldsymbol{X}_u \rightarrow \boldsymbol{X}_v$ and $\boldsymbol{F} : \boldsymbol{X}_v \rightarrow \boldsymbol{X}_u$, which are reverse of each other, and adding a cycle consistency loss into the adversarial loss that yields $\boldsymbol{F}(\boldsymbol{G}(\boldsymbol{x}_u)) \approx \boldsymbol{x}_u$, and $\boldsymbol{G}(\boldsymbol{F}(\boldsymbol{x}_v)) \approx \boldsymbol{x}_v$. The deep generative model performs both hash function learning and regeneration inputs from binary codes. See text for details.

Manuscript received April 21, 2018; revised August 19, 2018 and October 11, 2018; accepted October 28, 2018. Date of publication October 31, 2018; date of current version November 28, 2018. The work of Y. Wang was supported by NSFC under Grant 61806035. The work of L. Shao was supported in part by the Shenzhen Government under Grant GJHZ20180419190732022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yun Fu. *(Corresponding author: Yang Wang.)*

L. Wu is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230000, China, and also with The University of Queensland, St Lucia, QLD 4072, Australia (e-mail: jolin.lwu@gmail.com).

Y. Wang is with the Faculty of Electronic Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: yang.wang@dlut.edu.cn).

L. Shao is with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates (e-mail: ling.shao@ieee.org).

Digital Object Identifier 10.1109/TIP.2018.2878970

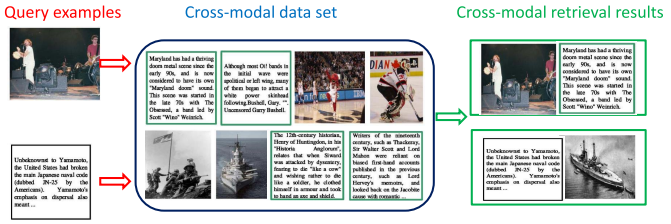


Fig. 2. The illustration on cross-modal retrieval with images and texts.

realistic cases, paired training data are available. Even labeling is feasible, deep hashing models trained on limited amount of labeled samples are inclined to be over-fitting, and thus the generalization is not guaranteed. Second, transforming an input which typically in high-dimension into its binary codes will inevitably cause information loss. Existing hashing methods unidirectionally learn hash functions from inputs to hash codes. However, the hash codes can also be used to regenerate the inputs [16], [17], which should be exploited to characterize the goodness of hash codes and regenerating inputs through hash codes provides a principle to minimize the information loss during the hash embedding.

Recent works have shown that generative adversarial networks combined with cycle-consistency constraints are surprisingly effective at mapping images between domains, even without the use of aligned image pairs [18], [19]. We therefore seek an algorithm that can learn to translate the binary codes between modalities without paired image-text examples. As shown in [18], there is some underlying relationship between the two modalities, for example, they are two different renderings of the same underlying scene and the relationship can be learnt by exploiting the supervision at the level of the sets: an adversary objective can be trained to induce an output distribution over the target modality that matches its empirical distribution. The optimal mapping thereby translates the source modality to a new modality distributed identically to the new one, whereas a cycle-consistent constraint is needed to guarantee an individual input and output are paired up in a meaningful way.

A. Our Approach: CYC-DGH

For addressing the two key issues in large-scale cross-modal retrieval, we propose Cycle-Consistent Deep Generative Hashing (CYC-DGH), which aims to produce hash embedding in the absence of paired training correspondence. The basic idea of the proposed approach is shown in Fig.1.

Specifically, our goal is to learn a couple of hash mappings that can translate between domains/modalities without paired input-output examples. We assume there is some underlying relationship between the domains, for example, they are images and texts of the same semantic meaning, and seek that relationship. Although we lack supervision in the form of paired examples, we can exploit supervision at the level of sets: given one set of images in modality $\boldsymbol{X}_u$ and a different set of texts in modality $\boldsymbol{X}_v$. We may train a mapping $\boldsymbol{G} : \boldsymbol{X}_u \rightarrow \boldsymbol{X}_v$ such that the output $\hat{\boldsymbol{x}}_v = \boldsymbol{G}(\boldsymbol{x}_u)$, $\boldsymbol{x}_u \in \boldsymbol{X}_u$, is indistinguishable from texts $\boldsymbol{x}_v \in \boldsymbol{X}_v$ by an adversary trained to classify $\hat{\boldsymbol{x}}_v$ apart from $\boldsymbol{x}_v$. The optimal $\boldsymbol{G}$ thereby translates the modality $\boldsymbol{X}_u$ into a modality $\hat{\boldsymbol{X}}_v$ distributed identically to $\boldsymbol{X}_v$. As a result, the modality gap can be reduced effectively. However, such a translation is highly under-constrained, and does not guarantee that an individual input $\boldsymbol{x}_u$ and an output $\boldsymbol{x}_v$ are matched in a meaningful way. In fact, there could be infinitely mappings $\boldsymbol{G}$ that will induce the same distribution over $\hat{\boldsymbol{x}}_v$. Towards this end, we exploit the property that the translation should be "cycle-consistent" [18], in the sense that if we translate, e.g., a sentence from English to French, and then translate it back from French to English, we should arrive back at the original sentence. In the case of cross-modality, if we have a domain translator $\boldsymbol{G} : \boldsymbol{X}_u \rightarrow \boldsymbol{X}_v$, and another translator $\boldsymbol{F} : \boldsymbol{X}_v \rightarrow \boldsymbol{X}_u$, then both $\boldsymbol{G}$ and $\boldsymbol{F}$ should be reverse of each other, and thus bijections. Hence, we train the mappings $\boldsymbol{G}$ and $\boldsymbol{F}$ simultaneously by combining a cycle consistency loss [20] with adversarial losses on modalities $\boldsymbol{X}_u$ and $\boldsymbol{X}_v$ that encourages $\boldsymbol{F}(\boldsymbol{G}(\boldsymbol{x}_u)) \approx \boldsymbol{x}_u$, and $\boldsymbol{G}(\boldsymbol{F}(\boldsymbol{x}_v)) \approx \boldsymbol{x}_v$.

To allow the regeneration from hash codes to the inputs so as to minimize the information loss, we decompose the mappings into: $\boldsymbol{G} : \boldsymbol{x}_u \rightarrow \boldsymbol{H}_u \rightarrow \boldsymbol{P}_u \rightarrow \boldsymbol{x}_v$, where $\boldsymbol{H}$ denotes the binary code learning and $\boldsymbol{P}$ is the reverse process of regenerating inputs from binary codes. Finally, the cycle consistent training can bring each input $\boldsymbol{x}_u$ back to itself through the generative model: $\boldsymbol{x}_u \rightarrow \boldsymbol{G}(\boldsymbol{x}_u) \rightarrow \boldsymbol{F}(\boldsymbol{G}(\boldsymbol{x}_u)) \approx \boldsymbol{x}_u$. The similar translation can also be applied on $\boldsymbol{F}$, which is omitted for simplicity. The proposed generative model which captures both the encoding of binary codes from the input and the decoding of input from binary codes, provides a principled hash learning framework, where the information loss during hash embedding is minimized. Therefore, the generated codes can compress the input data with maximum preserving of its information on its own domain as well as the relationship of samples from different modalities. And also the modality gap between the hash functions are reduced. Prior works on binary auto-encoders [11] and deep generative models [9], [21] also take a generative view of hashing but still require the correlation from paired samples. Generative hashing is introduced in [16] where the hash functions can be learned through minimum description length. However, their algorithm is limited in single-modality setting.

B. Contributions

The main contributions can be summarized as follows.

- A cycle-consistent based deep generative hashing approach for cross-modal retrieval, namely CYC-DGH, is presented to perform hash function learning without pair level training samples in correspondence.
- The formulation of cycle transformation is introduced into the cross-modal adversarial training to optimize the mappings that can be translated across modalities with the underlying semantics between the input and the output maximally guaranteed.
- A deep generative model is proposed to learn to regenerate the input from binary codes, which is coupled with hash function learning, and thus demonstrated to be able to reconstruct the inputs so as to minimize the information loss during the hash embedding.

The rest of this paper is organized as follows: We briefly introduce the related works on generative adversarial networks, cross-modal hashing methods as well as cycle consistency literature in Section II. Section III presents our proposed CYC-DGH approach. Section IV includes the experiments of cross-modal retrieval conducted on three cross-modal data sets with result analyses. Finally Section V concludes the paper.

II. Related Work

A. Generative Adversarial Networks

Recently, generative adversarial networks (GANs) have been proposed to estimate a generative model by an adversarial training process [22], and GANs-based networks can be used to generate new data such as image synthesis [23] and video prediction [24]. In [25], semi-supervised GANs are exploited to deal with few labeled training data by producing synthetic examples conditioning on class labels. Due to the strong ability of GANs in modeling data distribution, GANs have been utilized to model the joint distribution over the heterogenous data of different modalities [26]. However, Peng *et al.* [26] use GANs to learn the common representation and boost the cross-modal correlation learning. This is a different goal of this paper which is to learn hash functions without the dependence on paired training samples across modalities. Some advanced models are developed to transfer cross-view sentence-image retrieval problem into a single-view problem. For instance, an end-to-end differentiable architecture from the character level to pixel level is proposed by using a model conditioned on text descriptions to achieve sentences to corresponding image synthesis [27]. Inspired by this idea, Zhao *et al.* [21] turn cross-view hashing into single-view by generating fake images from text feature and jointly learn hash functions.

