Multi-pathway Generative Adversarial Hashing for Unsupervised Cross-modal Retrieval

Jian Zhang and Yuxin Peng

***Abstract*—Cross-modal hashing aims to map heterogeneous cross-modal data into a common Hamming space, which can realize fast and flexible retrieval across different modalities. Unsupervised cross-modal hashing is more flexible and applicable than supervised methods, since no intensive labeling work is involved. However, existing unsupervised methods learn the hashing functions by preserving inter and intra correlations, while ignoring the underlying manifold structure across different modalities, which is extremely helpful to capture the meaningful nearest neighbors of different modalities for cross-modal retrieval. Furthermore, existing works mainly focus on pairwise relation modeling, while ignoring the correlations within multiple modalities. To address the above problems, in this paper we propose a *multi-pathway generative adversarial hashing (MGAH)* approach for unsupervised cross-modal retrieval, which makes full use of GAN's ability for unsupervised representation learning to exploit the underlying manifold structure of cross-modal data. The main contributions can be summarized as follows: (1) We propose a multi-pathway generative adversarial network to model cross-modal hashing in an unsupervised fashion. In the proposed network, given a data of one modality, the generative model tries to fit the distribution over the manifold structure, and select informative data of other modalities to challenge the discriminative model. The discriminative model learns to distinguish the generated data and the true positive data sampled from the correlation graph to achieve better retrieval accuracy. These two models are trained in an adversarial way to improve each other and promote hashing function learning. (2) We propose a correlation graph based approach to capture the underlying manifold structure across different modalities, so that data of different modalities but within the same manifold can have smaller Hamming distance to promote retrieval accuracy. Extensive experiments compared with state-of-the-art methods on 3 widely-used datasets verify the effectiveness of our proposed approach.**

***Index Terms*—Cross-modal hashing, generative adversarial networks, manifold structure.**

I. Introduction

Multimedia retrieval has become an important application over the past decades, which can retrieve multimedia contents that users have interests in. However, it is a big challenge to retrieve multimedia data efficiently from large scale databases, due to the explosive growth of multimedia information. To address this issue, there are many hashing methods [1]–[3] proposed to accomplish efficient retrieval. The goal of hashing methods is to map high dimensional representations in the original space to short binary codes in the Hamming space. By using these binary hash codes, fast Hamming distance computation is applied based on bit operations that can be implemented efficiently. Moreover, binary codes take much less storage compared with original high dimensional representations.

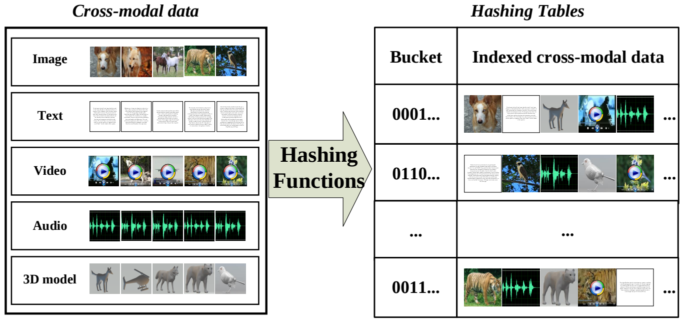


Fig. 1. Examples of cross-modal hashing, which can map data of different modalities (e.g. image, text, video, audio and 3D model) in to binary codes, and realize fast retrieval across different modalities by a query of any modality.

There are large numbers of hashing methods applied to single-modal retrieval [1], by which users can only retrieve data by a query with the same modality, such as text retrieval [4] and image retrieval [1]. Nevertheless, single-modal retrieval can not meet users' increasing demands, due to the different modalities of multimedia data. For example, by single modality retrieval, it is impracticable to search an image by using a textual sentence that describes the semantic content of the image. Therefore, cross-modal hashing has been proposed to meet this kind of retrieval demands in large scale cross-modal databases. Owing to the effectiveness and flexibility of cross-modal hashing, as shown in Fig. 1, users can submit whatever they have to retrieve whatever they want [5], [6].

"Heterogeneous gap" is the key challenge of cross-modal hashing, which means the similarity between different modalities cannot be measured directly. Consequently, some cross-modal hashing methods [7]–[12] have been proposed to bridge this gap. Existing cross-modal hashing methods can be categorized into traditional methods and Deep Neural Networks (DNN) based methods. Moreover, traditional methods can be divided into unsupervised methods and supervised methods based on whether semantic information is leveraged.

Unsupervised cross-modal hashing methods usually project data from different modalities into a common Hamming space to maximize their correlations, which hold the similar idea with Canonical Correlation Analysis (CCA) [13]. Representative unsupervised methods include Inter-Media

This work was supported by the National Natural Science Foundation of China under Grant 61771025.

The authors are with the Institute of Computer Science and Technology, Peking University, Beijing 100871, China. Corresponding author: Yuxin Peng (e-mail: pengyuxin@pku.edu.cn).

Hashing (IMH) [14], Cross-view Hash (CVH) [7], Predictable Dual-view Hashing (PDH) [8], Collective Matrix Factorization Hashing (CMFH) [9] and Composite Correlation Quantization (CCQ) [15]. Compared with unsupervised methods, ***supervised cross-modal hashing methods*** further utilize labeled semantic information to learn hashing functions. Representative supervised cross-modal hashing methods include Cross-Modality Similarity Sensitive Hashing (CMSSH) [16], Heterogeneous Translated Hashing (HTH) [17], Semantic Correlation Maximization (SCM) [10] and Semantics-Preserving Hashing (SePH) [18]. ***DNN based methods*** are inspired by the successful applications of deep learning, such as image classification [19]. Representative deep learning based methods include Deep and Bidirectional Representation Learning Model (DBRLM) [20], Deep Visual-semantic Hashing (DVH) [21] and Cross-Media Neural Network Hashing (CMNNH) [11].

Compared with unsupervised paradigm, supervised methods use labeled semantic information that requires massive labor to collect, resulting in a high labor cost in real world applications. On the contrary, unsupervised cross-modal hashing methods can leverage unlabeled data to realize efficient cross-modal retrieval, which are more flexible and applicable in real world applications. However, most unsupervised methods learn hashing functions by preserving inter-modal and intra-modal correlations, while ignoring the underlying manifold structure across different modalities, which is extremely helpful to capture meaningful nearest neighbors of different modalities. Furthermore, existing works mainly focus on pairwise relation modeling, while ignore the correlations within multiple modalities. To address this problem, in this paper we exploit correlation information from underlying manifold structure of unlabeled data across different modalities to enhance cross-modal hashing learning.

Inspired by recent progress of Generative Adversarial Network (GAN) [22]–[25], which has shown its ability to model the data distribution in an unsupervised fashion. In this paper, we propose multi-pathway generative adversarial hashing (MGAH) approach for unsupervised cross-modal retrieval. We design a graph-based unsupervised correlation method to capture the underlying manifold structure across different modalities, and a multi-pathway generative adversarial network to learn the manifold structure and further enhance the performance by an adversarial boosting paradigm. The main contributions of this paper can be summarized as follows:

- We propose a **multi-pathway generative adversarial network** to model cross-modal hashing in an unsupervised fashion. The proposed network has a multi-pathway structure that can support up to 5 modalities simultaneously. In the proposed network, given the data of any modality, the generative model tries to fit the distribution over the manifold structure, and selects informative data of other modalities to challenge the discriminative model. While the discriminative model learns to distinguish the generated data and the true positive data sampled from the correlation graph to achieve better retrieval accuracy.
- We propose a **correlation graph** based learning approach to capture the underlying manifold structure across different modalities, so that the data of different modalities but within the same manifold can have smaller Hamming distance and promote retrieval accuracy. We also integrate the proposed correlation graph into the proposed generative adversarial network to provide manifold correlation guidance to promote the cross-modal retrieval accuracy.

Compared with our previous conference paper [26], the new contributions of this paper can be summarized as follows: (1) The network structure in [26] has only two pathways, so the cross-modal hashing is limited to 2 modalities (image and text). However, the correlations among five modalities are much more complex than two modalities. Thus in this paper, we further extend the network structure to a five-pathway structure, which can support cross-modal retrieval up to 5 modalities simultaneously. We also extend the loss function and reward function to support multiple modalities at the same time. It is critical to model different modalities at the same time, which can take the advantage of the hidden complementary information. (2) In this paper, we propose a new hyper correlation graph to jointly model the manifold structure across up to 5 modalities. which is vital for only one model to support cross-modal retrieval across 5 modalities. While the correlation graph in [26] only supports to model 2 modalities. Extensive experiments compared with 6 state-of-the-art methods on 3 widely-used datasets verify the effectiveness of our proposed approach. It is noted that we also add an experiment on PKU XMedia dataset [5] to verify the effectiveness of our proposed approach on 5 modalities.

The rest of this paper is organized as follows. In section II, we introduce some related works. In section III, we present our proposed MGAH approach in detail. The experimental results and analyses are reported in section IV. Finally, we conclude this paper in section V.

II. Related work

In this section, we briefly review some related works from two aspects: cross-modal hashing and generative adversarial network.

A. Cross-modal Hashing

Hashing methods for single modality retrieval have been extensively studied in the past decades [1], [2], [27]–[34], and cross-modal hashing methods are receiving increasing attention in recent years. Generally speaking, most cross-modal hashing methods project data of different modalities into a common Hamming space to perform fast retrieval, and they can be divided into traditional methods and DNN based methods. Traditional methods can be further divided into unsupervised and supervised methods based on whether semantic information is involved. We will briefly review some representative works of cross-modal hashing.

