# Adversarial Tri-Fusion Hashing Network for Imbalanced Cross-Modal Retrieval

Xin Liu, *Member, IEEE*, Yiu-ming Cheung, *Fellow, IEEE*, Zhikai Hu, Yi He, and Bineng Zhong

***Abstract*—Cross-modal retrieval has received increasing attentions for efficient retrieval across different modalities, and hashing technique has made significant progress recently due to its low storage cost and high query speed. However, most existing cross-modal hashing works still face the challenges of narrowing down the semantic gap between different modalities and training with imbalanced multi-modal data. This article presents an efficient Adversarial Tri-Fusion Hashing Network (ATFH-N) for cross-modal retrieval, which lies among the early attempts to incorporate adversarial learning for working with imbalanced multi-modal data. Specifically, a triple fusion network associated with zero padding operation is proposed to adapt either balanced or imbalanced multi-modal training data. At the same time, an adversarial training mechanism is leveraged to maximally bridge the semantic gap of the common representations between balanced and imbalanced data. Further, a label prediction network is utilized to guide the feature learning process and promote hash code learning, while additionally embedding the manifold structure to preserve both inter-modal and intra-modal similarities. Through the joint exploitation of the above, the underlying semantic structure of multimedia data can be well preserved in Hamming space, which can benefit various cross-modal retrieval tasks. Extensive experiments on three benchmark datasets show that the proposed ATFH-N method yields the comparable performance in balanced scenario and brings substantial improvements over the state-of-the-art methods in imbalanced scenarios.**

***Index Terms*—Cross-modal hashing, imbalanced multi-modal data, adversarial tri-fusion hashing, manifold structure.**

Manuscript received March 6, 2020; accepted June 24, 2020. This work was supported in part by the National Science Foundation of China under Grants 61673185, 61672444 and 61972167, in part by Quanzhou City Science & Technology Program of China under Grant 2018C107R, in part by the State Key Laboratory of Integrated Services Networks of Xidian University under Grant ISN20-11, in part by Hong Kong Baptist University, Research Committee, Initiation Grant-Faculty Niche Research Areas (IG-FNRA) 2018/19 under Grant RC-FNRA-IG/18-19/SCI/03, in part by the project funded by HKBU Interdisciplinary Research Clusters Matching Scheme under Grant RC-IRCMs/18-19/SCI/01, and in part by ITF of ITC of Hong Kong SAR under Project ITS/339/18. *(Corresponding author: Xin Liu.)*

Xin Liu is with the Department of Computer Science, Huaqiao University, Xiamen 361021, China, and with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China, and also with the Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong (e-mail: xliu@hqu.edu.cn).

Yiu-ming Cheung is with the Department of Computer Science and Institute of Research and Continuing Education, Hong Kong Baptist University, Kowloon, Hong Kong (e-mail: ymc@comp.hkbu.edu.hk).

Zhikai Hu is with the Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong (e-mail: cszkhu@comp.hkbu.edu.hk).

Yi He and Bineng Zhong are with the Fujian Key Laboratory of Big Data Intelligence and Security, Xiamen, China, and also with the Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Huaqiao University, Xiamen 361021, China (e-mail: yhe@hqu.edu.cn; bnzhong@hqu.edu.cn).

Digital Object Identifier 10.1109/TETCI.2020.3007143

## I. Introduction

THE last decade has witnessed unprecedented growth of multimedia data on the Internet, and such explosion has significantly increased the demand for more sophisticated multimedia retrieval technologies. In order to maximally benefit from the richness of multimedia data, cross-modal retrieval is becoming more appealing since it enables the similarity search across different modalities, e.g., a user can specify a query item of one modality to retrieve semantically relevant items of another modality. Nevertheless, the heterogeneous data of different modalities often reside in different feature spaces, and such heterogeneity has been widely considered as a great challenge to cross-modal retrieval.

In recent years, a great deal of research has been devoted to bridge the semantic gap between different modalities. Intuitively, a common way is to learn a shared latent subspace to minimize the heterogeneity such that the mapping features in this shared subspace can be directly measured [1], [2]. Nevertheless, these subspace methods are computationally inefficient to process a large scale of high dimensional multi-modal data. To tackle this problem, cross-modal hashing, favored for its low storage cost and high query speed, has attracted considerable attention for efficient cross-modal retrieval on a very large-scale multimedia data [3]–[5]. More specifically, it aims to learn a series of hash functions from the training set to map the heterogeneous multimedia data into a common Hamming space, whose main challenge is to learn the compact binary codes that can construct the underlying correlations between different modalities [6]. It is noted that most existing cross-modal hashing methods mainly focus on dealing with the balanced multi-modal data collections and highly depend on the pairwise relationships to explore the semantical correlation between them. Intuitively, they may not generalize well on a more practical cross-modal retrieval scenario, i.e., the heterogeneous data may be practically imbalanced (e.g., one text document describes multiple pictorial examples), and little attention has been paid to handle this challenging scenarios.

In practice, the numbers of relevant multimedia data from different modalities may vary considerably, and there always exist the imbalanced relationships among these multi-modal data. That is, the data items from different modalities are not always paired. Taking bimodal data (i.e., image and text) for illustration, their relationships can be further divided into four branches: 1) one-to-one balanced (i.e., paired) data (Fig. 1(a), where there is one-to-one correspondence between the data of two modalities; 2) one-to-many imbalanced data (Fig. 1(b), where there is only

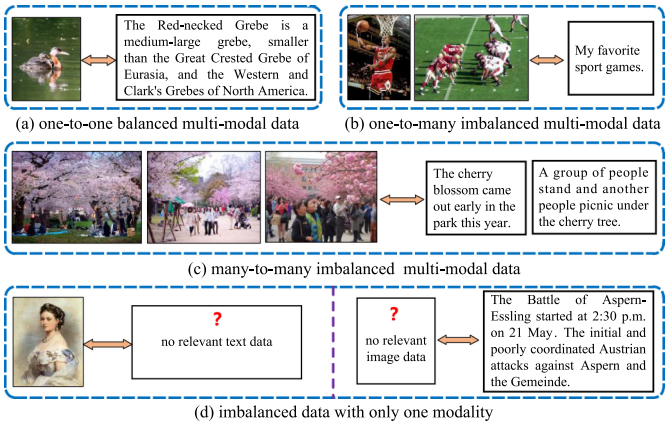



Fig. 1. Balanced and imbalanced relationships between the multi-modal data.

one example in one modality and multiple relevant samples in another modality; 3) many-to-many imbalanced data (Fig. 1(c)), where there are many-to-many correspondences between the data of two modalities, and their data numbers are different; 4) imbalanced data of only one modality (Fig. 1(d)), where the data is collected from only one modality and there is no corresponding data item in another modality. Since there is no relevant connection between the multi-modal data within the fourth case, this scenario is not generally considered for cross-modal analysis.

In the literature, most existing cross-modal hashing methods [7]–[11] mainly focus on dealing with the first kind of balanced multi-modal data, and these methods utilize the pairwise relationships to explore the semantic correlation across different modalities. Nevertheless, very few works [12], [13] have been designed to handle the imbalanced scenario. By class-wise pairing the data of one modality to match its corresponding data of another modality, one-to-many or even many-to-many imbalanced data may be potentially extended to a large number of one-to-one pairs. Nevertheless, this simple matching operation may mistakenly group the uncorrelated samples from heterogeneous modalities. For examples, some image examples belonging to only one semantic label may have significantly different appearances, e.g., an art image and an old building share the same semantic history, but they differ a lot in appearance. Accordingly, the textual description mismatched by this simple operation will fail to depict the correlated image example. For multi-label dataset, some imbalanced examples of one modality annotated with multiple labels may not have the exact one-to-one correspondence in another modality. Therefore, it is really difficult and even impossible to exactly correlate these imbalanced examples from a practical viewpoint. Therefore, the widespread existence of these complex multi-modal data has significantly increased the demand of more effective cross-modal hashing technologies to tackle these challenging scenarios.

In this paper, we address an efficient Adversarial Tri-Fusion Hashing (ATFH-N) Network for cross-modal retrieval, which seamlessly treats the balanced or imbalanced multi-modal data in an integrated way (extension of [14]). Specifically, a triple fusion network with sharing weights is proposed to adapt either balanced or imbalanced multi-modal training data, while an adversarial training mechanism is leveraged to maximally bridge the semantic gap of the common representations between balanced and imbalanced data. At the same time, a label prediction network is utilized to guide the feature learning process and promote hash code learning, while additionally embedding the inherent manifold structure to preserve both inter-modal and intra-modal similarities. Consequently, the semantic structure of multimedia data can be well preserved in Hamming space, and the derived hash codes are semantically meaningful for benefiting various cross-modal retrieval tasks. The major contributions are highlighted as follows:

- A novel adversarial tri-fusion hashing network is proposed to generalize cross-modal retrieval, which lies among the early attempts to incorporate adversarial learning for working with imbalanced multi-modal data.
- An efficient semantic preserving network associated with manifold embedding is proposed to guide the feature learning process and promote high-level semantic hash code learning, which can well narrow the semantic gap between heterogeneous data samples.
- Extensive experiments conducted on three benchmarks demonstrate the advantages of ATFH-N under various cross-modal retrieval tasks, and show its outstanding performance in both balanced and imbalanced datasets.

The remaining part of this paper is structured as follows: Section II surveys the existing cross-modal hashing methods, and Section III introduces the proposed ATFH-N algorithm in detail. The experimental results are provided in Section IV. Finally, we draw a conclusion in Section V.

## II. Related Work

Cross-modal hashing has received a lot of attention for similarity retrieval on large-scale datasets, and existing approaches roughly fall into training with balanced or imbalanced multi-modal data. This section surveys the related works concerning to these two different aspects.

Learning with balanced multi-modal data is an intuitive way to semantically correlate the data samples from the heterogeneous modalities, and different kinds of cross-modal hashing works have been proposed, either in unsupervised manner or supervised manner. The former approaches generally utilize the pairwise relationships to learn the binary codes by mapping the original feature space to Hamming space. Along this line, Inter-media Hashing (IMH) [12] obtains a common hamming space by preserving the inter-view and intra-view consistency, while Collective Matrix Factorization Hashing (CMFH) [9] jointly learns the common hash codes and modality-specific hash projection functions by collective matrix factorization. Similarly, Latent Semantic Sparse Hashing (LSSH) [15] utilizes sparse coding and matrix factorization to extract latent semantic features, while Fusion Similarity Hashing (FSH) [16] maintains the fusion similarity within the multiple modalities and learns the semantically correlated hash codes to represent heterogeneous data items. It is found that these methods are able to correlate the semantic relationships between heterogeneous modalities.

