Triplet Fusion Network Hashing for Unpaired Cross-Modal Retrieval

Zhikai Hu
Department of Computer Science, Huaqiao University & State Key Laboratory of Integrated Services Networks, Xidian University & Provincial Key Laboratory for Computer Information Processing Technology, Soochow University
Xiamen, China
zkhu@hqu.edu.cn

Xin Liu*
Department of Computer Science, Huaqiao University & State Key Laboratory of Integrated Services Networks, Xidian University
Xiamen, China
xliu@hqu.edu.cn

Xingzhi Wang
Department of Computer Science, Huaqiao University
Xiamen, China
xzwang@hqu.edu.cn

Yiu-ming Cheung
Department of Computer Science and Institute of Research and Continuing Education, Hong Kong Baptist University
Hong kong SAR, China
ymc@comp.hkbu.edu.hk

Nannan Wang
State Key Laboratory of Integrated Services Networks and School of Telecommunications Engineering, Xidian University
Xi'an, China
nnwang@xidian.edu.cn

Yewang Chen
Depart. of CS, Huaqiao University & Provincial Key Laboratory for Computer Information Processing Technology, Soochow University
Xiamen, China
ywchen@hqu.edu.cn

ABSTRACT

With the dramatic increase of multi-media data on the Internet, cross-modal retrieval has become an important and valuable task in searching systems. The key challenge of this task is how to build the correlation between multi-modal data. Most existing approaches only focus on dealing with paired data. They use pairwise relationship of multi-modal data for exploring the correlation between them. However, in practice, unpaired data are more common on the Internet but few methods pay attention to them. To utilize both paired and unpaired data, we propose a one-stream framework triplet fusion network hashing (TFNH), which mainly consists of two parts. The first part is a triplet network which is used to handle both kinds of data, with the help of zero padding operation. The second part consists of two data classifiers, which are used to bridge the gap between paired and unpaired data. In addition, we embed manifold learning into the framework for preserving both inter and intra modal similarity, exploring the relationship between unpaired and paired data and bridging the gap between them in learning process. Extensive experiments show that the proposed approach outperforms several state-of-the-art methods on two datasets in paired scenario. We further evaluate its ability of handling unpaired scenario and robustness in regard to pairwise constraint. The results show that even we discard 50% data under the setting in [19], the performance of TFNH is still better than that of other unpaired approaches and that only 70% pairwise relationships are preserved, TFNH can still outperform almost all paired approaches.

ICMR '19, June 10–13, 2019, Ottawa, ON, Canada

ACM ISBN 978-1-4503-6765-3/19/06…$15.00
https://doi.org/10.1145/3323873.3325041

CCS CONCEPTS

• **Information systems → Information retrieval**;

KEYWORDS

cross-modal retrieval; hashing technique; unpaired data; triplet network

ACM Reference Format:
Zhikai Hu, Xin Liu*, Xingzhi Wang, Yiu-ming Cheung, Nannan Wang, and Yewang Chen. 2019. Triplet Fusion Network Hashing for Unpaired Cross-Modal Retrieval. In *International Conference on Multimedia Retrieval (ICMR '19), June 10–13, 2019, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3323873.3325041

1 INTRODUCTION

As the explosion of multimedia data, more and more research interest has been transferred from single-modal retrieval task [5, 8, 11, 15, 18, 23, 28, 35, 36] to multi-modal retrieval task [1–3, 13, 14, 24–26, 30–34, 37–41]. Different from single-modal retrieval, cross-modal retrieval allows users to search information of one modality by using instances of another modality. For example, one can retrieve texts by using images. Because data of different modalities usually have inconsistent representation, discovering the correlation and bridging the gap across these modalities become the key challenges. Recently, more and more studies concentrated on common subspace learning and tried to introduce hashing technique to the learning process.

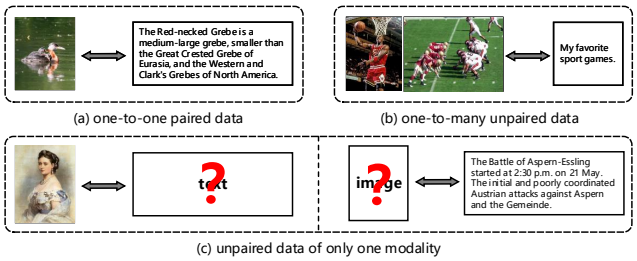

Figure 1: Three different kinds of multi-modal data.

In order to overcome these challenges, it is necessary to analyse the properties of multi-modal data. From a huge amount of multi-modal data, we noticed that more often than not the numbers of data of different modalities vary considerably with the level of difficulty to get them. As a result, there are several different relationships between multi-modal data. In this paper, these data are roughly categorized into three relationships (taking bimodal data images and texts for example):

(1) **one-to-one paired data (Figure 1(a))** In this case, there are one image and its corresponding only text. The connection between them is the most direct and the correlation between them is the strongest as well. Thus, they are of the greatest value to exploring the correlation between multi-modal data and it is relatively easy to utilize them.
(2) **one-to-many unpaired data (Figure 1(b))** In this case, there are one text and its corresponding many images or vice versa. The connection between them is relatively complex and the correlation between them is ambiguous, because it is hard to know whether the information provided by text links to which one of these images or all of them. They usually can not be utilized directly and some pretreatments for them are essential.
(3) **unpaired data of only one modality (Figure 1(c))** In this case, there is only one image or text without corresponding data of another modality. There is no obvious connection between these data and they provide very little information about the relationship between two modalities. Thus, it is extremely hard to use them as a base for establishing the correlation between multi-modal data.

It is easy to find that there are evident differences among these three kinds of data, so filling the data gap between them is necessary. Most existing hashing methods [2, 3, 14, 16, 25, 29, 33] focus on dealing with the first kind of data. They utilize the pairwise relationship of them for exploring the correlation across modalities. Specifically, their aim is that the hash codes learned from each image-text pair can be as similar as possible. The second kind of data are also suitable for them after pretreatment. There are two different pretreatments. One is simply discarding some features, then we can make one-to-many relationship become one-to-one relationship. Another is copying the only feature, then we can get several one-to-one paired data. However, the third kind of data can not be used in these methods and few approaches [19, 20, 22, 24] focus on dealing with the third kind of data.