Unsupervised cross-modal hashing methods have the similar idea with Canonical Correlation Analysis (CCA) [13], which maps heterogeneous data into a common Hamming space to maximize the correlation. Inter-Media Hashing (IMH) [14] is proposed to learn a common Hamming space to preserve both inter-media and intra-media consistency. Cross-view Hashing (CVH) [7] extends image hashing method

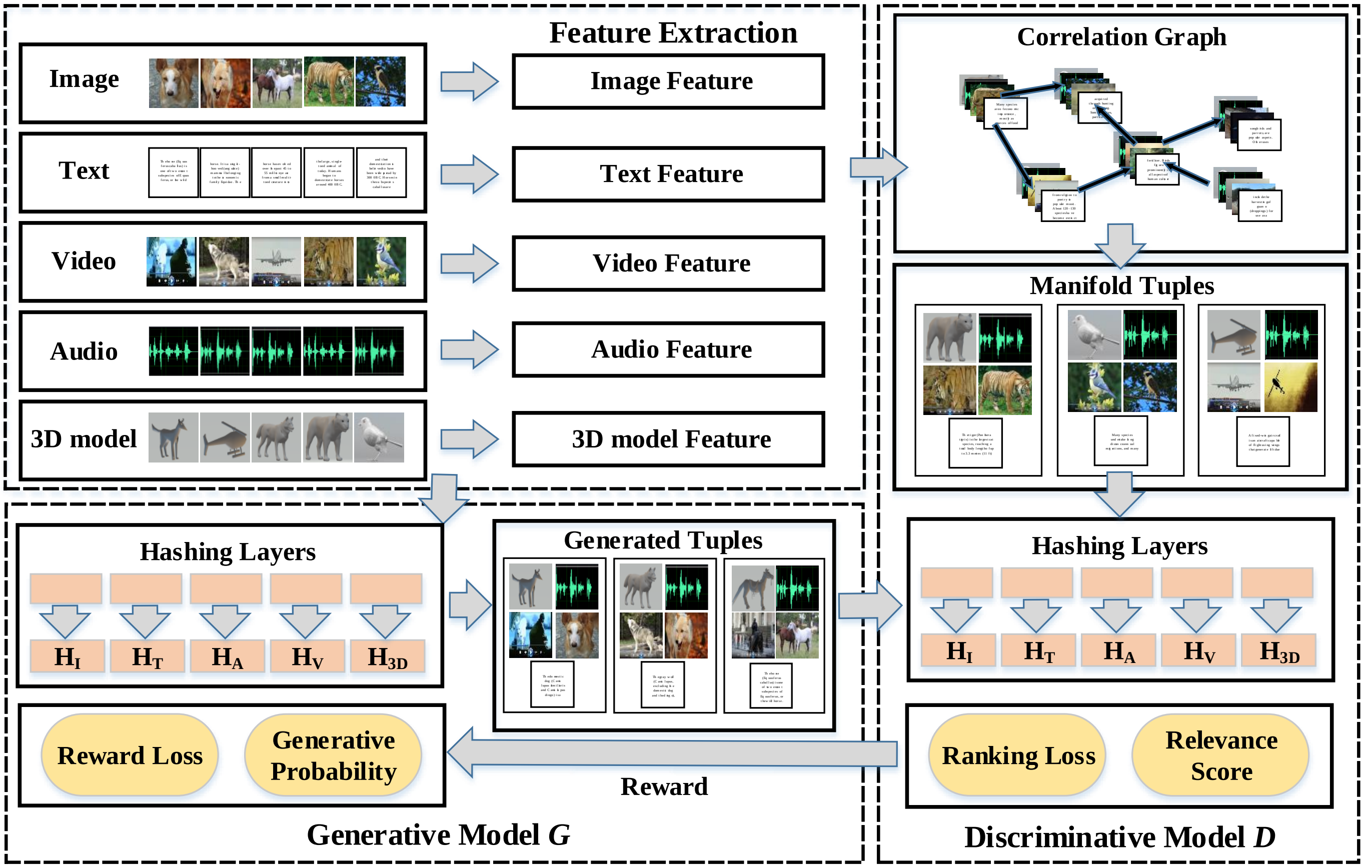


Fig. 2. The overall framework of our proposed multi-pathway generative adversarial hashing (MGAH) approach, which consists of feature extraction part, generative model G and discriminative model D. Both the generative and discriminative models are multi-pathway structures that can support up to 5 modalities.

Spectral Hashing (SH) [31] to consider both intra-view and inter-view similarities with a generalized eigenvalue formulation. Rastegari et al. propose Predictable Dual-View Hashing (PDH) [8], which proposes an objective function to preserve the predictability of pre-generated binary codes, and optimizes the objective function by an iterative method based on block coordinate descent. Hu et al. [35] propose termed collective reconstructive embeddings (CRE) aim to address the problem of heterogeneity and integration complexity in multimodal data. Ding et al. propose Collective Matrix Factorization Hashing (CMFH) [9], which learns unified hash codes from different modalities of one instance by collective matrix factorization with a latent factor model. Latent Semantic Sparse Hashing (LSSH) [36] is proposed to use sparse coding and matrix factorization to learn separate semantic features for images and text, and then map them into a joint abstract space to reduce semantic difference. Wang et al. propose Semantic Topic Multimodal Hashing (STMH) [37], which models texts as semantic topics while images as latent semantic concepts, then maps the learned semantic features for different modalities into a common semantic space, and finally generates hash codes by predicting whether topics or concepts are available in the original data.

Supervised cross-modal hashing methods leverage semantic information obtained from labels of training data, which achieve better retrieval accuracy than unsupervised methods. Cross-modality Similarity Sensitive Hashing (CMSSH) [16] is proposed to model hashing learning as a classification problem, and can be learned in a boosting manner. Zhen et al. propose Co-Regularized Hashing (CRH) [38], which learns hash functions of each bit sequentially so that the bias introduced by each hash function can be minimized. Hu et al. propose Iterative Multi-view Hashing (IMVH) [39], which tries to learn hashing functions by preserving both within-view similarities and between-view correlations. Heterogeneous Translated Hashing (HTH) [17] is proposed to learn different Hamming spaces for different modalities, and then learn translators to align these spaces to perform cross-modal retrieval. Zhang et al. propose Semantic Correlation Maximization (SCM) [10], which constructs semantic similarity matrix based on labels and learns hashing functions to preserve the constructed matrix. Wu et al. propose Quantized Correlation Hashing (QCH) [40] to simultaneously optimize cross-modal correlation and quantization error, which is also considered in many single modality hashing methods. Lin et al. propose Semantics-Preserving Hashing (SePH) [18], which is a two-step supervised hashing method. SePH firstly transforms the given semantic matrix of training data into a probability distribution and approximates it with learned hash codes in Hamming space via minimizing the KL-divergence.

DNN based methods are inspired by recent advance of deep neural networks, which have been applied in many

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2922128, IEEE Transactions on Multimedia

computer vision problems, such as image classification, [19] and object recognition [41]. Zhuang et al. propose Cross-Media Neural Network Hashing (CMNNH) [11], which learns hashing functions by preserving intra-modal discriminative capability and inter-modal pairwise correspondence. Wang et al. propose Deep Multimodal Hashing with Orthogonal Regularization (DMHOR) [42], which learns hashing functions by preserving intra-modal and inter-modal correlation, as well as reducing redundant information between hash bits. Cao et al. propose Cross Autoencoder Hashing (CAH) [43], which maximizes the feature correlation of bimodal data and the semantic correlation provided by similarity labels, and CAH is based on a deep autoencoder structure. Deep Visual-semantic Hashing (DVH) [21] is an end-to-end framework that integrates both feature learning and hashing function learning. Jiang et al. propose Deep Cross-modal Hashing (DCMH) [44], which performs feature learning and hashing function learning simultaneously. Yang et al. propose Pairwise Relationship Guided Deep Hashing (PRDH) [45], which integrates different pairwise constraints to guide hash code learning. Deng et al. propose Triplet-Based Deep Hashing (TDH) [46], which utilizes triplet labels to describe the relative relationship among three instances. Yang et al. propose Semantic Structure based unsupervised Deep Hashing (SSDH) [47], which tries to capture and utilize the semantic relationships between points in unsupervised settings. Besides hashing based approaches, Yang et al. propose Shared Predictive Deep Quantization (SPDQ) [48], which tries to exploit the intrinsic correlations among multiple modalities and learn compact codes of higher quality in a joint deep network architecture.

Compared with unsupervised paradigm, supervised methods rely on large amount of labeled training data which are labor intensive to obtain. It is even harder to label cross-modal hashing training data, since multiple modalities are involved. On the contrary, unsupervised cross-modal hashing methods can leverage unlabeled data to realize efficient cross-modal retrieval, which are more flexible and applicable in real world applications. However, most unsupervised methods ignore the underlying manifold structure across different modalities, which is extremely helpful to capture meaningful nearest neighbors of different modalities. To address this problem, in this paper we exploit correlation information from underlying manifold structure of unlabeled data across different modalities to enhance cross-modal hashing learning.