However, it should be noted that the binary hash codes produced in an unsupervised fashion are not discriminative enough such that their retrieval performances require further improvements.

With the label supervision, supervised cross-modal hashing methods utilize the semantic information to mitigate the semantic gap between heterogeneous modalities. For instance, Supervised Matrix Factorization Hashing (SMFH) [8] utilizes the label supervision to maintain the semantic consistency, while Co-Regularized Hashing (CRH) [17] learns hash functions of each bit sequentially to supervise the hash code learning. In addition, Semantic Correlation Maximization (SCM) [3] employs the semantic label information to correlate the heterogeneous modalities, while Semantic Preserved Hashing (SePH) [7] and Generalized Semantic Preserving Hashing (GSePH) [13] construct an affinity matrix in a supervised manner to approximate hash codes. Recently, Discrete Cross-modal Hashing (DCH) [18] and Discrete Latent Semantic Hashing (DLSH) [19] directly learn the hash codes while retaining the discrete constraints to learn more compact hash codes. Inspiring from the advancement of deep learning, deep-networks-based cross-modal hashing methods integrate feature learning and hash code learning into end-to-end trainable frameworks, which can handle the insufficient representation of the hand-crafted features more effectively. For instance, deep cross-modal hashing (DCMH) [20] utilizes an end-to-end deep learning framework to learn the feature representations and hash codes synchronously, while self-supervised adversarial hashing [21] proposes a self-supervised adversarial learning framework to promote the hash code learning, which can well narrow the semantic gap between the learned representations of different modalities. With label information embedding, these supervised methods can well bridge the heterogeneity gap between different modalities and achieve impressive performance. Nevertheless, there still exists a serious limitation for them. That is, these supervised methods only can deal with balanced multi-modal data.

Learning with imbalanced multi-modal data is more challenging because of its relatively complex data connections. To adapt imbalanced multi-modal data, Inter-media Hashing (IMH) [12] exploits two selective matrices to tackle imbalanced data and utilizes inter-view and intra-view consistency to generate the hash codes. However, this method generally ignores nonlinear structure embedded in real-world data, which often degrades its performance in practice. Differently, GSePH method [13] utilizes the semantic affinity matrix to learn the modality-specific hash codes of training instances, and designs a generalized hashing scheme to handle the unpaired multi-modal data. Although this approach is able to handle the imbalanced multi-modal data, it also fails to reduce the semantic gap between the feature vectors and hash codes, and its performance needs further improvement.

In a nutshell, on the one hand, most existing cross-modal hashing methods are mainly designed to deal with the balanced multi-modal data, but which may not be easily generalized to handle the imbalanced multi-modal data. On the other hand, the only generalized methods cannot well preserve the semantic consistency between the feature vector and hash code. Therefore, there is still a need to investigate a flexible and generalized cross-modal hash algorithm.

## III. Proposed Method

Without loss of generality, the proposed framework mainly studies on bimodal data for cross-modal hashing, and the proposed framework can be easily extended to other kinds of multi-modal data. Fig. 2 shows the schematic flows of the proposed cross-modal hashing framework, which mainly consists of two subnetworks: a triple fusion (tri-fusion) network to adapt different organizations of multi-modal training data, and followed by a semantic preserving network with manifold embedding to promote high-level semantic hash code learning.

Suppose the imbalanced multi-modal training data consists of image data $\mathbf{X}$ and text data $\mathbf{Y}$. More specifically, the image training dataset $\mathbf{X} = [\mathbf{X_p}, \mathbf{X_u}]$ contains two parts: image samples $\mathbf{X_p} \in \mathbb{R}^{n \times d_1}$ which have corresponding paired texts and image samples $\mathbf{X_u} \in \mathbb{R}^{n_1 \times d_1}$ without corresponding texts, where $n$ and $n_1$ respectively represent the sample numbers within these two parts and $d_1$ denotes the dimension of image feature. Similarly, the text training dataset $\mathbf{Y} = [\mathbf{Y_p}, \mathbf{Y_u}]$ also includes two parts: text samples $\mathbf{Y_p} \in \mathbb{R}^{n \times d_2}$ with paired images and text samples $\mathbf{Y_u} \in \mathbb{R}^{n_2 \times d_2}$ without corresponding images, where $n_2$ is the number of texts without corresponding images and $d_2$ is the dimension of text feature. The provided semantic labels of data $\mathbf{X}$ and $\mathbf{Y}$ are characterized by $\mathbf{L_x} \in \{0,1\}^{(n+n_1)\times c}$ and $\mathbf{L_y} \in \{0,1\}^{(n+n_2)\times c}$ respectively, where $c$ is the number of the semantic categories. The goal of the proposed cross-modal hashing method is to learn a common Hamming space for both balanced and imbalanced multi-modal data, and each instance can be represented as a compact binary vector $\mathbf{b} \in \{0,1\}^{1\times q}$, where $q$ represents the bit length of hash code. Evidently, the correlations between imbalanced multi-modal data are very complex and it is imperative to fill the data gap for efficient cross-modal analysis.

### A. Triple Fusion Network

For cross-modal retrieval, most existing methods [9], [21] often employ the two-stream structure to explore the common Hamming space for heterogeneous data representation. That is, as shown in Fig. 3(a), two separate learning branches are selected to map different modalities into Hamming space. It can be found that such two-stream learning network highly relies on the pairwise constraint, which are unsuitable for processing the imbalanced multi-modal data. In addition, the constraints imposed at the end of this model are the only tie between these two branches, which may not well narrow down the semantic gap between different modalities.

Motivated by the property of GAN [22], some approaches [23], [24] select the adversarial network to enhance the relationship between different modalities. However, these adversarial learning methods are not designed for cross-modal hashing works, while failing to handle the imbalanced training data. To tackle these problems, we propose an one-stream fusion structure to correlate the heterogeneous modalities, which can evolve the network structure implicitly to learn the semantically consistent feature representations. As shown in Fig. 3(b), two learning networks are twisted into a four-layer one-stream fusion network $\mathbf{G_{xy}}(\cdot;\theta_{\mathbf{xy}})$ $(d_1 + d_2 \to 1024 \to 512 \to q)$ to fuse the features from image and text modalities, where $\theta_{\mathbf{xy}}$ is the network

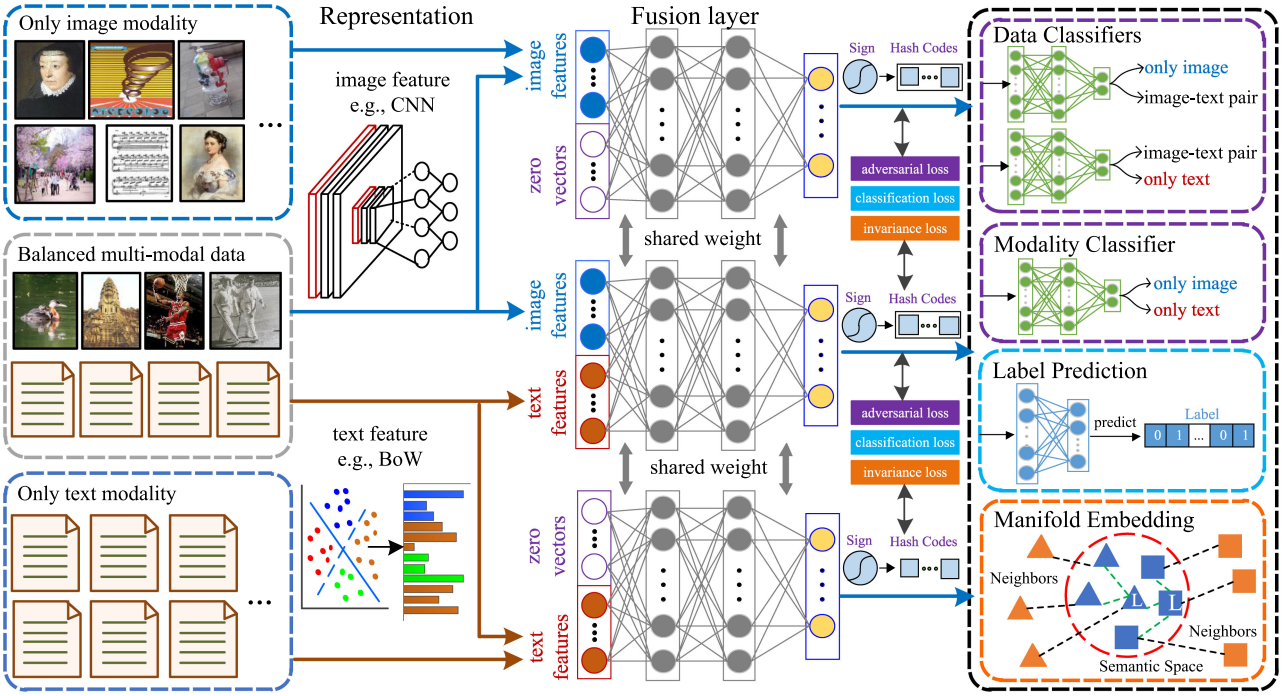



Fig. 2. The schematic pipeline of the proposed ATFH-N framework, which can handle both balanced and imbalanced multi-modal data.

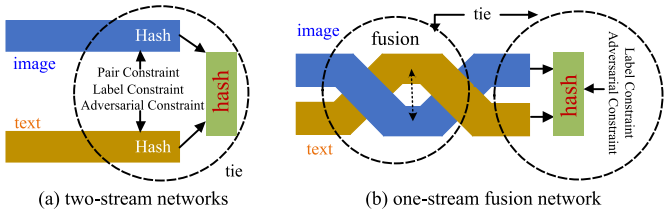



Fig. 3. The structure differences between the traditional two-stream learning network and the proposed one-stream fusion learning network.

parameters of $\mathbf{G_{xy}}$. Consequently, the relationships between heterogeneous modalities can be well correlated by using both the fusion network and other semantic constraints.

As shown in Fig. 1, the multi-modal dataset often consists of balanced and imbalanced instances. For each balanced image-text samples (i.e., paired) $\{\mathbf{X_p}, \mathbf{Y_p}\}$, we concatenate their features directly to form the fused feature set $\mathbf{Z_{xy}} \in \mathbb{R}^{n\times(d_1+d_2)}$:

$$\mathbf{Z}_{\mathbf{xy}}^{(t)} = [\mathbf{X}_{\mathbf{p}}^{(t)}, \mathbf{Y}_{\mathbf{p}}^{(t)}], \quad t = 1, 2, \ldots, n \tag{1}$$

where $\mathbf{Z}_{\mathbf{xy}}^{(t)}$, $\mathbf{X}_{\mathbf{p}}^{(t)}$ and $\mathbf{Y}_{\mathbf{p}}^{(t)}$ are the $t$-th instances of $\mathbf{Z_{xy}}$, $\mathbf{X_p}$ and $\mathbf{Y_p}$, respectively. Accordingly, $\mathbf{Z_{xy}}$ can be fed into network $\mathbf{G_{xy}}$ to get their high-level semantic representations.

