Multi-Task Consistency-Preserving Adversarial Hashing for Cross-Modal Retrieval

De Xie, Cheng Deng, *Member, IEEE*, Chao Li, Xianglong Liu, and Dacheng Tao, *Fellow, IEEE*

***Abstract*—Owing to the advantages of low storage cost and high query efficiency, cross-modal hashing has received increasing attention recently. As failing to bridge the inherent modality gap between modalities, most existing cross-modal hashing methods have limited capability to explore the semantic consistency information between different modality data, leading to unsatisfactory search performance. To address this problem, we propose a novel deep hashing method named Multi-Task Consistency-Preserving Adversarial Hashing (CPAH) to fully explore the semantic consistency and correlation between different modalities for efficient cross-modal retrieval. First, we design a consistency refined module (CR) to divide the representations of different modality into two irrelevant parts, *i.e.*, modality-common and modality-private representations. Then, a multi-task adversarial learning module (MA) is presented, which can make the modality-common representation of different modalities close to each other on feature distribution and semantic consistency. Finally, the compact and powerful hash codes can be generated from modality-common representation. Comprehensive evaluations conducted on three representative cross-modal benchmark datasets illustrate our method is superior to the state-of-the-art cross-modal hashing methods.**

***Index Terms*—Cross-modal retrieval, hashing, consistency-preserving, adversarial, multi-task.**

I. Introduction

With the explosive development of Internet and big data, large-scale and high-dimensional multi-modal data has been indwelt in social network and storage media. In order to deal with such massive data, cross-modal retrieval, which aims to search semantically similar instances in one modality by using a query from another modality, is increasingly significant and becomes a fundamental subject in computer vision applications [1], [2]. Actually, due to the data structure and feature distribution of different modalities are inconsistent, semantic consistency between a pair of relevant instances is difficult to being captured. How to effectively build the semantic consistency between the different modalities remains a challenging issue.

Most of existing cross-modal retrieval methods, including traditional statistical correlation [3], graph regularization [4], and dictionary learning [5], following the vein of subspace learning, which map different modality data into a common subspace and measure the similarities in this space. Therefore, these methods always suffer from high computation cost and low search accuracy. To tackle these drawbacks, hashing-based methods [6], [7], [8], [9], [10], project high-dimensional data from each modality into compact hash codes and preserve similar instances with similar hash codes. As such, the similarity is computed via fast bit wise XOR operation in Hamming space, which can save the memory storage cost and speed the query efficiency.

Traditional cross-modal hashing almost rely on shallow models, which can be divided into unsupervised [11], [12] and supervised settings [5], [13], [14], [15]. Compared with unsupervised counterpart, supervised cross-modal hashing methods can achieve better performance by exploiting semantic labels to build cross-modal correlation. Unfortunately, these shallow cross-modal hashing methods depend on hand-crafted features, which greatly limit the discriminative representation of instances and thus degrade the search performance. To take advantages of the recent progress of deep learning [16], [17], [18], deep cross-modal hashing methods are proposed as they are able to learn more discriminative representations and thus capture cross-modal correlations more effectively [19], [20], [21], [22], [23], [24], [25], [26], [27]. However, almost all of these deep models fail to distinguish modality-common and modality-private representation, making the learning of semantic consistency between these modalities inaccurate. Domain separation networks (DSN) [28] designs an encoder-decoder architecture to extract image representation that contains a common subspace across domains and a private subspace for each domain. Inspired by this common-private component analysis, SPDQ [29] projects multi-modal instances into a shared subspace and two private subspaces to learn modality-common and modality-private representations by multiple kernel maximum mean discrepancy (MK-MMD). ADAH [30] proposes an attention-aware method to generate attention mask for modality-common and modality-private representations extraction, simultaneously. In fact, the modality-common and modality-private representations are entangled with each other, which are difficult to be separated by these models. The more explicit guidance for disentangling modality-common and modality-private representations is necessary.

In this paper, we propose a novel deep architecture for

Manuscript received April 11, 2019; revised October 17, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61572388 and 61703327, in part by the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2018ZDXM-GY-176 and 2019ZDLGY03-02-01, and in part by the National Key R&D Program of China under Grant 2017YFE0104100. *(Corresponding authors: Cheng Deng and Xianglong Liu)*

De Xie, Cheng Deng and Chao Li are with the School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: xiede.xd@gmail.com; chdeng.xd@gmail.com; li_chao@stu.xidian.edu.cn).

X. Liu is with the State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China (e-mail: xlliu@nlsde.buaa.edu.cn

D. Tao is with the UBTECH Sydney Artificial Intelligence Centre and the School of Information Technologies, the Faculty of Engineering and Information Technologies, The University of Sydney, Darlington, NSW 2008, Australia (e-mail: dacheng.tao@sydney.edu.au).

cross-modal hashing retrieval, named multi-task consistency-preserving adversarial hashing (CPAH), where the modality-common representation, modality-private representation and compact hash codes for different modalities are learned in a unified framework. Specifically, we firstly designed a consistency refined module (CR) to divide each modality representation into modality-common and modality-private representation, respectively, which are irrelevant to each other. Subsequently, in order to bridge the modality gap and capture semantic consistency between different modalities effectively, we propose multi-task adversarial learning module (MA), which can make the feature distribution of modality-common representation from different modalities more close to each other and capture semantic consistency information between different modalities more effectively in an adversarial way. Finally, we explore each modality-common representation to generate discriminative and compact hash codes for cross-modal retrieval task. The main contributions of the proposed CPAH can be summarized as follows:

- We propose a multi-task consistency-preserving adversarial hashing for cross-modal retrieval, where multi-modal semantic consistency learning and hash learning are seamlessly incorporated in an end-to-end framework.
- We design consistency refined module (CR) for modality representation separation, which can divide modality representation into two incompatible parts, *i.e.*, modality-common and modality-private representations. Moreover, we propose multi-task adversarial learning module (MA), which can preserve the semantic consistency information between different modalities effectively by adversarial learning strategy.
- Extensive experiments conducted on three benchmark datasets demonstrate that the proposed CPAH significantly outperforms the state-of-the-art cross-modal hashing methods, including traditional as and as deep learning based methods.