B. Generative Adversarial Network

Generative Adversarial Network (GAN) [22] is first proposed to estimate generative model by an adversarial process. GAN consists of two models: a generative model G that captures the data distributions, and a discriminative model D that estimates the probability that a sample comes from real data rather than G. These two models are trained in an adversarial way so that they compete with each other, and both of them can learn better representations of the data. Inspired by the ability of modeling data distributions of GAN, many works have attempted to apply GAN in various computer vision problems. The most popular one is image synthesis. Radford et al. propose Deep Convolutional GAN (DCGAN) [49], which adopts convolutional decoder with batch normalization and achieves better image synthesis results. Mirza et al. proposed Conditional GAN (CGAN) [50], which provides side information for both generative and discriminative model to control the generated data. Inspired by CGAN, many works extend its idea to image synthesis problem, Reed et al. propose text-conditional GAN [23] which can generate images conditioned by textual descriptions. Odena et al. propose auxiliary classifier GAN (AC-GAN) [51] that generates images conditioned by class labels. Besides image synthesis, GAN is also applied to video prediction [52] and object detection [53]. Wang et al. propose Adversarial Cross-modal Retrieval (ACMR) [54] to learn common representation subspace for different modalities.

Inspired by the ability of GAN to model data distributions, in this paper we propose a multi-pathway generative adversarial hashing (MGAH) approach for unsupervised cross-modal retrieval. It aims to design a generative model to fit the distributions of different modalities, and a discriminative model to maintain the manifold structures across different modalities. These two models play a minimax game to iteratively optimize each other and boost cross-modal retrieval accuracy.

III. The proposed approach

Figure 2 presents the overview of our proposed approach, which consists of three parts, namely feature extraction, generative model G and discriminative model D. The feature extraction part employs different features to represent unlabeled data of different modalities as original features. The detailed implementation of this part will be described in section IV. Given a data of one modality, G attempts to select informative data from other modalities to generate a tuple of data that contains all modalities, and send them to D. In D, we construct a correlation graph, which can capture the manifold structure among the original features. D receives the generated tuples as inputs, and also samples positive data from the constructed graph to form a true manifold tuple. Then D tries to distinguish the manifold and the generated tuples in order to get better discriminate ability. These two models play a minimax game to boost each other, and the finally trained D can be used as the cross-modal hashing model. It is noted that both the generative and discriminative models are multi-pathway structures, which can support to model multiple modalities at the same time.

We denote the cross-modal dataset as $D_a = \{D^1, \cdots, D^s\}$, in which s denotes the number of modalities. In this paper $D_a = \{D^1, \cdots, D^s\}$ is further split into a retrieval database $D_{db} = \{D^1_{db}, \cdots, D^s_{db}\}$ and a query set $D_q = \{D^1_q, \cdots, D^s_q\}$. In the retrieval database D_{db}, $D^i_{db} = \{x^i_p\}^n_{p=1}$, where n is the number of data points in the retrieval database, i denotes the i-th modality. In the query set D_q, $D^i_q = \{x^i_p\}^t_{p=1}$, where t is the number of data points in the query set. The aim of cross-modal hashing is to learn functions to generate hash codes for data of different modalities, so that different modal data that has similar semantics are close in the common Hamming space. In addition, we denote the hash code length as l. By the generated hash codes, we can retrieve the relevant data by a query of any modality from the database of other modalities efficiently.

A. Generative Model

The network of generative model has a multi-pathway architecture, which receives the original features of each modality as inputs. Each pathway consists of a common representation layer and a hashing layer, whose implementations are two fully-connected layers. The first layer can convert the modality specific features to common representations, which make the instances of different modalities measurable in a common space. The representation produced by this layer can be denoted as follows:

$$\phi_c(x) = tanh(W_c x + b_c) \tag{1}$$

where x denotes the original features of different modalities, W_c is the weight parameter of the common representation layer, and b_c is the bias parameter.

The hashing layer can map the common representations into binary hash codes, so that the similarity between different modalities can be measured by fast Hamming distance calculation. The continuous real values of hash code is defined as:

$$h(x) = sigmoid(W_h \phi_c(x) + b_h) \tag{2}$$

where W_h is the weight parameter and b_h is the bias parameter. Then we can get the binary codes by a thresholding function:

$$b(x) = sgn(h_k(x) - 0.5), \quad k = 1, 2, \cdots, l \tag{3}$$

where l denotes the hash code length. Considering that it is hard to optimize binary codes directly, we use relaxed continuous real valued hash codes $h(x)$ in the training process.

Given a data of one modality, the goal of the generative model G is to fit the distribution over the manifold structure and select informative data of other modalities to challenge the discriminative model. The generative probability of G is defined as $p_\theta(x_U^i|q)$, which is the foundation to select relevant instance of other modalities when given a query of one modality. For example, given a image query q_i, the generative model tries to select relevant text t^U from T_{db}.

The generative probability $p_\theta(x_U^i|q)$ is defined as a softmax function:

$$p_\theta(x_U^i|q) = \frac{\exp(-\|h(q) - h(x_U^i)\|^2)}{\sum_{x_U^i} \exp(-\|h(q) - h(x_U^i)\|^2)} \tag{4}$$

where i denotes the $i-th$ modality. Given a query of one modality, we can use equation (4) to calculate the probability of each candidate of any modality, which indicates the possibility of becoming a relevant sample. Then for each query, we generate a tuple that contains all the modalities as the input of the discriminative model.

B. Discriminative Model

The network structure of discriminative model is the same as the generative model. The input of this network is the generated tuples by the generative model, and the manifold tuples provided by the proposed correlation graph. The goal of the discriminative model is to distinguish whether an input tuple is generated or from the correlation graph.

First of all, we introduce the correlation graph in the discriminative model. We propose a correlation graph to guide the training of discriminative model. The correlation graph can capture the underlying manifold structure across different modalities, so that data of different modalities but within the same manifold can have small Hamming distance and promote retrieval accuracy.

Specifically, we first construct undirected graphs $Graph_i = (V, W_i)$ for each modality, where i denotes the $i{-}th$ modality, V denotes the vertices and W_i is the similarity matrix. W_i is defined as follows:

$$w(p,q) = \begin{cases} 1: & x_p^i \in NN_k(x_q^i) \\ 0: & otherwise \end{cases} \tag{5}$$

where $NN_k(x_q^i)$ denotes the k-nearest neighbors of x_q^i in the training set. It is noted that cross-modal data naturally has the coexisting information, for example, text and image are often occurred in the same page, thus if the corresponding text t_k is within the same manifold with text query q^j, the coexisted image i_k is also in the same manifold with q^j and vice versa. Then we use the coexisting information to combine each $Graph_i$, which forms a *hypergraph*. In the constructed correlation graph, each node contains data of all modalities. Then we sample the data of the true distributions based on the constructed graph. Given a query q, we select x_k^i as the true relevant instance of q^j and combine them to form the manifold tuple $p_{manifold}(x_k^i|q)$, where i denotes the $i - th$ modality. By this definition, we intend to utilize the underlying manifold structure of different modality to guide the training of discriminative model. Intuitively, we want the data of different modality but within the same manifold to have small Hamming distance (e.g. small Hamming distance between text query q^j and image i_k).

After receiving the generated and manifold tuples, the discriminative model predicts a relevance score between each tuple as the judgment result. So the relevance score between instance x and its query q is defined as $f_\phi(x, q)$. The goal of discriminative model is to distinguish the true relevant data (manifold tuples) and non-relevant data (generated tuples) for a query q accurately.

The relevance score of $f_\phi(x^G, q)$ is defined by triplet ranking loss as follows:

$$\begin{aligned} f_\phi(x^G, q) = \max(0, m + \|h(q) - h(x^M)\|^2 \\ - \|h(q) - h(x^G)\|^2) \end{aligned} \tag{6}$$

where x^M is a manifold tuple with query q selected from the correlation graph, x^G is a generated tuple by the generative model, m is a margin parameter which is set to be 1 in our proposed approach, $\|\cdot\|^2$ means the average distance between query q and the data of different modality in tuple X^M or X^G. The above equation means that we want the distance between manifold tuple (q, x^M) smaller than that of generated tuple (q, x^G) by a margin m, so that the discriminative model can draw a clear distinguishing line between the manifold and generated tuples.

Then discriminative model D uses the relevance score to produce predicted probability of tuple x by a sigmoid function:

$$D(x|q) = sigmoid(f_\phi(x,q)) = \frac{\exp(f_\phi(x,q))}{1+\exp(f_\phi(x,q))} \tag{7}$$

The generative model tries to select informative data to challenge the discriminative model, which limits its capability to perform cross-modal retrieval. By contrast, the discriminative model is suitable for retrieving data across different modalities, after being promoted greatly by the generative model. Therefore after the proposed MGAH is trained, we use the discriminative model to perform cross-modal retrieval via produced hash codes.

C. Training algorithm

Given the definitions of the generative and discriminative models, we can conduct a minimax game for training them. However, before the adversarial training, the discriminative model should have some level of discriminative ability to guide the adversarial training. Thus the training algorithm contains two phases: we first pretrain the discriminative model by the constructed correlation graph, and then we conduct adversarial training between the generative model and the discriminative model.

1) Pretraining phase: We pretrain the discriminative model to preserve the manifold structure provided by the correlation graph. Specifically, in equation (6), instead of using the generated tuples, we randomly choose the negative tuples X^N from the correlation graph that X^N is not in the same manifold with query q. Then the pretraining loss function can be defined as:

$$\begin{aligned}\mathcal{J}(q,x^G,X^N) = \max(0, m + \|h(q)-h(x^M)\|^2 \\ - \|h(q)-h(x^N)\|^2)\end{aligned} \tag{8}$$

The loss function is similar with triplet loss [55], we can use back propagation algorithm to update the parameters of the discriminative model.