For the imbalanced multi-modal dataset, some examples of one modality may have no corresponding data items in another modality. To solve this problem, another two networks $\mathbf{G_{ox}}(\cdot;\theta_{\mathbf{ox}})$ and $\mathbf{G_{oy}}(\cdot;\theta_{\mathbf{oy}})$ are exploited to individually process only image or text samples, where $\theta_{\mathbf{ox}}$ and $\theta_{\mathbf{oy}}$ respectively denote the parameters of these two networks. For semantic consistency mining, these two networks designed for dealing with imbalanced image and text examples share the same network structure with $\mathbf{G_{xy}}(\cdot;\theta_{\mathbf{xy}})$. Nevertheless, the learning structure of $\mathbf{G_{xy}}$ is a fusion network with balanced multi-modal features, and the imbalanced samples $\mathbf{X_u}$ and $\mathbf{Y_u}$ can not be fed into $\mathbf{G_{ox}}$ and $\mathbf{G_{oy}}$ directly. To handle this problem, a multi-modal zero padding operation is introduced to satisfy the input requirement of fusion network, and the detail of this operation is given below.

### B. Multi-Modal Zero Padding Mechanism

In some feature augmentation methods [25], [26], the utilization of zero-padding operation is proved to be effective for making the dimensions of the data from two domains become the same. For simplicity, $\mathbf{0}$ denotes an all-zero vector or a zero matrix. For the missing modalities, there features are padded with zeros to balance the demand of input dimension. That is, $\mathbf{X_{ox}}$ and $\mathbf{Y_{oy}}$ are respectively concatenated with $\mathbf{0}$ vector to form longer representations $\mathbf{Z_{ox}} \in \mathbb{R}^{n_1\times(d_1+d_2)}$ and $\mathbf{Z_{oy}} \in \mathbb{R}^{n_2\times(d_1+d_2)}$ as follows:

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

where $\mathbf{X}_{\mathbf{ox}}^{(i)}$ and $\mathbf{Y}_{\mathbf{ox}}^{(j)}$ are respectively the $i$-th and $j$-th instance within dataset $\mathbf{X_{ox}}$ and $\mathbf{Y_{oy}}$. In practice, both balanced and imbalanced examples can be considered as imbalanced data for enriching the training samples and bridging the semantic gap between them. Therefore, the imbalanced samples can be enlarged as $\bar{\mathbf{Z}}_{\mathbf{ox}} \in \mathbb{R}^{(n+n_1)\times(d_1+d_2)}$ and $\bar{\mathbf{Z}}_{\mathbf{oy}} \in \mathbb{R}^{(n+n_2)\times(d_1+d_2)}$:

$$\begin{aligned} \bar{\mathbf{Z}}_{\mathbf{ox}}^{(i)} &= [\mathbf{X}^{(i)}, \mathbf{0}^{1\times d_2}], \quad i = 1, 2, \ldots, n+n_1 \\ \bar{\mathbf{Z}}_{\mathbf{oy}}^{(j)} &= [\mathbf{0}^{1\times d_1}, \mathbf{Y}^{(j)}], \quad j = 1, 2, \ldots, n+n_2 \end{aligned} \tag{3}$$

where the representation of $\bar{\mathbf{Z}}_{\mathbf{ox}}$ and $\bar{\mathbf{Z}}_{\mathbf{oy}}$ can be well fed into networks $\mathbf{G}_{\mathbf{ox}}(\cdot;\theta_{\mathbf{ox}})$ and $\mathbf{G}_{\mathbf{oy}}(\cdot;\theta_{\mathbf{oy}})$, respectively. Inspired by the network architecture in [27], the shared parameters $\theta_{\mathbf{g}}$ are utilized in these triple networks, i.e., $\theta_{\mathbf{xy}} = \theta_{\mathbf{ox}} = \theta_{\mathbf{oy}} = \theta_{\mathbf{g}}$, and allow the back-propagation algorithm to update the model with regard to such tri-fusion network. For simplicity, such tri-fusion network is denoted as $\mathbf{G}(\cdot;\theta_{\mathbf{g}})$. By feeding the training data into this framework, the learned high-level semantic feature representation can be obtained as:

$$\mathbf{H}_{\mathbf{xy}} = \mathbf{G}(\mathbf{Z}_{\mathbf{xy}}),\ \mathbf{H}_{\mathbf{ox}} = \mathbf{G}(\mathbf{Z}_{\mathbf{ox}}),\ \mathbf{H}_{\mathbf{oy}} = \mathbf{G}(\mathbf{Z}_{\mathbf{oy}}). \quad (4)$$

Accordingly, the corresponding hash codes can be derived by:

$$\mathbf{B}_{\mathbf{xy}} = sign(\mathbf{H}_{\mathbf{xy}}),\ \mathbf{B}_{\mathbf{ox}} = sign(\mathbf{H}_{\mathbf{ox}}),\ \mathbf{B}_{\mathbf{oy}} = sign(\mathbf{H}_{\mathbf{oy}}) \quad (5)$$

Since the zero-padding mechanism is applied in both modalities and this type of method is also named as multi-modal zero-padding mechanism, which can make neural networks more flexible in processing the imbalanced multi-modal data. In addition, there often exist a dominant domain problem in multi-modal feature fusion strategy. That is, if the features of one modality are more semantically meaningful than that of another modality, the corresponding weights of networks assigned to the former modality would be greater than that allocated to the latter one. Consequently, the modality with larger weights shall become the dominant domain and its features play an important role in decision task, while the features of another modality make less sense to the final decision. Fortunately, the proposed multi-modal zero-padding scheme is able to well solve such dominant domain problem in heterogeneous feature fusion. The main reason lies that the zero-padding scheme is applied in both modalities, and the tri-fusion networks with shared parameters can well balance the importance of heterogeneous modalities. To be specific, if the weights allocated to image sample are much larger than that to text samples in $\mathbf{G}_{\mathbf{ox}}$, the embedding of data $\mathbf{Z}_{\mathbf{oy}}$ would disable the networks. That is because the weights belonging to image modality will multiply zero vector in $\mathbf{G}_{\mathbf{oy}}$, which makes less contribution to the fusion network. On the contrary, if the text modality becomes the dominant domain, the embedding of data $\mathbf{Z}_{\mathbf{ox}}$ will disable the networks as well. Therefore, the proposed triple fusion network associated with multi-modal zero-padding mechanism is able to automatically adjust the weight values allocated to the fused features and therefore balance the importance between different modalities. Consequently, the high-level representations with semantic consistency can be well obtained for heterogeneous data representation.

### C. Adversarial Learning Mechanism

*1) Data Classifiers:* The goal of the proposed framework is to learn the similar representations between the balanced and imbalanced data, whereby the semantic consistency between different modalities can be well maintained. To this end, two data classification networks $\mathbf{D}_1(\cdot;\theta_{\mathbf{d}_1})$ ($l\to64\to32\to2$) and $\mathbf{D}_2(\cdot;\theta_{\mathbf{d}_2})$ ($l\to64\to32\to2$) are defined to act as the discriminator, where $\theta_{\mathbf{d}_1}$ and $\theta_{\mathbf{d}_2}$ are the parameters of these two networks respectively. The former discriminator $\mathbf{D}_1(\cdot;\theta_{\mathbf{d}_1})$ is designed to discriminate the data of only image samples from the balanced image-text instances. Under such circumstances, the former data can be regarded as the fake samples while the latter data as real samples. Therefore, the following adversarial loss is obtained:

$$\mathcal{L}_{adv}^{(1)} = \sum_{\mathbf{h}\in\vec{\mathbf{H}}_{\mathbf{xy}}} \sum_{\mathbf{h}_1\in\vec{\mathbf{H}}_{\mathbf{ox}}} (log\mathbf{D}_1(\mathbf{h};\theta_{\mathbf{d}_1}) + log(1-\mathbf{D}_1(\mathbf{h}_1;\theta_{\mathbf{d}_1}))) \quad (6)$$

where $\vec{\mathbf{H}}_{\mathbf{xy}}$ and $\vec{\mathbf{H}}_{\mathbf{ox}}$ respectively denote all row vectors of matrices $\mathbf{H}_{\mathbf{xy}}$ and $\mathbf{H}_{\mathbf{ox}}$, i.e., $\vec{\mathbf{H}}_{\mathbf{xy}} = \{\mathbf{H}_{\mathbf{xy}}^{(i)} | i = 1, 2, \ldots, n\}$, $\vec{\mathbf{H}}_{\mathbf{ox}} = \{\mathbf{H}_{\mathbf{ox}}^{(j)} | j = 1, 2, \ldots, n_1\}$.

Further, $\mathbf{D}_2(\cdot;\theta_{\mathbf{d}_2})$ is utilized to discriminate the data of only text modality from the balanced image-text instances, where the former data is regarded as the fake samples while the latter is considered as real samples. Therefore, the following adversarial loss is obtained:

$$\mathcal{L}_{adv}^{(2)} = \sum_{\mathbf{h}\in\vec{\mathbf{H}}_{\mathbf{xy}}} \sum_{\mathbf{h}_2\in\vec{\mathbf{H}}_{\mathbf{oy}}} (log\mathbf{D}_2(\mathbf{h};\theta_{\mathbf{d}_2}) + log(1-\mathbf{D}_2(\mathbf{h}_2;\theta_{\mathbf{d}_2}))) \quad (7)$$

Since the networks $\mathbf{G}_{\mathbf{ox}}$ and $\mathbf{G}_{\mathbf{oy}}$ play an equal role to the feature learning, the overall adversarial loss can now be modeled as a combination of $\mathcal{L}_{adv}^{(1)}$ and $\mathcal{L}_{adv}^{(2)}$:

$$\mathcal{L}_{adv}^{\mathbf{D}} = \mathcal{L}_{adv}^{(1)} + \mathcal{L}_{adv}^{(2)} \quad (8)$$

Within this minimax game, the adversarial loss is utilized for narrowing the gap between balanced and imbalanced data.