B. Deep Cross-Modal Hashing

Cross-modal multimedia retrieval performs the task of retrieving text documents by using a given query image and vice versa. The objective for many cross-modal methods is learn a common subspace between images and texts to model the correlations [28]–[33]. For example, in the literature canonical component analysis (CCA) is used to map both text documents and images into a latent space [28]. Wang *et al.* [29] learn a coupled feature space to select the most relevant and discriminative features for cross-modal matching.

In order to suit large-scale search, cross-modal hashing becomes a more desirable choice for efficiency [2]. The majority of cross-modal hash methods can be classified into two types: unsupervised (CVH [34], LSSH [35]) and supervised (SCM [15], SePH [1]). Unsupervised methods utilize co-occurrence information such that only the image-text pairs which occur in the same article are considered to be of similar semantic. For example, Zhou *et al.* [35] obtain a unified binary from a latent space learning method by using sparse coding and matrix factorization in the common space. On the other hand, supervised methods utilize semantic labels to enhance the correlation of cross-modal pairs. For instance, a semantic correlation maximization (SCM) is performed to use label information to learn a modality-specific transformation which can maximize the correlation between modalities. However, these studies are in shallow form in the sense that they only perform a single-layer of linear or non-linear transformation.

While there are a large body of methods that perform deep learning for cross-modal retrieval by integrating feature learning and hashing coding into end-to-end trainable frameworks [8], [9], [36], [37]. For instance, Cao *et al.* [8] learn a visual semantic fusion network with cosine hinge loss to obtain the binary codes and learn modality-specific networks to obtain the hash functions. To make the hashing network suitable for out-of-sample extension, a cross-modal deep variational hashing method (CMDVH) [9] is proposed to reformulate the modality-specific hashing networks into a generative form. They introduce a set of latent variables that are modeled similar to the inferred unified binary codes for each paired image/text sample through a plausible log likelihood criterion. However, this work still performs binary code inference from labeled training pairs to seek a common hamming space. In contrast, our approach is effective in reducing the modality gap in the absence of pairs by the merit of cycle-consistency loss in adversarial training to encourage the modality alignment. HashGAN [38] is to propose an adversarial hashing network with attention mechanism to enhance the measurement of content similarities by selectively focusing on informative parts of multi-modal data. Shen *et al.* [39] present a zero-shot sketch-image hashing (ZSIH) model to address the never-seen observation in training. The modality gap and semantic correlation between sketch-image can be enhanced by using a Kronecker fusion layer and graph convolution. They also formulate a generative hashing scheme in reconstructing semantic knowledge representations for zero-shot retrieval. However, the ZSIH model is dependent on paired sketch-image for training which is somewhat limited in practical situations.

C. Cycle-Consistency Constraints and Applications

The idea of using transitivity as a way to regularize structured data is receiving increasingly attention. One application is the task of image-to-image translation which is to translate one possible representation of a scene into another, given sufficient training data. More recent approaches use a collection of input-output examples to learn parametric translation function based on CNNs [40]. For instance, the "pix2pix" framework [41] investigates a conditional generative adversarial network to learn a mapping from input to output images. The idea of conditional adversarial nets is a general-purpose solution which has been applied into various tasks such as generating photographs from sketches [42]. However, these works learn the mapping with dependency on paired training examples. To tackle with unpaired setting, some methods are proposed to relate two data domains. For instance, CoGAN [43] and cross-modal scene networks [44] use a weight-sharing strategy to learn a common representation across domains. Recent studies show that higher-order cycle consistency can be used in depth estimation [45], co-segmentation [46] and domain adaptation [19] in which a cycle consistency loss can be used as a way of using transitivity to supervise the CNN training. For instance, Zhu *et al.* [18] introduce a similar loss to push the mappings to be consistent with each other

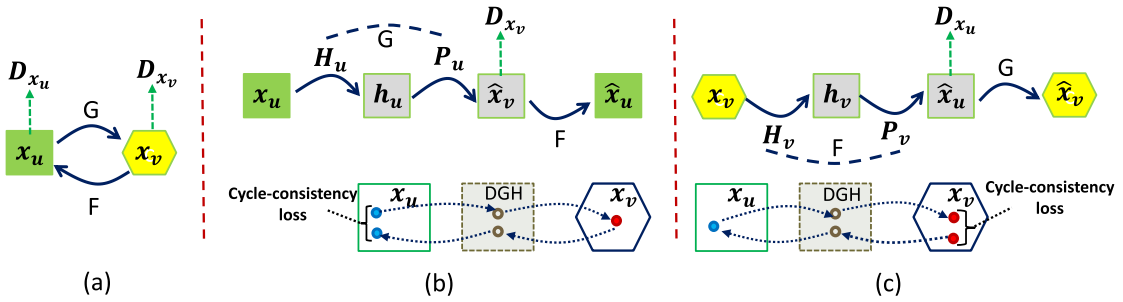


Fig. 3. The proposed cycle-consistent deep generative hashing (CYC-DGH) for cross-modal retrieval. (a) The model of CYC-DGH couples two mappings: $\boldsymbol{G}: \boldsymbol{x}_u \rightarrow \boldsymbol{x}_v$ and $\boldsymbol{F}: \boldsymbol{x}_v \rightarrow \boldsymbol{x}_u$ as well as associated adversarial discriminators $\boldsymbol{D}_{\boldsymbol{x}_v}$ and $\boldsymbol{D}_{\boldsymbol{x}_u}$. The two mappings are decomposed into the binary code generation and the reverse process of regenerating inputs from binary codes: $\boldsymbol{G}: \boldsymbol{x}_u \rightarrow \boldsymbol{H}_u \rightarrow \boldsymbol{P}_u \rightarrow \boldsymbol{x}_v$ and $\boldsymbol{F}: \boldsymbol{x}_v \rightarrow \boldsymbol{H}_v \rightarrow \boldsymbol{P}_v \rightarrow \boldsymbol{x}_u$. To regulate the mappings, two cycle-consistent losses are introduced: (b) forward $\boldsymbol{x}_u \rightarrow \boldsymbol{G}(\boldsymbol{x}_u) \rightarrow \boldsymbol{F}(\boldsymbol{G}(\boldsymbol{x}_u)) \approx \hat{\boldsymbol{x}}_u$, and (c) backward $\boldsymbol{x}_v \rightarrow \boldsymbol{F}(\boldsymbol{x}_v) \rightarrow \boldsymbol{G}(\boldsymbol{F}(\boldsymbol{x}_v)) \approx \hat{\boldsymbol{x}}_v$.

and demonstrate its superiority in image-to-image translation. Concurrent with [18], Yi *et al.* [47] use a similar objective for unpaired image-to-image translation, inspired by the dual learning in machine translation. Also, the cycle-consistency constraints can be combined with generative adversarial networks and shown to be surprisingly effective at mapping images between domains [19]. Our work is clearly inspired by the cycle-consistency loss, which is introduced into the cross-modal hashing learning to alleviate the pairing on training samples. In addition, we extend the cycle-consistency pipeline by enforcing the hash mappings to be translated without paired training samples.

III. Cycle-Consistent Deep Generative Hashing

In this section, we propose an end-to-end deep architecture for cross-modal hashing such that we are able to maximize the correlation between the two modalities in the absence of paired training examples. The network composes of generative hash functions in regards to two modalities for generating the binary codes from input as well as reversely generating input from binary codes, and mapping functions between two modalities without paired correspondence.

Let $\boldsymbol{X}_u = [\boldsymbol{x}_{u1}, \boldsymbol{x}_{u2}, \cdots, \boldsymbol{x}_{uN}] \in \mathbb{R}^{d_u \times N}$ and $\boldsymbol{X}_v = [\boldsymbol{x}_{v1}, \boldsymbol{x}_{v2}, \cdots, \boldsymbol{x}_{vN}] \in \mathbb{R}^{d_v \times M}$ be the training samples from different modalities, where u and v denote two different modalities, and each sample $\boldsymbol{x}_u$ ($\boldsymbol{x}_v$) is produced by the neural network with parameters $f_u(\cdot)$ ($f_v(\cdot)$). Our objective contains three types of terms: adversarial losses [22] for matching the distribution of generated samples to the data distribution in the target modality; the cycle-consistency losses [18] to present the learned mappings $\boldsymbol{G}$ and $\boldsymbol{F}$ from contradicting each other; the reconstruction loss of reconstructing the input from the binary codes. In the following, we present the respective losses and show how to perform optimizations in the proceeding subsections.