2) Adversarial training phase: Given a query, the generative model attempts to generate a tuple which is close to the manifold tuple to fool the discriminative model. The discriminative model tries to distinguish between manifold tuple sampled from the correlation graph and the generated tuple, which forms an adversarial training process. Inspired by the GAN [22], this adversarial process can be defined:

$$\begin{aligned}\mathcal{V}(G,D) = &\min_\theta \max_\phi \sum_{j=1}^{n} (E_{x\sim p_{true}(x^M|q^j)}[log(D(x^M|q^j))] \\ &+ E_{x\sim p_\theta(x^G|q^j)}[log(1-D(x^G|q^j))])\end{aligned} \tag{9}$$

The generative and discriminative models can be learned iteratively by maximizing and minimizing the above object function. As general training process, the discriminative model tries to *maximize* equation (9), while the generative model attempts to *minimize* equation (9) and fit the distribution over the manifold structure. We fix the discriminative model when training the generative model by the following equation:

$$\begin{aligned}\theta^* = &\arg\min_\theta \sum_{j=1}^{n} (E_{x\sim p_{true}(x^G|q^j)}[log(sigmoid(f_{\phi^*}(x^M,q^j)))] \\ &+ E_{x\sim p_\theta(x^G|q^j)}[log(1-sigmoid(f_{\phi^*}(x^G,q^j)))])\end{aligned} \tag{10}$$

where f_{ϕ^*} denotes the discriminative model at the previous iteration. The traditional GAN uses continuous noise vector to generate new data and is trained by stochastic gradient descent algorithm. By contrast, the generative model of our proposed MGAH selects data from unlabeled data to generate tuples and can not be optimized continuously due to the discrete selective strategy. We utilize reinforcement learning based parameters update policy to train the generative model as follows:

$$\begin{aligned}&\nabla_\theta E_{x\sim p_\theta(x^G|q^j)}[log(1+\exp(f_\phi(x^G,q^j)))] \\ &= \sum_{k=1}^{m} \nabla_\theta p_\theta(x_k^G|q^j) log(1+\exp(f_\phi(x_k^G,q^j))) \\ &= \sum_{k=1}^{m} p_\theta(x_k^U|q^j) \nabla_\theta log p_\theta(x_k^G|q^j) log(1+\exp(f_\phi(x_k^G,q^j))) \\ &= E_{x\sim p_\theta(x^G|q^j)}[\nabla_\theta log p_\theta(x^G|q^j) log(1+\exp(f_\phi(x^G,q^j)))] \\ &\simeq \frac{1}{m}\sum_{k=1}^{m} \nabla_\theta log p_\theta(x_k^G|q^j) log(1+\exp(f_\phi(x_k^G,q^j)))\end{aligned} \tag{11}$$

where k denotes the k-th instance selected by generative model according to a query q^j. From the perspective of reinforcement learning, according to the environment q^k, x_k^G is the action taken by policy $logp_\theta(x_k^G|q^j)$, and $log(1+\exp(f_\phi(x_k^G,q^j)))$ acts as the reward, which encourages the generative model to select data close to the distribution over manifold structure. We summarize the overall training process in Algorithm (1).

Finally the trained discriminative model can be used to generate binary codes for any input data of any modality, and cross-modal retrieval can be performed by fast Hamming distance computation between query and each data in the database.

IV. Experiments

In this section, we present the experimental results of our proposed MGAH approach. We first introduce the datasets, evaluation metrics and implementation details. Then we compare and analyze the results of MGAH with 6 state-of-the-art methods and 2 baseline methods.

A. Datasets

In the experiments, we conduct cross-modal hashing on 3 widely-used datasets: NUS-WIDE [56], MIRFLICKR [57] and PKU XMedia [5].

- **NUS-WIDE**[1] dataset [56] is a relatively large-scale image/tag dataset with 269498 images. Each image has corresponding textual tags, which are regarded as the text modality in our experiments. NUS-WIDE dataset has 81 categories, but there are overlaps among the categories.

[1] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

TABLE I
THE MAP SCORES OF TWO RETRIEVAL TASKS ON NUS-WIDE DATASET WITH DIFFERENT LENGTHS OF HASH CODES.

Methods	image→text				text→image			
	16	32	64	128	16	32	64	128
CVH [7]	0.458	0.432	0.410	0.392	0.474	0.445	0.419	0.398
PDH [8]	0.475	0.484	0.480	0.490	0.489	0.512	0.507	0.517
CMFH [9]	0.517	0.550	0.547	0.520	0.439	0.416	0.377	0.349
CCQ [15]	0.504	0.505	0.506	0.505	0.499	0.496	0.492	0.488
CMSSH [16]	0.512	0.470	0.479	0.466	0.519	0.498	0.456	0.488
SCM_orth [10]	0.389	0.376	0.368	0.360	0.388	0.372	0.360	0.353
SCM_seq [10]	0.517	0.514	0.518	0.518	0.518	0.510	0.517	0.518
Baseline	0.540	0.537	0.573	0.598	0.554	0.555	0.583	0.608
Baseline-GAN	0.575	0.594	0.602	0.623	0.580	0.609	0.617	0.629
MGAH (Ours)	**0.613**	**0.623**	**0.628**	**0.631**	**0.603**	**0.614**	**0.640**	**0.641**

TABLE II
THE MAP SCORES OF TWO RETRIEVAL TASKS ON MIRFLICKR DATASET WITH DIFFERENT LENGTHS OF HASH CODES.

Methods	image→text				text→image			
	16	32	64	128	16	32	64	128
CVH [7]	0.602	0.587	0.578	0.572	0.607	0.591	0.581	0.574
PDH [8]	0.623	0.624	0.621	0.626	0.627	0.628	0.628	0.629
CMFH [9]	0.659	0.660	0.663	0.653	0.611	0.606	0.575	0.563
CCQ [15]	0.637	0.639	0.639	0.638	0.628	0.628	0.622	0.618
CMSSH [16]	0.611	0.602	0.599	0.591	0.612	0.604	0.592	0.585
SCM_orth [10]	0.585	0.576	0.570	0.566	0.585	0.584	0.574	0.568
SCM_seq [10]	0.636	0.640	0.641	0.643	0.661	0.664	0.668	0.670
Baseline	0.619	0.631	0.633	0.646	0.625	0.635	0.634	0.649
Baseline-GAN	0.630	0.643	0.651	0.664	0.660	0.657	0.670	0.688
MGAH (Ours)	**0.685**	**0.693**	**0.704**	**0.702**	**0.673**	**0.676**	**0.686**	**0.690**

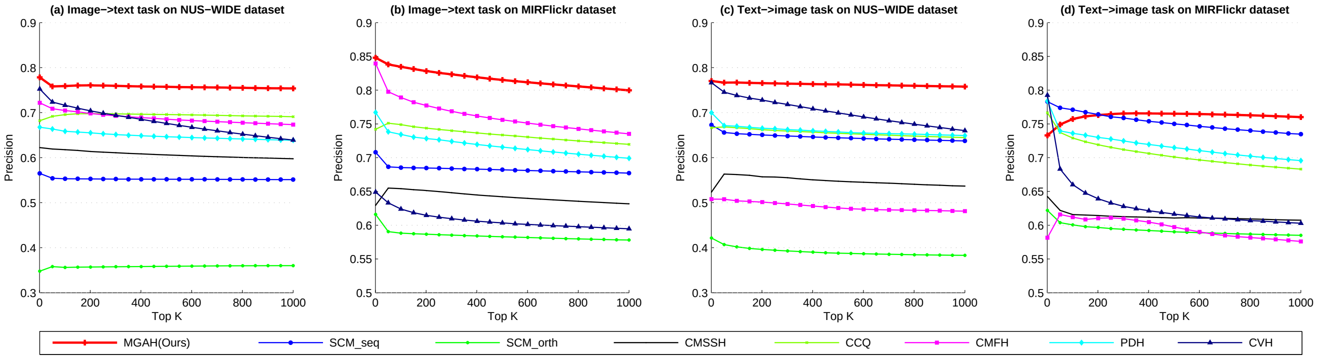


Fig. 3. The top$K$-precision curves with 128 bit hash codes. The left two figures present the result of image→text task on NUS-WIDE and MIRFlickr datasets, while the right two figures show the result of text→image task.

Following [18], we select the 10 largest categories and the corresponding 186557 images. We take 1% data of NUS-WIDE dataset as the query set, and the rest as the retrieval database. We randomly selected 5000 images as training set for the supervised methods. We represent each image by 4096 dimensional deep features extracted from 19-layer VGGNet, and each text by 1000 dimensional BoW.