*2) Modality Classifier:* Further, a modality discriminator $\mathbf{D}_3(\cdot;\theta_{\mathbf{d}_3})$ with parameters $\theta_{\mathbf{d}_3}$ is difined to discriminate the data of only image modality from the text modality, which also acts as an adversary. Therefore, the former image data can be regarded as the fake samples while the latter text data as real samples. Similarly, the following adversarial loss is obtained:

$$\mathcal{L}_{adv}^{\mathbf{M}} = \sum_{\mathbf{h}_1\in\vec{\mathbf{H}}_{\mathbf{ox}}} \sum_{\mathbf{h}_2\in\vec{\mathbf{H}}_{\mathbf{oy}}} (log\mathbf{D}_3(\mathbf{h}_1;\theta_{\mathbf{d}_3}) + log(1-\mathbf{D}_3(\mathbf{h}_2;\theta_{\mathbf{d}_3}))) \quad (9)$$

Within such minimax game, the adversarial loss is utilized for bridging the semantic gap between different modality.

### D. Label Prediction

The semantic label information not only can promote the hash code learning, but also could mitigate the semantic gap between heterogeneous modalities. In order to ensure that the discrimination in data representation is preserved after feature projection, it is reasonable to assume that the semantic categories can be directly predicted from the high-level feature representations. Accordingly, a three-layer classification network $\mathbf{C}(\cdot;\theta_{\mathbf{c}})$ ($p\to p/2\to c$) is utilized to predict the semantic label of each instance, where $\theta_{\mathbf{c}}$ is the network parameter. To main the semantic consistency, both balanced data and imbalanced data should share the similar representations under the supervision of data classifiers. Therefore, the long representations of data items $\mathbf{Z}_{\mathbf{xy}}$, $\mathbf{Z}_{\mathbf{ox}}$ and $\mathbf{Z}_{\mathbf{oy}}$ are all fed into the same classifier $\mathbf{C}(\cdot;\theta_{\mathbf{c}})$

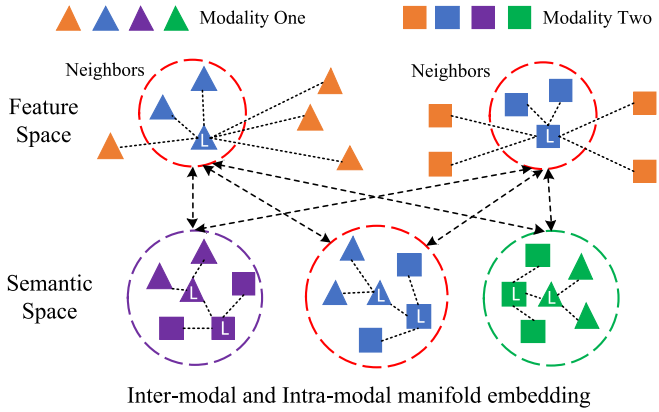



Fig. 4. Illustration of manifold embedding for hash code learning.

for semantic label prediction. Accordingly, the following classification loss is obtained for minimization:

$$\mathcal{L}_{class} = \sum_{\mathbf{h}\in\vec{\mathbf{H}}_{\mathbf{xy}}} ||\mathbf{C}(\mathbf{h};\theta_{\mathbf{c}}) - l_{\mathbf{h}}||_2^2 + \sum_{\mathbf{h}_1\in\vec{\mathbf{H}}_{\mathbf{ox}}} ||\mathbf{C}(\mathbf{h}_1;\theta_{\mathbf{c}}) - l_{\mathbf{h}_1}||_2^2 + \sum_{\mathbf{h}_2\in\vec{\mathbf{H}}_{\mathbf{oy}}} ||\mathbf{C}(\mathbf{h}_2;\theta_{\mathbf{c}}) - l_{\mathbf{h}_2}||_2^2 \quad (10)$$

where $l_{\mathbf{h}}$, $l_{\mathbf{h}_1}$ and $l_{\mathbf{h}_2}$ are the corresponding true labels of $\mathbf{h}$, $\mathbf{h}_1$ and $\mathbf{h}_2$, respectively.

### E. Manifold Embedding

Recent studies have demonstrated that it is beneficial for a retrieval model to exploit the manifold structure embedded in multimedia data, and many methods leverage the manifold learning to generate the hash codes. The goal of the proposed framework is to minimize the semantic gap among the high-level representations of all semantically similar samples from heterogeneous modalities. Accordingly, the intrinsic manifold structure residing in different modalities is able to promote the hash code learning. To embed the manifold structure into the binary codes learning process, two manifold regularizers that model the manifold structure of both the inter-modal similarity and inter-modal similarity are seamlessly integrated to the network, featuring on preserving the geometric structures of the training data in different modalities. More specifically, the preservation of inter-modal similarity ensures that the representations of multi-modal data with the same semantical category should be close to each other, while the preservation of intra-modal similarity is often utilized to maintain the neighborhood relationships within the training data points. To this end, as shown in Fig. 4, the balanced data is utilized to preserve the inter-modal similarity and the semantic labels are selected to construct the semantic affinity matrix $\mathbf{S}_{\mathbf{xy}}$:

$$\mathbf{S}_{\mathbf{xy}}^{(ij)} = \begin{cases} 1, & \text{if } \mathbf{X}_{\mathbf{p}}^{(i)} \text{ and } \mathbf{Y}_{\mathbf{p}}^{(j)} \text{ have the same category} \\ 0. & \text{otherwise} \end{cases} \quad (11)$$

where $\mathbf{X}_{\mathbf{p}}^{(i)}$ and $\mathbf{Y}_{\mathbf{p}}^{(j)}$ are respectively the $i$-th and $j$-th instance of $\mathbf{X}_{\mathbf{p}}$ and $\mathbf{Y}_{\mathbf{p}}$. Accordingly, the following objection function is obtained for minimization:

$$\begin{aligned} \mathcal{L}_{inter} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{S}_{\mathbf{xy}}^{(ij)} ||\mathbf{H}_{\mathbf{xy}}^{(i)} - \mathbf{H}_{\mathbf{xy}}^{(j)}||_2^2 \\ &= \mathrm{Tr}(\mathbf{H}_{\mathbf{xy}}(\mathbf{W}_{\mathbf{xy}} - \mathbf{S}_{\mathbf{xy}})\mathbf{H}_{\mathbf{xy}}^{\mathrm{T}}) \end{aligned} \quad (12)$$

where $\mathbf{H}_{\mathbf{xy}}^{(i)}$ denotes the $i$-th instance of $\mathbf{H}_{\mathbf{xy}}$, $\mathbf{W}_{\mathbf{xy}} \in \mathbb{R}^{n\times n}$ is a diagonal matrix whose entries are the column sum of $\mathbf{S}_{\mathbf{xy}}$, and $\mathrm{Tr}(\cdot)$ is the trace norm. Accordingly, Eq. (12) implicitly reflects the manifold structure of semantic embedding space.

For imbalanced data of only one modality, if two instances are close on the intrinsic data manifold, their semantic categories should be close as well. Under the manifold assumption, the geometric structure of one instance can be modeled by a nearest neighbor graph in the instance space. For imbalanced data $\mathbf{X}_{\mathbf{u}}$, the local similarity metric is utilized to model the intra-modal similarity:

$$\mathbf{S}_{\mathbf{ox}}^{(ij)} = \begin{cases} 1, & \text{if } \mathbf{X}_{\mathbf{u}}^{(i)} \in \mathbf{N}_k(\mathbf{X}_{\mathbf{u}}^{(j)}) \text{ or } \mathbf{X}_{\mathbf{u}}^{(j)} \in \mathbf{N}_k(\mathbf{X}_{\mathbf{u}}^{(i)}) \\ 0. & \text{otherwise} \end{cases} \quad (13)$$

where $\mathbf{N}_k(\cdot)$ is the top-$k$ nearest neighbor set. Similarly, for the imbalanced data $\mathbf{Y}_{\mathbf{u}}$, the local similarity metric is employed to model the intra-modal similarity:

$$\mathbf{S}_{\mathbf{oy}}^{(ij)} = \begin{cases} 1, & \text{if } \mathbf{Y}_{\mathbf{u}}^{(i)} \in \mathbf{N}_k(\mathbf{Y}_{\mathbf{u}}^{(j)}) \text{ or } \mathbf{Y}_{\mathbf{u}}^{(j)} \in \mathbf{N}_k(\mathbf{Y}_{\mathbf{u}}^{(i)}) \\ 0. & \text{otherwise} \end{cases} \quad (14)$$

It is noted that Eq. (13) and Eq. (14) implicitly reflect the manifold structure of feature embedding space. To preserve the manifold structure in mapping features, both balanced and imbalanced examples can be regarded as imbalanced data for enriching the training samples and bridging the semantic gap between them. Therefore, the whole data $\mathbf{X}$ and $\mathbf{Y}$ are respectively utilized to replace $\mathbf{X}_{\mathbf{u}}$ and $\mathbf{Y}_{\mathbf{u}}$ to construct $\mathbf{S}_{\mathbf{ox}}$ and $\mathbf{S}_{\mathbf{oy}}$, and the following objective function is obtained to preserve the manifold structure embedded in the data:

$$\begin{aligned} \mathcal{L}_{intra} &= \sum_{i=1}^{n+n_1}\sum_{j=1}^{n+n_1} \mathbf{S}_{\mathbf{ox}}^{(ij)} ||\mathbf{H}_{\mathbf{ox}}^{(i)} - \mathbf{H}_{\mathbf{ox}}^{(j)}||_2^2 \\ &\quad + \sum_{i=1}^{n+n_2}\sum_{j=1}^{n+n_2} \mathbf{S}_{\mathbf{oy}}^{(ij)} ||\mathbf{H}_{\mathbf{oy}}^{(i)} - \mathbf{H}_{\mathbf{oy}}^{(j)}||_2^2 \\ &= \mathrm{Tr}(\mathbf{H}_{\mathbf{ox}}(\mathbf{W}_{\mathbf{ox}} - \mathbf{S}_{\mathbf{ox}})\mathbf{H}_{\mathbf{ox}}^{\mathrm{T}}) \\ &\quad + \mathrm{Tr}(\mathbf{H}_{\mathbf{oy}}(\mathbf{W}_{\mathbf{oy}} - \mathbf{S}_{\mathbf{oy}})\mathbf{H}_{\mathbf{oy}}^{\mathrm{T}}) \end{aligned} \quad (15)$$

where $\mathbf{W}_{\mathbf{ox}} \in \mathbb{R}^{(n+n_1)\times(n+n_1)}$ is a diagonal matrix whose entities are the column sum of $\mathbf{S}_{\mathbf{ox}}$ and $\mathbf{W}_{\mathbf{oy}} \in \mathbb{R}^{(n+n_2)\times(n+n_2)}$ is a diagonal matrix whose entities are the column sum of $\mathbf{S}_{\mathbf{oy}}$. By minimizing Eq. (12) and Eq. (15), the networks can well preserve both inter-modal similarity and intra-modal similarity. For cross-modal retrieval task, the embedded inter-modal and intra-modal manifold structure should be exploited in an integrated way. That is, through the joint exploitation of the Eq. (12) and Eq. (15), the underlying manifold structures of multimedia data are well preserved in Hamming space.