The rest of this paper is organized as follows. We introduce some related works pertaining to cross-modal hashing and adversarial learning in Section II. Section III presents our proposed CPAH and optimization with theoretical analysis in detail. Section IV illustrates the experimental results and analysis. Finally, Section V concludes our work.

II. Related Work

A. Cross-modal Hashing

Cross-modal hashing can be categorized into two groups: unsupervised and supervised methods. The unsupervised methods only use co-occurrence information to learn hash functions for multi-modal data. For example, CVH [31] extends spectral hashing from uni-modal to multi-modal scenarios. LSSH [32] jointly learns latent features from images and texts with sparse coding. On the other hand, the supervised methods exploit label information to learn more discriminative common representation. CMSSH [33] regards each hash function as a binary classification problem and uses a boosting algorithm in the learning process. SCM [34] utilizes non-negative matrix factorization and neighbor preserving algorithm to maintain inter-modal and intra-modal semantic correlations.

Recently, deep learning based cross-modal hashing has shown powerful ability to exploit nonlinear correlations across different modalities. DCMH [20] combines feature learning and hash learning into a unified framework. PRDH [21] also adopts deep CNN models to learn feature representations and hash codes simultaneously. DVSH [35] utilizes convolutional neural network (CNN) and long short-term memory (LSTM) to separately learn the common representations. TDH [24] adopts deep neural networks with triplet label to capture more general semantic correlations between modalities.

B. Adversarial learning

With the evolution of generative adversarial networks (GANs) [36], adversarial learning has received increasing attention and achieved considerable progress in image generation [37], [38] and representation learning [39], [40]. Meanwhile, adversarial learning have also been employed in retrieval task. HashGAN [41] is a typical adversarial hash-based retrieval method, which can learn compact binary hash codes from both real images and diverse images synthesized by generative models. Nevertheless, HashGAN belongs to uni-modal retrieval, which only utilizes generative adversarial learning as a means of data augmentation. In contrast, for real-valued based cross-modal retrieval, ACMR [42] adopts adversarial learning to discover an effective common subspace and generate modality-invariant representations. Li et al. [22] proposed a self-supervised adversarial hashing (SSAH), which leverages two adversarial networks to capture the semantic correlation between different modalities. Zhang et al. [30] presented attention-aware deep adversarial hashing (ADAH) to generate an attention mask for attended and unattended feature representations learning. In comparison to these methods, our CPAH utilizes multi-task adversarial learning for explicitly modeling modality-common and modality-private representations learning, which can maximumly preserve semantic consistency between modalities.

III. Multi-Task Consistency-Preserving Adversarial Hashing

In this section, we present the details about our CPAH model, including model formulation and optimization algorithm. Fig. 1 shows the whole flowchart of the CPAH, which is an end-to-end framework with a couple of networks and contains two main steps, *i.e.*, consistency learning and hash learning.

A. Problem Formulation

The notations and the problem formulation is first introduced for this paper. Our CPAH can be expanded to multiple modalities, such as image, audio, and video. For simplicity, we only use image and text to explain our method. Overall, matrices and vectors are written in boldface with uppercase and lowercase letters respectively. To denote the i-th element of a vector $\mathbf{m}$, we use the notation m_i. When it comes to a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2020.2963957, IEEE Transactions on Image Processing

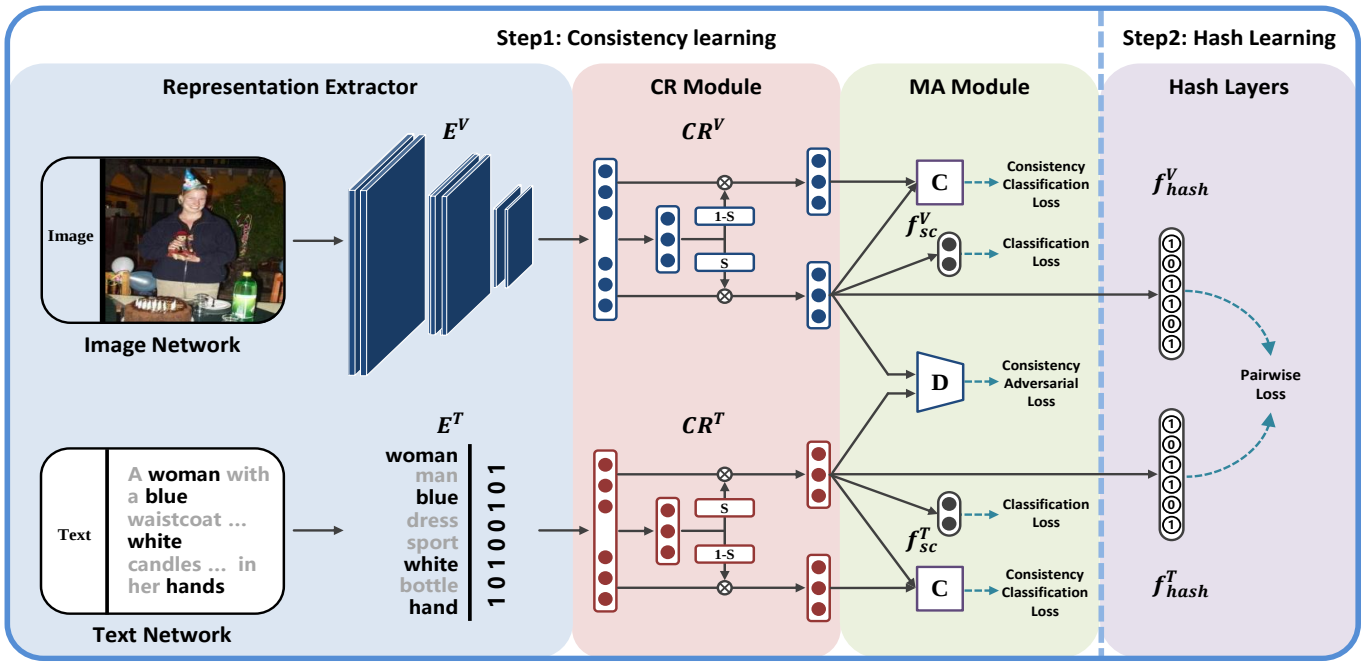


Fig. 1: The framework of the proposed CPAH. CPAH is an end-to-end framework with a couple of networks, *i.e.*, image network and text network, and contains two main steps. Step 1 is consistency learning, which can extract semantic consistency information through consistency refined module (CR) and multi-task adversarial learning module (MA). Step 2 is hash learning, which utilizes modality-common representation to yield compact hash codes.

matrix $\mathbf{M}$, the form of M_{ij} is adopted to denote the i-th row and j-th column element. Furthermore, the superscript V/T of a variable indicate that it belongs to the image/text modality.