- **MIRFlickr**[2] dataset [57] has 25000 images collected from Flickr, which has 24 categories. Each image is also associated with text tags. Following [18], we take 5% of the dataset as the query set and the remaining as the retrieval database. We also randomly select 5000 images as training set for supervised methods. Similarly, we represent each image by 4096 dimensional deep features extracted from 19-layer VGGNet, and each text by 1000 dimensional BoW.
- **PKU XMedia**[3] dataset [5] is the first publicly available cross-modal dataset with up to 5 modalities (text, image, video, audio and 3D model), for comprehensive evaluation of cross-modal retrieval. There are totally 20 categories in PKU XMedia dataset, which are specific objects such as insect, bird, wind, dog, and elephant. There are 250 texts, 250 images, 25 videos, 50 audio clips and 25 3D models for each category, and the total number of instances is 12,000. For this dataset, we randomly select 2400 instances as query set, while the remaining 9600 instances as retrieval database. For PKU XMedia dataset,

[2] http://http://press.liacs.nl/mirflickr

[3] http://www.icst.pku.edu.cn/mipl/PKU XMedianet

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2922128, IEEE Transactions on Multimedia

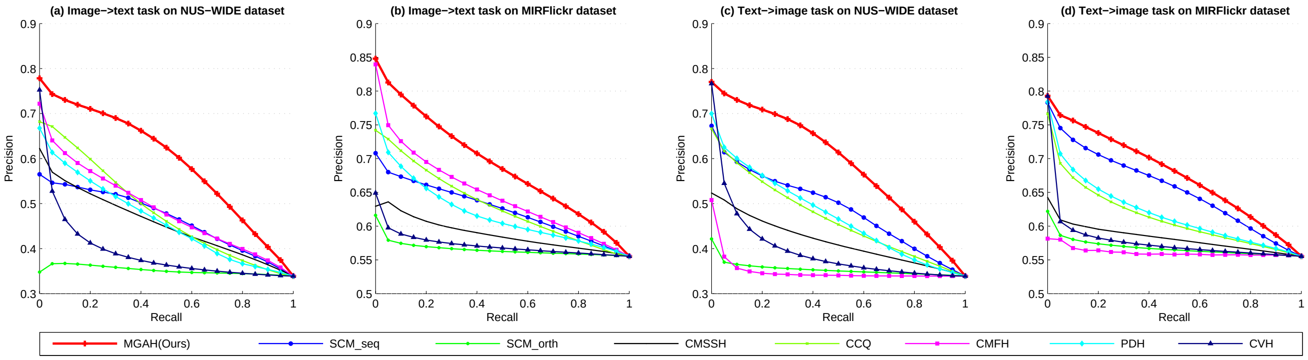


Fig. 4. The precision-recall curves with 128 bit hash codes. The left two figures present the result of image→text task on NUS-WIDE and MIRFlickr datasets, while the right two figures show the result of text→image task.

Algorithm 1 Training algorithm of the proposed MGAH

```
Input: The generative model $G$, the discriminative model $D$,
    the training data $D_{db}$
Randomly initialize the parameters of $G$ and $D$
repeat
    Sample $m$ manifold tuples and negative tuples based on
    correlation graph
    Train the discriminative model based on equation (8)
until The discriminative model $D$ converges
repeat
    for d-step do
        Generate $m$ tuples by $p_{\theta^*}(x_U|q)$ given query $q$
        Sampled $m$ manifold tuples based on correlation
        graph
        Train the discriminative model $D$ by equation (6)
    end for
    for g-step do
        Generate $m$ tuples by $p_\theta(x_U|q, r)$ given query $q$
        Calculate reward by $log(1 + \exp(f_\phi(x_U, q)))$
        Update parameters of generative model $G$ by equation (11)
    end for
until MGAH converges
Output: Optimized generative model $G$ and discriminative
    model $D$
```

we represent texts by 3000-dimensional BoW features, images by 4096 dimensional deep features extracted from 19-layer VGGNet, videos by 4096 dimensional deep features extracted from C3D model [58] pre-trained on Sports1M [59], audios by the 78-dimensional features extracted by jAudio [60] using its default setting, and 3D models by the concatenated 4700-dimensional vectors of a LightField descriptor set [61].

B. Compared Methods

In order to verify the effectiveness of our proposed approach, there are 4 unsupervised methods and 2 supervised methods compared in the experiment, including unsupervised methods CVH [7], PDH [8], CMFH [9] and CCQ [15], and

TABLE III
THE MAP SCORES OF IMAGE→ALL TASK ON PKU XMEDIA DATASET WITH DIFFERENT LENGTHS OF HASH CODES.

Methods	image→all			
	16	32	64	128
CVH [7]	0.070	0.065	0.061	0.055
PDH [8]	0.253	0.297	0.312	0.245
CMFH [9]	0.174	0.160	0.135	0.108
CCQ [15]	0.056	0.057	0.080	0.194
CMSSH [16]	0.079	0.085	0.088	0.085
SCM_orth [10]	0.125	0.101	0.075	0.060
SCM_seq [10]	0.070	0.069	0.100	0.337
Baseline	0.272	0.306	0.351	0.380
Baseline_GAN	0.277	0.317	0.372	0.416
MGAH(Ours)	**0.290**	**0.333**	**0.389**	**0.422**

TABLE IV
THE MAP SCORES OF TEXT→ALL TASK ON PKU XMEDIA DATASET WITH DIFFERENT LENGTHS OF HASH CODES.

Methods	text→all			
	16	32	64	128
CVH [7]	0.248	0.127	0.094	0.072
PDH [8]	0.201	0.232	0.287	0.304
CMFH [9]	0.074	0.071	0.066	0.062
CCQ [15]	0.213	0.246	0.267	0.289
CMSSH [16]	0.079	0.084	0.084	0.085
SCM_orth [10]	0.186	0.176	0.151	0.072
SCM_seq [10]	0.217	0.199	0.261	0.206
Baseline	0.213	0.223	0.262	0.298
Baseline_GAN	0.231	0.245	0.284	0.321
MGAH(Ours)	**0.237**	**0.262**	**0.319**	**0.334**

TABLE V
THE MAP SCORES OF AUDIO→ALL TASK ON PKU XMEDIA DATASET WITH DIFFERENT LENGTHS OF HASH CODES.

Methods	audio→all			
	16	32	64	128
CVH [7]	0.062	0.076	0.076	0.051
PDH [8]	0.098	0.111	0.123	0.118
CMFH [9]	0.063	0.064	0.064	0.064
CCQ [15]	0.086	0.083	0.091	0.180
CMSSH [16]	0.086	0.080	0.080	0.080
SCM_orth [10]	0.125	0.132	0.144	0.051
SCM_seq [10]	0.082	0.077	0.066	0.113
Baseline	0.103	0.120	0.133	0.136
Baseline_GAN	0.119	0.127	0.143	0.152
MGAH(Ours)	**0.137**	**0.142**	**0.154**	**0.172**

TABLE VI
THE MAP SCORES OF VIDEO→ALL TASK ON PKU XMEDIA DATASET WITH DIFFERENT LENGTHS OF HASH CODES.

Methods	video→all			
	16	32	64	128
CVH [7]	0.143	0.125	0.086	0.069
PDH [8]	0.242	0.234	0.265	0.285
CMFH [9]	0.100	0.111	0.112	0.118
CCQ [15]	0.109	0.116	0.147	0.176
CMSSH [16]	0.075	0.083	0.086	0.088
SCM_orth [10]	0.213	0.130	0.088	0.061
SCM_seq [10]	0.258	0.256	0.263	0.107
Baseline	0.232	0.229	0.306	0.334
Baseline_GAN	0.242	0.231	0.323	0.367
MGAH(Ours)	**0.269**	**0.268**	**0.347**	**0.376**

TABLE VII
THE MAP SCORES OF 3D→ALL TASK ON PKU XMEDIA DATASET WITH DIFFERENT LENGTHS OF HASH CODES.

Methods	3D→all			
	16	32	64	128
CVH [7]	0.073	0.058	0.056	0.055
PDH [8]	0.196	0.204	0.236	0.257
CMFH [9]	0.103	0.107	0.117	0.126
CCQ [15]	0.120	0.122	0.130	0.165
CMSSH [16]	0.084	0.084	0.087	0.082
SCM_orth [10]	0.156	0.146	0.129	0.078
SCM_seq [10]	0.103	0.111	0.213	0.255
Baseline	0.195	0.196	0.234	0.252
Baseline_GAN	0.203	0.207	0.260	0.290
MGAH(Ours)	**0.219**	**0.224**	**0.291**	**0.306**

supervised methods CMSSH [16] and SCM [10]. We briefly introduce those compared methods as follows:

- CVH [7] extends Spectral Hashing (SH) [31] to considers both intra-view and inter-view similarities with a generalized eigenvalue formulation.
- PDH [8] proposes to preserve the predictability of pre-generated binary codes, and optimize the objective function by an iterative method based on block coordinate descent algorithm.
- CMFH [9] learns unified hash codes from different modalities of one instance by collective matrix factorization with a latent factor model.
- CCQ [15] jointly learns the correlation-maximal mappings that transform different modalities into isomorphic latent space, and learns composite quantizers that convert the isomorphic latent features into compact binary codes.
- CMSSH [16] models hashing learning as a classification problem, and it is learned in a boosting manner.
- SCM [10] constructs semantic similarity matrix based on labels and learns hashing functions to preserve the constructed matrix.

Besides state-of-the-art methods, we also compare our MGAH approach with 2 baseline methods to verify the effectiveness of our contributions.

- *Baseline*: We design a baseline method without the correlation graph and adversarial training denoted as *Baseline*. It is implemented by training the discriminative model alone with a triplet ranking loss in equation (8). However, instead of using the proposed correlation graph, the positive pairs are provided by cross-modal datasets, while the negative pairs are randomly selected from unpaired data. Thus this *Baseline* method is without correlation graph and adversarial learning.
- *Baseline-GAN*: We add the adversarial training to *Baseline*, which means that we further promote discriminative model in *Baseline* by adversarial training defined in equation (9). Thus *Baseline-GAN* is without correlation graph but with adversarial training.

Comparing *Baseline-GAN* with *Baseline*, we can verify the effectiveness of our proposed generative adversarial network for cross-modal hashing. Since the proposed approach MGAH is with both correlation graph and adversarial training, comparing MGAH with Baseline-GAN can demonstrate the effectiveness of the proposed correlation graph.