### *F. Optimization*

By integrating the tri-fusion network $\mathbf{G}(\cdot;\theta_{\mathbf{g}})$, semantic prediction network $\mathbf{C}(\cdot;\theta_{\mathbf{c}})$, adversarial learning networks $\mathbf{D}_1(\cdot;\theta_{\mathbf{d}_1})$, $\mathbf{D}_1(\cdot;\theta_{\mathbf{d}_2})$ and $\mathbf{D}_1(\cdot;\theta_{\mathbf{d}_3})$, the process of learning the high-level representations and optimal hash codes can be conducted by jointly minimizing the label prediction loss, manifold embedding loss and adversarial loss. Since the optimization goal between the adversarial loss and other two embedding loss are different, the learning process runs as a minimax game between these concurrent sub-processes:

$$\mathcal{L}_{all} = \mathcal{L}_{class} + \beta\mathcal{L}_{inter} + \gamma\mathcal{L}_{intra} - \mu_1\mathcal{L}^{\mathbf{D}}_{adv} - \mu_2\mathcal{L}^{\mathbf{M}}_{adv} \tag{16}$$

where $\beta$, $\gamma$, $\mu_1$ and $\mu_2$ are balance parameters. Consequently, both the manifold structure and semantic information are seamlessly embedded into the final hash codes, and such optimization problem can be efficiently solved by an iterative updating scheme.

$$\theta_{\mathbf{g}} = \arg\min_{\theta_{\mathbf{g}}} \mathcal{L}_{all} \tag{17}$$

$$\theta_{\mathbf{c}} = \arg\min_{\theta_{\mathbf{c}}} \mathcal{L}_{all} = \arg\min_{\theta_{\mathbf{c}}} \mathcal{L}_{class} \tag{18}$$

$$(\theta_{\mathbf{d}_1}, \theta_{\mathbf{d}_2}) = \arg\max_{\theta_{\mathbf{d}_1},\theta_{\mathbf{d}_2}} \mathcal{L}_{all} = \arg\min_{\theta_{\mathbf{d}_1},\theta_{\mathbf{d}_2}} \mathcal{L}^{\mathbf{D}}_{adv} \tag{19}$$

$$\theta_{\mathbf{d}_3} = \arg\max_{\theta_{\mathbf{d}_3}} \mathcal{L}_{all} = \arg\min_{\theta_{d_3}} \mathcal{L}^{\mathbf{M}}_{adv} \tag{20}$$

This minimax game can be efficiently implemented using a stochastic gradient descent (SGD) optimization solver, e.g., Adam optimizer [28] (the default value of $\alpha = 0.001$ is utilized for all experiments), and these optimization problems can be iteratively solved until the convergence is reached. Let $\alpha$ be the learning rate in Adam optimizer, the optimal parameters can be well obtained via Algorithm 1. It is noted that the adversarial loss, label prediction loss and manifold embedding loss are integrated in Hamming space to guide the high-level feature learning process, and the proposed optimization process can well preserve the semantic consistency between continues features and discrete hash codes. As a result, in Eq. (5) the quantitation errors resulted by performing sign function between features and hash codes are not significant, and the derived hash codes are semantically meaningful for benefiting various cross-modal retrieval tasks.

**Algorithm 1:** Optimization Pseudocode for ATFH-N.

**input:** Image data $\mathbf{X}$ with labels $\mathbf{L_x}$ and text data $\mathbf{Y}$ with labels $\mathbf{L_y}$, parameters $\beta$, $\gamma$, $\mu_1$ and $\mu_2$;
**output:** Network parameters $\theta_{\mathbf{g}}$, $\theta_{\mathbf{c}}$, $\theta_{\mathbf{d}_1}$, $\theta_{\mathbf{d}_2}$, $\theta_{\mathbf{d}_3}$

Initialize $\theta_{\mathbf{g}}$, $\theta_{\mathbf{c}}$, $\theta_{\mathbf{d}_1}$, $\theta_{\mathbf{d}_2}$, $\theta_{\mathbf{d}_3}$ with random values;
Extract representative image features and text features;
**repeat**
  update $\theta_{\mathbf{t}}$ as: $\theta_{\mathbf{t}} \leftarrow \theta_{\mathbf{d}_1} - \alpha\frac{\partial\mathcal{L}^{(1)}_{adv}}{\partial\theta_{\mathbf{t}}}$;
  update $\theta_{\mathbf{c}}$ as: $\theta_{\mathbf{c}} \leftarrow \theta_{\mathbf{c}} - \alpha\frac{\partial\mathcal{L}_{class}}{\partial\theta_{\mathbf{c}}}$;
  update $\theta_{\mathbf{d}_1}$ as $\theta_{\mathbf{d}_1} \leftarrow \theta_{\mathbf{d}_1} - \alpha\frac{\partial\mathcal{L}^{(1)}_{adv}}{\partial\theta_{\mathbf{d}_1}}$;
  update $\theta_{\mathbf{d}_2}$ as: $\theta_{\mathbf{d}_2} \leftarrow \theta_{\mathbf{d}_2} - \alpha\frac{\partial\mathcal{L}^{(2)}_{adv}}{\partial\theta_{\mathbf{d}_2}}$;;
  update $\theta_{\mathbf{d}_3}$ as $\theta_{\mathbf{d}_3} \leftarrow \theta_{\mathbf{d}_3} - \alpha\frac{\partial\mathcal{L}^{\mathbf{M}}_{adv}}{\partial\theta_{\mathbf{d}_3}}$;
**until** (convergency or reaching maximum iterations)
**return** $\theta_{\mathbf{g}}$, $\theta_{\mathbf{c}}$, $\theta_{\mathbf{d}_1}$, $\theta_{\mathbf{d}_2}$, $\theta_{\mathbf{d}_3}$.

## IV. Experiment

This section conducts a series of quantitative experiments to verify the robustness of the proposed framework on both balanced and imbalanced multi-modal datasets.

### *A. Datasets and Protocol*

Three public available multi-modal datasets, i.e., Wiki, MIR-Flickr and NUS-WIDE, are chosen in the experiments, and their main descriptions are briefly described as follows:

**Wiki** [29] consists of 2,866 image-text pairs from ten classes, where each image is represented by 128-d SIFT descriptor and each text by a 10-dimensional Latent Dirichlet Allocation (LDA) vector. The whole dataset is divided into a training set of 2,173 instances and a test set of 693 instances.

**MIRFlickr** [30] includes 25,000 image-text pairs from Flickr website, and each instance is annotated with at least one of 24 categories. Specifically, the samples without labels or textual tags appearing less than 20 times are removed. Each image is characterized by a 150-dimensional edge histogram descriptor while the text is annotated by a 500-dimensional tagging vector. We randomly choose 5% of the data pairs as the query set and the remaining parts as the training set.

**NUS-WIDE-100 k** [31] consists of 269,548 image-text pairs from 81 concepts. As a large part of concepts contain little samples, we randomly select 100,000 labeled image-text pairs from the top 10 most frequent concepts for evaluation. Specifically, every image is described by a 500-dimensional SIFT feature vector and each text by a 1000-dimensional BoW feature vector. We randomly select 5% of the pairs as the query set and the remaining pairs as the training set.

It should be noted that the deep cross-modal hashing works learn the high-level feature representations and hash codes together [20], [21], and the proposed framework is significantly different from those works. In that sense, it is inappropriate to perform a relatively fair and meaningful comparison with these deep approaches. In particular, we compare the proposed ATFH-N method with state-of-the-art competing methods, i.e., CCA [1], CMFH [9], SCM [3], SePH [7], SMFH [8] and GSePH [13]. For most baselines, the source codes kindly provided by respective authors are selected for implementation, and the training samples are initialized as the same as the description of dataset. Since SMFH and SePH are very computationally intensive, it is impossible for these two methods to perform the training process on very large datasets. Therefore, as suggested in their original works [8], [13], we randomly choose a subset of 5000 instances, respectively from the larger MIRFlickr and NUS-WIDE-100 k datasets, to form the training sets. Meanwhile, the popular mean average precision (mAP) [9] score is selected as evaluation metric to validate the cross-modal

TABLE I
QUANTITATIVE EVALUATIONS OF BALANCED SCENARIO ON DIFFERENT DATASETS, AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

| Task | Method | Wiki | | | | MIRFlickr | | | | NUS-WIDE-100k | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 16 | 32 | 64 | 128 | 16 | 32 | 64 | 128 | 16 | 32 | 64 | 128 |
| I→T | CCA [1] | 0.1743 | 0.1645 | 0.1584 | 0.1584 | 0.5819 | 0.5756 | 0.5710 | 0.5679 | 0.3848 | 0.3713 | 0.3607 | 0.3536 |
| | CMFH [9] | 0.2172 | 0.2231 | 0.2316 | 0.2395 | 0.5683 | 0.5684 | 0.5687 | 0.5693 | 0.3428 | 0.3434 | 0.3433 | 0.3432 |
| | SMFH [8] | 0.2698 | 0.2900 | 0.2929 | 0.3009 | 0.5913 | 0.5997 | 0.5956 | 0.5986 | 0.3612 | 0.3613 | 0.3628 | 0.3635 |
| | FSH [16] | 0.2235 | 0.2316 | 0.2408 | 0.2474 | 0.5893 | 0.6027 | 0.6006 | 0.6022 | 0.4927 | 0.4986 | 0.5015 | 0.5057 |
| | SCM [3] | 0.2341 | 0.2410 | 0.2445 | 0.2569 | 0.6280 | 0.6345 | 0.6385 | 0.6490 | 0.5275 | 0.5414 | 0.5481 | 0.5498 |
| | SePH [7] | 0.2770 | 0.2964 | 0.3153 | 0.3138 | 0.6736 | 0.6789 | 0.6822 | 0.6851 | 0.5381 | 0.5517 | 0.5556 | 0.5654 |
| | GSePH [13] | 0.2778 | 0.2882 | 0.3044 | 0.3040 | 0.6460 | 0.6649 | 0.6725 | 0.6835 | 0.5018 | 0.5370 | 0.5595 | 0.5715 |
| | ATFH-N | **0.3166** | **0.3209** | **0.3326** | **0.3314** | **0.7337** | **0.7479** | **0.7328** | **0.7216** | **0.6151** | **0.6114** | **0.6196** | **0.6054** |
| T→I | CCA [1] | 0.1611 | 0.1436 | 0.1293 | 0.1233 | 0.5803 | 0.5750 | 0.5708 | 0.5679 | 0.3811 | 0.3687 | 0.3592 | 0.3572 |
| | CMFH [9] | 0.4902 | 0.5077 | 0.5173 | 0.5348 | 0.5646 | 0.5652 | 0.5649 | 0.5653 | 0.3464 | 0.3472 | 0.3473 | 0.3474 |
| | SMFH [8] | 0.6085 | 0.6274 | 0.6308 | 0.6445 | 0.5890 | 0.5909 | 0.5915 | 0.5954 | 0.3524 | 0.3524 | 0.3529 | 0.3538 |
| | FSH [16] | 0.4805 | 0.4804 | 0.5127 | 0.5182 | 0.5865 | 0.5970 | 0.5965 | 0.5969 | 0.4751 | 0.4785 | 0.4822 | 0.4879 |
| | SCM [3] | 0.2257 | 0.2459 | 0.2494 | 0.2535 | 0.6176 | 0.6234 | 0.6285 | 0.6369 | 0.4952 | 0.5076 | 0.5157 | 0.5174 |
| | SePH [7] | 0.6402 | 0.6543 | 0.6585 | 0.6674 | 0.7313 | 0.7320 | 0.7381 | 0.7442 | 0.6310 | 0.6546 | 0.6628 | 0.6702 |
| | GSePH [13] | 0.6445 | 0.6639 | 0.6683 | 0.6755 | 0.6663 | 0.7113 | 0.7269 | 0.7441 | 0.5595 | 0.6379 | 0.6593 | 0.6764 |
| | ATFH-N | **0.6957** | **0.6945** | **0.7030** | **0.7014** | **0.7903** | **0.8030** | **0.7943** | **0.7681** | **0.6951** | **0.7135** | **0.7218** | **0.6992** |

