Unsupervised Deep Imputed Hashing for Partial Cross-modal Retrieval

Dong Chen, Miaomiao Cheng, Chen Min, Liping Jing*
Beijing Key Lab of Traffic Data Analysis and Mining
Beijing Jiaotong University
Beijing, China
{chendong, 15112085, minchen, lpjing}@bjtu.edu.cn

***Abstract*—Cross-modal retrieval, given the data of one specific modality as a query, aims to search the relevant data in other modalities. Recently, cross-modal hashing has attracted much attention due to its high efficiency and low storage cost. Its main idea is to approximate the cross-modality similarity via binary codes. This kind of method works well when the cross-modal data is completely observed. However, the real-world application usually avoids this situation, where part of the information is unobserved in some modality. Such partial multimodal data will result in the lack of pairwise information and then destroy the performance of cross-modal hashing. In this paper, we proposed a novel unsupervised cross-modal hashing approach, named as Unsupervised Deep Imputed Hashing (UDIH). It is a two-stage learning strategy. Firstly, the unobserved pairwise data is imputed by the proposed generators. Then a neural network with weighted triplet loss is applied on the correlation graph to learn the hashing code in the Hamming space for each modality, where the correlation graph is constructed with the aid of augmented data. UDIH has the ability to preserve the semantic consistency and difference among data objects. The extensive experimental results have shown that the proposed method outperforms the state-of-the-art methods on two benchmark datasets (MIRFlickr and NUS-WIDE). The source code could be available at https://github.com/AkChen/UDIH**

***Index Terms*—cross-modal retrieval, partial multimodal data,cross-modal hashing, imputation, unsupervised learning**

I. Introduction

Multimodal data has grown dramatically. For example, in the recommendation system, each product is demonstrated from multiple views such as pictures, textual description, and even video, so that users can sufficiently understand its characteristics. To flexibly meet the users' requirements, cross-modal retrieval, instead of single-modal retrieval plays a more and more important role in real-world applications. Given the data of one specific modality as a query, cross-modal retrieval has the ability to search the relevant data in other modalities. Recently, cross-modal hashing (CMH) becomes one of the most popular cross-modal retrieval strategies due to its high efficiency and low storage cost. Most existing CMH methods are designed by assuming that the multimodal data is completely observed. However, in practical applications, it is hard to collect data with full modalities because of various unpredictable reasons, such as unforeseeable malfunction, collection mode limitation, etc.

Fig. 1. The illustration of partial multimodal data. There existing both paired data objects and unpaired data objects. The blank areas indicate that the corresponding information is missing due to some unpredictable reasons, such as unforeseeable malfunction, collection mode limitation, etc.

The other challenging issue in cross-modal retrieval is that the data of different modalities may be represented in different feature spaces, which is usually called as 'heterogeneity gap' [16]. To implement cross-modal retrieval, it is necessary to collect sufficient **pairwise information** (the one-to-one correspondence of paired data) so that the consistent semantic information can be captured from multiple modalities.

However, in real-world applications, partial multi-modal data without label always exists, as shown in Figure 1. This situation makes cross-modal retrieval much more challenging because there is no knowledge to supervise the learning process. In literatures, researchers began to focus on unsupervised cross-modal hashing (CMH) [1], [5], [7], [8], [19]–[21], but only few methods are designed to handle incomplete multimodal data, called as unsupervised partial cross-modal hashing (PCMH) [10], [11], [13], [14], [17]. The main purpose of existing PCMH approaches is to preserve the pairwise consistency among data objects. In fact, cross-modal retrieval aims to identify the semantic relations among multi-modal data objects which are not the same with the pairwise consistency [3], [22]. Therefore, it is necessary to consider the semantic consistency and difference among modalities when learning the hash codes. Meanwhile, their performance significantly depends on the amount of pairwise information, which limits their application on partial cross-modal retrieval.

To address the above drawbacks, we propose the Unsupervised Deep Imputed Hashing (UDIH) method from image-text cross-modal retrieval. It aims to learn the cross-modal hashing code by sufficient capturing the multi-level features of multi-

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

modal data. Specifically, UDIH consists of two parts. The first part tries to fill in the incomplete pairwise information via two generators, one for text data generation from image information, and the other for image data generation from text information with the aid of cascaded residual autoencoder [15]. The augmented pairwise data is used to construct the correlation graph, on which a deep neural network with a weighted triplet loss function is designed to learn the hashing code. Thus, the proposed UDIH is expected to sufficiently preserve the semantic consistency and difference among multi-modal data objects.

We summarize the main contributions of this paper as follows.

- A new unsupervised partial cross-modal hashing (UDIH) framework is proposed to learn the hashing codes of multimodal data objects by sufficiently mine the relative semantic similarity among multiple modalities.
- To make up for the lack of pairwise information, UDIH imputes the incomplete multimodal data according to the estimated data distribution, so that the original limited pairwise information can be sufficiently augmented.
- UDIH constructs a weighted graph to reasonably exploit the original pairwise information and imputed pairwise information to effectively capture the cross-modal correlation.
- To correctly represent the multimodal data in Hamming space, UDIH learns the hashing codes of the augmented multi-modal data via a neural network with weighted triplet loss. It has the ability to preserve semantic similarity among multiple modalities.
- The performance of UDIH is thoroughly investigated on two widely-used datasets, indicating its advantage over the state-of-the-art baselines.

The remainder of the paper is organized as follows. We briefly introduce the related works on partial cross-modal retrieval methods in Section II. The proposed UDIH framework is described in Section III. Extensive experimental results are listed and discussed in Section IV to verify the performance and the motivation of UDIH. Lastly, we draw a brief conclusion in Section V.

II. RELATED WORK

In this section, we review related works from the following aspects: matrix factorization (MF)-based and graph-based unsupervised PCMH methods.