A. Adversarial Loss

We denote the data distribution as $\boldsymbol{x}_u \sim p_{data}(\boldsymbol{x}_u)$ and $\boldsymbol{x}_v \sim p_{data}(\boldsymbol{x}_v)$. As shown in Fig. 3 (a), for the cross-modal correlation, our model includes two mappings $\boldsymbol{G}: \boldsymbol{x}_u \rightarrow \boldsymbol{x}_v$ and $\boldsymbol{F}: \boldsymbol{x}_v \rightarrow \boldsymbol{x}_u$. In addition, we introduce two adversarial discriminators: $\boldsymbol{D}_{\boldsymbol{x}_u}$ and $\boldsymbol{D}_{\boldsymbol{x}_v}$, where $\boldsymbol{D}_{\boldsymbol{x}_u}$ aims to distinguish between images $\{\boldsymbol{x}_u\}$ and translated texts $\{\boldsymbol{F}(\boldsymbol{x}_v)\}$; in the same way, $\boldsymbol{D}_{\boldsymbol{x}_v}$ aims to discriminate between $\{\boldsymbol{x}_v\}$ and $\{\boldsymbol{G}(\boldsymbol{x}_u)\}$. For the mapping function $\boldsymbol{G}: \boldsymbol{x}_u \rightarrow \boldsymbol{x}_v$ and its discriminator $\boldsymbol{D}_{\boldsymbol{x}_v}$, the objective can be expressed as:

$$\begin{aligned}\mathcal{L}_{GAN}(\boldsymbol{G}, \boldsymbol{D}_{\boldsymbol{x}_v}, \boldsymbol{x}_u, \boldsymbol{x}_v) = \mathbb{E}_{\boldsymbol{x}_v \sim p_{data}(\boldsymbol{x}_v)}[\log \boldsymbol{D}_{\boldsymbol{x}_v}(\boldsymbol{x}_v)] \\ + \mathbb{E}_{\boldsymbol{x}_u \sim p_{data}(\boldsymbol{x}_u)}[\log(1 - \boldsymbol{D}_{\boldsymbol{x}_v}(\boldsymbol{G}(\boldsymbol{x}_u)))], \quad (1)\end{aligned}$$

where $\boldsymbol{G}$ attempts to generate images $\boldsymbol{G}(\boldsymbol{x}_u)$ that look similar to items from domain $\boldsymbol{X}_v$, while $\boldsymbol{D}_{\boldsymbol{x}_v}$ aims to distinguish between translated samples $\boldsymbol{G}(\boldsymbol{x}_u)$ and real samples $\boldsymbol{x}_v$. $\boldsymbol{G}$ aims to minimize the objective against an adversary $\boldsymbol{D}_{\boldsymbol{x}_v}$ that tries to maximize it, i.e., $\min_{\boldsymbol{G}} \max_{\boldsymbol{D}_{\boldsymbol{x}_v}} \mathcal{L}_{GAN}(\boldsymbol{G}, \boldsymbol{D}_{\boldsymbol{x}_v}, \boldsymbol{x}_u, \boldsymbol{x}_v)$. Similarly, an adversarial loss is introduced for the mapping function $\boldsymbol{F}: \boldsymbol{x}_v \rightarrow \boldsymbol{x}_u$ and its discriminator $\boldsymbol{D}_{\boldsymbol{x}_u}$: i.e., $\min_{\boldsymbol{F}} \max_{\boldsymbol{D}_{\boldsymbol{x}_u}} \mathcal{L}_{GAN}(\boldsymbol{F}, \boldsymbol{D}_{\boldsymbol{x}_u}, \boldsymbol{x}_v, \boldsymbol{x}_u)$.

B. Cycle-Consistency Loss

One characteristic of cross-modal retrieval is to minimize the modality discrepancy, and thus the semantics can be consented and represented by objects in different modalities. Existing methods typically deploy the paired correspondence to supervise the learning of a common subspace between images and text [2], [29] or the translation between the two domains [21]. However, manually labelling pairs is not viable in web-scale multimedia retrieval. In addition, learning a mapping between a number of specific image-text pairs is limited in producing a general-purpose solution of capturing high-level correspondences in many vision tasks. To this end, cycle-consistency loss [18] is introduced to learn mappings translated across two domains without paring constraint. This consistent loss is motivated to regulate the adversarial training to guarantee the learned function can map an individual input $\boldsymbol{x}_{ui}$ to a desired output $\boldsymbol{x}_{vi}$.

In our case, the two mappings $\boldsymbol{G}$ and $\boldsymbol{F}$ are composed of binary code generation ($\boldsymbol{H}_* : \boldsymbol{x}_* \rightarrow \boldsymbol{h}_* \in \{0, 1\}^K$, where $* = \{u, v\}$, $\hat{*} \neq *$) and the reverse process of generating inputs from binary codes ($\boldsymbol{P}_* : \boldsymbol{h}_* \rightarrow \boldsymbol{x}_{\hat{*}}$, where $* = \{u, v\}$), that are, $\boldsymbol{G}: \boldsymbol{x}_u \rightarrow \boldsymbol{H}_u \rightarrow \boldsymbol{P}_u \rightarrow \boldsymbol{x}_v$ and $\boldsymbol{F}: \boldsymbol{x}_v \rightarrow \boldsymbol{H}_v \rightarrow \boldsymbol{P}_v \rightarrow \boldsymbol{x}_u$, as shown in Fig.3 (b) and (c), respectively. Thus, the learned mapping functions should be cycle-consistent: as shown in Fig. 3 (b), for each input $\boldsymbol{x}_u$ from modality $\boldsymbol{X}_u$,

the input translation cycle should be able to bring $\boldsymbol{x}_u$ back to the original input through the generative hamming space ($\{\boldsymbol{H}_*, \boldsymbol{P}_*\}$), i.e., $\boldsymbol{x}_u \rightarrow \boldsymbol{G}(\boldsymbol{x}_u) \rightarrow \boldsymbol{F}(\boldsymbol{G}(\boldsymbol{x}_u)) \approx \hat{\boldsymbol{x}}_u$. Similarly, as illustrated in Fig. 3 (c), for each input $\boldsymbol{x}_v$ from modality $\boldsymbol{X}_v$, $\boldsymbol{G}$ and $\boldsymbol{F}$ should also satisfy: $\boldsymbol{x}_v \rightarrow \boldsymbol{F}(\boldsymbol{x}_v) \rightarrow \boldsymbol{G}(\boldsymbol{F}(\boldsymbol{x}_v)) \approx \hat{\boldsymbol{x}}_v$. And we have

$$\begin{aligned}\mathcal{L}_{cyc}(\boldsymbol{G}, \boldsymbol{F}) = \mathbb{E}_{\boldsymbol{x}_u \sim p_{data}(\boldsymbol{x}_u)} [||\boldsymbol{F}(\boldsymbol{G}(\boldsymbol{x}_u)) - \boldsymbol{x}_u||_1] \\ + \mathbb{E}_{\boldsymbol{x}_v \sim p_{data}(\boldsymbol{x}_v)} [||\boldsymbol{G}(\boldsymbol{F}(\boldsymbol{x}_v)) - \boldsymbol{x}_v||_1]. \quad (2)\end{aligned}$$

C. Deep Generative Hashing

1) Generative Model: In our case, we introduce a generative model that defines the likelihood of generating input $\boldsymbol{x}_v$ given the binary code in its correspondence $\boldsymbol{h}_u$, that is, $\boldsymbol{P}_u : \boldsymbol{h}_u \rightarrow \boldsymbol{x}_v$, denoted as $p(\boldsymbol{x}_v|\boldsymbol{h}_u)$; and we also have $\boldsymbol{P}_v : \boldsymbol{h}_v \rightarrow \boldsymbol{x}_u$, denoted as $p(\boldsymbol{x}_u|\boldsymbol{h}_v)$. This cross-modal generative hashing is to ensure objects in different modalities are translated through their hashing codes, and thus semantic consistence is achieved even without the paired constraint. $\boldsymbol{P}_*$ are also referred as decoding functions where $* = \{u, v\}$ denotes the modality. Inspired by the recent stochastic generative hashing [16], we use the simple Gaussian distribution to model the generation of $\boldsymbol{x}$ given $\boldsymbol{h}$ (for simplicity, we omit the subscripts), which is defined as:

$$p(\boldsymbol{x}, \boldsymbol{h}) = p(\boldsymbol{x}|\boldsymbol{h})p(\boldsymbol{h}), \quad p(\boldsymbol{x}|\boldsymbol{h}) = \mathcal{N}(\boldsymbol{U}\boldsymbol{h}, \rho^2\boldsymbol{I}), \quad (3)$$

where $\boldsymbol{U} = \{\boldsymbol{u}_i\}_{i=1}^K$, $\boldsymbol{u}_i \in \mathbb{R}^d$ is a code book with K code words. The prior $p(\boldsymbol{h}) \sim \mathcal{B}(\theta) = \Pi_{i=1}^K \theta_i^{\boldsymbol{h}_i}(1-\theta_i)^{1-\boldsymbol{h}_i}$ is modeled as the multivariate Bernoulli distribution on the hash codes, where $\theta = [\theta_i]_{i=1}^K \in [0,1]^K$. This intuition can be interpreted as an additive model which reconstructs the input $\boldsymbol{x}$ by summing the selected columns of $\boldsymbol{U}$ given $\boldsymbol{h}$, with a Bernoulli prior on the distribution of hash codes. The joint distribution can be formulated as:

$$p(\boldsymbol{x}, \boldsymbol{h}) \propto \exp\left(\frac{1}{2\rho^2}||\boldsymbol{x} - \boldsymbol{U}^T\boldsymbol{h}||_2^2 - (\log\frac{\theta}{1-\theta})^T\boldsymbol{h}\right). \quad (4)$$

It has been shown that the Gaussian reconstruction error $||\boldsymbol{x} - \boldsymbol{U}^T\boldsymbol{h}||_2^2$ is a surrogate for Euclidean neighborhood preservation [16], and thus this generative model is able to preserve the local neighborhood structure of the input $\boldsymbol{x}$ when the Frobenius norm of $\boldsymbol{U}$ is bounded, that is, minimizing the Gaussian reconstruction error $-\log p(\boldsymbol{x}|\boldsymbol{h})$ will leads to the Euclidean neighborhood preservation. This property is critical to cross-modal retrieval in the sense that the modality-specific local neighborhood structure of data objects can be well characterized.