performance, indexing the relevant text samples by given image query (I→T) and vice versa (T→I). In the experiments, $\beta = 0.1$, $\gamma = 0.3$, $\mu_1 = 2$ and $\mu_2 = 0.1$, are empirically set for implementation, while top-25, top-10 and top-25 nearest neighbors are chosen for Wiki, MIRFlickr and NUS-WIDE-100 k datasets, respectively. In the experiment, we perform five runs for each algorithm and take the average performance for illustration.

### *B. Results of Balanced Scenario*

For the balanced scenario, we set $n_1 = n_2 = 0$ to equalize the training numbers of both modalities and utilize all the paired dataset for testing. That is, each data instance of one modality has the corresponding paired data item in another modality. Accordingly, the mAP scores are recorded on all three benchmark datasets and the baseline methods, CCA [1], CMFH [9], SMFH [8], SCM [3], FSH [16], SePH [7], and GSePH [13] are selected for comparison. The I→T and T→I retrieval performances tested with different hash lengths (i.e., 16 bits, 32 bits, 64 bits and 128 bits) are shown in Table. I, it can be observed that the proposed ATFH-N method has achieved very competitive cross-modal retrieval performances in different hash length settings, and generally performed better than the selected baselines.

For the small Wiki dataset, each instance is annotated with a single label and some examples sharing the same semantic category may have significantly diverse feature representations. For instance, an artist image and a building image share the same semantic category 'art,' but there appearances are totally different. Consequently, most existing cross-modal hashing methods often degrade their retrieval performance to some degree. For instance, SePH [7] and GSePH [13] utilize the semantic affinity matrix to produce the hash codes, and the derived hash codes are not discriminative enough for measuring the cross-modal similarity. For instance, SePH [7] and GSePH [13] have yielded a bit lower mAP score in I→T task (i.e., 16 and 32 bits). In contrast to this, the proposed ATFH-N framework has yielded very competitive cross-modal retrieval performance when tested on Wiki dataset, and mAP scores obtained from 64 and 128 bits respectively reach up to 0.7030 and 0.7014 when tested on T→I task.

For the larger MIRFlickr and NUS-WIDE-100 k, each instance is annotated with multiple labels. Although each instance can be well described by the semantic labels, the hash codes learned only from such supervised information often fail to correlate the heterogeneous samples. In contrast to this, the hash codes derived from the proposed ATFH-N approach are more semantically meaningful than those generated from SePH and GSePH. As a result, the proposed ATFH-N method has yielded the best retrieval performance on the larger datasets. For instance, when the hash length is set at 64, the mAP values obtained by baseline approaches are respectively less than 0.73 and 0.66, respectively tested on the MIRFlickr and NUS-WIDE-100 k datasets, in T→I task. By contrast, the mAP scores obtained by the proposed ATFH-N approach are higher than 0.79 and 0.72, respectively evaluated on the MIRFlickr and NUS-WIDE-100 k datasets. Meanwhile, the precision-recall curves are recorded in Fig. 5, it can be observed that the proposed ATFH-N approach always yields the highest precision scores than those baselines under the similar recall values. This indicates that the proposed ATFH-N framework has strong ability to return much more similar samples in the retrieval results, which plays an important role for a practical retrieval system. The main superiorities contributed to these very competitive performances are two-fold: 1) ATFH-N learns the common embedding from the fused features of two modalities, whereby the semantic consistency between heterogeneous modalities can be well exploited. 2) The joint exploitation of label prediction module and manifold embedding module is able to well guide the high-level feature learning process and promote compact hash code learning, which can well preserve both the inter-modal and intra-modal similarities. Consequently, the hash codes learned by the proposed ATFH-N framework are more semantically meaningful for efficient cross-modal retrieval.

Further, it has been demonstrated that the visual features obtained from the pretrained or fine-tuned CNN models have

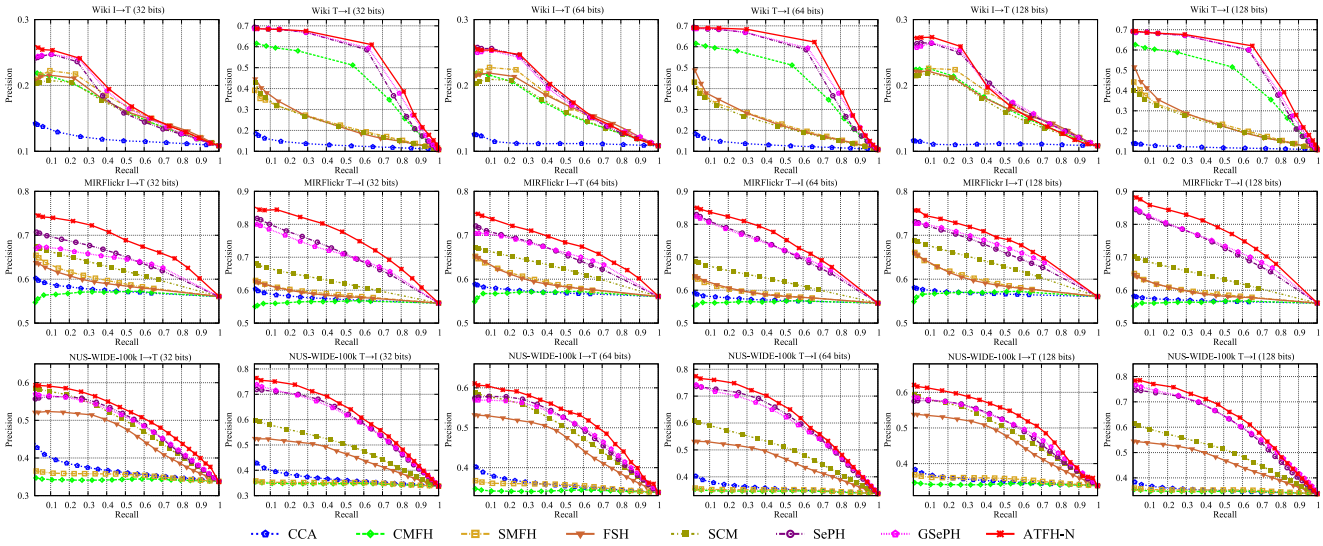



Fig. 5. Precision-recall curves generated by different approaches and tested with different code lengths (32, 64 and 128 bits).

TABLE II
RESULTS (MAP) OF CROSS-MODAL RETRIEVAL ON CNN VISUAL FEATURES, AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

| Method | Wiki dataset with CNN visual features | | | | | |
|---|---|---|---|---|---|---|
| | 32 bits | | | 128 bits | | |
| | I→T | T→I | Average | I→T | T→I | Average |
| CMFH [9] | 0.184 | 0.265 | 0.449 | 0.187 | 0.325 | 0.256 |
| SePH [7] | 0.476 | 0.734 | 0.605 | 0.52 | 0.774 | 0.647 |
| GSePH [13] | 0.494 | 0.762 | 0.628 | 0.508 | 0.777 | 0.643 |
| ASPH-N | **0.514** | **0.798** | **0.656** | **0.512** | **0.799** | **0.656** |

demonstrated to be effective for cross-modal retrieval [32], and the improved performance can be obtained by traditional cross-modal retrieval methods. Accordingly, we further extract the 4096-d CNN visual features from the last fully connected layer by VGG19 model, and compare the proposed ATFH-N method with state-of-the-art competing methods. The representative cross-modal retrieval performances evaluated on the fine-tuned CNN visual features and tested on Wiki dataset are shown in Table II, it can be found that the competing SePH [7], GSePH [13]) methods and the proposed ATFH-N approach have yielded the better retrieval performances than the results produced by hand-craft features. Remarkably, the proposed ATFH-N method with CNN visual features often boosts the retrieval performances in different hash length settings, and significantly outperforms the competing state-of-the-art baselines. That is, the proposed learning framework is adaptive to various kinds of visual features, and the experimental results have shown its outstanding performances.

### C. Evaluation of Imbalanced Scenario

As discussed in Section I, the imbalanced multi-modal samples are public available, where the data of two modalities does not exist one-to-one correspondence. For instance, e.g., ten images and five text documents share the same semantic tag 'history'. In the literature, except for IMH [12] and GSePH [13], most existing cross-modal hashing algorithms developed for balanced multi-modal collections are not applicable to handle this imbalanced scenario. Fortunately, the proposed ATFH-N method is able to well handle such imbalanced data collections. Similar to [13], we keep the text dataset unchanged and randomly choose 90% of images as 'imbalanced-1' and vice verse as 'imbalanced-2'. Specifically, the training set itself serves as the retrieval set while the query set is kept unchanged as in the balanced cases. It is noted that GSePH cannot handle all the training samples of a larger dataset. Therefore, Wiki and MIRFlickr datasets are selected for evaluations. To maintain the consistency with balanced scenario, mAP@all is selected to evaluate the imbalanced cross-modal retrieval performance.