A. MF-based unsupervised PCMH methods

Matrix factorization models map data to a joint latent factor space of dimensionality. [6] MF-based methods attempt to learn an optimal linear matrix to map data into common space and to learn Hamming representation lastly. For each individual modality, $\mathrm{PM}^2\mathrm{H}$ [17] learns a modality-specific basis matrix to learn the latent representation of instances by minimizing the reconstruction error. To break the $heterogeneity\ gap$, $\mathrm{PM}^2\mathrm{H}$ [17] shares the latent representation of all available paired data. For preserving intral-modality similarity, graph $Laplacian$ [18] is exploited to preserve the local structure in the common Hamming space for each modality. Therefore, by sharing paired data, the relationship between modalities is established. CCQ [10] is another MF based method in this area. In order to efficiently reduce quantization error during hashing learning, CCQ [10] adopts a similarity-preserving codebook [10] to replace the traditional mapping matrix. By selecting optimal code from codebooks, data could be directly mapped to common Hamming space. Similar to $\mathrm{PM}^2\mathrm{H}$ [17], CCQ [10] breaks the heterogeneity gap by requiring condition the codes of the inter-modal pairs as consistent as possible.

B. Graph-based unsupervised PCMH methods

Graph-based methods attempt to construct a similarity graph in the level of original feature space to constraint the hashing learning. Based on the graph, the similarity among different modalities can be directly measured. IMH [14] constructs affinity matrices for each modality to preserve intra-modality consistency. Then, to make the Hamming representation of paired data consistent, IMH [14] minimizes the distance of the hash codes of paired data. Inspired by anchor graph for hashing learning [9], SPDH [13] constructs an novel anchor graph based on anchors [9] to measure the cross-modal similarity. A graph Laplcian [18], constructed by the anchor graph, will be exploited to map different data to common space. Moreover, SPDH learns hash codes bit by bit, which could also reduce quantization error.

In general, these methods have achieved good results, but there are still some problems. These methods work well on preserving the data similar. However, the difference between dissimilar data is not fully considered. Meanwhile, to handle the situation with limited pairwise information, and the limitation of the only constraint on consistency, we propose graph-based UDIH in this paper.

III. PROPOSED METHOD

In this section, an Unsupervised Deep Imputed Hashing (**UDIH**) is proposed, as shown in Fig.2. **UDIH** contains two parts, one for generating pairwise information and the other one for determining the discrete hamming space by exploiting the constructed weighted cross-modal correlation graph. In the pairwise generation, the partial objects will be completed by exploiting the specific-modal imputation generator which is trained by the corrupted complete cases. In the discrete hamming space, the objects' binary codes will be characterized by simultaneously preserving the semantic consistency and difference.

A. Problem Statement

In partial cross modal retrieval, the retrieval database and query usually consist of objects from different modalities. Here we use image and text as two modalities to explain our method. Let $\mathbf{S}^{(v)} = [\mathbf{X}^{(v)}, \mathbf{Y}^{(v)}] \in \mathbb{R}^{d_v \times n_v} (v \in \{1, 2\})$ indicate the training set in the v-th modality, where $\mathbf{X}^{(v)} = \{\mathbf{x}_i^{(v)}\}_{i=1}^{c}$ and $\mathbf{Y}^{(v)} = \{\mathbf{y}_i^{(v)}\}_{i=1}^{n_v - c}$. $\mathbf{X}^{(v)}$ and $\mathbf{Y}^{(v)}$ denote the complete cases and partial cases. d_v and n_v are the dimensionality and