To take advantage of all kinds of data, we propose a novel approach and fill the gap between them, Triplet Fusion Network Hashing (TFNH), to deal with both the first and third kinds of data simultaneously under the concept of adversarial learning. In this paper, we simply regard the second kind of data as unpaired data of only one modality and discard their one-to-many relationship. As illustrated in Figure 2, there are two primary components, triplet network, which consists of three networks that share the same weights and data classifiers, which play an adversarial role, in our framework. Triplet network is used to receive different kinds of data and generate corresponding representations and hash codes. It aims to confuse two data classifiers. Data classifiers try to distinguish between paired and unpaired data. By playing this minimax game, the learned representations of unpaired data can be as effective as that of paired data. In addition, we use manifold learning for strengthening the relationship between paired and unpaired data and exploring the value of the latter. The main contributions of this paper are

- We propose a triplet fusion network and introduce the zero padding operation to handle both paired and unpaired data simultaneously, which can easily tackle the dominant domain problem as well.
- Two data classifiers are designed to narrow the gap between paired and unpaired data. By playing the minimax game, it is ensured that learned representations of the latter are as effective as that of the former.
- Extensive experiments are conducted to verity the ability of our proposed approach about handling both paired and unpaired scenario and its robustness in regard to pairwise constraint.

The remainder of this paper is organized as follows. We overview the related work on cross-modal retrieval methods in Section 2. Section 3 elaborates our proposed method TFNH. Section 4 provides extensive experimental validation on two datasets and proves the robustness of method in regard to pairwise constraint. The conclusions are made in Section 5.

2 RELATED WORK

In this section, we provide pointers to some of the related work on cross-modal retrieval methods that focus on paired and unpaired data.

Approaches focusing on paired data can be roughly divided into unsupervised [3, 4, 13, 27, 39, 40] and supervised ones [2, 12, 14, 16, 19, 25–27, 29, 30, 33, 39, 40]. Unsupervised methods rely only on one-to-one relationship of data in learning process. KSH-CV [40] adopts an Adaboost framework to learn kernel hash functions for multi-modal data, preserving inter-view similarities simultaneously. Latent semantic sparse hashing (LSSH) [39] first uses sparse coding and matrix factorization for learning representation of multi-modal data, and then projects them into a joint abstract space to generate hash codes. Collective matrix factorization hashing (CMFH) [3] uses collective matrix factorization to learn unified hash codes for features of different views.

Supervised methods always perform better because of the usage of label information. Semantic correlation maximization (SCM)

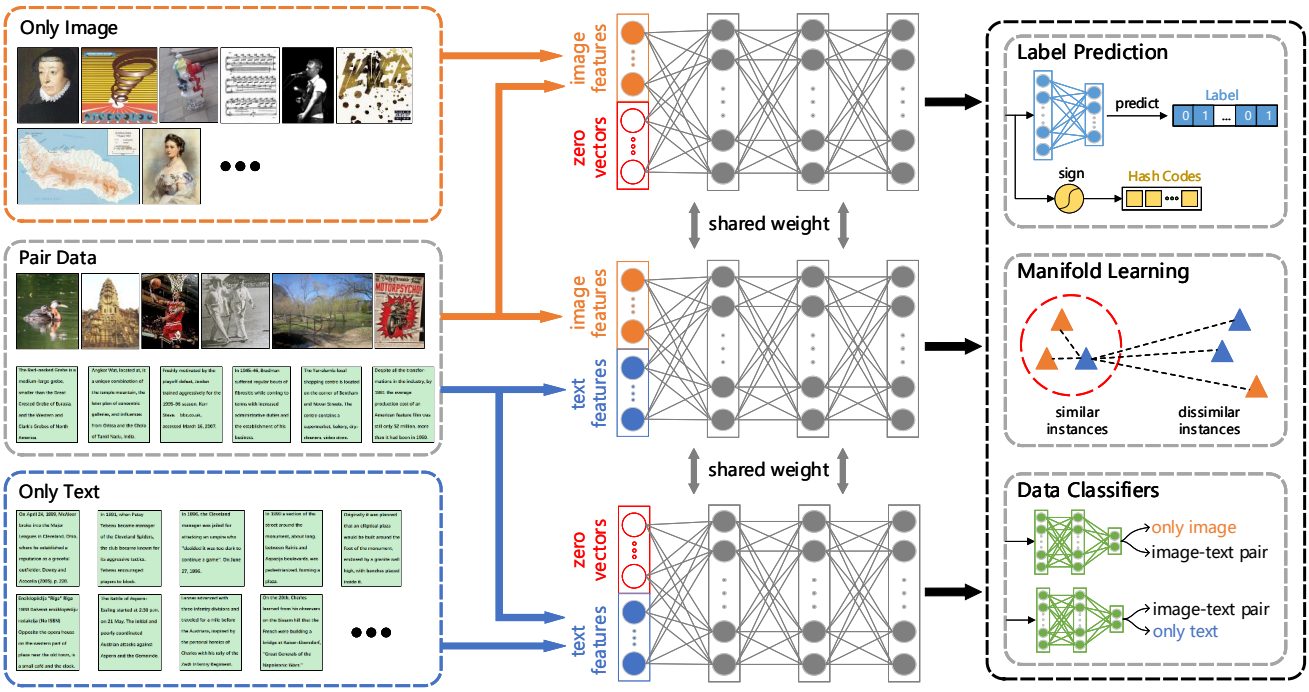

Figure 2: Framework of our proposed method.

[33] uses label information for preserving the maximal correlation between modalities. Semantic-preserving hashing (SePH) [16] first generates hash codes from the semantic affinity matrix by minimizing KL divergence, and then learns two hash function via kernel logistic regression. Supervised matrix factorization hashing (SMFH) [25] extends CMFH and introduces manifold learning to learn more effective hash codes, preserving both inter-modal and intra-modal similarity. Deep cross-modal hashing (DCMH) [12] attempts to combine deep learning with hashing technique and proposes an end-to-end framework to learning features of raw data and hash codes simultaneously. It outperforms several traditional methods which use hand crafted features. Inspired by the idea of GAN [6], adversarial cross-modal retrieval (ACMR) [26] uses a modality classifier to narrow the gap between learned features of different modalities, but it is not a hashing method. Self-supervised adversarial hashing [14] further combines adversarial learning with hash technique and proposes a self supervised learning framework, where hash codes are derived from a lab-net with concurrent supervision of training other two networks.

Only few approaches [19, 20, 22, 24] focus on unpaired data. Inter-media hashing (IMH) [24] defines two selective matrices to handle unpaired data and uses inter-view and intra-view consistency to learn hash functions. However, its performance is not very satisfying, this is probably because that label information is not used in its framework and the learned hash functions are linear. Mandal et al. [19] are the first people who divide the cross-modal retrieval task into four categories: single label paired (SL-P), multi label paired (ML-P), single label unpaired (SL-U) and multi label unpaired (ML-U). They propose a generalized hashing scheme which can seamlessly handle all these scenarios. The performance of this method is much better than IMH, but the ability of handling unpaired scenario was not fully presented in that paper because they only abandon 10% data of one modality to form a new dataset, which is not so different with the original one.