2) Encoding Model ($\boldsymbol{H}_* : \boldsymbol{x}_* \rightarrow \boldsymbol{h}_*$): Directly computing the posterior $p(\boldsymbol{h}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{h})}{p(\boldsymbol{x})}$, and seeking the maximum a posterior (MAP) solution to the posterior involves solving an expensive integer programming subproblem. Some recent studies on variational auto-encoder [48], [49] show that this difficulty can be avoided by fitting an approximate inference model, $q(\boldsymbol{h}|\boldsymbol{x})$, to approximate the exact posterior of the encoding function $p(\boldsymbol{h}|\boldsymbol{x})$. Thus, the encoding function can be re-parameterized as

$$q(\boldsymbol{h}|\boldsymbol{x}) = \Pi_{k=1}^K q(\boldsymbol{h}_k = 1|\boldsymbol{x})^{\boldsymbol{h}_k} q(\boldsymbol{h}_k = 0|\boldsymbol{x})^{1-\boldsymbol{h}_k}, \quad (5)$$

where $\boldsymbol{h} = [\boldsymbol{h}_k]_{k=1}^K \sim \mathcal{B}(\sigma(\boldsymbol{W}^T\boldsymbol{x}))$ is the linear parametrization where $\boldsymbol{W} = [\boldsymbol{w}_k]_{k=1}^K$. Hence, we can have

$$p(\boldsymbol{h}|\boldsymbol{x}) = \arg\max_{\boldsymbol{h}} q(\boldsymbol{h}|\boldsymbol{x}) = \frac{sign(\boldsymbol{W}^T\boldsymbol{x}) + 1}{2}. \quad (6)$$

D. Training Objective

Our full training objective is formulated to be:

$$\begin{aligned}&\mathcal{L}(\boldsymbol{G}, \boldsymbol{F}, \boldsymbol{D}_{\boldsymbol{x}_v}, \boldsymbol{D}_{\boldsymbol{x}_u}, \boldsymbol{H}_*) \\ &= \mathcal{L}_{GAN}(\boldsymbol{G}, \boldsymbol{D}_{\boldsymbol{x}_v}, \boldsymbol{x}_u, \boldsymbol{x}_v) \\ &\quad + \mathcal{L}_{GAN}(\boldsymbol{F}, \boldsymbol{D}_{\boldsymbol{x}_u}, \boldsymbol{x}_v, \boldsymbol{x}_u) + \lambda\mathcal{L}_{cyc}(\boldsymbol{G}, \boldsymbol{F}) \\ &\quad + D_{KL}(q(\boldsymbol{h}_*|\boldsymbol{x}_*)||p(\boldsymbol{h}_*|\boldsymbol{x}_*)) \quad (7)\end{aligned}$$

where λ controls the relative importance of the two sub-objectives in cycle-consistence training. The set of parameters are denoted as $\Theta_* = \{\boldsymbol{W}_*, \boldsymbol{U}_*, \rho_*, \beta_* := \log\frac{\theta}{1-\theta}\}$. $\boldsymbol{U}_*, \rho_*, \beta_* := \log\frac{\theta}{1-\theta}$ are parameters of the generative model $p(\boldsymbol{x}_*, \boldsymbol{h}_*)$ and $\boldsymbol{W}_*$ are from the function Eq. (5).

Our model can be intuitively viewed as training two auto-encoders in which we learn one auto-encoder in the form of $\boldsymbol{F} \circ \boldsymbol{G} : \boldsymbol{X}_u \rightarrow \boldsymbol{X}_u$, jointly with another $\boldsymbol{G} \circ \boldsymbol{F} : \boldsymbol{X}_v \rightarrow \boldsymbol{X}_v$. One difference is our auto-encoder training has an internal structure, that is, it maps an image/text to itself via an intermediate representation that is a translation of the image/text to another domain. This is similar to the adversarial auto-encoders [50] where an adversarial loss is used to train an auto-encoder to match an arbitrary target distribution, whereas in our case the target distribution for the auto-encoder $\boldsymbol{X}_u \rightarrow \boldsymbol{X}_u$ is the modality of $\boldsymbol{X}_v$. Thus, we aim to solve:

$$\boldsymbol{G}, \boldsymbol{F}, \boldsymbol{W} = \arg\min_{\boldsymbol{G},\boldsymbol{F},\boldsymbol{W}} \max_{\boldsymbol{D}_{X_v}, \boldsymbol{D}_{X_u}} \mathcal{L}(\boldsymbol{G}, \boldsymbol{F}, \boldsymbol{D}_{\boldsymbol{x}_v}, \boldsymbol{D}_{\boldsymbol{x}_u}, \boldsymbol{H}_*), \quad (8)$$

where $\boldsymbol{H}$ is linear parametrization of $\boldsymbol{W}$. In what follows, we will describe the optimization and inference on the generators/discriminators and hash functions.

a) Network architectures with optimization and inference: We adapt our generator and the discriminator architectures from those in [23] where both the generator and discriminator use the modules of convolution-BatchNorm-ReLU [51]. The details of the architectures are described below. Let C_b denote a Convolution-BatchNorm-ReLU layer with b filters. CD_b denotes a Convolution-BatchNorm-Dropout-ReLU layer with a dropout rate 0.5. All convolutions are 4×4 spatial filters applied with stride 2. Convolutions in the encoder and the discriminator are down-sampled by a factor of 2 whereas in the decoder they are up-sampled by a factor of 2.

- The generator is designed with an encoder-decoder network [52] in which the input is passed through a series of layers that progressively down-sample until a bottleneck layer, at which point the process is reversed. Our generator adopts the encoder-decoder architecture that consists of: *encoder*: $C_{64} - C_{128} - C_{256} - C_{512} - C_{512} - C_{512} - C_{512} - C_{512}$ and *decoder*: $CD_{512} - CD_{512} - CD_{512} - CD_{512} - CD_{512} - CD_{256} - CD_{128} - CD_{64}$.
- The discriminator is adopted from the PatchGAN [41] that only penalizes structure as the scale of patches. This discriminator will classify if each $N \times N$ patch in an

image is real or fake. We run this discriminator convolutionally across the image, and average all responses to provide the ultimate output of discriminator. The N is much smaller than full size of the image and very advantageous because a smaller PatchGAN has fewer parameters, and can be applied on arbitrary large images. It is noted that when we perform the optimization of discriminator on the text domain, the N is set to be 1.

Following the [22], we alternate between one gradient descent step on the generator and then one step on the discriminator. However, further closer examination on the training objective suggests that it is unable to directly compute the gradients w.r.t parameters of $p(\boldsymbol{x}|\boldsymbol{h})$. In particular, it is difficult to compute the stochastic gradients w.r.t $\boldsymbol{W}$ because it depends on the stochastic binary variables $\boldsymbol{h}$. In order to back-propagate the discrete stochastic variables, we follow the SGH [16] to adopt an approximation to the gradient w.r.t $\boldsymbol{W}$, which is derived based on distributional derivatives. Specifically, a stochastic neuron is introduced for re-parameterizing the Bernoulli variable $\boldsymbol{h}_k(z)$ with $z \in (0, 1)$, and defined as

$$\hat{\boldsymbol{h}}(z, \xi) = \begin{cases} 1 & \text{if } z \geq \xi \\ 0 & \text{if } z < \xi \end{cases} \tag{9}$$