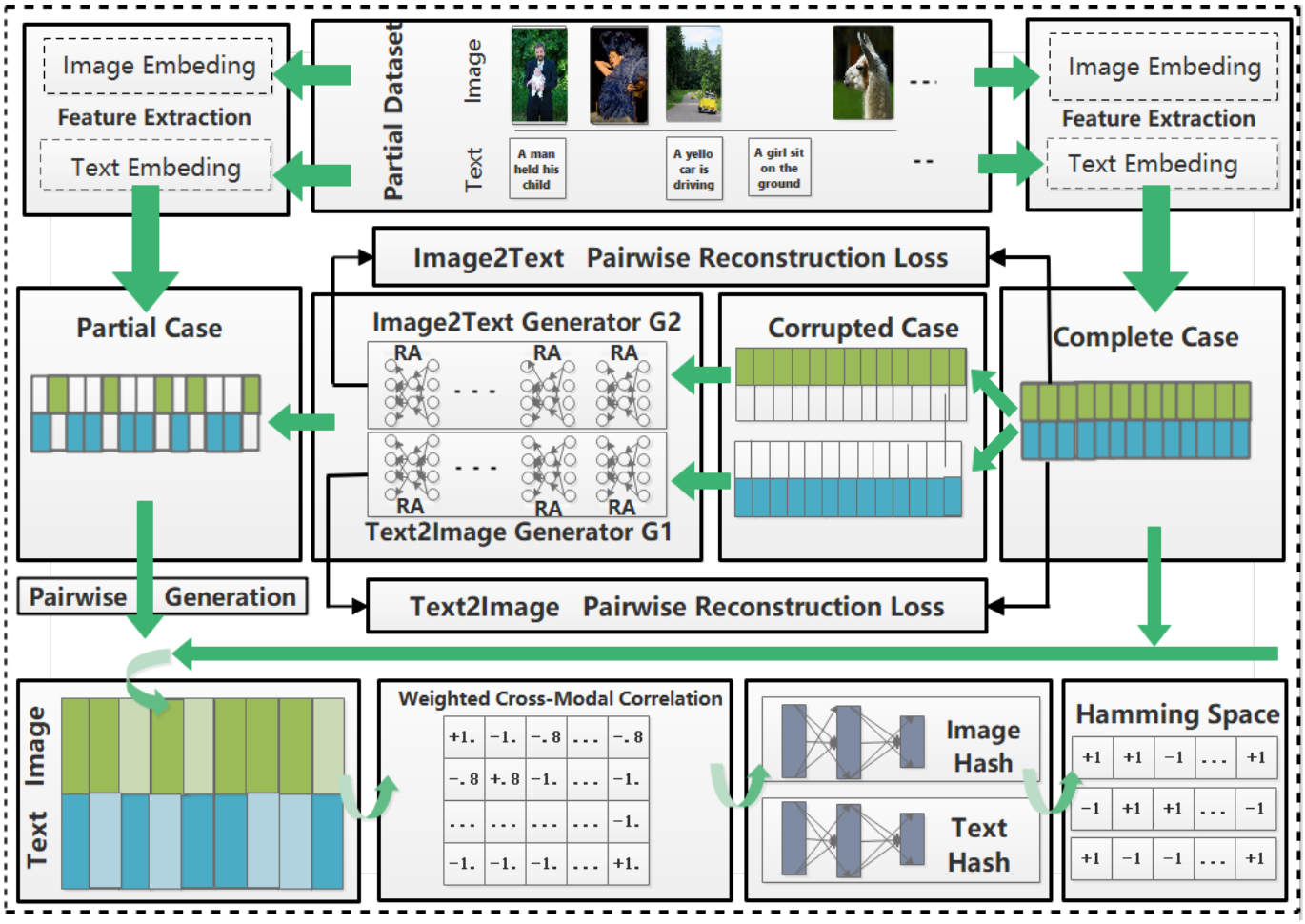

Fig. 2. The overall framework of UDIH.

the numbers of samples in the v-th modality, respectively. c is the number of complete cases. For complete cases, $[\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}]$, the objects have one-to-one correspondence among different modalities. For partial cases, $[\mathbf{y}_i^{(1)}, \mathbf{o}^{(2)}]$ or $[\mathbf{o}^{(1)}, \mathbf{y}_i^{(2)}]$, the objects are only partially provided. Here $\mathbf{o}^{(v)} = [0, 0, \cdots, 0] \in \mathbb{R}^{d_v}$ represents the unobserved elements in the v-modality.

Our goal is to learn modal-specific hash function to determine the common Hamming space. In the common binary space, image and text can be easily comparable such that cross-modal retrieval can be readily supported, i.e., given a query text can efficiently retrieval the relevant images.

Throughout the paper, vectors and matrices are denoted by lowercase bolded letters (e.g.,$\mathbf{a}$)and uppercase bolded letters (e.g., $\mathbf{A}$), respectively. $\mathbf{a}_i$ is the i-th column of $\mathbf{A}$ and a_{ij} is the j-th entry of $\mathbf{a}_i$. The Frobenius norm of a matrix is defined as $\|\mathbf{A}\|_F^2 = \sum_{ij} a_{ij}^2$.

B. Pairwise Information Generation

The main idea of **UDIH** is to impute the unobserved pairwise data to determine the common Hamming space with the constructed correlation graph. To achieve the first mission via the complex relatedness among different modalities, a two-pathway architecture generator is designed by exploiting the complete cases to impute the unobserved data. Each pathway is composed of a set of stacked residual auto-encoder (RAs) [12] that iteratively model the residual. For the first RA, it takes the artificially corrupted complete cases $\{\overline{\mathbf{x}}_i\}_{i=1}^c$ as the input. This problem can be formulated as

$$\overline{\mathbf{X}}_{i1} = \begin{cases} [\mathbf{x}_i^1; \mathbf{o}^2] & : \textit{text is corrupted} \\ [\mathbf{o}^1; \mathbf{x}_i^2] & : \textit{image is corrupted} \end{cases} \tag{1}$$

Its desired output is the difference between the input data sample and the complete data sample, i.e., $\Delta\overline{\mathbf{X}}_{i1} = \mathbf{x}_i - \overline{\mathbf{x}}_i$. RA aims to make the estimated output to be close to the desired output as possible. Thus, its loss function can be calculated as

$$\mathcal{L}_{\mathcal{R}\mathcal{A}1} = \|\Delta\overline{\mathbf{X}}_{i1} - \Delta\mathbf{X}_{i1}\|_F^2, \tag{2}$$

where $\Delta\mathbf{X}_{i1}$ is the output of RA.

To refine the estimation, the input of the remaining RAs is the summation of the input of the last RA and the output of the last RA. Specifically, the input of the k-th ($k \geq 2$) RA can be represented as $\overline{\mathbf{X}}_{ik} = \overline{\mathbf{X}}_{ik-1} + \Delta\mathbf{X}_{ik-1}$, where $\Delta\mathbf{X}_{ik-1}$ is the output of the last RA. Each RA will be learnt via minimizing the difference between current estimation and the complete cases, and therefore the loss function of the k-th RA can be formulated as

$$\mathcal{L}_{\mathcal{R}\mathcal{A}k} = \|\Delta\overline{\mathbf{X}}_{ik} - \Delta\mathbf{X}_{ik}\|_F^2 \tag{3}$$

To estimate a function well approximating the complete data, the imputation generator can be trained with the corrupted data by a forward and layer fashion. More specifically, each additional RA is trained to further minimize the reconstruction error of the current RA, so that the estimation can be refined by combining all the outputs of RAs. Mathematically, the estimation data sample can be formulated as

$$\tilde{\mathbf{x}}_i = G(\overline{\mathbf{X}}_{i1}) = \overline{\mathbf{X}}_{i1} + \sum_{k=1}^{t} \Delta \mathbf{X}_{ik},$$

where t is the number of RA, $G(\cdot)$ is the imputation generator function.