Given a dataset $\mathbf{O} = \{o_i\}_{i=1}^{N}$ including image and text modalities, $o_i = \{v_i, t_i, l_i\}$, where v_i and t_i are raw image and text data for i-th instance, and $l_i = \{l_{i1}, l_{i2}, \ldots, l_{iL}\}$ is the multi-label annotations assigned to o_i. L is the class number. If o_i belongs to j-th class, $l_{ij} = 1$, otherwise $l_{ij} = 0$. The pairwise multi-label similarity matrix $\mathbf{S}$ is used to describe semantic similarity between each two instances, where $S_{ij} = 1$ means o_i is semantically similar with o_j, otherwise $S_{ij} = 0$. For multi-label setting, two instances o_i and o_j are annotated with multiple labels. Thus, we define $S_{ij} = 1$, if o_i and o_j share at least one label, otherwise $S_{ij} = 0$.

For an image (text) query, the goal of cross-modal hashing is to learn two hash functions for the two modalities: $B^V \in \{-1, +1\}^K$ for the image modality and $B^T \in \{-1, +1\}^K$ for the text modality, where K is the length of hash code. The similarity between two hash codes is measured by Hamming distance. The relationship between their Hamming distance $dis_H(B_i^V, B_j^T)$ and their inner product $\langle B_i^V, B_j^T \rangle$ can be formulated as:

$$dis_H(B_i^V, B_j^T) = \frac{1}{2}(K - \langle B_i^V, B_j^T \rangle), \tag{1}$$

thus we can use inner product to evaluate the similarity of two hash codes. Given $\mathbf{S}$, the probability of $\mathbf{S}$ under the condition B^* $(* = V, T)$ can be expressed as:

$$p(S_{ij}|B^*) = \begin{cases} \sigma(\theta_{ij}), & S_{ij} = 1 \\ 1 - \sigma(\theta_{ij}), & S_{ij} = 0 \end{cases} \tag{2}$$

where $\sigma(\theta_{ij}) = \frac{1}{1+e^{-\theta_{ij}}}$, and $\theta_{ij} = \frac{1}{2}\langle B_i^V, B_j^T \rangle$. Therefore, two instances with larger inner produce should be similar with a high probability. The problem of quantifying the similarity between binary codes in Hamming space can be transformed into calculating the inner product of their original features.

B. Framework

As shown in Fig. 1, we exploit two deep neural networks, *i.e.*, image network and text network, for image modality V and textual modality T, respectively. For image network, we extract image representation via universal convolutional neural network such as CNN-F [43], which is a popular framework used in cross-modal hashing. The original CNN-F [43] consists of five convolutional layers ($conv_1 - conv_5$) and three fully-connected layers ($fc_6 - fc_8$). We utilize the output of fc_7 as the representation of image modality. Then a consistency refined module (CR) is designed for image modality to generate two incompatible representations, *i.e.*, modality-common representation and image modality-private representation. For text network, we first transform each text instance into a one-hot vector using bag-of-words (BOWs) representation. The BOWs representation is used as the input to two convolutional layers ($conv_1 - conv_2$) to extract text representation. Then, we also apply a consistency refined module (CR) to divide text representation into two incompatible representations, *i.e.*, modality-common representation and textual modality-private representation. After that, we design a multi-task adversarial learning module (MA), which contains semantic consistency discriminator, semantic consistency classifier and semantic classifier. These three components cooperate each other in a adversarial way, which can generate more effective modality-common representation. Finally, to map the learned modality-common representation into Hamming space directly, we design two fully-connected hash layers, fc_h^V and fc_h^T, with

K hidden nodes, as in [24], to yield compact binary codes for cross-modal hashing retrieval.

1) Consistency refined module: In cross-modal retrieval setting, the essential problem is to find the semantic consistency information between the instances of different modalities. In fact, for a pair of relevant instances, each instance not only has their common semantic consistency information, but also has it own private information, *e.g.*, the semantically irrelevant background information in images. Although the two kinds of information are mutually exclusive, their extraction process can promote each other. Therefore, for image modality and text modality, we design consistency refined module to divided image representation and text representation into image/text modality-common representation and image/text modality-private representation. Suppose $\boldsymbol{r}^V$ and $\boldsymbol{r}^T$ are image and text representations, we input them into two different convolutional layers f_{cr}^V and f_{cr}^T, where kernel size is 1×1 and activation function is $\mathrm{sigmoid}$, to generation information selection mask for each modality. Note $\boldsymbol{m}_c^V$, $\boldsymbol{m}_p^V$ and $\boldsymbol{m}_c^T$, $\boldsymbol{m}_p^T$ are the information selection masks of image modality and text modality respectively, which can be formulated as:

$$\begin{aligned} \boldsymbol{m}_c^V &= \mathrm{sigmoid}(f_{cr}^V(\boldsymbol{r}^V)), \\ \boldsymbol{m}_p^V &= 1 - \mathrm{sigmoid}(f_{cr}^V(\boldsymbol{r}^V)), \\ \boldsymbol{m}_c^T &= \mathrm{sigmoid}(f_{cr}^T(\boldsymbol{r}^T)), \\ \boldsymbol{m}_p^T &= 1 - \mathrm{sigmoid}(f_{cr}^T(\boldsymbol{r}^T)). \end{aligned} \tag{3}$$

After that, the modality-common representations $\boldsymbol{r}_c^V$, $\boldsymbol{r}_c^T$ and modality-private representations $\boldsymbol{r}_p^V$, $\boldsymbol{r}_p^T$ can be obtained through these masks as:

$$\begin{aligned} \boldsymbol{r}_c^V &= \boldsymbol{r}^V \cdot \boldsymbol{m}_c^V, \\ \boldsymbol{r}_p^V &= \boldsymbol{r}^V \cdot \boldsymbol{m}_p^V, \\ \boldsymbol{r}_c^T &= \boldsymbol{r}^T \cdot \boldsymbol{m}_c^T, \\ \boldsymbol{r}_p^T &= \boldsymbol{r}^T \cdot \boldsymbol{m}_p^T. \end{aligned} \tag{4}$$

Since the modality-common representation and modality-private representation are mutually exclusive, learning modality-common and modality-private representations simultaneously is conducive to obtain more discriminative modality-common representation.