In this paper, the problem is simplified, we only discuss paired and unpaired scenarios, although our framework in fact have abilities to cope with both single and multi label scenarios.

3 PROPOSED METHOD

Without losing generality, we focus on the discussion of bimodal data, specifically images and texts, in this paper. It is easy to extend our method to multi-modal data.

Assume that data of image modality $X = [X_P, X_U]$ consist of two parts: images with corresponding texts and images without corresponding texts. They are denoted as $X_P \in R^{n \times d_1}$ and $X_U \in R^{n_1 \times d_1}$ respectively, where n and n_1 are the numbers of instances of these two parts respectively and d_1 is the dimension of image feature. Similarly, data of text modality $Y = [Y_P, Y_U]$ consist of two parts: texts with corresponding images and texts without corresponding images. They are denoted as $Y_P \in R^{n \times d_2}$ and $Y_U \in R^{n_2 \times d_2}$ respectively, where n_2 is the number of texts without corresponding images and d_2 is the dimension of text feature. Labels of data X and Y are denoted as $L_X \in \{0,1\}^{(n+n_1) \times c}$ and $L_Y \in \{0,1\}^{(n+n_2) \times c}$ respectively, where c is the number of the category.

Our aim is to learn a common Hamming semantic subspace for bimodal data. In this space, every image or text instance is represented as a binary vector $b \in \{0,1\}^{1 \times l}$, where l is the dimension of the learned Hamming space.

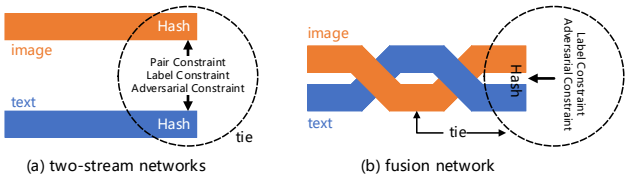

Figure 3: Differences between two-stream structure and one-stream fusion network.

3.1 Fusion Network

Most existing methods (both deep learning based methods [14, 26, 30] and matrix factorization based methods [2, 3, 25]) employed two-stream structure to explore the Hamming subspace. That is, they use two separated networks (deep learning based methods) or matrices (matrix factorization based methods) to learn projection functions that map image and text features into Hamming space respectively, usually under the supervision of label information.

However, there are are two problems in this structure. Taking deep learning based methods for example, firstly, it highly relies on pairwise constraint. That is, it deeply depends on the pairwise relationship of data to explore the correlation between two modalities. This structure fails to build the semantic gap between modalities if there is no one-to-one or one-to-many relationship. In addition, label constraint imposed on the end of this structure is the only tie between two separated networks, as shown in Figure 3(a). This means that the correlation between modalities hardly explored by these networks themselves, it depends more on pairwise and label constraints. Recently, inspired by the idea of GAN [6], some methods [14, 26] introduce adversarial network to strengthen the relationship between modalities, but these adversarial networks are all imposed on the same position where label information is imposed. Thus, the second problem still exists.

To tackle the second problem, we adopt a new one-stream structure in this paper. Firstly, we twist the two separate networks into a four-layer fusion network $G_{xy}(\cdot;\theta_{xy})$ $(d_1+d_2 \rightarrow 1024 \rightarrow 512 \rightarrow l)$ to handle image and text data simultaneously, where θ_{xy} denotes the parameters of fusion network G_{xy}. As shown in Figure 3(b), in our one-stream structure, we can explore the correlation between modalities by both fusion network and other constraints, since the fusion network becomes a new strong tie between two modalities.

As mentioned before, data consist of paired and unpaired ones. For image-text pair X_P and Y_P, we concatenate them to form longer features $F_{xy} \in R^{n\times(d_1+d_2)}$. As a result, they are represented as

$$F_{xy}^{(k)} = [X_P^{(k)}, Y_P^{(k)}] \quad , \quad k = 1, 2, ..., n \tag{1}$$

where $F_{xy}^{(k)}$, $X_P^{(k)}$ and $Y_P^{(k)}$ are the k-th instances of data F_{xy}, X_P and Y_P respectively. These new features are fed into fusion network G_{xy} to get their corresponding representations and binary codes.

To tackle the first problem, we construct another two networks $G_{ox}(\cdot;\theta_{ox})$ and $G_{oy}(\cdot;\theta_{oy})$ that share the same structure with G_{xy} to handle unpaired image and text data separately, where θ_{ox} and θ_{oy} are the parameters of these two networks respectively. These two networks are used for dealing with unpaired images and texts. It is easy to find that because there is only one modality, X_U and Y_U can not be fed into $G_{ox}(\cdot;\theta_{ox})$ and $G_{oy}(\cdot;\theta_{oy})$ directly. To meet the need of input dimension, we introduce an operation zero padding. The detail of this operation is given in the section 3.2.

3.2 Zero Padding

In the rest of this section, $\mathbf{0}$ denotes an all-zero vector or a matrix. To meet the demand of input dimension, we concatenate data X and Y with $\mathbf{0}$ to get longer representations $F_{ox} \in R^{(n+n_1)\times(d_1+d_2)}$ and $F_{oy} \in R^{(n+n_2)\times(d_1+d_2)}$. In this paper, both paired and unpaired images or texts are regarded as unpaired data for increasing the training samples sand eliminating the gap between them. As a result, the new longer features are represented as

$$\begin{aligned} F_{ox}^{(i)} &= [X^{(i)}, \mathbf{0}^{1\times d_2}] \quad , \quad i = 1, 2, ..., n+n_1 \\ F_{oy}^{(j)} &= [\mathbf{0}^{1\times d_1}, Y^{(j)}] \quad , \quad j = 1, 2, ..., n+n_2 \end{aligned} \tag{2}$$

where $F_{ox}^{(i)}$ and $X^{(i)}$ are the i-th instances of data F_{ox} and X respectively, and $F_{oy}^{(j)}$ and $Y^{(j)}$ are the j-th instances of data F_{oy} and Y respectively. Then, they will be fed into $G_{ox}(\cdot;\theta_{ox})$ and $G_{oy}(\cdot;\theta_{oy})$ respectively.