The cross-modal retrieval performances tested on imbalanced datasets are shown in Table III, it can be observed that IMH and GSePH methods have delivered relatively lower mAP scores, for reason that the correlation between the imbalanced data is relatively complex. Specifically, IMH generally ignores the nonlinear structure embedded in real-world data, which often degrades its performance in practice. Although the performance delivered by GSePH is much better than that obtained by IMH, there still exist a huge gap between the performance within the balanced and imbalanced scenarios, especially for T→I task tested on Wiki dataset. For instance, the mAP scores obtained by GSePH methods are only equal to 0.438 and 0.456, respectively, tested on 32 and 64 hash bits, which are significantly lower than that obtained within the balanced case. The main reason lies that GSePH utilizes two-stage learning schemes to produce the hash codes, which may result a bit large semantic gap between the feature vectors and hash codes. Accordingly, its performance is a bit poor. Comparatively speaking, our proposed ATFH-N method delivers almost the similar retrieval performances with the balanced scenario and considerably outperforms these two baselines. For T→I task, the mAP scores obtained by the the proposed ATFH-N approach and tested on MIRFlickr datasets reach up to 0.789 and 0.783, respectively evaluated on 'imbalanced-1'

TABLE III
QUANTITATIVE EVALUATIONS OF IMBALANCED SCENARIO ON WIKI AND MIRFLICKR DATASETS. BEST RESULTS ARE MARKED IN BOLD

| Method | | Wiki dataset | | | | | | MIRFlickr dataset | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | imbalanced-1 | | | imbalanced-2 | | | imbalanced-1 | | | imbalanced-2 | | |
| | | I→T | T→I | average | I→T | T→I | average | I→T | T→I | average | I→T | T→I | average |
| IMH [12] | | 0.176 | 0.156 | 0.166 | 0.178 | 0.154 | 0.166 | 0.581 | 0.579 | 0.580 | 0.581 | 0.579 | 0.580 |
| GSePH [13] | 16 | 0.257 | 0.453 | 0.355 | 0.268 | 0.422 | 0.345 | 0.651 | 0.631 | 0.641 | 0.653 | 0.645 | 0.649 |
| | 32 | 0.273 | 0.477 | 0.375 | 0.279 | 0.438 | 0.358 | 0.648 | 0.633 | 0.641 | 0.658 | 0.635 | 0.647 |
| | 64 | 0.283 | 0.483 | 0.383 | 0.298 | 0.456 | 0.377 | 0.665 | 0.665 | 0.665 | 0.675 | 0.663 | 0.669 |
| ATFH-N | 16 | 0.308 | 0.677 | 0.493 | 0.314 | 0.689 | 0.502 | 0.732 | 0.776 | 0.754 | 0.735 | 0.774 | 0.755 |
| | 32 | 0.315 | 0.695 | 0.505 | 0.318 | 0.691 | 0.505 | **0.743** | **0.789** | **0.766** | **0.749** | **0.783** | **0.766** |
| | 64 | **0.328** | **0.702** | **0.515** | **0.334** | **0.708** | **0.521** | 0.731 | 0.788 | 0.759 | 0.721 | 0.771 | 0.746 |

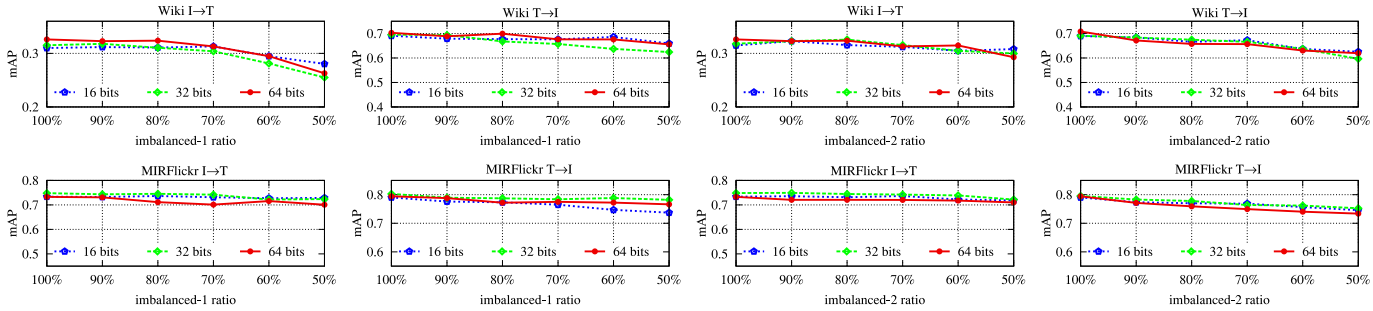



Fig. 6. Retrieval performances obtained by ATFH-N method and tested on Wiki dataset with different ratios of balanced data.

and 'imbalanced-2' cases. That is, our proposed ATFH-N method can not only handle the imbalanced multi-modal data collections, but also could produce relatively stable retrieval performance on different retrieval tasks.

In above imbalanced experiments, 10% data of one modality are discarded and data imbalances between different modalities are slight. Evidently, it is difficult to correlate the multi-modal data if the imbalanced ratio between heterogeneous data is very large. To show the flexibility of the proposed framework to process the imbalanced multi-modal data, we increase the ratio of imbalanced data and further evaluate the effectiveness of the proposed framework. The representative results with ratio ranging from 50% to 100% are presented in Fig. 6, it can be observed that the mAP scores obtained by the proposed ATFH-N method slightly decline as the ratio of imbalanced data increase, this is because the lack of balanced data naturally makes it difficult to correlate the heterogeneous data points from different modalities. Fortunately, the mAP scores obtained by ATFH-N does not substantially change even if the ratios between the imbalanced multi-modal data are very large. For instance, the mAP scores obtained by ATFH-N almost remain the same with the increase of imbalanced-2 ratio. That is, our proposed ATFH-N method has achieved very stable performance on various imbalanced data collections, and it indicates that the proposed ATFH-N method is effective to handle the imbalanced retrieval task.

### D. Evaluation of Different Pairwise Constraints

It is noted that most cross-modal hashing methods highly depend on the pairwise relationships to explore the correlation between multi-modal data. That is, there is one-to-one correspondence between the data of two modalities. In practice, the data from the heterogeneous modalities may be collected with none pairwise constraint, and this kind of data can be considered as imbalanced data as well. Fortunately, the proposed ATFH-N method is able to handle different kinds of multi-modal data. To verify the flexibility of the proposed ATFH-N framework, we maintain the number of training data unchanged, and gradually reduce the percentage of pairwise relationship ($\frac{n}{n+n_1}$ and $\frac{n}{n+n_2}$). More specifically, the whole dataset is selected as the training dataset and some pairwise relationships (e.g., 10%) are randomly removed to form the new training dataset. Accordingly, the data without pairwise relationship can be reasonably regarded as imbalanced data.

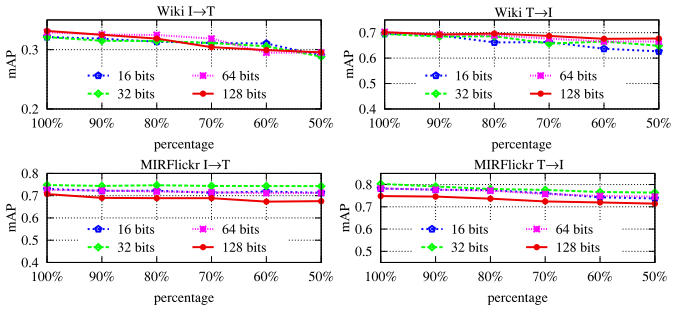



Fig. 7. Evaluation of pairwise constraint with different percentages.

The representative cross-modal retrieval performances evaluated on different pairwise constraints are shown in Fig. 7, it can be observed that the mAP scores drop only a little with the decreasing of pairwise constraint. It is noted that such decline is interpretable because the pairwise constraint is of crucial importance to the correlation mining between the multi-modal

TABLE IV
PERFORMANCE OF ATFH-N UNDER DIFFERENT LEARNING COMBINATIONS (HASH LENGTH: 64)

| Method | Wiki dataset | | | | | |
|---|---|---|---|---|---|---|
| | 32 bits | | | 64 bits | | |
| | I→T | T→I | avg | I→T | T→I | avg |
| N-L.+ATFH-N | 0.1233 | 0.1491 | 0.1362 | 0.1249 | 0.1283 | 0.1266 |
| N-A.+ATFH-N | 0.1891 | 0.4216 | 0.3054 | 0.1993 | 0.4251 | 0.3122 |
| N-$A^{DC}$.+ATFH-N | 0.1996 | 0.4632 | 0.3314 | 0.2012 | 0.4718 | 0.3365 |
| N-$A^{MC}$.+ATFH-N | 0.2015 | 0.4923 | 0.3469 | 0.2032 | 0.4953 | 0.3493 |
| N-M.+ATFH-N | 0.2695 | 0.6431 | 0.4563 | 0.2687 | 0.6486 | 0.4587 |
| N-$M^{inter}$.+ATFH-N | 0.2886 | 0.6443 | 0.4664 | 0.3089 | 0.6553 | 0.4821 |
| N-$M^{intra}$.+ATFH-N | 0.2817 | 0.6503 | 0.4660 | 0.2878 | 0.6678 | 0.4778 |
| Ours | **0.3209** | **0.6945** | **0.5077** | **0.3226** | **0.7030** | **0.5128** |

data. Fortunately, the retrieval performance has no substantial changes, and the mAP scores almost resides in the same interval even a large part of pairwise relationships are removed. For instance, if the percentage of pairwise constraint is set at 70%, the proposed ATFH-N method has produced a relatively stable I→T performance when tested on MIRFlickr dataset, and the mAP values derived from different hash lengths do not change significantly. Further, as shown in Table. I, it is noted that the proposed ATFH-N method with less pairwise relationships has delivered the comparable or even better retrieval performances in comparison with the competing baselines. That is, the proposed approach is insensitive to the pairwise relationships among the multi-modal data, which is adaptive to more challenging multi-modal datasets.