2) Multi-Task Adversarial learning: For learning modality-common and modality-private representations effectively, we design multi-task adversarial learning module that contains three tasks: consistency adversarial learning, information classification and semantic classification.

Actually, adversarial learning can be regarded as distribution approximator, which is able to make the distribute of the generated data similar with the groundtruths. Therefore, we design consistency adversarial learning task, aiming to make the distributions of each modality-common representation close to each other, which contains two steps: discriminative step and generative step. Let Θ_c^V, Θ_c^T, Θ_p^V and Θ_p^T be the parameters of G_c^V, G_c^T, G_p^V and G_p^T networks, respectively, which generate modality-common representations $\boldsymbol{r}_c^V$, $\boldsymbol{r}_c^T$ and modality-private representations $\boldsymbol{r}_p^V$, $\boldsymbol{r}_p^T$. Θ_D is the parameter of discriminator D. For discriminative step, we train D to classify the modality-common representation into "True" and classify the modality-private representation into "Fake". The object function of the discriminator can be defined as:

$$\min_{\Theta_D} \mathcal{L}_{cal}^d = \sum_{i=1}^{N} \|D(r_{ci}^V) - 1\|^2 + \|D(r_{ci}^T) - 0\|^2, \tag{5}$$

where r_{ci}^V and r_{ci}^T are modality-common representations of the i-th pairwise instances of image and text modalities, v_i and t_i, respectively. On the contrary, in generative step, we train G_c^V and G_c^T to classify the modality-common representation into "Fake" and classify the modality-private representation into "True". The object function of the generator is defined as:

$$\min_{\Theta_c^V, \Theta_c^T} \mathcal{L}_{cal}^g = \sum_{i=1}^{N} \|D(G_c^V(v_i)) - 0\|^2 + \|D(G_c^T(t_i)) - 1\|^2. \tag{6}$$

The generator G_c^V, G_c^T and the discriminator D are trained adversarially with each other using a adversarial learning strategy. By this strategy, the distributions of modality-common representation $\boldsymbol{r}_c^V$ and $\boldsymbol{r}_c^T$ are able to close to each other gradually.

In order to effectively separate modality-common and modality-private representations, we apply information classification to learn modality-common representation, image modality-private representation and text modality-private representation respectively in a weakly supervised manner. Specifically, we design a information classifier C and regard information classification as a multi-class classification task. The label of different representation are one-hot vector, which can be defined as $\gamma \in \mathbb{R}^3$. The image modality-private representation label γ_p^V is "0", the modality-common representation labels γ_c^V and γ_c^T are "1", and the text modality-private representation label γ_p^T is "2". Above all, the objective function of information classification is expressed as follows:

$$\min_{\Theta_c^V, \Theta_c^T, \Theta_p^V, \Theta_p^T, \Theta_C} \mathcal{L}_{ic} = -\sum_{i=1}^{N} \gamma \cdot \log(C(r_i)), \tag{7}$$

where $\gamma \in \{\gamma_p^V, \gamma_c^V, \gamma_c^T, \gamma_p^T\}$, $r_i \in \{r_{pi}^V, r_{ci}^V, r_{ci}^T, r_{pi}^T\}$ and Θ_C is the parameter of C.

We also adopt semantic classification to preserve the semantically discriminative capability of image and text modality-common representations. To this end, two semantic classifier layers fc_{sc}^V and fc_{sc}^T are designed to capture a fine-grained cross-modal semantic consistency. The objection function of the semantic classification can be denoted as:

$$\begin{aligned} \min_{\Theta_c^V, \Theta_c^T, \Theta_{sc}^V, \Theta_{sc}^T} \mathcal{L}_{sc} = -\sum_{i=1}^{N} [& l_i \cdot \log(fc_{sc}^V(r_{ci}^V)) \\ & + l_i \cdot \log(fc_{sc}^T(r_{ci}^T)) \\ & + (1 - l_i) \log(1 - fc_{sc}^V(r_{ci}^V)) \\ & + (1 - l_i) \log(1 - fc_{sc}^T(r_{ci}^T))], \end{aligned} \tag{8}$$

where l_i are the corresponding one-hot label vector of the i-th pairwise instances v_i and t_i. Θ_{sc}^V and Θ_{sc}^T are the parameters of fc_{sc}^V and fc_{sc}^T, respectively. Combining Eq. (5), Eq. (6) Eq. (7) and Eq. (8), we can obtain the object of multi-task adversarial learning as:

$$\mathcal{L}_{mtal} = \mathcal{L}_{cal}^d + \mathcal{L}_{cal}^g + \alpha(\mathcal{L}_{ic} + \mathcal{L}_{sc}), \tag{9}$$

where α is a hyper-parameter. Through these three tasks, we can disentangle modality-common and modality-private representations, then further obtain a pair of uniform distributed modality-common representations, which contain effective semantic consistency information.