The joint loss function of modal-specific generator $G^{(v)}$ is defined as

$$\mathcal{L}_{re}^{(v)} = \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|_F^2 + \lambda_{G^{(v)}} \|\theta_{G^{(v)}}\|_F^2, v \in \{1, 2\}, \tag{4}$$

where $\lambda_{G^{(v)}}$ and $\theta_{G^{(v)}}$ are the weight decay parameter and trained parameters in the v-th imputation generator, respectively. Once achieving the generator, the missing data sample in each modal $\{\mathbf{o}^{(v)}\}_{v=1}^2$ can be imputed. It is worth noting that, unlike the existing methods, we only randomly impute p_i image data objects and p_t text data objects to avoid introducing more noise data, which results in the augmented feature matrix, $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_{c+p_i+p_t}] \in \mathbb{R}^{(d_1+d_2)\times(c+p_i+p_t)}$, where c is the number of complete cases ($\mathbf{z}_i = [\mathbf{z}_1^{(1)}, \mathbf{z}_2^{(2)}]$).

C. Discrete Hamming Space Identification

To significantly improve the cross-modal retrieval speed and storage, **UDIH** adopts two-pathway deep layers to determine the common hamming space. The input of each pathway is the respective features. Because networks are independent of each other, inter-modal relationships, correlation graph $\mathbf{Z}$, are needed to constrain them. The correlation graph aims to capture the underlying structure across different modalities so that the data of different modalities but relevant can have small hamming distance and promote retrieval performance.

As the correlation graph is used to guide the training Hamming space identification model, it [13] is crucial for hashing to achieve good performance. However, it is very challenging to directly analyze the correlation among partial cross-modal data, as they belong to different modalities and partial pairwise information is also not available. In most existing methods [9], [13], [21], K-nearest neighbors graph is used to model the correlation among partial cross-modal data. However, this suffers from two drawbacks. One is how to set the hyperparameter K in an unsupervised manner. The other one is the semantic similarity calculated by distance can not be well matched.

To address these two issues, we design a simple yet effective weighted correlation graph construction approach with the help of the augmented feature matrix, to uncover similarities among partial cross-modal data. The main idea is utilizing the relevant among paired case and the irrelevant among unpaired case. Then the cross-modal correlation between the i-th image and the j-th text can be calculated as

$$\mathbf{W}(i,j) = \begin{cases} +1.0 * \mathbf{Pair}(i,j): \ i = j \\ -1.0 * \mathbf{Pair}(i,j): \ i \neq j \end{cases}, \tag{5}$$

where

$$\mathbf{Pair}(i,j) = \begin{cases} \alpha \ : \ \mathbf{Z}_i^1 \ or \ \mathbf{Z}_j^2 \ is \ imputed \\ 1.0: \ otherwise \end{cases}, \tag{6}$$

α denotes the reliability of imputed data.

Once receiving the correlation graph, we intend to utilize the underlying data manifold of different modalities to determine the hamming space. Intuitively, the data of different modalities but relevant is desired to have small hamming distance, while the irrelevant data of different modalities is desired to large hamming distance, and therefore the retrieval performance can be improved. **UDIH** adopts two pathway deep layers to learn the modality-specific binary codes. It consists of two fully connected layers.

The first layer severs an intermediate layer that maps the original modality specific feature into a hidden representation space with *tanh* activation function. Mathematically, it can be formulated as

$$\boldsymbol{\rho}^v(\mathbf{z}_i^{(v)}) = tanh(\mathbf{w}_c^{(v)} \mathbf{z}_i^{(v)} + \mathbf{b}_c^{(v)}), v \in \{1, 2\}, \tag{7}$$

where $\mathbf{z}_i^{(v)}$ is i-th feature data sample of v-th modality, $\mathbf{w}_c^{(v)}$ denotes the weights parameters in the hidden representation learning layer and $\mathbf{b}_c^{(v)}$ is the bias parameter in v-th pathway.

The other layer serves as common representation learning, which maps the intermediate feature into common representation

$$\boldsymbol{h}^{(v)}(\mathbf{z}_i^{(v)}) = sigmoid(\mathbf{w}_h^{(v)} \boldsymbol{\rho}^{(v)}(\mathbf{z}_i^{(v)}) + \mathbf{b}_h^{(v)}) \ \in (0,1)^r, \ v \in \{1, 2\}, \tag{8}$$

where $\mathbf{w}_h^{(v)}$ denotes the weights parameters in the common representation learning layer and $\mathbf{b}_h^{(v)}$ is the bias parameter in v-th pathway.

To improve the modality-specific hash function, the similarity information among original data is desired to preserve as much as possible, i.e., the related data as consistent as possible, while the irrelevant is as inconsistent as possible. Thus the weighted triplet loss function can be defined as

$$\begin{aligned} \mathcal{L}_{sim}^{(v)} = max(0, \delta + \mathbf{W}(i,j) \|\mathbf{h}^{(v)}(\mathbf{z}_i^{(v)}) - \mathbf{h}^{(!^*v)}(\mathbf{z}_j^{(!^*v)})\|_F^2 \\ + \mathbf{W}(i,k) \|\mathbf{h}^{(v)}(\mathbf{z}_i^{(v)}) - \mathbf{h}^{(!^*v)}(\mathbf{z}_k^{(!^*v)})\|_F^2) + \lambda_{H^{(v)}} \|\theta_{H^{(v)}}\|_F^2, \\ v \in \{1, 2\}, \end{aligned} \tag{9}$$

where data sample i and j is relevant, data sample i and k is irrelevant via the constructed correlation graph $\mathbf{W}$, $\theta_{H^{(v)}}$ is the parameters of the v-th hashing network, δ is a bias parameter. $!^*v = !(2 - v) + 1$, where $!1 = 0$ and $!0 = 1$.