Inspired by [9], we let these three networks share the same parameters θ_g, that is, let $\theta_{xy} = \theta_{ox} = \theta_{oy} = \theta_g$. Then, the triplet network is denoted as $G(\cdot;\theta_g)$ briefly in the rest of this paper. The features learned from triplet fusion network by feeding different data are represented as

$$H_{xy,ox,oy} = G(F_{xy,ox,oy};\theta_g) \tag{3}$$

And their corresponding hash codes can be derived from

$$B_{xy,ox,oy} = sign(H_{xy,ox,oy}) \tag{4}$$

By introducing these two operations, we can surprisingly tackle a thorny issue dominant domain. Dominant domain is a common problem when fusing multi-modal data. That is, if data of one modality are more discriminative than that of other modalities, the weights assigned to the former would be greater than that assigned to the latter. As a result, this modality become the dominant domain and data of this modality play a more crucial role in classification or retrieval task, while other data tend to become a kind of noise, which makes less sense of results.

By adopting zero padding strategy and letting triplet network share the same parameters, the dominant domain problem can be easily solved. To be specific, in the training process, if weights assigned to image modality are far greater than to text modality, data F_{oy} would make the networks dysfunctional, because when F_{oy} are fed into network, weights of image modality would multiply zero vector and then make no contribution to the result. Similarly, if the text modality become the dominant domain, data F_{ox} would make the networks dysfunctional as well. As a result, our triplet fusion networks would adjust the weights assigned to different modalities automatically and reach a balance finally.

3.3 Data Classifiers

Our aim is that the distribution of representations learned from unpaired data and that from paired data are as same as possible. To reach this goal, we defined two data classifiers $D_1(\cdot;\theta_{d_1})$ $(l \rightarrow 64 \rightarrow 32 \rightarrow 2)$ and $D_2(\cdot;\theta_{d_2})$ $(l \rightarrow 64 \rightarrow 32 \rightarrow 2)$, which act as the role of discriminator in GAN, where θ_{d_1} and θ_{d_2} denote the parameters

of two classifiers respectively. The first discriminator $D_1(\cdot;\theta_{d_1})$ is used to discriminate between data from only image modality and that from image-text pairs. We regard the former as fake samples and the latter as real samples. Then, following the formulation in [6], we can get the adversarial loss

$$\mathcal{L}_{adv}^{(1)} = \sum_{h_1 \in \hat{H}_{xy}} \sum_{h_2 \in \hat{H}_{ox}} (logD_1(h_1;\theta_{d_1}) + log(1 - D_1(h_2;\theta_{d_1}))) \quad (5)$$

where $\hat{A}$ denotes the set consisting of all vectors of matrix A. For example, $\hat{H}_{xy} = \{H_{xy}^{(i)} | i = 1, 2, ..., n\}$.

Similarly, the second discriminator $D_2(\cdot;\theta_{d_2})$ is used to discriminate between data from only text modality and that from image-text pairs. We regard the former as fake samples and the latter as real samples and get the adversarial loss

$$\mathcal{L}_{adv}^{(2)} = \sum_{h_1 \in \hat{H}_{xy}} \sum_{h_2 \in \hat{H}_{oy}} (logD_2(h_1;\theta_{d_2}) + log(1 - D_2(h_2;\theta_{d_2}))) \quad (6)$$

We combine these two parts and define the overall adversarial loss

$$\mathcal{L}_{adv} = \mathcal{L}_{adv}^{(1)} + \mathcal{L}_{adv}^{(2)} \quad (7)$$

In this minimax game, this term is used for narrowing the data gap mentioned before.

3.4 Label Prediction

To make the learned feature more discriminative, we construct a three-layer classification network $C(\cdot;\theta_c)$ $(l \to l/2 \to c)$ to predict the label of each instance, where θ_c denotes the parameters of this network. We assume that all features exhibit the same distribution under the supervision of data classifiers. Thus, H_{xy}, H_{ox} and H_{oy} are all fed into the same classifier $C(\cdot;\theta_c)$ to predict their labels. Then, we can minimize the following objection

$$\mathcal{L}_{class} = \sum_{h \in \hat{H}_{xy} \cup \hat{H}_{ox} \cup \hat{H}_{oy}} ||C(h;\theta_c) - lab||_2^2 \quad (8)$$

where lab is the label of h.

3.5 Manifold Learning

We utilize paired data for preserving the inter-modal similarity and unpaired data for preserving intra-modal similarity. Thus, we construct two kinds of similarity matrix for them respectively.

For paired data, we hope that they should be as close as possible after being projected into common semantic space if they share the same label. Thus, we use the label information for constructing the semantic affinity matrix S_{xy}. The item in i-th row and j-th column of S_{xy} is defined as

$$S_{xy}^{(ij)} = \begin{cases} 1 & , \quad if\ X_P^{(i)}\ and\ Y_P^{(j)}\ have\ the\ same\ category \\ 0 & , \quad otherwise \end{cases} \quad (9)$$

where $X_P^{(i)}$ and $Y_P^{(j)}$ are the i-th instance of X_P and j-th instance of Y_P. Then, to preserve the inter-modal similarity, we can minimize the following objection

$$\mathcal{L}_{inter} = \sum_{i=1}^{n} \sum_{j=1}^{n} S_{xy}^{(ij)} ||H_{xy}^{(i)} - H_{xy}^{(j)}||_2^2 \quad (10)$$

where $H_{xy}^{(i)}$ is the i-th instance of H_{xy}.

For unpaired data of only one modality, we hope that they should be still near after projecting into the common semantic space if they are near in the original space. Thus, we use the distances of them for defining the similarity matrix S_{ox} of unpaired data X. The item in i-th row and j-th column of S_{ox} is defined as

$$S_{ox}^{(ij)} = \begin{cases} 1 & , \quad if\ dist(X^{(i)}, X^{(j)}) \le threshhold \\ 0 & , \quad otherwise \end{cases} \quad (11)$$

where $threshold = max(dist(X^{(i)}, X^{(j)}))/20$ and $X^{(i)}$ is the i-th instance of X. In the same way, we can define the similarity matrix S_{oy} of unpaired data Y. The item in i-th row and j-th column of S_{oy} is defined as

$$S_{oy}^{(ij)} = \begin{cases} 1 & , \quad if\ dist(Y^{(i)}, Y^{(j)}) \le threshhold \\ 0 & , \quad otherwise \end{cases} \quad (12)$$

where $threshold = max(dist(Y^{(i)}, Y^{(j)}))/20$ and $SY^{(i)}$ is the i-th instance of Y. Then, to preserve the intra-modal similarity, we can minimize the following objection

$$\begin{aligned} \mathcal{L}_{intra} = & \sum_{i=1}^{n+n_1} \sum_{j=1}^{n+n_1} S_{ox}^{(ij)} ||H_{ox}^{(i)} - H_{ox}^{(j)}||_2^2 \\ & + \sum_{i=1}^{n+n_2} \sum_{j=1}^{n+n_2} S_{oy}^{(ij)} ||H_{oy}^{(i)} - H_{oy}^{(j)}||_2^2 \end{aligned} \quad (13)$$

We take advantage of the whole data X and Y instead of X_U and Y_U to construct S_{ox} and S_{oy} out of two considerations. Firstly, more information is helpful to train networks. More importantly, doing so is beneficial to bridge the data. To be specific, we can preserve the similarity between X_P and X_U or Y_P and Y_U by S_{ox} or S_{oy}, thereby construct the relationship between X_U and Y_U by X_P, Y_P and S_{oy}.