3) Hash learning: After multi-task adversarial learning, the compact and discriminative hash codes of image and text can be generated from modality-common representation $\boldsymbol{r}_c^V$, $\boldsymbol{r}_c^T$. To this end, we design two hash layers f_h^V and f_h^T for image and text hash codes generation. We utilize pairwise loss function to measure the similarity between their hash codes, which can be devised as:

$$\min_{\Theta_c^V,\Theta_c^T,\Theta_h^V,\Theta_h^T} \mathcal{L}_p = -\sum_{i,j=1}^{N}(S_{ij}\theta_{ij} - \log(1+e^{\theta_{ij}})). \quad (10)$$

Suppose that $\boldsymbol{h}^V = f_h^V(\boldsymbol{r}_c^V)$, $\boldsymbol{h}^T = f_h^T(\boldsymbol{r}_c^T)$, $\theta_{ij} = \frac{1}{2}h_{\star i}^\top h_{\star j}$, where Θ_h^V and Θ_h^T are the parameters of hash layers f_h^V and f_h^T respectively. It is easy to find that minimizing this loss function is equivalent to maximizing the likelihood, which can make the similarity (inner product) between $h_{\star i}$ and $h_{\star j}$ large when $S_{ij}=1$ and small when $S_{ij}=0$. Hence, optimizing $\mathcal{L}_p$ can preserve the cross-modal similarity in $\mathbf{S}$ with the image hashing layer output $\boldsymbol{h}^V$ and text hashing layer output $\boldsymbol{h}^T$. Furthermore, since hash codes are discrete, the quantization error is unavoidable in hash codes learning procedure. In order to reduce quantization error, the quantization loss is designed as:

$$\begin{aligned}\min_{\Theta_c^V,\Theta_c^T,\Theta_h^V,\Theta_h^T} \mathcal{L}_q &= \|\boldsymbol{b}^V - \boldsymbol{h}^V\|_F^2 + \|\boldsymbol{b}^T - \boldsymbol{h}^T\|_F^2 \\ &= \|\boldsymbol{b} - \boldsymbol{h}^V\|_F^2 + \|\boldsymbol{b} - \boldsymbol{h}^T\|_F^2 \\ &s.t. \quad \boldsymbol{b} \in \{-1,+1\}^{K\times N},\end{aligned} \quad (11)$$

where $\boldsymbol{b}^V = \mathrm{sign}(\boldsymbol{h}^V)$ and $\boldsymbol{b}^T = \mathrm{sign}(\boldsymbol{b}^T)$. We keep the modalities share hash codes $\boldsymbol{b}$ in training and consider $\boldsymbol{h}^V$, $\boldsymbol{h}^T$ to be the continuous surrogate of $\boldsymbol{b}$, respectively. Because $\boldsymbol{h}^V$ and $\boldsymbol{h}^T$ can preserve the cross-modal similarity in $\mathbf{S}$, the binary hash codes $\boldsymbol{b}^V$ and $\boldsymbol{b}^T$ can also be expected to preserve the cross-modal similarity in $\mathbf{S}$, which exactly matches the goal of cross-modal hashing. Combing Eq. (10) with Eq. (11), we obtain the objective function of the hash learning as follow:

$$\mathcal{L}_{hash} = \mathcal{L}_p + \beta\mathcal{L}_q, \quad (12)$$

where β is a hyper-parameter.

4) Objective Function: By merging the above two parts together (*i.e.*, the multi-task adversarial learning loss $\mathcal{L}_{mtal}$ and the hash learning loss $\mathcal{L}_{hash}$), we can obtain the whole objective function as the follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{mtal} + \mathcal{L}_{hash} \\ &= \mathcal{L}_{cal}^d + \mathcal{L}_{cal}^g + \mathcal{L}_p + \alpha(\mathcal{L}_{ic} + \mathcal{L}_{sc}) + \beta\mathcal{L}_q.\end{aligned} \quad (13)$$

In oder to yield powerful hash codes, we minimize the Eq. (13) with an alternative manner.

C. Learning Algorithm

Since the objective function is not convex, We adopt an alternative learning strategy to train our CPAH modal. Each time we learn one parameter with the other parameters are fixed. The model is optimized iteratively until the parameters are converged or the preset maximum number of epochs is reached. The whole alternative learning algorithm is briefly outlined in Algorithm 1.

Algorithm 1 The optimization algorithm for our CPAH.

Training Stage

Input: Image set V; Text set T; Label set L;
Output: Optimal code matrix B
Initialization
Initialize network parameters Θ_c^V, Θ_c^T, Θ_p^V, Θ_p^T, Θ_D, Θ_C, Θ_{sc}^V, Θ_{sc}^T, Θ_h^V, Θ_h^T;
mini-batch size: $N^{V,T}$, learning rate: μ;
iteration number for one epoch: $t = N/N^{V,T}$;
initial epoch number: e, maximal epoch number: E;
repeat
for iter $= 1, 2, \ldots,$ t **do**
Update Θ_D with Eq. (5) by BP algorithm.
Update Θ_c^V, Θ_c^T,Θ_p^V. Θ_p^T, Θ_C, Θ_{sc}^V, Θ_{sc}^T with Eq. (6), Eq. (7), Eq. (8) by BP algorithm.
end for
for iter $= 1, 2, \ldots,$ t **do**
Update Θ_c^V, Θ_h^V with Eq. (12) by BP algorithm.
Update Θ_c^T, Θ_h^T with Eq. (12) by BP algorithm.
end for
Update the parameter $\boldsymbol{b}$ by $\boldsymbol{b} = \mathrm{sign}(\frac{1}{2}(\boldsymbol{h}^V + \boldsymbol{h}^T))$.
$e = e + 1$.
until e = E.

Testing Stage

Input: Query instance for different modalities: Q^V or Q^T;
Output: Hashing codes for query: $\boldsymbol{b}_q^V$ or $\boldsymbol{b}_q^T$.
Procedure
Calculate $\boldsymbol{h}_q^V$: Given Q^V, directly feed the raw image into image network;
Calculate $\boldsymbol{h}_q^T$: Given Q^T, directly feed the bag-of-words vector into text network;
Calculate $\boldsymbol{b}_q^V$ or $\boldsymbol{b}_q^T$: $\boldsymbol{b}_q^V = \mathrm{sign}(\boldsymbol{h}_q^V)$, $\boldsymbol{b}_q^T = \mathrm{sign}(\boldsymbol{h}_q^T)$.