C. Retrieval Tasks and Evaluation Metrics

The retrieval tasks in our paper belong to information retrieval tasks, which are similar with existing works, such as CVH [7], PDH [8] and CMFH [9]. Specifically, for the NUS-WIDE and MIRFlickr datasets, two retrieval tasks are performed: retrieving text by image query (image→text) and retrieving images by text query (text→image). We first obtain the hash codes for the images and texts in the query and retrieval database with our MGAH approach and all the compared methods. Then we take one of the images as query, compute the Hamming distance with all text in database. While for the PKU XMedia dataset, five retrieval tasks are performed, where we use a query of one modality to retrieve all the modalities in the retrieval database. We denote these five tasks as image→all, text→all, audio→all, video→all and 3d→all. It is noted that for all the tasks, the category labels are only used for evaluation.

We use 3 evaluation metrics to measure the retrieval effectiveness: Mean Average Precision (MAP), precision recall curve (PR-curve) and precision at top k returned results (top$K$-precision), which are defined as follows:

- The **MAP scores** are computed as the mean of average precision (AP) for all queries, and AP is computed as:

$$AP = \frac{1}{R} \sum_{k=1}^{n} \frac{k}{R_k} \times rel_k \tag{12}$$

  where n is the size of database, R is the number of relevant instances in database, R_k is the number of relevant images in the top k returns, and $rel_k = 1$ if the image ranked at k-th position is relevant and 0 otherwise.
- Precision recall curve (**PR-curve**): The precision at certain level of recall of the retrieved ranking list, which is widely used to measure the retrieval performance.
- Precision at top k returned results (**top$K$-precision**): The precision with respect to different numbers of retrieved samples from the ranking list.

It should be noted that the MAP score is computed for all the retrieval results with 4 different lengths of hash codes, while PR-curve and top$K$-precision are evaluated on 128 bit hash codes.

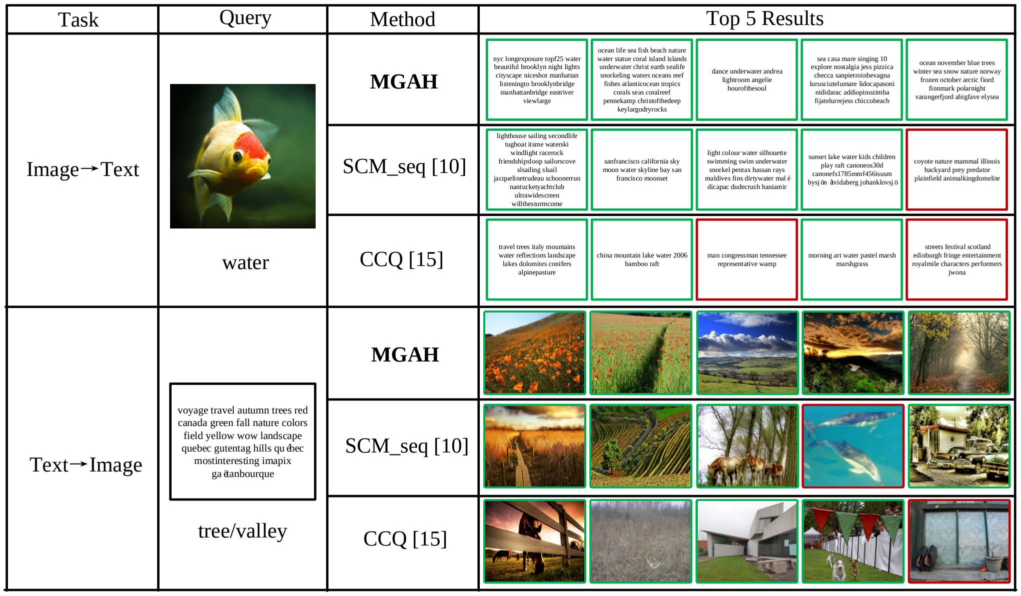

Task
Query
Method
Top 5 Results
Image→Text
water
MGAH
nyc longexposure topf25 water beautiful brooklyn night lights cityscape niceshot manhattan listeningto brooklynbridge manhattanbridge eastriver viewlarge
ocean life sea fish beach nature water statue coral island islands underwater christ earth sealife snorkeling waters oceans reef fishes atlanticocean tropics corals seas coralreef pennekamp christofthedeep keylargodryrocks
dance underwater andrea lightroom angelie hourofthesoul
sea casa mare singing 10 explore nostalgia jess pizzica checca sanpietroinbevagna lurusciutelumare lidocapasoni nididarac addiopinozimba fijatelurrejess chiccobeach
ocean november blue trees winter sea snow nature norway frozen october arctic fiord finnmark polarnight varangerfjord abigfave elysea
SCM_seq [10]
lighthouse sailing secondlife tugboat itsme waterski windlight racerock friendshipsloop sailorscove slsailing slsail jacquelinetrudeau schoonerrun nantucketyachtclub ultrawidescreen willthestormcome
sanfrancisco california sky moon water skyline bay san francisco moonset
light colour water silhouette swimming swim underwater snorkel pentax hassan rays maldives fins dirtywater mal é dicapac dudecrush haniamir
sunset lake water kids children play raft canoneos30d canonefs1785mmf456isusm bysj ön åvidaberg johanklovsj ö
coyote nature mammal illinois backyard prey predator plainfield animalkingdomelite
CCQ [15]
travel trees italy mountains water reflections landscape lakes dolomites conifers alpinepasture
china mountain lake water 2006 bamboo raft
man congressman tennessee representative wamp
morning art water pastel marsh marshgrass
streets festival scotland edinburgh fringe entertainment royalmile characters performers jwona
Text→Image
voyage travel autumn trees red canada green fall nature colors field yellow wow landscape quebec gutentag hills qu ébec mostinteresting imapix ga ëtanbourque
tree/valley
MGAH
SCM_seq [10]
CCQ [15]

Fig. 5. The qualitative results of our proposed MGAH and compared methods on NUSWIDE dataset. The green rectangles denote the correct retrieval results. The red rectangles indicate wrong retrieval results.

D. Implementation Details

In this section, we present the implementation details of our MGAH in the experiments. We implement the proposed MGAH by tensorflow[4]. The dimension of common representation layer is set to be 4096, while the hashing layer's dimension is set to be the same as the hash code length.

Moreover, we train the proposed MGAH in a mini-batch way and set the batch size as 64 for both the discriminative and generative models. We train the proposed MGAH iteratively. After the discriminative model is trained in 1 epoch, the generative model respectively will be trained in 1 epoch. The learning rate of MGAH is decreased by a factor of 10 every two epochs, while it is initialized as 0.01.

For the compared methods, we apply the implementations provided by their authors, and follow their best settings to preform the experiments. And it is noted that for fair comparison between different methods, we use the same image and text features for all compared methods. And for PKU XMedia dataset, since all compared methods are designed for only two modalities, we train a model between any two modalities, then we combine the results between two modalities to obtain the final results. For example, for the image→all task, we train 4 models between image and text, audio, video, 3D model, then the results are combined by using the 4 trained models. *It is noted that we have to train 20 models for the compared methods to support the retrieval, while our proposed MGAH only needs one model to support all modalities.*

E. Experiment Results

1) Results on two modalities: Tables I and II show the MAP scores of our MGAH and the compared methods on NUS-WIDE and MIRFlickr datasets. Compared with state-of-the-art methods, it can be seen that our proposed MGAH approach achieves the best retrieval accuracy on all 2 datasets. For convenience, we categorize these result tables into three parts: unsupervised compared methods, supervised compared methods and baseline methods. On NUS-WIDE dataset, our proposed MGAH keeps the best average MAP score of 0.624 on image$\rightarrow$text task and 0.625 on text$\rightarrow$image task. Compared with the best unsupervised methods CCQ [15], our MGAH achieves an inspiring accuracy improvement from 0.505 to 0.624 on image$\rightarrow$text task, and improves the average MAP score from 0.494 to 0.625 on text$\rightarrow$image task. Even compared with supervised methods SCM_seq [10], our MGAH also improves average MAP scores from 0.517 to 0.624 on image$\rightarrow$text task, and from 0.516 to 0.625 on text$\rightarrow$image task. We can observe the similar trends on MIRFlickr dataset from Tables II.

Figures 3 and 4 show the top$K$-precision and precision-recall curves on the two datasets with 128 bit code length. We can observe that on both image$\rightarrow$text and text$\rightarrow$image tasks, MGAH achieves the best accuracy among all compared unsupervised methods. MGAH even achieves better retrieval accuracy than compared supervised methods on most of the evaluation metrics, which further demonstrates the effectiveness of our proposed approach.