3.6 Optimization

The overall loss function, consisting of the label prediction term $\mathcal{L}_{class}$ in Eq.(8), the pair relationship term $\mathcal{L}_{inter}$ in Eq.(10), the local structure term $\mathcal{L}_{intra}$ in Eq.(13) and the adversarial term $\mathcal{L}_{adv}$ in Eq.(7), is given as follow

$$\mathcal{L}_{all} = \mathcal{L}_{class} + \beta\mathcal{L}_{inter} + \gamma\mathcal{L}_{intra} - \mu\mathcal{L}_{adv} \quad (14)$$

where β, γ and μ are the hyper-parameters that balance four items. In this paper, we set $\beta = 0.1$, $\gamma = 0.3$ and $\mu = 2$.

$$(\theta_g, \theta_c) = \arg\min_{\theta_g, \theta_c} \mathcal{L}_{all} \quad (15)$$

$$(\theta_{d_1}, \theta_{d_2}) = \arg\max_{\theta_{d_1}, \theta_{d_2}} \mathcal{L}_{all} = \arg\min_{\theta_{d_1}, \theta_{d_2}} \mathcal{L}_{adv} \quad (16)$$

Based on Eq.(15) and Eq.(16) we can update the parameter as follows:

$$\begin{aligned} \theta_g \leftarrow \theta_g - \alpha\frac{\partial\mathcal{L}_{all}}{\partial\theta_g}\ ,\ \theta_c \leftarrow \theta_c - \alpha\frac{\partial\mathcal{L}_{class}}{\partial\theta_c} \\ \theta_{d_1} \leftarrow \theta_{d_1} - \alpha\frac{\partial\mathcal{L}_{adv}^{(1)}}{\partial\theta_{d_1}}\ ,\ \theta_{d_2} \leftarrow \theta_{d_2} - \alpha\frac{\partial\mathcal{L}_{adv}^{(2)}}{\partial\theta_{d_2}} \end{aligned} \quad (17)$$

where α is learning rate. To solve this problem, we use stochastic gradient descent (SGD) strategy and Adam optimizer and set $\alpha = 0.001$ in all experiments.

Table 1: Comparison of cross-modal retrieval performance (MAP) of the proposed approach with the state-of-the-art on Wiki dataset for paired scenario with different hash code lengths. Best results are marked in bold.

Method	I→T			T→I		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
IMH	0.1176	0.1167	0.1188	0.1096	0.1118	0.1139
CMFH	0.2172	0.2231	0.2316	0.4902	0.5077	0.5173
LSSH	0.1541	0.1546	0.1544	0.2641	0.2723	0.2795
SCM	0.2257	0.2459	0.2461	0.2341	0.2410	0.2445
SePH	0.2562	0.2654	0.2793	0.6276	0.6324	0.6513
SMFH	0.2507	0.2646	0.2715	0.4481	0.4827	0.4920
DCMH	0.2798	0.2809	0.2910	0.6292	0.6524	0.6674
GSePH	0.2778	0.2882	0.3044	0.6445	0.6639	0.6683
MCTD	0.2919	0.3048	0.3068	0.6482	0.6832	0.6898
TFNH	**0.3158**	**0.3248**	**0.3368**	**0.6813**	**0.6908**	**0.7044**

4 EXPERIMENT

In this section, extensive experiments are conducted to evaluate the effectiveness of our method TFNH. Firstly, we compared the proposed method with several state-of-art methods on two public cross-modal datasets. Then, we conduct a series of experiments to verify the robustness of our method of dealing with unpaired data. Finally, we evaluate the contribution of each part of our object function.

4.1 Datasets and Protocol

Here we briefly introduce the two datasets mentioned before.

Wiki [21] consists of 2,866 image-text pairs, which are categorized into 10 classes (single label). Each image is represented by 128-d SIFT descriptor and each text is represent by 10-d LDA feature. It was split into a training set of 2,173 instances and a test set of 693 instances.

MIRFlickr [10] consists of 25,000 image-text pairs, each of which is annotated with at least one of 24 labels (multi label). For each instance, its image and text are represented by 150-d edge histogram and 500-d PCA feature respectively. Follow the pretreatment in [16], we firstly removed the instances without labels or textual tags appearing less than 20 times. Then, we took 95% of the remainder as training set and 5% as test set.

For fair comparison, the widely used mean average precision (mAP) score is employed as evaluation metric to measure the performance of both TFNH and compared methods. We perform two cross-modal retrieval tasks: searching relevant text by given image query(I→T) and vice versa (T→I). In our experiments, we repeat three times for every different settings and report the mean mAP score.

4.2 Evaluation of Paired Scenario

In this scenario, we set $n_1 = n_2 = 0$, that is, there are only paired data and no unpaired data. We first evaluate TFNH with different length of hash codes (16 bits, 32 bits and 64 bits) on Wiki datase, a single label dataset that has been widely used as a benchmark dataset and compare it with both unsupervised methods IMH [24], CMFH [3], LSSH [39] and state-of-the-art supervised methods SCM [33], SePH [16], SMFH [25], DCMH [12], GSePH [19], MCTD [2]. The performance (mAP scores) of them are reported in Table 1 and the best results are marked in bold.

Table 2: Comparison of cross-modal retrieval performance (MAP) of the proposed approach with the state-of-the-art on MIRFlickr dataset for paired scenario with different hash code lengths. Best results are marked in bold.