Finally, the corresponding hash function of v-th modality is defined as:

$$\mathbf{H}^{(v)}(\mathbf{x}_i^{(v)}) = sign(\mathbf{h}^{(v)}(\mathbf{z}_i^{(v)}) - 0.5) \ \in \{-1, +1\}^r, \tag{10}$$

$$v \in \{1, 2\}, \tag{11}$$

The learned hash function $\mathbf{H}^{(v)}(\mathbf{x}_i^{(v)})$ could map different data to common hamming space so that the efficiently cross-modal retrieval task can be available.

IV. Experiment

In this section, a series of experiments are conducted to validate the performance of the proposed model UDIH by comparing with the-state-of-the-art unsupervised partial cross-modal hashing methods.

A. Datasets

Two kinds of widely-used cross-modality datasets are adopted to evaluate the performance.

The **MIRFlickr** [4] dataset contains 25, 000 instances. Following the experimental protocols in SPDH [13], we select 20, 015 instances for our experiment. The texts are expressed as 500-D feature vector derived from PCA on the bag of words vector. And we represent each image as 150-D edge histogram. We take only 10% of the dataset as our query set and the rest as the database and training set. The $\mathbf{NUS-WIDE}$ [2] dataset is a public web image dataset and it originally contains 269,648 instances. Following the experimental protocols in [13], only the top ten most frequent labels and the corresponding 186, 577 image-text pairs are kept. The images are represented by 500-D bag-of-visual-words and tags are represented by 1000-D tag occurrence vectors. We randomly select 1% of the dataset as our query set and the rest as the database and training set.

B. Methodology

To illustrate the effectiveness of our proposed UDIH, there are five state-of-the-art unsupervised PCMH baselines compared in experiments, including Cluster-CCA [11], CCQ [10], IMH [14], $\mathrm{PM^2H}$ [17], and SPDH [13]. We also design a simple method UDIH-W that do not use pairwise augmentation. Parameters of all the other methods are carefully tuned according to the corresponding literature, and their best performances are reported here. To make the experimental results more realistic, the average results of 10 runs are recorded for all the experiments. In our experiments, Image query Text ($I2T$) denotes retrieving text by image query, and vice verse. The MAP score is the mean of average precision (AP) for all queries, and AP is computed as follows:$AP = \frac{1}{R}\sum_{k=1}^{T}\frac{k}{R_k} \times rel_k$, where T denotes the top T results, R denotes number of relevant samples, R_k is the number of relevant samples in the top k retrieved results, and rel_k = 1 if the k-th retrieved result is relevant to the query set and 0 otherwise. In our experiments, we set T = 50. The $Partial\ Data\ Ratio$ is defined as: PDR = $(n-c)/n$, where n is the number of instances and c denotes the count of complete cases. To mimic the real situation, we use PDR to randomly select samples from training set as partial cases. Network structure is defined as: RA structure = [d_1+d_2, 512, 256, d_1+d_2], hashing network structure = [d_m,1024, r]. The source code could be available at https://github.com/AkChen/UDIH

C. Results and Discussions

In this section, we first analyze the effectiveness of imputation generators. Then we analyze the effect of parameters. To prove the effectiveness of proposed UDIH, we compare our method with several state-of-the-art methods under two different experiment scenarios. The first one is to compare them using fixed PDR. Due to the imbalance of fixed PDR, the second one is to compare them under varying PDR.

1) Ablation study of UDIH: To verify the impact of the pairwise information generation, we design a UDIH without imputation(UDIH-W). As is shown in Figure 3, UDIH performs much better than UDIH-W which demonstrates the importance of the imputation. Without more pairwise information, the UDIH-W could only use complete cases to break the 'heterogeneity gap' and learn hash codes. By comparing UDIH-W and UDIH, it is demonstrated that the imputed data improves the performance of hash learning during training. It is worth noting that the improvement of imputation on the NUS-WIDE dataset is not as good as that on MIRFlickr. This may be due to NUS-WIDE already having more pairwise information on the same PDR. Therefore, the retrieval performance is more obvious when the paired information is less.

2) Comparison with fixed PDR: We compare the MAP results of the five methods and our UDIH on different datasets with the same and different training setting, which is shown in Table I, Table II, Table III, and Table IV. We conduct an extreme PDR value of 0.9 to validate the performance of different methods. Firstly, all partial cases will be exploited by compared methods, consistent with the original settings of these methods, while UDIH only uses p_i image missed partial cases and p_t text missed partial cases. It can be seen that our UDIH outperforms the other five methods significantly. Cluster-CCA [11] has the worst performance among these methods due to the lack of sufficient pairwise information. IMH [14] outperforms $\mathrm{PM^2H}$ [17] on MIRFlickr and NUS-WIDE while their results worse than SPDH [13], CCQ [10], and our UDIH. CCQ [10] has promising results due to its outstanding quantization strategy. However, CCQ [10] only uses pairwise information to connect the two modalities, and does not retain intra-modal similarity. Therefore, in the case of few pairwise information, the effect of reconstruction error will cover the preserving of consistency, which is opposit of cross-modal retrieval. Due to the leveraging of cross-modality similarity, SPDH [13] has secondary performance. Secondly, we design an additional experiment setting that using the same p_i and p_t. As shown in Table III and Table IV, the performance of most of the compared methods has decreased, but CCQ [10] has a strange improvement or almost no degradation. Because the effect of reconstruction errors is reduced, and with the proposed quantization strategy, the only existing pairwise information may be more effective for PCMH. Therefore, UDIH-W also works well without any imputation. Motivated by this, UDIH augments more pairwise information insdead of directly learning on all data objects.