2) Results on multiple modalities: Tables III, IV, V, VI and VII demonstrate the MAP scores of our proposed MGAH and the compared methods on the five retrieval tasks on PKU XMedia dataset. From these tables we can observe that the proposed MGAH achieves the best map scores on all the five retrieval tasks. This is because two advantages of the proposed MGAH: (1) Our proposed MGAH models the manifold structure by a multi-pathway generative adversarial

[4] https://www.tensorflow.org

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2922128, IEEE Transactions on Multimedia

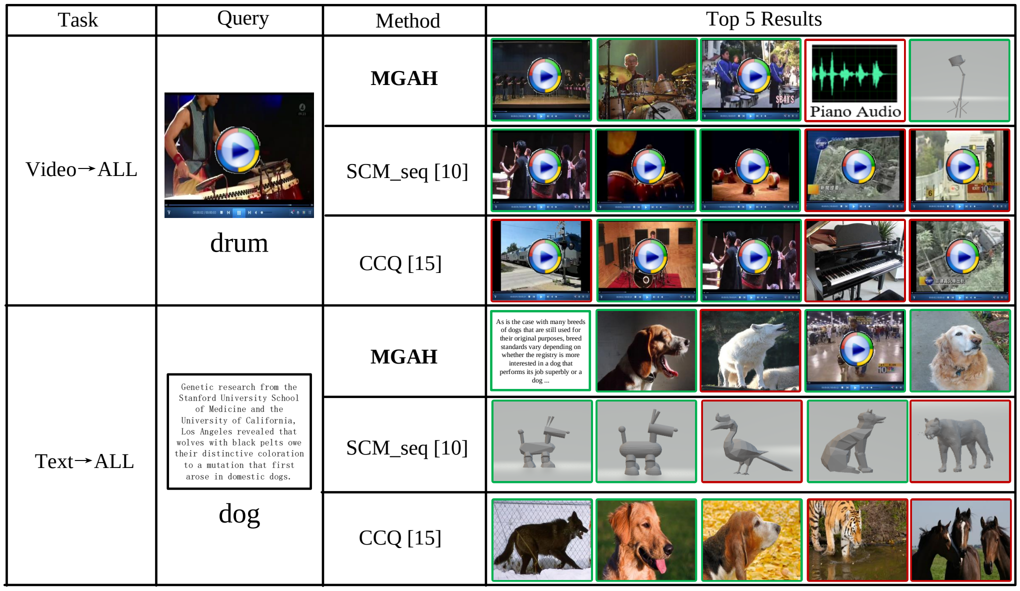

Task
Query
Method
Top 5 Results
Video→ALL
drum
MGAH
Piano Audio
SCM_seq [10]
CCQ [15]
Text→ALL
Genetic research from the Stanford University School of Medicine and the University of California, Los Angeles revealed that wolves with black pelts owe their distinctive coloration to a mutation that first arose in domestic dogs.
dog
MGAH
As is the case with many breeds of dogs that are still used for their original purposes, breed standards vary depending on whether the registry is more interested in a dog that performs its job superbly or a dog ...
SCM_seq [10]
CCQ [15]

Fig. 6. The qualitative results of our proposed MGAH and compared methods on PKU XMedia dataset, where we demonstrate the retrieval results of video→all and text→all. The green rectangles denote the correct retrieval results. The red rectangles indicate wrong retrieval results.

networks, which improves the retrieval accuracy. (2) Because this dataset contains more than 2 modalities, the proposed MGAH models multiple modalities at the same time, which can make full use of the intrinsic correlations contained between multiple modalities. While the compared methods only model two modalities, which ignore the correlations hidden within multiple modalities. Figures 7 and 8 demonstrate the top$K$-precision and precision-recall curves of the five retrieval tasks on the PKU XMedia datasets with 128 bit code length. We can observe that our proposed MGAH achieves the best results compared with both unsupervised and supervised methods, which further shows the advantages of our proposed approach.

Finally, we demonstrate the qualitative results of our proposed approach and compared deep cross-modal hashing methods. As shown in Figure 5 and 6, our proposed MGAH achieves the best results on both NUSWIDE and PKU XMedia datasets. We also have two extra observations: (1) For the PKU XMedia dataset, the ranking results of our proposed MGAH are more diverse than compared methods. It is because the proposed MGAH models all modalities simultaneously, while the compared methods can only model two modalities each time, which causes this phenomenon. (2) In some few cases, the correlation graph is not accurate enough to capture the neighborhood structure. For example, in Figure 6, our proposed approach returns a wolf image for a query of dog, which are similar but not in the same category. However, our approach still achieves the best retrieval accuracy, which indicates the solid effectiveness of our approach.

3) Retrieval Efficiency Comparison: Besides the retrieval performance, we also compare the computation time cost of the proposed SCH-GAN approach with other compared methods. All the experiments conducted on a PC with NVIDIA Titan X GPU, Intel Core i7-5930k 3.50GHz CPU and 64 GB memory. Hashing based image retrieval process generally consists of three parts: Feature extraction, hash codes generation and cross-modal retrieval among database. For a fair comparison between deep hashing methods and traditional methods, we run the traditional methods with deep features input and sum the time cost of three parts as final retrieval time cost.The average computation time of different methods is shown in table VIII. We can observe that, our proposed MGAH achieves comparable retrieval speed with other hashing methods. All the methods take the deep features as inputs and use the hash codes in same bit length for retrieval, thus the difference is mostly in the hash codes generation part. While for the compared methods and our proposed MGAH, the hash code generation is mostly a matrix multiplication operation.

TABLE VIII
COMPARISON OF THE AVERAGE RETRIEVAL TIME COST (MILLISECOND PER IMAGE) ON MIRFLICKR DATASET BY FIXING THE CODE LENGTH AS 64.

Methods	Time cost (ms)
CVH [7]	11.789
PDH [8]	11.404
CMFH [9]	11.398
CCQ [15]	11.672
CMSSH [16]	11.370
SCM_orth [10]	11.393
SCM_seq [10]	11.607
MGAH (Ours)	**11.307**

4) Baseline experiment results: Compared with 2 baseline methods on NUS-WIDE dataset, we can observe that *Baseline-GAN* has an improvement of 0.037 and 0.034 on two tasks, which demonstrates the effectiveness of our proposed generative adversarial network for cross-modal hashing. Comparing

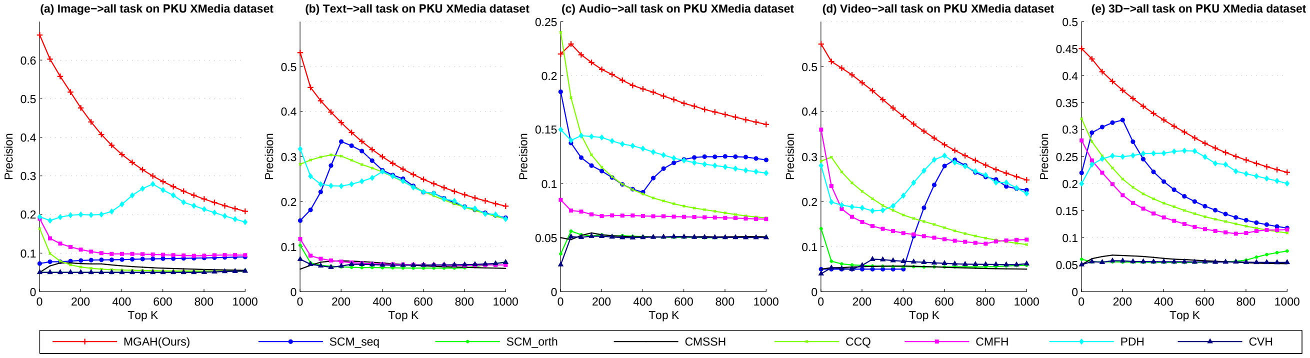


Fig. 7. The top$K$-precision curves with 128 bit hash codes of the five retrieval tasks on PKU XMedia dataset.

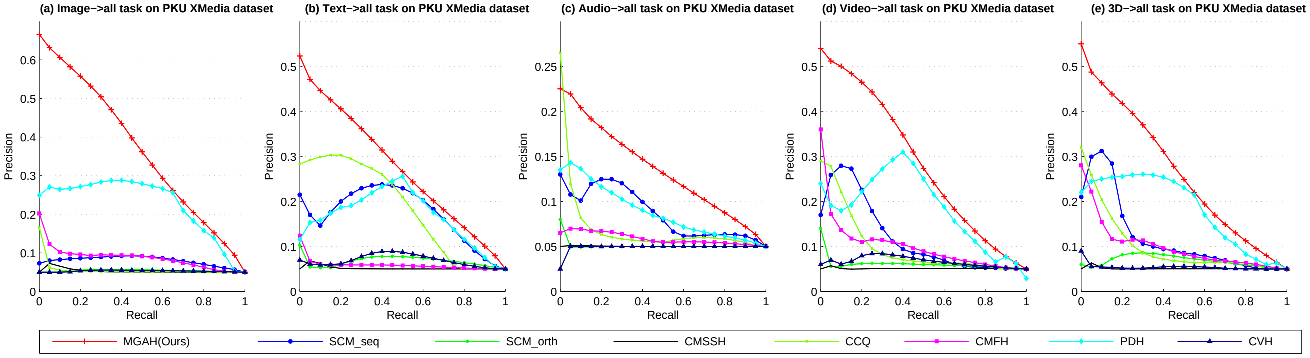


Fig. 8. The precision-recall curves with 128 bit hash codes of the five retrieval tasks on PKU XMedia dataset.

our proposed MGAH with *Baseline-GAN*, we can observe an improvement of 0.025 and 0.016 on two tasks, which demonstrates the effectiveness of our proposed correlation graph. Similar trends can be also observed on MIRFlickr dataset and PKU XMedia dataset.

Since the manifold structure is captured by the correlation graph, we also conduct experiment to demonstrate the impact of k in the correlation graph. Specifically, we set k ranging from 1 to 8, and calculate the MAP scores on NUS-WIDE with 16 bit code length. The results are shown in Figure 9, where we can observe that the parameter k does not influence the results, since the neighborhood graph will be accurate within this range and can capture the underlying structure of cross-modal data.