1) Multi-Task Adversarial Learning: Firstly, the discriminator $\boldsymbol{D}$ is trained by Eq. (5). The parameter Θ_D is updated while Θ_c^V and Θ_c^T are fixed:

$$\Theta_D = \arg\min_{\Theta_D} \mathcal{L}_{cal}^d. \quad (14)$$

Then, G_c^V, G_c^T, G_p^V, G_p^T, C, fc_{sc}^V and fc_{sc}^T are trained by updating $\Theta_G = \{\Theta_c^V, \Theta_c^T, \Theta_p^V, \Theta_p^T, \Theta_C, \Theta_{sc}^V, \Theta_{sc}^T\}$ with Θ_D is fixed:

$$\Theta_G = \arg\min_{\Theta_G} \mathcal{L}_{cal}^g + \alpha(\mathcal{L}_{ic} + \mathcal{L}_{sc}). \quad (15)$$

2) Hash learning: For hash learning optimization, we firstly optimize the image hash codes generation by training G_c^V and f_h^V, while the parameters of G_c^T and f_h^T are fixed:

$$\Theta_c^V, \Theta_h^V = \arg\min_{\Theta_c^V,\Theta_h^V} \mathcal{L}_{hash}. \quad (16)$$

After that, we optimize the parameters of G_c^T and f_h^T for text hash codes generation with the parameters of G_c^V and f_h^V are fixed:

$$\Theta_c^T, \Theta_h^T = \arg \min_{\Theta_c^T, \Theta_h^T} \mathcal{L}_{hash}. \tag{17}$$

3) Out-of-Sample Extension: Any new instance that is not from the training data can be represented as a specific hash code as long as one of its modality (image or text) is observed. For example, given one instance $\boldsymbol{q}^V$ from image modality, we can generate hash codes by forward propagating the image network, as follows:

$$\boldsymbol{b}_q^V = \text{sign}(f_h^V(G_c^V(\boldsymbol{q}^V))). \tag{18}$$

Similarly, given an instance $\boldsymbol{q}^T$ from text modality, we can also generate the hash codes by text modality, as follows:

$$\boldsymbol{b}_q^T = \text{sign}(f_h^T(G_c^T(\boldsymbol{q}^T))). \tag{19}$$

By this means, it can be seen that our CPAH can be employed for cross-modal retrieval where the query data and the result data are from different modalities.

4) Multiple Modalities Extension: By extending the networks to multiple networks, the proposed CPAH can be extended to multiple modalities. The extension of CPAH in Eq. (13) from bimodal to multiple modalities is quite straightforward as:

$$\begin{aligned} \min \mathcal{L}_{multi} = & \sum_{i,j=1,i\neq j}^{M} \mathcal{L}_{mtal}^{ij} + \sum_{i,j=1,i\neq j}^{M} \mathcal{L}_{p}^{ij} \\ & + \sum_{i=1}^{M} (\beta \|\boldsymbol{b}^i - \boldsymbol{h}^i\|_F^2) \\ s.t. \quad & \boldsymbol{b} = \boldsymbol{b}^1 = \cdots = \boldsymbol{b}^M \in \{-1, +1\}^{K\times N}, \end{aligned} \tag{20}$$

where M is the number of modalities, $\mathcal{L}_{mtal}^{ij}$ is the multi-task adversarial learning loss function between i-th modality and j-th modality, $\mathcal{L}_p^{ij}$ represents the pairwise loss function of i-th modality and j-th modality, $\boldsymbol{b}^i$ is the hash codes of i-th modality, and $\boldsymbol{h}^i \in \mathbb{R}^i$ denotes the output i-th modality from i-th network. It is straightforward to adjust the iterative algorithm presented above to solve the new optimization problem.

TABLE I: Configuration of the proposed architecture.

Module		Layers	Configuration
E^V	CNN-F	$conv_1$	f. 64×11×11; st. 4×4;pad 0; LRN; ×2 pool
		$conv_2$	f. 256×5×5; st. 1×1;pad 2; LRN; ×2 pool
		$conv_3$	f. 256×3×3; st. 1×1; pad 1;
		$conv_4$	f. 256×3×3; st. 1×1; pad 1;
		$conv_5$	f. 256×3×3; st. 1×1; pad 1; ×2 pool
		fc_6	4096;
		fc_7	4096;
	VGG16	$conv_1$	f. 64×3×3; st. 1×1; pad 1; f. 64×3×3; st. 1×1; pad 1; ×2 pool
		$conv_2$	f. 128×3×3; st. 1×1; pad 1; f. 128×3×3; st. 1×1; pad 1; ×2 pool
		$conv_3$	f. 256×3×3; st. 1×1; pad 1; f. 256×3×3; st. 1×1; pad 1; f. 256×3×3; st. 1×1; pad 1; ×2 pool
		$conv_4$	f. 512×3×3; st. 1×1; pad 1; f. 512×3×3; st. 1×1; pad 1; f. 512×3×3; st. 1×1; pad 1; ×2 pool
		$conv_5$	f. 512×3×3; st. 1×1; pad 1; f. 512×3×3; st. 1×1; pad 1; f. 512×3×3; st. 1×1; pad 1; ×2 pool
		fc_6	4096
		fc_7	4096
E^T		$conv_1$	f. 4096×1×1; st. 1×1; pad 0
		$conv_2$	f. 4096×1×1; st. 1×1; pad 0
CR^V		fc_1	512
		fc_2	512
CR^T		fc_1	512
		fc_2	512
D		fc_1	64
		fc_2	1
C		fc_1	64
		fc_2	3
f_{sc}		fc_{sc}^V	The number of classes: L
		fc_{sc}^T	The number of classes: L
f_{hash}		fc_h^V	The length of hash codes: K
		fc_h^T	The length of hash codes: K

IV. Experiments and Discussions

A. Datasets and Settings

1) Datasets: We evaluate our proposed method on three popular benchmark cross-modal datasets: MIRFLICKR-25K [44], IAPRTC-12 [45] and NUS-WIDE [46].

MIRFLICKR-25K: This dataset consists of 25,000 images collected from Flickr website. Each image is associated with several textual tags. Hence, each instance is a image-text pair. Totally, we select 20,961 instances in our experiment. The text for each instance is represented as a 2029-dimensional bag-of-words vector, and each instance is manually annotated with at least one of the 24 unique labels.

IAPRTC-12: This dataset consists of 20,000 image-text pairs which presents various semantics such as landscape, action and people categories. We use 18,584 instances in our experiment. The text for each instance is represented as a 4027-dimensional bag-of-words vector. The image-text pairs belong to the 22 frequent labels from the 275 semantic concepts.

NUS-WIDE: This dataset contains 269,648 web images, and some images are associated with textual tags, where each pair is annotated with multiple labels among 81 semantic concepts. We select 188,321 image-text pairs that belong 21 most frequent concepts. The text for each instance is represented as an 1000-dimensional bag-of-words vector.