TABLE I
THE MAP SCORES OF TWO CROSS-MODAL RETRIEVAL TASKS WITH DIFFERENT p_i, p_t ON MIRFLICKR. PDR = 0.9

	Image query Text			Text query Image				
Methods	16(BIT)	32(BIT)	64(BIT)	16(BIT)	32(BIT)	64(BIT)	p_i	p_t
Cluster-CCA [11]	0.6124	0.6044	0.6125	0.6019	0.6003	0.6125	ALL	ALL
CCQ [10]	0.6223	0.6152	0.591	0.6053	0.5910	0.5819	ALL	ALL
IMH [14]	0.6217	0.6209	0.6314	0.6309	0.6282	0.6301	ALL	ALL
PM^2H [17]	0.6223	0.6195	0.5944	0.6013	0.6119	0.5925	ALL	ALL
SPDH [13]	0.6402	0.6384	0.6401	0.6381	0.64054	0.6362	ALL	ALL
UDIH-W	0.6424	0.6431	0.6484	0.6439	0.6448	0.6435	0.0	0.0
UDIH	**0.6624**	**0.6601**	**0.6668**	**0.6648**	**0.6541**	**0.6550**	0.02*(n-c)	0.01*(n-c)

TABLE II
THE MAP SCORES OF TWO CROSS-MODAL RETRIEVAL TASKS WITH DIFFERENT p_i, p_t ON NUS-WIDE. PDR = 0.9

	Image query Text			Text query Image				
Methods	16(BIT)	32(BIT)	64(BIT)	16(BIT)	32(BIT)	64(BIT)	p_i	p_t
Cluster-CCA [11]	0.4535	0.4235	0.4566	0.4231	0.4175	0.4302	ALL	ALL
CCQ [10]	0.5029	0.5131	0.5236	0.5280	0.5426	0.5420	ALL	ALL
IMH [14]	0.4651	0.475	0.4921	0.5016	0.5231	0.5281	ALL	ALL
PM^2H [17]	0.5013	0.4732	0.4651	0.4474	0.4362	0.4012	ALL	ALL
SPDH [13]	0.5325	0.5199	0.5218	0.5465	0.5421	0.5435	ALL	ALL
UDIH-W	0.5408	0.5519	0.5671	0.5421	0.5537	0.5679	0.0	0.0
UDIH	**0.5608**	**0.5709**	**0.5971**	**0.5603**	**0.5772**	**0.5898**	0.02*(n-c)	0.01*(n-c)

TABLE III
THE MAP SCORES OF TWO CROSS-MODAL RETRIEVAL TASKS WITH THE SAME p_i AND p_t ON MIRFLICKR. PDR = 0.9

	Image query Text			Text query Image				
Methods	16(BIT)	32(BIT)	64(BIT)	16(BIT)	32(BIT)	64(BIT)	p_i	p_t
Cluster-CCA [11]	0.5675	0.5585	0.5705	0.5755	0.5721	0.5767	0.02*(n-c)	0.01*(n-c)
CCQ [10]	0.6191	0.6153	0.6113	0.6243	0.6214	0.6235	0.02*(n-c)	0.01*(n-c)
IMH [14]	0.5806	0.5895	0.5877	0.5821	0.5835	0.5814	0.02*(n-c)	0.01*(n-c)
PM^2H [17]	0.5872	0.5877	0.5831	0.5129	0.5134	0.5158	0.02*(n-c)	0.01*(n-c)
SPDH [13]	0.6124	0.6137	0.6156	0.6201	0.6232	0.6185	0.02*(n-c)	0.01*(n-c)
UDIH-W	0.6424	0.6431	0.6484	0.6439	0.6448	0.6435	0.0	0.0
UDIH	**0.6624**	**0.6601**	**0.6668**	**0.6648**	**0.6541**	**0.6550**	0.02*(n-c)	0.01*(n-c)

TABLE IV
THE MAP SCORES OF TWO CROSS-MODAL RETRIEVAL TASKS WITH THE SAME p_i AND p_t ON NUS-WIDE. PDR = 0.9

	Image query Text			Text query Image				
Methods	16(BIT)	32(BIT)	64(BIT)	16(BIT)	32(BIT)	64(BIT)	p_i	p_t
Cluster-CCA [11]	0.4425	0.4401	0.4351	0.4025	0.4175	0.4201	0.02*(n-c)	0.01*(n-c)
CCQ [10]	0.5101	0.5111	0.5153	0.5266	0.5316	0.5333	0.02*(n-c)	0.01*(n-c)
IMH [14]	0.4452	0.4425	0.4429	0.4785	0.4823	0.4833	0.02*(n-c)	0.01*(n-c)
PM^2H [17]	0.4355	0.4337	0.4425	0.4425	0.4197	0.4042	0.02*(n-c)	0.01*(n-c)
SPDH [13]	0.5123	0.5079	0.5110	0.5157	0.5131	0.5130	0.02*(n-c)	0.01*(n-c)
UDIH-W	0.5408	0.5519	0.5671	0.5421	0.5537	0.5679	0.0	0.0
UDIH	**0.5608**	**0.5709**	**0.5971**	**0.5603**	**0.5772**	**0.5898**	0.02*(n-c)	0.01*(n-c)

3) Comparison under varying PDR: Though previous works that could handle partial multimodal data have promising success, they will be greatly affected when partial data ratio(PDR) is too high. The reason why PDR effect so much on CMH is that CMH needs to leverage more pairwise information, the one-to-one correspondence, to dominantly learn shared common representation. As is shown in Figure 4, we can see that the different results on different PDR. Overall MAP decreases as PDR increases. With the increasing of PDR, IMH [14], CCQ [10], SPDH [13], and UDIH outperform other methods. When the PDR approximate to 0, SPDH [13] and UDIH have similar retrieval performance. When PDR is 0.9, UDIH always performs well than other methods. SPDH [13] has a secondary performance. CCQ [10] also has a promising result, which benefits from its quantization strategy. On the other hand, when PDR is too high, CCQ [10] and SPDH [13] has a larger decline rate than other methods while UDIH has the smallest decline rate. The reason is that UDIH generates more one-to-one information when more data is missing