where $\xi \sim \mathcal{U}(0, 1)$ are random variables. The stochastic neuron can be used to re-parameterize the binary variables $\boldsymbol{h}$ by replacing $[\boldsymbol{h}_k]_{k=1}^K(\boldsymbol{x}) \sim \mathcal{B}(\sigma(\boldsymbol{w}_k^T\boldsymbol{x}))$ with $[\hat{\boldsymbol{h}}_k(\sigma(\boldsymbol{w}_k^T\boldsymbol{x}), \xi_k)]_{k=1}^K$. Hence, we re-parameterize the training objective as $\hat{D}_{KL}(\Theta) = \sum_{\boldsymbol{x}} \hat{D}_{KL}(\Theta, \boldsymbol{x}) := \sum_{\boldsymbol{x}} \mathbb{E}_{\xi}\left[l(\hat{\boldsymbol{h}}, \boldsymbol{x})\right]$ where $l(\hat{\boldsymbol{h}}, \boldsymbol{x}) = q(\hat{\boldsymbol{h}}(\sigma(\boldsymbol{w}^T\boldsymbol{x}), \xi)|\boldsymbol{x})||p(\boldsymbol{x}, \hat{\boldsymbol{h}}(\sigma(\boldsymbol{w}^T\boldsymbol{x}), \xi))$ with $\xi \sim \mathcal{U}(0, 1)$. Due to the discontinuity of the stochastic neuron $\hat{\boldsymbol{h}}(z, \xi)$, a more generalized distributional derivative [53] can be computed instead of computing the standard Stochastic Gradient Descent. Specifically, given a sample $\boldsymbol{x}$, the distributional derivative of function $\hat{D}_{KL}$ w.r.t $\boldsymbol{W}$ is computed as

$$\begin{aligned} &\hat{D}_{KL}(\Theta; \boldsymbol{x}) \\ &= \mathbb{E}_{\xi}\left[\triangle_{\hat{\boldsymbol{h}}} l(\hat{\boldsymbol{h}}(\sigma(\boldsymbol{W}^T\boldsymbol{x}), \xi))\sigma(\boldsymbol{W}^T\boldsymbol{x}) \bullet (1 - \sigma(\boldsymbol{W}^T\boldsymbol{x}))\boldsymbol{x}^T\right], \end{aligned} \tag{10}$$

where $\bullet$ denotes the point-wise product and $\triangle_{\hat{\boldsymbol{h}}} l(\hat{\boldsymbol{h}})$ is further defined as $\left[\triangle_{\hat{\boldsymbol{h}}} l(\hat{\boldsymbol{h}})\right]_k = l(\hat{\boldsymbol{h}}_k^1) - l(\hat{\boldsymbol{h}}_k^0)$, where $[\hat{\boldsymbol{h}}_k^i]_m = \hat{\boldsymbol{h}}_m$ if $k \neq m$, otherwise $[\hat{\boldsymbol{h}}_k^i]_m = i$, $i \in \{0, 1\}$.

E. Training Details

Our architecture and experiments are implemented on the open-source Torch7 framework. We take the text encoding model [27] to extract text features. For the text features, we use a character-level text-based ConvNet [54] which can be viewed as a standard CNN for images, except that the image width is 1 and the number of channels is equal to the alphabet size. The 2D convolution and spatial max-pooling are replaced by temporal 1D convolution and temporal max-pooling. After each convolutional layer, the rectified linear activation unit (ReLU) is applied. The variable length of sequences is handled by zero-padding the input past the final input character. Thus, the text-based CNN is constructed with convolution, pooling and thresholding activation function layers, followed by fully-connected layers to project onto the embedding space. The fully-connected layers on top of text features are set to be [1000 $\rightarrow$ 500 $\rightarrow$ 200], [11500 $\rightarrow$ 500 $\rightarrow$ 200], and [10 $\rightarrow$ 100 $\rightarrow$ 200] for the COCO, IAPR TC-12, and Wiki, respectively. For the generator on images, we adopt the architecture from [55] which has shown promising results for style transfer. This generative network contains two stride-2 convolutions, several residual blocks, and two fractionally-stride convolutions with stride of $\frac{1}{2}$. We use 6 blocks for all training images. For the discriminator, following [18] we use 70×70 PatchGAN [41], which aims to classify whether 70×70 overlapping image patches are real or fake. The patch-level discriminator has fewer parameters than a full-image discriminator, and it can be applied to arbitrarily-sized images.

As suggested by [18], two strategies can employed to stabilize the network training. First, for the loss of $\mathcal{L}_{GAN}$, the negative log likelihood objective is replaced by a least-square loss. In specific, for a GAN loss $\mathcal{L}_{GAN}(\boldsymbol{G}, \boldsymbol{D}_{\boldsymbol{x}_v}, \boldsymbol{x}_u, \boldsymbol{x}_v)$, we train the $\boldsymbol{G}$ to minimize $\mathbb{E}_{\boldsymbol{x}_v \sim p_{data}(\boldsymbol{x}_v)}[(\boldsymbol{D}_{\boldsymbol{x}_v}(\boldsymbol{G}(\boldsymbol{x}_v)) - 1)^2]$, and train the $\boldsymbol{D}$ to minimize $\mathbb{E}_{\boldsymbol{x}_u \sim p_{data}(\boldsymbol{x}_u)}[(\boldsymbol{D}_{\boldsymbol{x}_v}(\boldsymbol{x}_u) - 1)^2] + \mathbb{E}_{\boldsymbol{x}_v \sim p_{data}(\boldsymbol{x}_v)}[(\boldsymbol{D}_{\boldsymbol{x}_v}(\boldsymbol{x}_v))]$. Second, to reduce the model oscillation, the discriminator is updated using a history of generated images/texts rather than the ones produced by the latest generative networks. In all experiments, we set $\lambda = 10$ in Eq.(7), and Adam solver [56] is uses with a batch size of 1. The networks are trained with a learning rate of 0.0002, which is maintained for the first 100 epoches, and linearly decayed the rate to zero over the next 100 epoches. At inference time, we employ the generator net in the same as we run during the training and the batch normalization is applied by using the statistics of testing batch rather than aggregated statistics of the training batch.

IV. Experiments

In this section, we conduct extensive experiments to evaluate the efficiency of the proposed CYC-DGH against state-of-the-arts on three widely-used benchmark data sets.

A. Data Preparation

- **Microsoft COCO** [57] The recent release of the data set contains 82,783 training images and 40,137 validation images. For each image, the data set provides at least five sentences annotations, belonging to 80 most frequent categories as ground truth labels. Some images that have no category information are removed from training set, and thus we get 82,081 training images.
- **IAPR TC-12** [58] This data set consists of 20,000 images collected from a wide variety of domains, such as sports and actions, people, animals, cities, landscapes, and so on. Each image has at least 1 sentence annotations as well as category annotations generated from segmentation with 275 concepts. Following the setting of TUCH [21], we select images with 22 most frequent concept tags, and thus a new training set with 18,673 images are formed.

- **Wiki**[1] This data set contains 2,866 Wikipedia documents, where each document contains a single image and a corresponding text of at least 70 words. These documents are categorized into 10 semantic classes, and each document is from one class. Each text is represented by a 10-dimensional feature vector computed from the Latent Dirichelet Allocation model. We randomly select 75% documents from this data set as database and the rest a query samples.

B. Competitors and Evaluation Setup

- TUCH [21]: A model that is able to turn cross-view hashing into single-view hashing without multi-view embedding such that the information loss is minimized.
- CMDVH [9]: A cross-modal deep variational hashing method that learns non-linear transformations from image-text input pairs so that unified binary codes can be obtained. They also learn model-specific hashing networks in generative form for the out-of-sample extension.
- DVSH [8]: The method performed end-to-end supervised metric-based training in the form of cosine hinge loss to obtain the binary codes, and learned modality-specific deep networks to obtain the hash functions.
- CorrAE [11]: The model is constructed by correlating hidden representations of two uni-modal auto-encoders, which is optimized by the correlation learning error between hidden representations of two modalities.
- CMNN [10]: The method is to learn a similarity preserving network for cross-modalities through a coupled Siamese network with hinge loss.
- CAH [12]: A model that is designed with a stacked auto-encoder architecture to jointly maximize the feature and semantic correlation across modalities.
- DCMH [13]: It is an end-to-end deep learning framework with a negative log likelihood criterion to preserve the similarity between real-value representations having the same class.
- HashGAN [38]: An adversarial hashing network with attention mechanism to enhance the measurement of content similarities.

Evaluation Metrics: For each data set, we perform two cross-modal retrieval tasks: image-to-text retrieval ($I \rightarrow T$) and text-to-image retrieval ($T \rightarrow I$), which search texts by a query image and search images by a query text, respectively. We use the mean average precision (mAP) to measure the performance of different retrieval methods. mAP is defined as the mean of all queries' average precision (AP), defined as $AP = \frac{1}{M}\sum_{r=1}^{R} prec(r) \odot rel(r)$ where M is the number of relevant instances in the retrieved set, $prec(r)$ denotes the precision of the top r retrieved set, and $rel(r)$ is an indicator of relevance of a given rank (which is set to 1 if relevant and 0 otherwise).