We follow dataset split similar as [20] to construct the test (query) sets, retrieval sets and training sets. For MIRFLICKR-25K and IAPRTC-12 datasets, 2,000 instances are randomly sampled as the test set and the remaining instances as the

retrieval set. Moreover, 10,000 instances are sampled from the retrieval set as training set for both datasets. For NUS-WIDE dataset, we take 2,100 instances as the test set and the rest as the retrieval set. In addition, we sample 10,500 instances from the retrieval set as training set.

2) Implementation details: We implement our CPAH based on open source deep toolbox *TensorFlow*. All the experiments are running on a server with one NVIDIA TITAN X GPU. In training procedure, both multi-task adversarial learning and hash learning are optimized in an alternative way through adaptive moment estimation optimizer (Adam). The batch size is 64, and the total epoch is 80. The initial learning rate is set to 0.0001 and decreases to its one-fifth every 30 epochs. In testing procedure, we only use the modality-common representations of image and text modalities to construct the binary codes.

For image modality, the input images are first resized to $224 \times 224 \times 3$. We use two CNN architectures serve as the extractor E^V to extract image representation: CNN-F [43] and VGG16 [47], which both consist of five convolutional blocks and three fully-connected layers. The output of their fc_7 layer is served as image representation. For text modality, the text representation extraction E^T is constructed by two 1×1 convolutional layers. After extracting the representations of image and text, we design two consistency refined module, *i.e.*, CR^V and CR^T, which consist by two different fully-connected layers with both 512 neural nodes, respectively. For multi-task adversarial learning module, we frame a discriminator C, an information classier C and two semantic classifier fc_{sc}^V and fc_{sc}^T, which are both consisted by fully-connected layers. For hash learning, two hash layers fc_h^V and fc_h^T are designed for hash codes generation, which are two fully-connected layers with K neural nodes. Table I shows the detailed configuration of different modules in the proposed CPAH. Specifically, "f" denotes convolution filters, "st." is the convolution stride, "pad" is the number of pixels to add to the input, "LRN" denotes Local Response Normalization, and "pool" denotes the down-sampling factor.

B. Evaluation

In this part, we perform two cross-modal retrieval tasks: image-to-text retrieval and text-to-image retrieval, which search texts by a query image and search images by a query text, respectively. To evaluate the performance of our method, we adopt three evaluation criteria: Mean average precision (MAP), the top$N$-precision curve (Top$N$ curve) and precision-recall curve (PR curve).

To calculate the MAP, we first evaluate the average precision (AP). Given a query instance, we obtain a ranking list of R retrieved results, and the value of AP is defined as:

$$AP = \frac{1}{N} \sum_{r=1}^{R} p(r)\delta(r) \tag{21}$$

where N is the number of relevant instances in the retrieved set, $p(r)$ denotes the precision of the top r retrieved instance, and $\delta(r) = 1$ if the r-th retrieved result is relevant to the query instance, otherwise, $\delta(r) = 0$. Here, we consider two samples similar as long as they share at least one similar label. Generally, MAP measures the discriminative learning ability of different cross-modal retrieval methods, where a higher MAP indicates better retrieval performance. The top$N$-precision curve reflects the changes in precision according to the number of retrieved instances and the precision-recall curve reflects the precision at different recall levels and can be obtained by varying the Hamming radius of the retrieved instances in a certain range in order to evaluate the precision and recall.

TABLE II: The MAP scores of two retrieval tasks on the MIRFLICKR-25K dataset with different hash code lengths. The baselines are based on CNN-F features and the best accuracy is shown in boldface.

TASK	Method	MIRFLICKR-25K		
		16 bits	32 bits	64 bits
Image Query v.s. Text Database	CVH	0.532	0.531	0.530
	STMH	0.584	0.584	0.595
	CMSSH	0.566	0.565	0.561
	SCM	0.672	0.680	0.685
	LSSH	0.557	0.572	0.566
	SePH	0.681	0.693	0.695
	DCMH	0.733	0.745	0.751
	Ours	**0.775**	**0.791**	**0.787**
Text Query v.s. Image Database	CVH	0.524	0.523	0.523
	STMH	0.585	0.588	0.599
	CMSSH	0.549	0.559	0.547
	SCM	0.699	0.707	0.710
	LSSH	0.557	0.559	0.561
	SePH	0.622	0.632	0.633
	DCMH	0.749	0.764	0.766
	Ours	**0.777**	**0.787**	**0.789**

TABLE III: The MAP scores of two retrieval tasks on the MIRFLICKR-25K dataset with different hash code lengths. The baselines are based on VGG16 features and the best accuracy is shown in boldface.

TASK	Method	MIRFLICKR-25K		
		16 bits	32 bits	64 bits
Image Query v.s. Text Database	CVH	0.534	0.533	0.533
	STMH	0.590	0.594	0.599
	CMSSH	0.548	0.563	0.556
	SCM	0.699	0.706	0.711
	LSSH	0.596	0.603	0.613
	SePH	0.729	0.736	0.740
	DCMH	0.729	0.758	0.760
	Ours	**0.789**	**0.796**	**0.795**
Text Query v.s. Image Database	CVH	0.535	0.534	0.534
	STMH	0.596	0.602	0.601
	CMSSH	0.527	0.557	0.540
	SCM	0.713	0.721	0.725
	LSSH	0.567	0.566	0.567
	SePH	0.639	0.645	0.648
	DCMH	0.754	0.764	0.770
	Ours	**0.778**	**0.786**	**0.785**

C. Comparison with State-of-the-art Methods

Our CPAH is compared with several state-of-the-art algorithms: CVH [31], STMH [48], CMSSH [33], SCM [34],

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2020.2963957, IEEE Transactions on Image Processing