Method	I→T			T→I		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
CCA	0.5819	0.5756	0.5710	0.5803	0.5750	0.5708
CMFH	0.5683	0.5684	0.5687	0.5646	0.5652	0.5649
SMFH	0.5913	0.5997	0.5956	0.5890	0.5652	0.5649
FSH	0.5893	0.6027	0.6006	0.5865	0.5970	0.5965
SCM	0.6280	0.6345	0.6385	0.6176	0.6234	0.6285
GSePH	0.6460	0.6649	0.6725	0.6663	0.7113	0.7269
SePH	0.6736	0.6789	0.6822	0.7313	0.7320	0.7381
TFNH	**0.6962**	**0.7066**	**0.6881**	**0.7378**	**0.7572**	**0.7613**

We have three observations from Table 1. Firstly, our TFNH achieves much better performance than all compared methods in the overall experiments, which demonstrates its superiority. To be specific, for I→T task, TFNH outperforms MCTD by 2.29%, 2% and 3% in cases that hash code length is 16 , 32, 64 bits. Whereas for T→I task, TFNH outperforms MCTD by 3.31%, 0.76% and 1.06%. It is worth noting that MCTD uses the combination of two kinds of features (128-d SIFT histogram and 128-d CNN feature) in its learning process, while our approach only uses 128-SIFT feature as input. In other words, TFNH uses less information but yields better results, which further verifies its advantage.

Secondly, the mAP score of TFNH in the case that hash code length is 16 bits is higher than that of all compared methods except MTCD in the case that hash code length is 64 bits. It not only is another evidence to prove the superiority of our approach but also brings another storage benefit. That is, TFNH can use less space for storing hash codes but reach comparable results. For example, the mAP score of TFNH is 0.3158 in 16 bits, which is higher than 0.3068 of MCTD in 64 bits. This means $(64 - 16) \times n$ bits space can be saved, where n is the size of datasets.

In addition, with the increase in the hash code length, the performance of TFNH monotonically increases, which indicates that the longer hash code length is, the more information hash codes preserve.

To further verify the ability of TFNH about handling multi label scenario, we conduct the same experiments on another multi label dataset MIRFlickr and compare it with both unsupervised methods CCA [7], CMFH [3] and state-of-the-art supervised methods SCM [33], FSH [17], SePH [16], SMFH [25], GSePH [19]. Since there are no available multi-view data of each modality, which are needed in MCTD, we do not report the result of this approach. The performance (mAP scores) of them are reported in Table 2 and the best results are marked in bold.

We have three observations from Table 2 as well. Firstly, as on Wiki dataset, TFNH yields higher mAP scores than all compared methods in the overall experiments. To be specific, for I→T task,

Table 3: Evaluation of the proposed algorithm in the unpaired scenario for the Wiki dataset. Best results are marked in bold.

Method		case1			case2		
		I→T	T→I	avg	I→T	T→I	avg
IMH		0.112	0.114	0.113	0.122	0.109	0.116
GSePH	16	0.257	0.453	0.355	0.268	0.422	0.345
	32	0.273	0.477	0.375	0.279	0.438	0.358
	64	0.283	0.484	0.383	0.298	0.456	0.377
TFNH	16	0.304	0.677	0.490	0.308	0.671	0.489
	32	0.308	0.682	0.495	0.319	0.676	0.497
	64	**0.325**	**0.695**	**0.510**	**0.340**	**0.698**	**0.519**

TFNH outperforms SePH by 2.26%, 2.77% and 0.59% in cases that hash code length is 16 , 32, 64 bits. Whereas for T→I task, TFNH outperforms SePH by 0.65%, 2.52% and 2.32%. Secondly, TFNH remains the second storage advantage mentioned before. The mAP score of TFNH in the case that hash code length is 16 bits is higher than that of all compared methods except SePH in the case that hash code length is 64 bits.

The last interesting finding is that when hash code length increases from 32 bits to 64 bits, for I→T task, the corresponding mAP score of TFNH drops by 1.85% (from 0.7066 to 0.6881). This phenomenon violates the third observation we find in Table 1. It is not peculiar to TFNH, the same changes are found in CCA, SMFH and FSH. However, the performance of these three approaches continuously increases with the hash code length increases in Table 1. Thus, we think this strange phenomenon is possibly due to the differences between these two datasets, such as the dimension of image and text features and the number of categories. We do not further investigate it because it is not the focus of this paper.

4.3 Evaluation of Unpaired Scenario

To evaluate the ability of TFNH about handling unpaired scenario, we compare it with two approaches IMH [24] and GSePH [19], following the same setting in [19]. In case 1, the data of text modality keep the same while only 90% data of image modality are retained and vice versa in case 2. In [19], mAP@all was used as metric in paired scenario while mAP@50 was used in unpaired scenario, which we think is not so reasonable. Thus, we adopt the the same metric mAP@all to evaluate the performance in both scenarios for fair comparison. The performance of them are reported in Table 3 and the best results are marked in bold.

We observe that almost all performance reported in Table 3 is worst than that in paired scenario. The performance of IMH remains the same in unpaired scenario but it is far from acceptable. The performance of GSePH is much better than IMH but there still is a huge gap between the performance in paired and unpaired scenarios, especially for T→I task. Specifically, in case 1, for both tasks, the mAP scores of 64 bits drop by 2.13% and 18.48% respectively; in case 2, those drop by 0.61% and 21.25%. TFNH considerably outperforms both IMH and GSepH and the difference between the performance in paired and unpaired scenarios is slight. Specifically, in case 1, for both tasks, the mAP score of 64 bits drop by 1.21% and 0.97% respectively; in case 2, for T→I task, the mAP score of 64 bits drops by 0.67%, whereas for I →T task, that increases by 0.29% beyond expectation.

In the above experiments, 90% data of one modality are retained. This change is slight, which we think can not fully verify the ability of handling unpaired scenario. Thus, we decrease this percentage to further evaluate this ability of TFNH. The results are presented in Figure 4.

From Figure 4 (a) and (b), we can observe that in case 1, as image data decrease, the mAP scores of I→T task decline gradually while that of T→I task almost remain the same. This is reasonable because the decrease of the number of image training data make the network disable to deal with image data as masterly as text data. The opposite happens in case 2. As shown in Figure 4 (c) and (d), with the decrease of text data, the mAP scores of I→T task almost remain the same while that of T→I task decline gradually. This phenomenon provides a support for our claim. In addition, even thought we discard 50% data, the performance of TFNH is still better than that of GSePH in Table 3 where only 10% data are discarded.

4.4 Robustness of TFNH

To verify the robustness of our method in regard to pairwise constraint, we relax the pairwise constraint to different extend and construct a series of experiment. Keep the number of training data unchanged, we gradually reduce the percentage of pairwise relationship ($\frac{n}{n+n_1}$ and $\frac{n}{n+n_2}$). Taking Wiki dataset for example, there are 2,173 image-text pairs. We select all of them as training set, but randomly remove some (for example 10%) pairwise relationship of these data and regarded them as unpaired data.