4) Effect of Parameters: We futher conduct a series of experiments to analyze different parameters in the proposed UDIH. All the experiments are done on both MIRFlickr and NUS-WIDE datasets with hash bits fixed on 16. In the proposed UDIH, p_i and p_t played a decisive role. That is

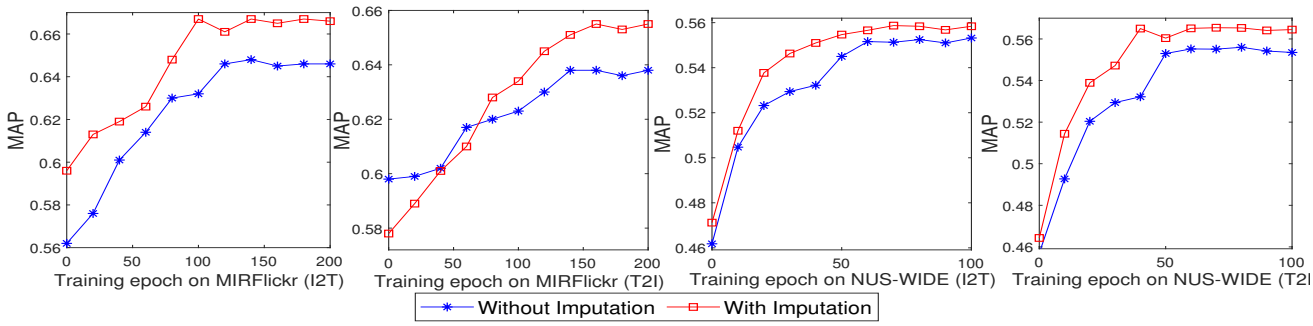

Fig. 3. The effect of imputation on MIRFlickr and NUS-WIDE at 16 bits

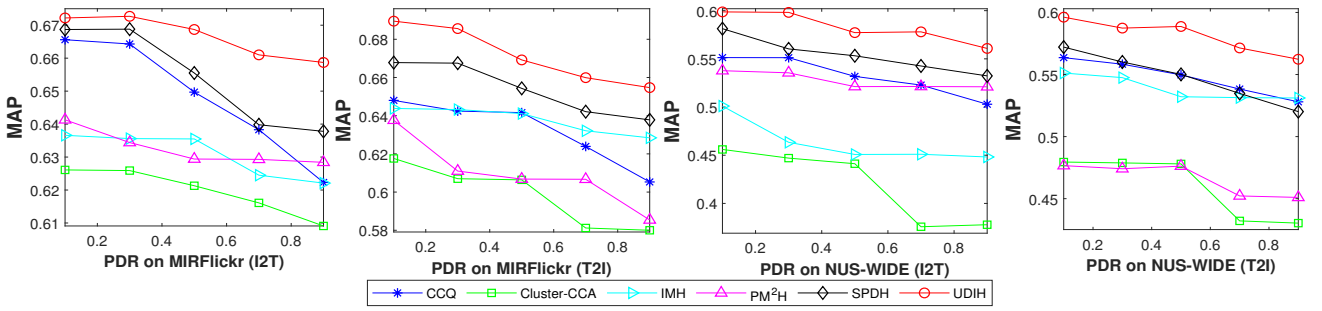

Fig. 4. Comparisons with different PDR on MIRFlickr and NUS-WIDE at 16 bits

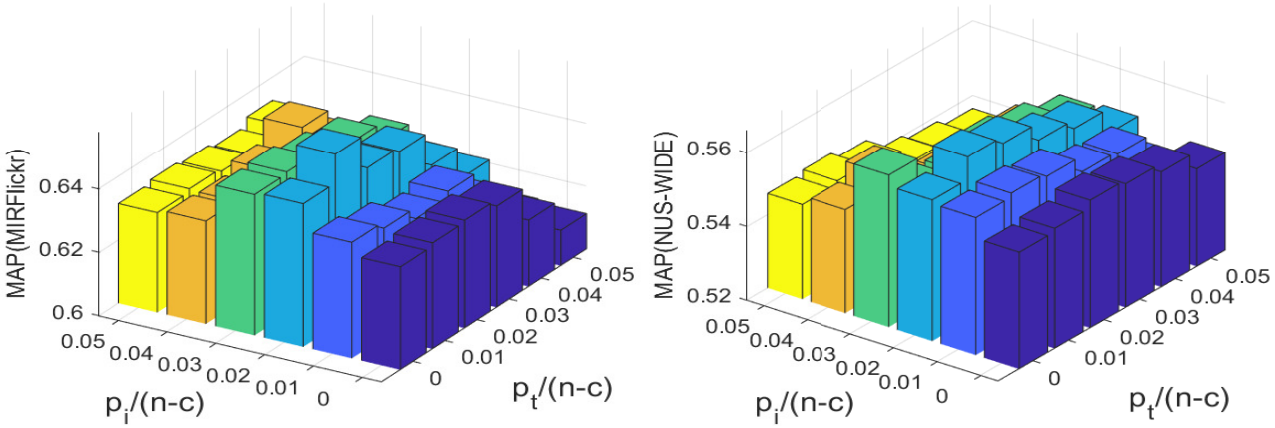

Fig. 5. The effect of p_i and p_t on MIRFlickr and NUS-WIDE at 16 bits

because p_i and p_t denote the number of augmented data objects. If too much data is augmented, more noisy data will be introduced into the training set. Obviously, the number of complete cases should be inversely proportional to the number of augmented data objects. Figure 5 shows the results of different p_i and p_t. The best p_i is around 0.02*$(n-c)$ while p_t is around 0.01*$(n-c)$. Note that this is only a empirical setting. Intuitively, properly small p_i and p_t will perform well. We then analyze the importance of the parameter α. The α defines the reliability of the generated data object. Figure 6 shows the results of different α. With the increasing of α, the MAP gradually increases. But when α is larger than 0.4, the MAP gradually becomes smaller. Because the generated data object has some differences from the real data distribution. Finally, we empirically select optimal parameters as follows. α = [0.4, 0.5]. δ = [2.0,4.0,8.0]. t = [4,5,6]. $p_i = 0.02*(n-c)$. $p_t = 0.01*(n-c)$. lr = 0.01 . λ_{G^1},λ_{G^2},λ_{H^1},λ_{H^2} = 0.001.

V. Conclusion

In this paper, we have proposed a UCMH method UDIH for partial cross-modal retrieval which is a challenging but common problem. UDIH can impute partial data to generate

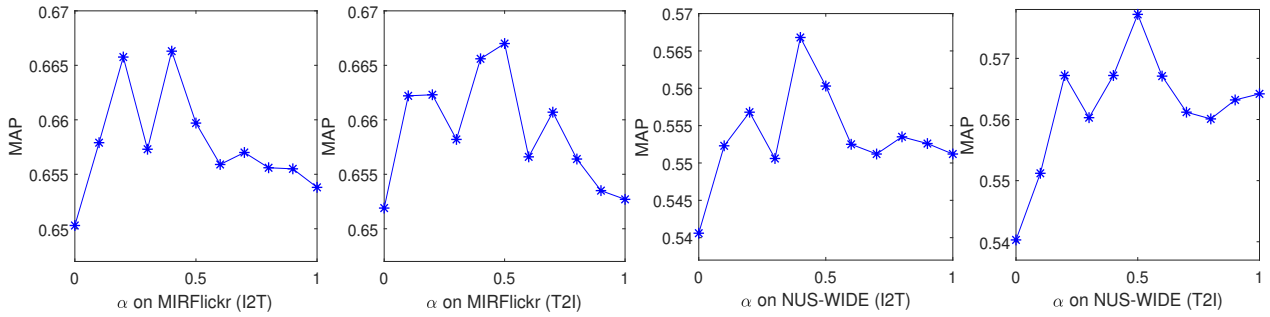

Fig. 6. The effect of α on MIRFlickr and NUS-WIDE at 16 bits

more pairwise information. The weighted cross-modal correlation graph and the two-pathway hashing scheme are exploited to efficiently learn hash functions with different weights of data. Both the consistency and difference will be preserved. Experiments on benchmark datasets verify the effectiveness of UDIH compared with five state-of-the-art approaches. We also explained the motivation of pairwise information generation in the comparison section. However, the main components of UDIH are learning separately. In the future, we will construct a jointly learning framework. Moreover, the extension of preserving intra-modality similarity will be considered in our future work.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 61822601, 61773050, and 61632004; the Beijing Natural Science Foundation under Grant Z180006; National Key Research and Development Program (2017YFC1703506) ; the Fundamental Research Funds for the Central Universities (2019JBZ110).

REFERENCES