C. Ablation Studies

In this experiment, we provide analysis on the proposed loss function by comparing against ablations of the full objective, including the adversarial loss $\mathcal{L}_{GAN}$ alone and the cycle consistency loss $\mathcal{L}_{cyc}$ alone. In this experiment, following the cycle-GAN [18], we adopt the FCN-score from [41] as the automatic quantitative measure to evaluate the $text \rightarrow$ image task on Microsoft-COCO data set. The metric of fully-convolutional network (FCN) [40] evaluates how interpretable the generated images are in accordance to an off-the-shelf semantic segmentation algorithm. The FCN predicts a label map over a generated image, which can be compared against the input ground truth labels using the semantic segmentation metrics: per-pixel accuracy and per-class accuracy. We also consider the mean average precision (MAP) metric to evaluate the overall performance of each loss function. The ablation studies on varied of our loss function are given in Table I. It can observed that muting either the GAN loss or the cycle-consistency loss can substantially degrade the accuracy results. We therefore conclude that both terms are critical to the regeneration results. This discovery is very consistent to that in cycle-GAN [18] which also claims that both cycle-consistent loss and GAN loss are indispensable in image translation.

TABLE I

FCN-SCORES AND MAP VALUES FOR VARIANTS OF CYC-DGH, EVALUATED ON MICROSOFT-COCO WITH 64 BITS TO REGENERATE THE IMAGES

Loss	Per-pixel accuracy	Per-class accuracy	MAP
Cycle alone	0.724	0.270	0.41
GAN alone	0.611	0.126	0.34
CYC-DGH	0.584	0.192	**0.57**

TABLE II

THE MAP VALUES FOR VARIED λ, EVALUATED ON MICROSOFT-COCO WITH 64 BITS TO REGENERATE THE IMAGES

	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 10$	$\lambda = 100$
MAP	0.27	0.30	0.44	0.57	0.49

In addition, we study the effect of varied balance parameter λ to empirically set its appropriate value in the objective function. As shown in Table II, the parameter λ is set to be 10 to account for its optimal empirical evaluation w.r.t the MAP values, which is also consistent with the setting in [18].

D. Experimental Results

In this section, we evaluate the proposed CYC-DGH by performing extensive comparisons with baselines.

1) Reconstruction Loss: To demonstrate the flexibility of generative modeling in reconstructing the inputs from the binary codes, we compare the L_2 reconstruction error to that of the ITQ [59] and the generative model of SGH [16], showing the benefits of regenerating the inputs under the cycle-consistent constraint. Recall that our method has a generative model $\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{h})$, we can compute the regenerated input via $\hat{\boldsymbol{x}} = \arg\max \boldsymbol{p}(\boldsymbol{x}|\boldsymbol{h})$, and then calculate the L_2 reconstruction loss of the regenerated input and the original $\boldsymbol{x}$ via $||\boldsymbol{x} - \hat{\boldsymbol{x}}||_2^2$. ITQ [59] trains by minimizing the binary quantization loss, that is, $\min ||B - XWR||_F^2$ (where $B = sign(XW)$, X is the data matrix and W is the encoding matrix), which is essentially L_2 reconstruction loss when the magnitude of the feature

[1] http://www.svcl.ucsd.edu/projects/crossmodal/

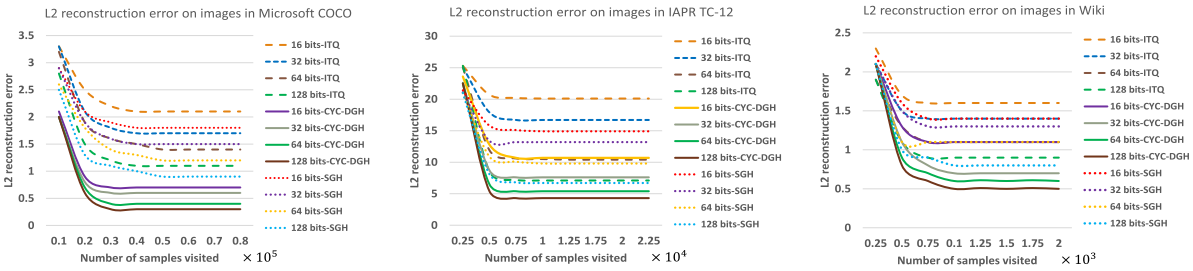


Fig. 4. The L_2 reconstruction error on images in three data sets.

TABLE III
TRAINING TIME COMPARISON ON MICROSOFT-COCO

Training time on Microsoft-COCO in seconds				
Method	16 bits	32 bits	64 bits	128 bits
CYC-DGH	4.23	6.38	9.71	12.35
ITQ [60]	22.74	38.36	51.91	67.23
SGH [16]	4.04	5.85	9.16	10.17

TABLE IV
TRAINING TIME COMPARISON ON IAPR TC-12

Training time on IAPR TC-12 in seconds				
Method	16 bits	32 bits	64 bits	128 bits
CYC-DGH	3.92	5.84	9.11	11.05
ITQ [60]	17.49	30.17	46.77	60.22
SGH [16]	3.17	5.25	7.91	9.86

vectors is compatible with the radius of the binary code. Then, the ITQ [59] uses the covariance matrix W formed by the eigenvalues and the orthogonal matrix R to reconstruction the inputs, that is, $\min_{R,b} ||\boldsymbol{x}_i - WRb_i||^2$. To have fair evaluation on the ITQ reconstruction capability for cross-modal retrieval, we use the covariance and orthogonal matrix learnt from the text domain to reconstruct the images: $||\boldsymbol{x}_u - WRb_v||_2$. The reconstruction loss computed for SGH [16] is the same as our method. We plot the L_2 reconstruction error of our method, ITQ [59] and SGH [16] on three data sets in Fig.4, where it shows the average L_2 reconstruction loss computed against the number of examples seen by the training process. The training time comparison with ITQ [59] and SGH [16] are given in Table III, Table IV, and Table V, respectively. Unlike ITQ [59] that iterative optimizes the quantization error under the orthogonal constraint, our proposed method is able to reconstruct the inputs through the Gaussian reconstruction without the orthogonality constraints, which can make the training more efficient. SGH [16] is more efficient than our approach in training time because they use the minimal length description as the objective optimization. However, the lower reconstruction loss compared to the SGH [16] demonstrates the flexibility of the proposed CYC-DGH afforded by the re-parameterizations via stochastic neurons with cycle-consistent constraint. This is the merit of cycle-consistent constraint which can ensure the minimal loss in reconstruction, and thus makes our method more apt to cross-modal retrieval and augment the correlation of image/text correspondence.

2) Comparison to Existing Cross-Modal Hashing Methods: We compare our approach with existing methods in terms of MAP, precision-recall curves and precision@

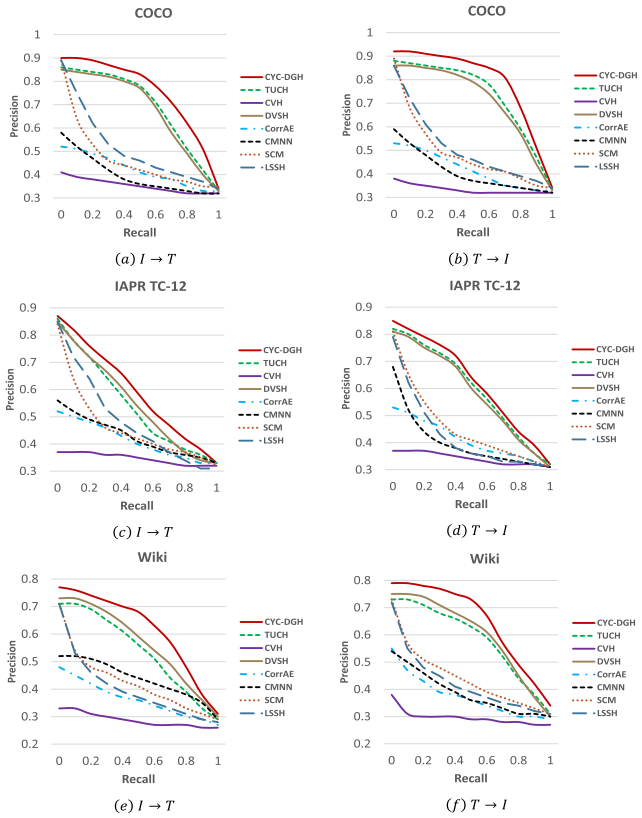


Fig. 5. The precision-recall curves of cross-modal retrieval on Microsoft COCO, IAPR TC-12, and Wiki @32 bits. (a) $I \rightarrow T$. (b) $T \rightarrow I$. (c) $I \rightarrow T$. (d) $T \rightarrow I$. (e) $I \rightarrow T$. (f) $T \rightarrow I$.