The results are reported in Figure 5. We can observe that as the number of pairwise relationship decreases, the performance of both I→T and T→I tasks declines gradually. This decline is accountable because pairwise constraint is propitious to explore the correlation of modalities. There are two obvious drop intervals 100%-90% and 70%-60%. In interval 100%-90%, for both tasks, performance of all bits decrease by about 2%. Similarly, in interval 70%-60%, for I→T task, performance of all bits decrease by about 2%, while for T→I task, they slump by at most 4%. Follow these two drops, performance almost remain the same within the intervals 70%-90% and 40%-60%.

It is worth noting that even if there are only 70% pairwise relationships, for I→T task, the performance of our approach (0.2999 of 16 bits, 0.3146 of 32 bits and 0.3222 of 64 bits) still outperforms all compared methods, whereas for T→I task, the performance of our method (0.6388 of 16 bits, 0.6743 of 32 bits and 0.6863 of 64 bits) is only second to MCTD, while still being better than the other techniques. This means that TFNH can use less information but achieve better performance. This is also a strong evidence to prove the robustness of our method in regard to pairwise constraint.

4.5 Evaluation of Each Term

To evaluate the contribution of the pair relationship term $\mathcal{L}_{inter}$, the local structure term $\mathcal{L}_{intra}$ and the adversarial term $\mathcal{L}_{adv}$, we set part of hyper-parameters to 0 and conduct a series of experiments. Table 4 reports the mAP scores of various combinations of them ($l = 64$).

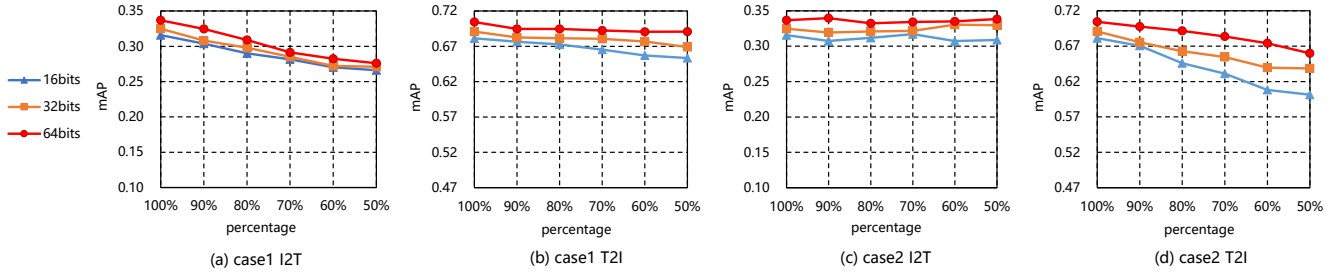

Figure 4: Performance of TFNH on Wiki dataset with different percentage of paired data.

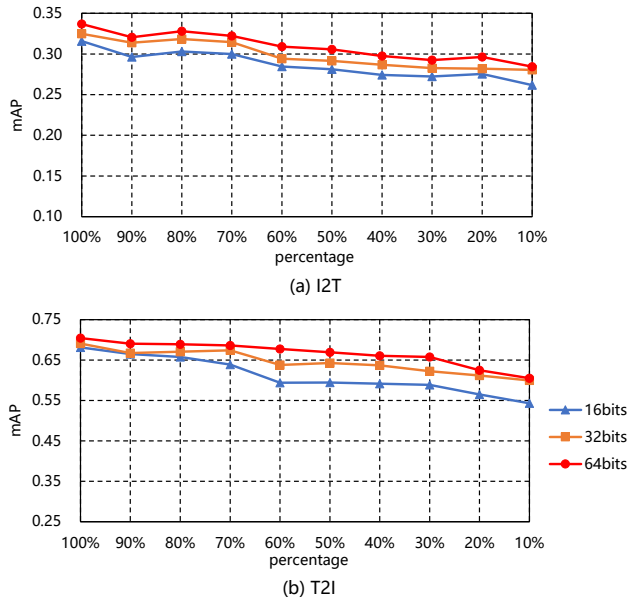

Figure 5: Performance of TFNH on Wiki dataset with different percentage of pairwise constraint.

Table 4: Performance of TFNH under different settings in case that hash code length is 64.

No.	hyper-parameters	I→T	T→I	avg
1	$\beta = 0$	0.3225	0.6965	0.5096
2	$\gamma = 0$	0.3088	0.6924	0.5006
3	$\mu = 0$	0.3133	0.6948	0.5041
4	$\beta = \gamma = 0$	0.2921	0.6971	0.4946
5	$\beta = \mu = 0$	0.3094	0.6926	0.5010
6	$\gamma = \mu = 0$	0.2899	0.6891	0.4895
7	$\beta = \gamma = \mu = 0$	0.2718	0.6877	0.4798
8	overall	**0.3368**	**0.7044**	**0.5206**

Comparing experiments No. 1-7 with No. 8, we find that the performance of all kinds of combination are inferior to that of overall loss defined in Eq.(14). When discarding only one term ($\beta = 0$ or $\gamma = 0$ or $\mu = 0$), the decline in mAP score is not so obvious. For example, setting $\beta = 0$, the scores of I→T and T→I tasks drop slightly by 1.43% and 0.79% respectively. When discarding two of them ($\beta = \gamma = 0$ or $\beta = \mu = 0$ or $\gamma = \mu = 0$), the mAP scores of I→T decline more evidently, all dropping at least by 3%. Without all these three terms ($\beta = \gamma = \mu = 0$), the performance of both task slump dramatically, especially of I→T task, dropping from 0.3368 to 0.2718. From the comparison between experiments No. 1-3, we find that the local structure term $\mathcal{L}_{intra}$ plays the most essential role among the three terms. Specifically, when $\mathcal{L}_{inter}$ is discarded ($\beta = 0$), the performance of both tasks drops by 1.43% and 0.79% respectively; when $\mathcal{L}_{intra}$ is discarded ($\gamma = 0$), that drops by 2.80% and 1.20% respectively; when $\mathcal{L}_{adv}$ is discarded ($\mu = 0$), that drops by 2.35% and 0.96% respectively. The decline in the second case is the most obvious, so we claim the importance of the local structure term. From the comparisons between experiments No.1,4,5 and No.3,5,6, the same conclusion can be made.

5 CONCLUSIONS

In this paper, we propose a one-stream framework, triplet fusion network hashing (TFNH), to handle both paired and unpaired data simultaneously. By introducing zero padding operation, both kinds of data can be fed into the triplet fusion network and the dominant domain problem can be tackled. Under the supervision of data classifiers, learned representations of unpaired data can be as effective as that of paired data. In paired scenario, TFNH outperforms all compared approached on two widely used datasets. In unpaired scenario, extensive experiments verify that our proposed TFNH has strong abilities of dealing with unpaired data and does not depend on pairwise constraint as seriously as other methods.