V. Conclusion

In this paper, we have proposed a multi-pathway generative adversarial hashing (MGAH) approach for unsupervised cross-modal retrieval, which intends to make full use of GAN's ability of unsupervised representation learning to exploit the underlying manifold structure of cross-modal data. On one hand, we propose a multi-pathway generative adversarial network to model cross-modal hashing in an unsupervised fashion. In the proposed MGAH, the generative model tries to fit the distribution over the manifold structure, and select informative data of other modalities to challenge the discriminative model. While the discriminative model learns to preserve traditional

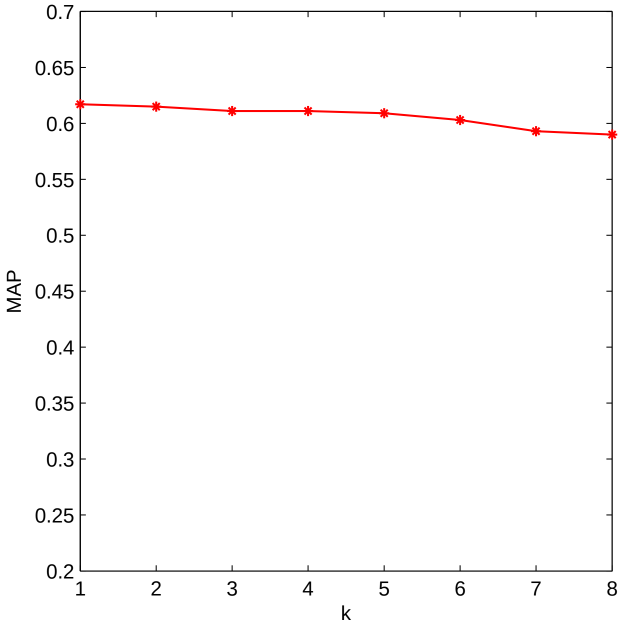


Fig. 9. The MAP scores w.r.t different value of k on NUS-WIDE dataset.

inter correlation, and the manifold correlations provided by generative model to achieve better retrieval accuracy. These two models are trained in an adversarial way to improve each other and achieve better retrieval accuracy. On the other hand, we propose a graph based correlation learning approach to capture the underlying manifold structure across different modalities, so that data of different modalities but within the same manifold can have smaller Hamming distance and promote retrieval accuracy. Experiments compared with 6 state-of-the-art methods on 3 widely-used datasets verify the effectiveness of our proposed approach.

The future works lie in two aspects. Firstly, we attempt to extend current framework to other scenarios such as image caption to verify its versatility. Secondly, we intend want to apply more advanced graph algorithms into current framework to better model the manifold structure across different modalities and promote retrieval accuracy.

REFERENCES

[1] J. Wang, W. Liu, S. Kumar, and S. F. Chang, "Learning to hash for indexing big data-a survey," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 34–57, 2016.

[2] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *VLDB*, vol. 99, 1999, pp. 518–529.

[3] J. Zhang and Y. Peng, "Ssdh: Semi-supervised deep hashing for large scale image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–14, 2017.

[4] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[5] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–14, 2017.

[6] Y. Peng, W. Zhu, Y. Zhao, C. Xu, Q. Huang, H. Lu, Q. Zheng, T. Huang, and W. Gao, "Cross-media analysis and reasoning: advances and directions," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 1, pp. 44–57, 2017.

[7] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *IJCAI proceedings-international joint conference on artificial intelligence*, vol. 22, 2011, p. 1360.

[8] M. Rastegari, J. Choi, S. Fakhraei, D. Hal, and L. Davis, "Predictable dual-view hashing," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1328–1336.

[9] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5427–5440, 2016.

[10] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *AAAI*, 2014, p. 7.

[11] Y. Zhuang, Z. Yu, W. Wang, F. Wu, S. Tang, and J. Shao, "Cross-media hashing with neural networks," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 901–904.

[12] C. Deng, X. Tang, J. Yan, W. Liu, and X. Gao, "Discriminative dictionary learning with common label alignment for cross-modal retrieval," *IEEE Transactions on Multimedia (TMM)*, vol. 18, no. 2, pp. 208–218, 2016.

[13] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[14] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, 2013, pp. 785–796.

[15] M. Long, Y. Cao, J. Wang, and P. S. Yu, "Composite correlation quantization for efficient multimodal retrieval," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2016, pp. 579–588.

[16] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3594–3601.

[17] Y. Wei, Y. Song, Y. Zhen, B. Liu, and Q. Yang, "Scalable heterogeneous translated hashing," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 791–800.

[18] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3864–3872.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[20] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1363–1377, 2016.

[21] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yuy, "Deep visual-semantic hashing for cross-modal retrieval," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, p. 1445.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[23] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, 2016, pp. 1060–1069.

[24] 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 *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 3511–3517.

[25] J. Wang, L. Yu, W. Zhang, Y. Gong, Y. Xu, B. Wang, P. Zhang, and D. Zhang, "Irgan: A minimax game for unifying generative and discriminative information retrieval models," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 515–524.

[26] J. Zhang, Y.-X. Peng, and M.-K. Yuan, "Unsupervised generative adversarial cross-modal hashing," in *32th AAAI Conference on Artificial Intelligence (AAAI)*, 2018, pp. 1–8.

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

[28] G. Irie, Z. Li, X.-M. Wu, and S.-F. Chang, "Locally linear hashing for extracting non-linear manifolds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2115–2122.

[29] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3270–3278.

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

[31] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Advances in neural information processing systems*, 2009, pp. 1753–1760.

[32] J. Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for scalable image retrieval," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3424–3431.

[33] Y. Yang, Y. Duan, X. Wang, Z. Huang, N. Xie, and H. T. Shen, "Hierarchical multi-clue modelling for poi popularity prediction with heterogeneous tourist information," *IEEE Transactions on Knowledge and Data Engineering*, doi:10.1109/TKDE.2018.2842190.

[34] M. Hu, Y. Yang, F. Shen, N. Xie, and H. T. Shen, "Hashing with angular reconstructive embeddings," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 545–555, 2018.

[35] M. Hu, Y. Yang, F. Shen, N. Xie, R. Hong, and H. T. Shen, "Collective reconstructive embeddings for cross-modal hashing," *IEEE Transactions on Image Processing (TIP)*, vol. 28, no. 6, pp. 2770–2784, 2019.

[36] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 415–424.

[37] D. Wang, X. Gao, X. Wang, and L. He, "Semantic topic multimodal hashing for cross-media retrieval," in *IJCAI*, 2015, pp. 3890–3896.

[38] Y. Zhen and D.-Y. Yeung, "Co-regularized hashing for multimodal data," in *Advances in neural information processing systems*, 2012, pp. 1376–1384.

[39] Y. Hu, Z. Jin, H. Ren, D. Cai, and X. He, "Iterative multi-view hashing for cross media indexing," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 527–536.

[40] B. Wu, Q. Yang, W.-S. Zheng, Y. Wang, and J. Wang, "Quantized correlation hashing for fast cross-modal search," in *IJCAI*, 2015, pp. 3946–3952.

[41] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[42] D. Wang, P. Cui, M. Ou, and W. Zhu, "Learning compact hash codes for multimodal representations using orthogonal deep structure," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1404–1416, 2015.

[43] Y. Cao, M. Long, J. Wang, and H. Zhu, "Correlation autoencoder hashing for supervised cross-modal search," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 2016, pp. 197–204.

[44] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3232–3240.

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

[46] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 8, pp. 3893–3903, 2018.

[47] E. Yang, C. Deng, T. Liu, W. Liu, and D. Tao, "Semantic structure-based unsupervised deep hashing." in *IJCAI*, 2018, pp. 1064–1070.

[48] E. Yang, C. Deng, C. Li, W. Liu, J. Li, and D. Tao, "Shared predictive cross-modal deep quantization," *IEEE transactions on neural networks and learning systems*, no. 99, pp. 1–12, 2018.

[49] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[50] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[51] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," *arXiv preprint arXiv:1610.09585*, 2016.

[52] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," in *Advances in Neural Information Processing Systems*, 2016, pp. 64–72.

[53] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1222–1230.

[54] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *ACM Multimedia*, 2017, pp. 154–162.

[55] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," in *Advances in Neural Information Processing Systems (NIPS)*, 2003, pp. 41–48.

[56] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*. ACM, 2009, p. 48.

[57] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 2008, pp. 39–43.

[58] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4489–4497.

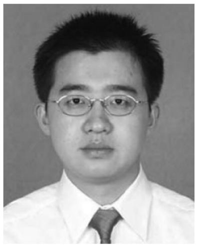
[59] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[60] C. McKay, I. Fujinaga, and P. Depalle, "jaudio: A feature extraction library," in *Proceedings of the International Conference on Music Information Retrieval*, 2005, pp. 600–3.

[61] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung, "On visual similarity based 3d model retrieval," in *Computer graphics forum*, vol. 22, no. 3. Wiley Online Library, 2003, pp. 223–232.

Jian Zhang received the B.S. degree in computer science and technology and the Ph.D. degree in computer application from Peking University, Beijing, China, in 2012 and 2018, respectively.

His current research interests include multimedia retrieval and machine learning.



Yuxin Peng received the Ph.D. degree in computer application technology from Peking University, Beijing, China, in 2003. He is currently a Professor with the Institute of Computer Science and Technology and the Chief Scientist with the National Hi-Tech Research and Development Program of China (863 Project). He has authored more than 140 papers in refereed international journals and conference proceedings, including IJCV, TIP, TMM, TCSVT, TCYB, TOMM, ACM MM, ICCV, CVPR, IJCAI, and AAAI. He has submitted 39 patent applications and received 24 of them. His current research interests mainly include cross-media analysis and reasoning, image and video analysis and retrieval, and computer vision. He led his team to win the first place in video instance search evaluation of TRECVID in the recent 7 years. He was the recipient of the First Prize of the Beijing Science and Technology Award in 2016 (technological invention, ranking first).