TABLE V
TRAINING TIME COMPARISON ON WIKI

Training time on Wiki in seconds				
Method	16 bits	32 bits	64 bits	128 bits
CYC-DGH	2.03	3.18	5.32	7.65
ITQ [60]	12.54	18.36	21.91	27.23
SGH [16]	1.89	2.92	4.91	6.99

top-R returned curves in two cross-modal retrieval tasks: image query against textual database ($I \rightarrow T$), and textual query against image database ($T \rightarrow I$). The methods include unsupervised (CVH [34], LSSH [35]), and supervised (SCM [15], SePH [1]). To have fair comparison with these non-deep-learning methods, we use the CNN features extracted at the FC7 layer for the images from the pre-trained

TABLE VI

MEAN AVERAGE PRECISION (MAP) COMPARISON OF EXISTING CROSS-MODAL HASHING METHODS ON THREE DATA SETS

Task	Method	Microsoft COCO				IAPR TC-12				Wiki			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
$I \rightarrow T$	CVH [34]	0.373	0.368	0.366	0.357	0.537	0.541	0.524	0.496	0.238	0.204	0.179	0.158
	SCM [15]	0.570	0.600	0.631	0.649	0.567	0.505	0.454	0.418	0.139	0.137	0.141	0.136
	LSSH [35]	-	-	-	-	0.544	0.577	0.596	0.599	0.364	0.371	0.378	0.358
	SePH [1]	0.581	0.613	0.625	0.634	0.618	0.645	0.650	0.678	0.414	0.435	0.437	0.447
	CYC-DGH	**0.722**	**0.754**	**0.781**	**0.780**	**0.771**	**0.815**	**0.832**	**0.831**	**0.794**	**0.811**	**0.813**	**0.820**
$T \rightarrow I$	CVH [34]	0.373	0.369	0.365	0.371	0.568	0.578	0.561	0.536	0.388	0.336	0.257	0.230
	SCM [15]	0.558	0.619	0.658	0.686	0.652	0.570	0.478	0.421	0.132	0.143	0.156	0.149
	LSSH [35]	-	-	-	-	0.487	0.526	0.555	0.572	0.606	0.626	0.638	0.638
	SePH [1]	0.613	0.650	0.672	0.693	0.610	0.634	0.640	0.673	0.701	0.699	0.710	0.715
	CYC-DGH	**0.761**	**0.796**	**0.834**	**0.859**	**0.772**	**0.798**	**0.837**	**0.842**	**0.811**	**0.823**	**0.826**	**0.822**

TABLE VII

MEAN AVERAGE PRECISION (MAP) COMPARISON OF STATE-OF-THE-ART DEEP CROSS-MODAL HASHING METHODS ON THREE DATA SETS

Task	Method	Microsoft COCO				IAPR TC-12				Wiki			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
$I \rightarrow T$	TUCH [21]	0.628	0.714	0.735	0.766	0.595	0.637	0.690	0.714	0.578	0.621	0.637	0.660
	CMDVH [9]	0.669	0.697	0.721	0.769	0.720	0.773	0.800	0.790	0.424	0.443	0.452	0.444
	DVSH [8]	0.587	0.713	0.739	0.755	0.570	0.632	0.696	0.723	-	-	-	-
	CMNN [10]	0.556	0.560	0.585	0.594	0.516	0.542	0.577	0.600	-	-	-	-
	CorrAE [11]	0.550	0.556	0.569	0.581	0.495	0.525	0.557	0.589	0.507	0.548	0.569	0.610
	CAH [12]	-	-	-	-	0.559	0.590	0.603	0.609	0.492	0.502	0.542	0.560
	DCMH [13]	-	-	-	-	0.624	0.635	0.676	0.671	0.569	0.588	0.631	0.633
	HashGAN [38]	-	-	-	-	0.529	0.528	0.544	0.546	0.756	0.772	0.772	0.774
	CYC-DGH	**0.722**	**0.754**	**0.781**	**0.780**	**0.771**	**0.815**	**0.832**	**0.831**	**0.794**	**0.811**	**0.813**	**0.820**
$T \rightarrow I$	TUCH [21]	0.649	0.760	0.786	0.812	0.624	0.656	0.688	0.709	0.598	0.627	0.655	0.679
	CMDVH [9]	0.691	0.735	0.768	0.765	0.735	0.774	0.804	0.811	0.727	0.733	0.738	0.737
	DVSH [8]	0.591	0.737	0.758	0.767	0.604	0.640	0.681	0.675	-	-	-	-
	CMNN [10]	0.579	0.598	0.620	0.645	0.512	0.540	0.549	0.565	-	-	-	-
	CorrAE [11]	0.559	0.581	0.611	0.626	0.498	0.519	0.533	0.549	0.523	0.561	0.582	0.617
	CAH [12]	-	-	-	-	0.582	0.621	0.638	0.647	0.527	0.534	0.570	0.596
	DCMH [13]	-	-	-	-	0.648	0.672	0.698	0.707	0.589	0.611	0.651	0.662
	HashGAN [38]	-	-	-	-	0.536	0.557	0.565	0.571	0.792	0.806	0.807	0.809
	CYC-DGH	**0.761**	**0.796**	**0.834**	**0.859**	**0.772**	**0.798**	**0.837**	**0.842**	**0.811**	**0.823**	**0.826**	**0.822**

model of CNN-F from [60]. Table VI shows the mAP performance by the hamming ranking. For unsupervised methods such as CVH [34] and LSSH [35], their performance are less competitive to supervised. This is mainly because the cross-modal correlation can be achieved without the aid of semantics, and thus making the hashing learning of functions not discriminative. During retrieval, the supervised method of SePH [1] can employ unified binary code learning across both the query set and the gallery set, and thus it achieves improved mAP values. However, SePH [1] still requires the training sample in the form of aligned pairs, which would limit its application in practice. Also, it is clear to observe that our approach provides the best performance compared to these shallow cross-modal hashing methods.

3) Comparison to State-of-the-Art Deep Cross-Modal Hashing Methods: In this experiment, we compare our approach with recent deep cross-modal hashing competitors and show results in Table VII. It can be seen that our method achieves the best results in terms of MAP values on different hashing bits over three data sets. This may be due to several reasons. First, the method of CAH [12] still uses handcrafted image features as the input to their deep neural networks whereas our model starts learning from raw images. In the methods of DVSH [8] and CMDVH [9], the modality-specific hash functions are learned such that the non-linear relationship of samples from different modalities are exploited. Furthermore, CMDVH [9] performed a shared binary code learning strategy to reduce the modality gap between the hash functions. However, explicitly learning modality-specific hash functions cannot render the hashing effective in cross-modal context. In other words, the learned hash functions are still limited in their specific modalities even in the joint learning with unified binary codes. In contrast, the proposed CYC-DGH eliminates the requirement of coupled training pairs in semantics, and the discriminative binary codes are produced through the enforced cycle-consistency loss. This cycle-consistent loss can uniquely correlate each pair with the same semantics and the modality heterogeneity can be addressed by the adversarial loss. The method of TUCH [21] utilizes the generative nets to convert one modality data into the target modality so as to minimize the information loss caused by the respective hashing embedding. However, converting data into a different modality is unable to maintain the relationship in the original source data and thus cannot achieve very comparable results to CMDVH [9] and our method. In contrast, the proposed CYC-DGH effectively address the information loss by proposing to regenerating the inputs through the binary codes, and the hash function learning as well as regeneration process are jointly achieved in the full loss function.

The precision-recall curves with 32 bits on for the two cross-modal tasks on three benchmarks are shown in Fig.5, respectively. It can be seen that CYC-DGH achieves the best

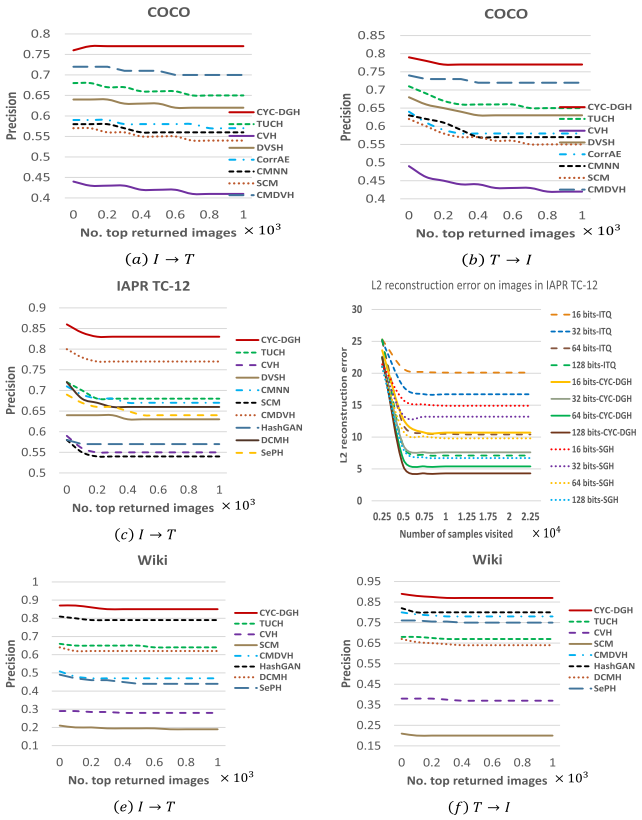


Fig. 6. The precision@top-R return curves of cross-modal retrieval on Microsoft COCO, IAPR TC-12, and Wiki @32 bits. (a) $I \rightarrow T$. (b) $T \rightarrow I$. (c) $I \rightarrow T$. (e) $I \rightarrow T$. (f) $T \rightarrow I$.

performance at two asks on all recall levels. This is mainly because the removal of paired training correspondence can still be supplemented by the effective cycle-consistency loss, and the adversarial training is able to reduce the modality heterogeneity gap effectively. Moreover, another major challenge in cross-modal retrieval is the information loss caused by the binary embedding which is combated by our approach with input regeneration from binary codes.