### *E. Ablation Studies and Parameter Analysis*

Within the proposed framework, the adversarial learning mechanism, label prediction module and manifold embedding module, are efficiently considered for achieving cross-modal hashing. Next, we further evaluate the effectiveness of each learning module and validate the performance of different learning combinations: 1) Without label prediction module, the adversarial learning mechanism and manifold embedding module are integrated to achieve cross-modal retrieval (abbreviated as N-L.+ATFH-N). 2) Without adversarial learning module, the label prediction module and manifold embedding module are integrated to achieve cross-modal retrieval (abbreviated as N-A.+ATFH-N). Meanwhile, we also evaluate the cross-modal hashing performance by respectively removing the data classifiers (abbreviated as N-$A^{DC}$.+ATFH-N) and the modality classifier (abbreviated as N-$A^{MC}$.+ATFH-N). 3) Without manifold embedding module, the adversarial learning mechanism and label prediction module are integrated to realize cross-modal retrieval (abbreviated as N-M.+ATFH-N). Meanwhile, we also report the cross-modal retrieval performance by respectively removing the inter-modal manifold embedding (abbreviated as N-$M^{inter}$.+ATFH-N) and intra-modal manifold embedding (abbreviated as N-$M^{intra}$.+ATFH-N). Accordingly, the mAP scores are recorded to validate these different learning combinations.

Table IV displays the cross-modal retrieval performances obtained by different learning combinations. It can be found that the label prediction module plays an important role for cross-modal retrieval tasks. For instance, the mAP scores of both I→T and T→I tasks derived from the N-L.+ATFH-N and tested with different hash lengths are lower than 0.2, which lead to very poor retrieval performance. In contrast to this, the mAP scores of T→I task derived from N-M.+ATFH-N reach up to 0.6431 and 0.6486, respectively tested on 32 and 64 hash bits. Meanwhile, adversarial learning mechanism also serves an important role within the proposed framework, and the adversarial loss generated by data classifiers or modality classifier often contributes to a higher retrieval performance than the model without adversarial learning module. For example, the average mAP scores derived from N-A.+ATFH-N, N-A.+ATFH-N and are lower than 0.5.

In addition, the utilization of manifold embedding module incorporating the inter-modal or intra-modal preservation also improves the cross-modal performance than that obtained by the non-manifold embedding module. For instance, the mAP scores of T→I task obtained by N-$M^{intra}$.+ATFH-N reach up to 0.6678, which is higher than that obtained from N-M.+ATFH-N. Importantly, the proposed framework performs better in different cross-modal retrieval tasks and generally outperforms these different learning combinations. For instance, the T→I retrieval scores (i.e., mAP values) obtained by the proposed method reach up to 0.6945 and 0.7030, respectively tested on 32 and 64 hash bits. That is, the proposed ATFH-N approach is capable of producing more effective hash codes for improving cross-modal retrieval performance.

Within the proposed ATFH-N learning framework, four parameters, i.e., $\beta$, $\gamma$, $\mu_1$ and $\mu_2$ are involved, where $\beta$ and $\gamma$ are utilized to balance two manifold embedding items in Eq. (16). On the one hand, a larger $\beta$ may emphasize more in preserving the inter-modal similarity, while a larger $\gamma$ shall focus more on preserving the intra-modal similarity. In the experiments, different values of $\beta$ and $\gamma$ are attempted, by varying the value of one parameter while fixing the another one. It is noted that the results perform well when $\beta$ is selected within the range of [0.05, 0.2] and $\beta$ is chosen within the range of [0.2, 0.4]. On the other hand, $\mu_1$ and $\mu_2$ balances two adversarial learning mechanism in Eq. (16). Since the proposed ATFH-N learning framework mainly utilizes tri-fusion networks to process both balanced and imbalanced multi-modal data, $\mu_1$ naturally plays an important role to discriminate the data of only modality from the paired image-text instances and it is generally set a larger $\mu_1$ value in comparison with $\mu_2$. As pointed in [23], $\mu_2$ is usually set at 0.1 in most cases because it is insensitive to the least square optimization. Since $\mu_1$ is generally larger than $\mu_2$, different $\mu_1$ values are experimented and it is found that the results perform well within the range of [1, 3]. In addition, we also assess the parameter $k$ that influences the manifold embedding module of nearest neighbors, and empirically find that the different settings of $k$ within the range of [10, 30] only induce a minor fluctuation to the retrieval performance. Therefore, these parameters are generally insensitive to the cross-modal retrieval performances within a wide range of values.

## V. CONCLUSION

This paper has proposed an Adversarial Tri-Fusion Hashing Network (ATFH-N) for efficient cross-modal retrieval, which

can seamlessly work in balanced or imbalanced multi-modal data collections. Within the proposed ATFH-N framework, a triple fusion network associated with multi-modal zero padding mechanism is exploited to adapt either balanced or imbalanced multi-modal training data. At the same time, an adversarial training mechanism is leveraged to maximally bridge the gap of the representations between balanced and imbalanced data, while a label prediction network is efficiently utilized to guide the feature learning process and promote hash code learning. By embedding the manifold structure within the learning process, the underlying semantic structure of multimedia data can be well preserved in Hamming space, which can benefit various cross-modal retrieval tasks. To the best of our knowledge, this work is the early attempt to incorporate adversarial learning for working with imbalanced multi-modal data. Extensive experiments on various kinds of retrieval tasks have shown its outstanding performance.

Further research is anticipated along the present lines of work in order to solve several problems. For example, if new multi-modal data samples of other semantics are added into the training database, then the proposed model will have to learn the network parameters again, which would be time-consuming. Therefore, it would be necessary to extend the network model so as to handle the new multi-modal data adaptively. In addition, questions like how to further increase the cross-modal retrieval performance, and how to handle extremely imbalanced data are yet to be solved.

## REFERENCES


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

[2] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. IEEE Int. Conf. Multimedia*, 2010, pp. 251–260.

[3] D. Zhang and W. J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 2177–2183.

[4] M. Yu, L. Liu, and L. Shao, "Binary set embedding for cross-modal retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 12, pp. 2899–2910, Dec. 2017.

[5] E. Yang, C. Deng, C. Li, W. Liu, J. Li, and D. Tao, "Shared predictive cross-modal deep quantization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 12, pp. 5292–5303, Nov. 2018.

[6] X. Liu, Z. Hu, H. Ling, and Y. M. Cheung, "MTFH: A matrix tri-factorization hashing framework for efficient cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi:10.1109/TPAMI.2019.2940446.

[7] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3864–3872.

[8] J. Tang, K. Wang, and L. Shao, "Supervised matrix factorization hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3157–3166, Jul. 2016.

[9] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 2075–2082.

[10] L. Cui, Z. Chen, J. Zhang, L. He, Y. Shi, and P. S. Yu, "Multi-view collective tensor decomposition for cross-modal hashing," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2018, pp. 73–81.

[11] M. Hu, Y. Yang, F. Shen, N. Xie, R. Hong, and H. T. Shen, "Collective reconstructive embeddings for cross-modal hashing," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2770–2784, Jun. 2019.

[12] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 785–796.

[13] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 102–112, Jan. 2019.

[14] Z. Hu, X. Liu, X. Wang, Y. M. Cheung, N. Wang, and Y. Chen, "Triplet fusion network hashing for unpaired cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2019, pp. 141–149.

[15] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 415–424.

[16] H. Liu, R. Ji, Y. Wu, F. Huang, and B. Zhang, "Cross-modality binary code learning via fusion similarity hashing," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6345–6353.

[17] Y. Zhen and D. Y. Yeung, "Co-regularized hashing for multimodal data," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1376–1384.

[18] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2494–2507, May 2017.

[19] X. Lu, L. Zhu, Z. Cheng, X. Song, and H. Zhang, "Efficient discrete latent semantic hashing for scalable cross-modal retrieval," *Signal Process.*, vol. 154, pp. 217–231, 2019.

[20] Q. Y. Jiang and W. J. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3232–3240.

[21] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 4242–4251.

[22] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

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

[24] X. Xu, J. Song, H. Lu, Y. Yang, F. Shen, and Z. Huang, "Modal-adversarial semantic learning network for extendable cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2018, pp. 46–54.

[25] W. Li, L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1134–1148, Jun. 2014.

[26] A. Wu, W. Zheng, H. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5390–5399.

[27] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop Similarity-Based Pattern Recognit.*, 2015, pp. 84–92.

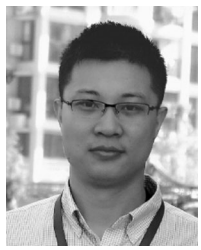
[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[29] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2010, pp. 251–260.

[30] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2008, pp. 39–43.

[31] 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 *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, pp. 1–9.

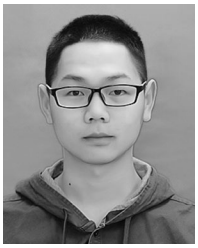
[32] Y. Wei *et al.*, "Cross-modal retrieval with cnn visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

**Xin Liu** (Member, IEEE) received the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, in 2013. He was a Visiting Scholar with Computer & Information Sciences Department, Temple University, Philadelphia, USA, from 2017 to 2018. Currently, he is an Associate Professor with the Department of Computer Science and Technology & Fujian Key Laboratory of Big Data Intelligence and Security, Huaqiao University, Xiamen, China, and also a Research Fellow with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China. His present research interests include multimedia analysis, computational intelligence, computer vision, pattern recognition and machine learning.

**Yiu-ming Cheung** (Fellow, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong. He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include machine learning, pattern recognition, visual computing, and optimization.

Prof. Cheung is the Founding Chair of the Computational Intelligence Chapter of the IEEE Hong Kong Section, and the Chair of the Technical Committee on Intelligent Informatics of the IEEE Computer Society. He serves as an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, *Pattern Recognition, Knowledge and Information Systems, Neurocomputing*, to name a few. He is a Fellow of IET, BCS and RSA, and Distinguished Fellow of IETI.



**Zhikai Hu** received his B.S. degree in computer science from China Jiliang University, Hangzhou, China, in 2015, and the M.S. degree in computer science from Huaqiao University, Xiamen, China, in 2019. He is currently pursuing the Ph.D. degree at the Department of Computer Science in Hong Kong Baptist University, Hong Kong. His present research interests include information retrieval, pattern recognition and deep learning.

**Yi He** received his B.S. degree in computer science from Qingdao University of Technology, Qingdao, China, in 2018. Currently, he is pursuing the M.S. degree in the Department of Computer Science and Technology, Huaqiao University, Xiamen, China. His present research interests include data mining, pattern recognition and deep learning.

**Bineng Zhong** received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2004, 2006, and 2010, respectively. From 2007 to 2008, he was a Research Fellow with the Institute of Automation and Institute of Computing Technology, Chinese Academy of Science. From September 2017 to September 2018, he is a visiting scholar in Northeastern University, Boston, MA, USA. Currently, he is a Professor with the School of Computer Science and Technology, Huaqiao University, Xiamen, China. His current research interests include pattern recognition, machine learning, and computer vision.