[1] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S Yu. Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1445–1454. ACM, 2016.

[2] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao 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*, page 48. ACM, 2009.

[3] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing*, 27(8):3893–3903, 2018.

[4] Mark J Huiskes, Bart Thomee, and Michael S Lew. New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. In *Proceedings of the international conference on Multimedia information retrieval*, pages 527–536. ACM, 2010.

[5] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3232–3240, 2017.

[6] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[7] Shaishav Kumar and Raghavendra Udupa. Learning hash functions for cross-view similarity search. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[8] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Semantics-preserving hashing for cross-view retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3864–3872, 2015.

[9] Wei Liu, Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. Hashing with graphs. 2011.

[10] Mingsheng Long, Yue Cao, Jianmin Wang, and Philip 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*, pages 579–588. ACM, 2016.

[11] Nikhil Rasiwasia, Dhruv Mahajan, Vijay Mahadevan, and Gaurav Aggarwal. Cluster canonical correlation analysis. In *Artificial Intelligence and Statistics*, pages 823–831, 2014.

[12] Patrick Royston. Multiple imputation of missing values. *The Stata Journal*, 4(3):227–241, 2004.

[13] Xiaobo Shen, Fumin Shen, Quan-Sen Sun, Yang Yang, Yun-Hao Yuan, and Heng Tao Shen. Semi-paired discrete hashing: Learning latent hash codes for semi-paired cross-view retrieval. *IEEE transactions on cybernetics*, 47(12):4275–4288, 2017.

[14] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 785–796. ACM, 2013.

[15] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1405–1414, 2017.

[16] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

[17] Qifan Wang, Luo Si, and Bin Shen. Learning to hash on partial multi-modal data. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[18] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In *Advances in neural information processing systems*, pages 1753–1760, 2009.

[19] Jianlong Wu, Zhouchen Lin, and Hongbin Zha. Joint dictionary learning and semantic constrained latent subspace projection for cross-modal retrieval. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1663–1666. ACM, 2018.

[20] Ting-Kun Yan, Xin-Shun Xu, Shanqing Guo, Zi Huang, and Xiao-Lin Wang. Supervised robust discrete multimodal hashing for cross-media retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1271–1280. ACM, 2016.

[21] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[22] Bohan Zhuang, Guosheng Lin, Chunhua Shen, and Ian Reid. Fast training of triplet-based deep binary embedding networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5955–5964, 2016.