Figure. 6 shows the precision@top-R return curves of all comparison methods with 32 bits on three data sets. It displays how the precision changes against the number of R of top-retrieved results. CYC-DGH outperforms all the competitors and shows consistent effectiveness against the increased number of top-retrieved items.

V. Conclusion and Future Work

In this paper, we present a novel deep generative hashing framework towards cross-modal retrieval that is able to produce hashing learning functions without the need of paired training samples. The proposed model, namely Cycle-consistent Deep Generative Hashing (CYC-DGH), builds adversarial training across modality to reduce the heterogeneity, which is augmented with cycle-consistent constraint to uniquely maximize the correlation of the input-output without the semantic labeling. Besides, we introduce the generative model into the hashing learning, which jointly performs binary code learning as well as input generation from binary codes so as to minimize the information loss to great extent. Extensive empirical evidences including comparison with competitors and ablation studies show that our CYC-DGH approach advance the state-of-the-arts on image to text (and text to image) retrieval tasks, over three benchmarks. In the future, we plan to explore more powerful generative models to further improve the generation capability of our method.

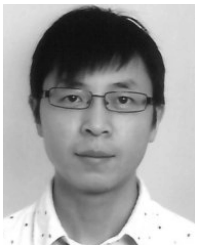
References

[1] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. CVPR*, Jun. 2015, pp. 3864–3872.
[2] Y. Wang, X. Lin, L. Wu, W. Zhang, and Q. Zhang, "LBMCH: Learning bridging mapping for cross-modal hashing," in *Proc. ACM SIGIR*, 2015, pp. 999–1002.
[3] F. Zheng, Y. Tang, and L. Shao, "Hetero-manifold regularisation for cross-modal hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1059–1071, May 2018.
[4] Y. Luo, Y. Yang, F. Shen, Z. Huang, P. Zhou, and H. T. Shen, "Robust discrete code modeling for supervised hashing," *Pattern Recognit.*, vol. 75, pp. 128–135, Mar. 2018.
[5] M. Hu, Y. Yang, F. Shen, L. Zhang, H. T. Shen, and X. Li, "Robust Web image annotation via exploring multi-facet and structural knowledge," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4871–4884, Oct. 2017.
[6] L. Wu, Y. Wang, and S. Pan, "Exploiting attribute correlations: A novel trace lasso-based weakly supervised dictionary learning method," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4497–4508, Dec. 2017.
[7] Y. Wang, X. Lin, L. Wu, and W. Zhang, "Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1393–1404, Mar. 2017.
[8] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *Proc. KDD*, 2016, pp. 1445–1454.
[9] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Cross-modal deep variational hashing," in *Proc. ICCV*, Oct. 2017, pp. 4097–4105.
[10] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, "Multimodal similarity-preserving hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 824–830, Apr. 2014.
[11] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimediaa*, 2014, pp. 7–16.
[12] Y. Cao, M. Long, J. Wang, and H. Zhu, "Correlation autoencoder hashing for supervised cross-modal search," in *Proc. ICMR*, 2016, pp. 197–204.
[13] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proc. CVPR*, Jul. 2017, pp. 3270–3278.
[14] J. T. Zhou, H. Zhao, X. Peng, M. Fang, Z. Qin, and R. S. M. Goh, "Transfer hashing: From shallow to deep," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.
[15] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. AAAI*, 2014, p. 7.
[16] B. Dai, R. Guo, S. Kumar, N. He, and L. Song, "Stochastic generative hashing," in *Proc. ICML*, 2017, pp. 913–922.
[17] L.-Y. Duan, Y. Wu, Y. Huang, Z. Wang, J. Yuan, and W. Gao, "Minimizing reconstruction bias hashing via joint projection learning and quantization," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3127–3141, Jun. 2018.
[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017, pp. 2242–2251.
[19] J. Hoffman *et al.* (2017). "CyCADA: Cycle-consistent adversarial domain adaptation." [Online]. Available: https://arxiv.org/abs/1711.03213
[20] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3D-guided cycle consistency," in *Proc. CVPR*, 2016, pp. 117–126.
[21] X. Zhao, G. Ding, Y. Guo, J. Han, and Y. Gao, "Tuch: Turning cross-view hashing into single-view hashing via generative adversarial nets," in *Proc. IJCAI*, 2017, pp. 3511–3517.
[22] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014.
[23] A. Radford, L. Metz, and S. Chintala. (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1511.06434
[24] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," in *Proc. NIPS*, 2016, pp. 64–72.

[25] Z. Qiu, Y. Pan, T. Yao, and T. Mei, "Deep semantic hashing with generative adversarial networks," in *Proc. SIGIR*, 2017, pp. 225–234.
[26] Y. Peng, J. Qi, and Y. Yuan. (2017). "CM-GANs: Cross-modal generative adversarial networks for common representation learning." [Online]. Available: https://arxiv.org/abs/1710.05106
[27] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. ICML*, 2016, pp. 1060–1069.
[28] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 251–260.
[29] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2088–2095.
[30] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.
[31] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, and X. Huang, "Robust subspace clustering for multi-view data by exploiting correlation consensus," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3939–3949, Nov. 2015.
[32] Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4833–4843, Oct. 2018.
[33] L. Wu, Y. Wang, J. Gao, and X. Li, "Deep adaptive feature embedding with local sample distributions for person re-identification," *Pattern Recog.*, vol. 73, pp. 275–288, Jan. 2018.
[34] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. IJCAI*, 2011, p. 1360.
[35] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proc. SIGIR*, 2014, pp. 415–424.
[36] Y. Shen, L. Liu, L. Shao, and J. Song, "Deep binaries: Encoding semantic-rich cues for efficient textual-visual cross retrieval," in *Proc. ICCV*, 2017, pp. 4117–4126.
[37] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: Fast free-hand sketch-based image retrieval," in *Proc. CVPR*, Jul. 2017, pp. 2298–2307.
[38] X. Zhang *et al.* (2017). "HashGAN: Attention-aware deep adversarial hashing for cross modal retrieval." [Onlile]. Available: https://arxiv.org/abs/1711.09347
[39] Y. Shen, L. Liu, F. Shen, and L. Shao, "Zero-shot sketch-image hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mar. 2018, pp. 3598–3607.
[40] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Jun. 2015, pp. 3431–3440.
[41] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017, pp. 5967–5976.
[42] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *Proc. CVPR*, 2017.
[43] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. NIPS*, 2016.
[44] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba. (2016). "Cross-modal scene networks," [Online]. Available: https://arxiv.org/abs/1610.09003
[45] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. CVPR*, 2017, p. 7.
[46] F. Wang, Q. Huang, and L. J. Guibas, "Image co-segmentation via consistent functional maps," in *Proc. ICCV*, Dec. 2013, pp. 849–856.
[47] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. ICCV*, 2017, pp. 2868–2876.
[48] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, 2014.
[49] A. Mnih and K. Gregor, "Neural variational inference and learning in belief networks," in *Proc. ICML*, 2014, pp. 1791–1799.
[50] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. (2016). "Adversarial autoencoders." [Online]. Available: https://arxiv.org/abs/1511.05644
[51] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167
[52] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[53] G. Grubb, *Distributions and Operators*, vol. 52. New York, NY, USA: Springer, 2008.
[54] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. CVPR*, 2016, pp. 49–58.
[55] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.
[56] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization," [Online]. Available:https: //arxiv.org/abs/1412.6980
[57] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.
[58] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The IAPR TC-12 benchmark—A new evaluation resource for visual information systems," in *Proc. Int. Workshop OntoImage*, 2006, pp. 13–23.
[59] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *Proc. CVPR*, Jun. 2011, pp. 817–824.
[60] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. BMVC*, 2014.

Lin Wu received the Ph.D. degree from The University of New South Wales, Sydney, Australia, in 2014. She has published a number of academic papers in CVPR, ACM Multimedia, IJCAI, ACM SIGIR, IEEE-TIP, IEEE-TNNLS, IEEE-TCYB, *Neural Networks*, and *Pattern Recognition*. She also served as the Program Committee Member for international conferences and an Invited Journal Reviewer for IEEE TIP, IEEE TNNLS, IEEE TCSVT, IEEE TMM, and *Pattern Recognition*.

Yang Wang received the Ph.D. degree from The University of New South Wales, Kensington, Australia, in 2015. He has published 50 research papers, most of which have appeared at the competitive venues, including IEEE TIP, IEEE TNNLS, IEEE TCYB, IEEE TKDE, IEEE TMM, *Pattern Recognition*, *Neural Networks*, ACM Multimedia, ACM SIGIR, IJCAI, IEEE ICDM, ACM CIKM, and VLDB Journal. He served as the Invited Journal Reviewer for more than 15 leading journals, such as IEEE TPAMI, IEEE TIP, IEEE TNNLS, IEEE TKDE, *Machine Learning* (Springer), and IEEE TMM.



Ling Shao is the CEO and the Chief Scientist of the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. His research interests include computer vision, machine learning, and medical imaging. He is a fellow of the IAPR, the IET, and the BCS. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and several other journals.