ACKNOWLEDGMENTS

The work was supported by National Science Foundation of China (Nos. 61673185, 61672444, 61876142 and 61876068), the National Science Foundation of Fujian Province (Nos. 2017J01112 and 2018J01094), State Key Laboratory of Integrated Services Networks of Xidian University (No. ISN20-11), Quanzhou City Science & Technology Program of China (No. 2018C107R), Innovation and Technology Fund (ITF) with Project Code: ITS/339/18, Initiation Grant for Faculty Niche Research Areas of Hong Kong Baptist University with Project Code: RC-FNRA-IG/18-19/SCI/03, the open project of Provincial Key Laboratory for Computer Information Processing Technology, Soochow University (NO. KJS1839) , and in part by CCF-Tencent Open Fund. Xin Liu is the corresponding author.

REFERENCES

[1] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios. 2010. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3594–3601.

[2] Limeng Cui, Zhensong Chen, Jiawei Zhang, Lifang He, Yong Shi, and Philip S Yu. 2018. Multi-view Collective Tensor Decomposition for Cross-modal Hashing. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. ACM, 73–81.

[3] Guiguang Ding, Yuchen Guo, and Jile Zhou. 2014. Collective matrix factorization hashing for multimodal data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2075–2082.

[4] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. ACM, 7–16.

[5] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. 2013. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 12 (2013), 2916–2929.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[7] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* 16, 12 (2004), 2639–2664.

[8] Kaiming He, Fang Wen, and Jian Sun. 2013. K-means hashing: An affinity-preserving quantization method for learning binary compact codes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2938–2945.

[9] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*. 84–92.

[10] Mark J Huiskes and Michael S Lew. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. ACM, 39–43.

[11] Go Irie, Zhenguo Li, Xiao-Ming Wu, and Shih-Fu Chang. 2014. Locally linear hashing for extracting non-linear manifolds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2115–2122.

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

[13] Shaishav Kumar and Raghavendra Udupa. 2011. Learning hash functions for cross-view similarity search. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Vol. 22. 1360.

[14] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. 2018. Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4242–4251.

[15] Guosheng Lin, Chunhua Shen, Qinfeng Shi, Anton Van den Hengel, and David Suter. 2014. Fast supervised hashing with decision trees for high-dimensional data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1963–1970.

[16] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. 2015. Semantics-preserving hashing for cross-view retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3864–3872.

[17] Hong Liu, Rongrong Ji, Yongjian Wu, Feiyue Huang, and Baochang Zhang. 2017. Cross-modality binary code learning via fusion similarity hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6345–6353.

[18] Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang. 2012. Supervised hashing with kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2074–2081.

[19] Devraj Mandal, Kunal N Chaudhury, and Soma Biswas. 2017. Generalized semantic preserving hashing for n-label cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2633–2641.

[20] Viresh Ranjan, Nikhil Rasiwasia, and CV Jawahar. 2015. Multi-label cross-modal retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. 4094–4102.

[21] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, and N. Vasconcelos. 2010. A New Approach to Cross-Modal Multimedia Retrieval. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. 251–260.

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

[23] Fumin Shen, Chunhua Shen, Qinfeng Shi, Anton Van Den Hengel, and Zhenmin Tang. 2013. Inductive hashing on manifolds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1562–1569.

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

[25] Jun Tang, Ke Wang, and Ling Shao. 2016. Supervised matrix factorization hashing for cross-modal retrieval. *IEEE Transactions on Image Processing* 25, 7 (2016), 3157–3166.

[26] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. ACM, 154–162.

[27] Di Wang, Xinbo Gao, Xiumei Wang, and Lihuo He. 2015. Semantic Topic Multimodal Hashing for Cross-Media Retrieval.. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 3890–3896.

[28] Yair Weiss, Antonio Torralba, and Rob Fergus. 2009. Spectral hashing. In *Advances in Neural Information Processing Systems*. 1753–1760.

[29] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li. 2017. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing* 26, 5 (2017), 2494–2507.

[30] Xing Xu, Jingkuan Song, Huimin Lu, Yang Yang, Fumin Shen, and Zi Huang. 2018. Modal-adversarial Semantic Learning Network for Extendable Cross-modal Retrieval. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. ACM, 46–54.

[31] Zhou Yu, Fei Wu, Yi Yang, Qi Tian, Jiebo Luo, and Yueting Zhuang. 2014. Discriminative coupled dictionary hashing for fast cross-media retrieval. In *Proceedings of the 37th international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 395–404.

[32] Deming Zhai, Hong Chang, Yi Zhen, Xianming Liu, Xilin Chen, and Wen Gao. 2013. Parametric Local Multimodal Hashing for Cross-View Similarity Search.. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 2754–2760.

[33] Dongqing Zhang and Wu-Jun Li. 2014. Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization.. In *Proceedings of AAAI Conference on Artificial Intelligence*, Vol. 1. 7.

[34] Dan Zhang, Fei Wang, and Luo Si. 2011. Composite hashing with multiple information sources. In *Proceedings of the 34th international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 225–234.

[35] Dell Zhang, Jun Wang, Deng Cai, and Jinsong Lu. 2010. Self-taught hashing for fast similarity search. In *Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 18–25.

[36] Peichao Zhang, Wei Zhang, Wu-Jun Li, and Minyi Guo. 2014. Supervised hashing with latent factor models. In *Proceedings of the 37th international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 173–182.

[37] Yi Zhen and Dit-Yan Yeung. 2012. Co-regularized hashing for multimodal data. In *Advances in Neural Information Processing Systems*. 1376–1384.

[38] Yi Zhen and Dit-Yan Yeung. 2012. A probabilistic model for multimodal hash function learning. In *Proceedings of the 18th international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 940–948.

[39] Jile Zhou, Guiguang Ding, and Yuchen Guo. 2014. Latent semantic sparse hashing for cross-modal similarity search. In *Proceedings of the 37th international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 415–424.

[40] Jile Zhou, Guiguang Ding, Yuchen Guo, Qiang Liu, and XinPeng Dong. 2014. Kernel-based supervised hashing for cross-view similarity search. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. IEEE, 1–6.

[41] Xiaofeng Zhu, Zi Huang, Heng Tao Shen, and Xin Zhao. 2013. Linear cross-modal hashing for efficient multimedia search. In *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 143–152.