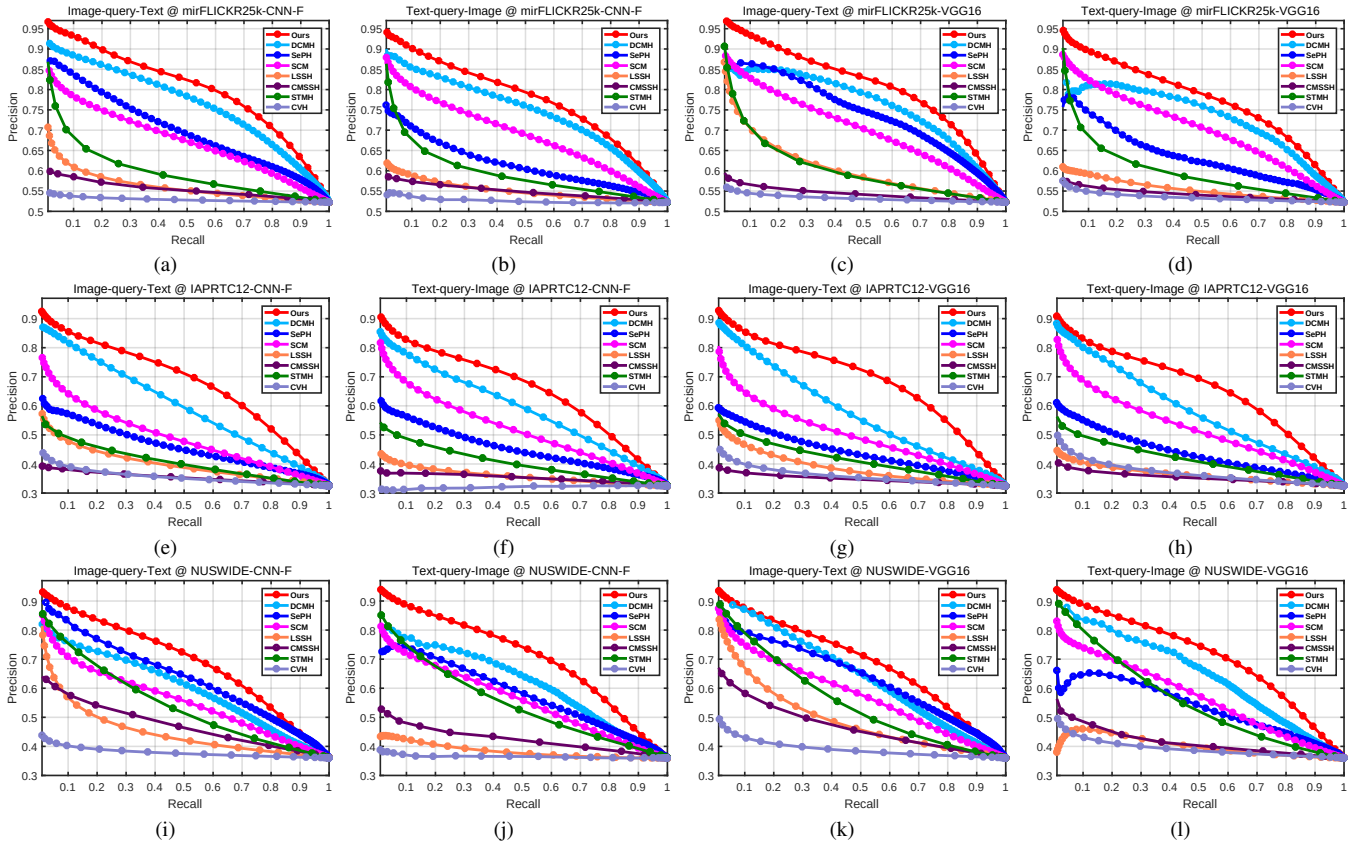


Fig. 2: The Precision-recall curves of different methods for image query text and text query image task on the MIRFLICKR-25K, IAPRTC-12 and NUS-WIDE datasets respectively (The code length is 64).

TABLE IV: The MAP scores of two retrieval tasks on the IAPRTC-12 dataset with different hash code lengths. The baselines are based on CNN-F features and the best accuracy is shown in boldface.

TASK	Method	IAPRTC-12		
		16 bits	32 bits	64 bits
Image Query v.s. Text Database	CVH	0.358	0.359	0.358
	STMH	0.370	0.383	0.398
	CMSSH	0.378	0.366	0.361
	SCM	0.475	0.493	0.503
	LSSH	0.404	0.412	0.414
	SePH	0.451	0.459	0.465
	DCMH	0.512	0.531	0.553
	Ours	**0.626**	**0.649**	**0.658**
Text Query v.s. Image Database	CVH	0.324	0.323	0.320
	STMH	0.375	0.403	0.426
	CMSSH	0.372	0.369	0.348
	SCM	0.504	0.527	0.539
	LSSH	0.371	0.378	0.379
	SePH	0.450	0.457	0.465
	DCMH	0.537	0.543	0.589
	Ours	**0.616**	**0.642**	**0.655**

TABLE V: The MAP scores of two retrieval tasks on the IAPRTC-12 dataset with different hash code lengths. The baselines are based on VGG16 features and the best accuracy is shown in boldface.

TASK	Method	IAPRTC-12		
		16 bits	32 bits	64 bits
Image Query v.s. Text Database	CVH	0.360	0.358	0.355
	STMH	0.389	0.398	0.412
	CMSSH	0.363	0.378	0.370
	SCM	0.476	0.498	0.509
	LSSH	0.394	0.418	0.407
	SePH	0.438	0.448	0.453
	DCMH	0.535	0.553	0.567
	Ours	**0.626**	**0.663**	**0.671**
Text Query v.s. Image Database	CVH	0.371	0.373	0.368
	STMH	0.395	0.415	0.438
	CMSSH	0.338	0.343	0.349
	SCM	0.503	0.527	0.542
	LSSH	0.376	0.382	0.385
	SePH	0.433	0.442	0.447
	DCMH	0.581	0.592	0.608
	Ours	**0.614**	**0.649**	**0.661**

LSSH [32], SePH [49] and DCMH [20] in all three datasets and the length of hash codes is presented by 16 bits, 32 bits and 64 bits. In order to make fair comparisons, we utilize deep features for image extracted from pre-trained CNN-F and VGG16 for all shallow baseline methods mentioned above.

Table II and Table III, Table IV and Table V, Table VI and Table VII, present the MAP values of our method and other methods with CNN-F features and VGG16 features for text query image task and image query text task for the MIRFLICKR-25K, IAPRTC-12, and NUS-WIDE datasets respectively. From these results, we find that our proposed CPAH is superior to all comparison methods on different

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2020.2963957, IEEE Transactions on Image Processing

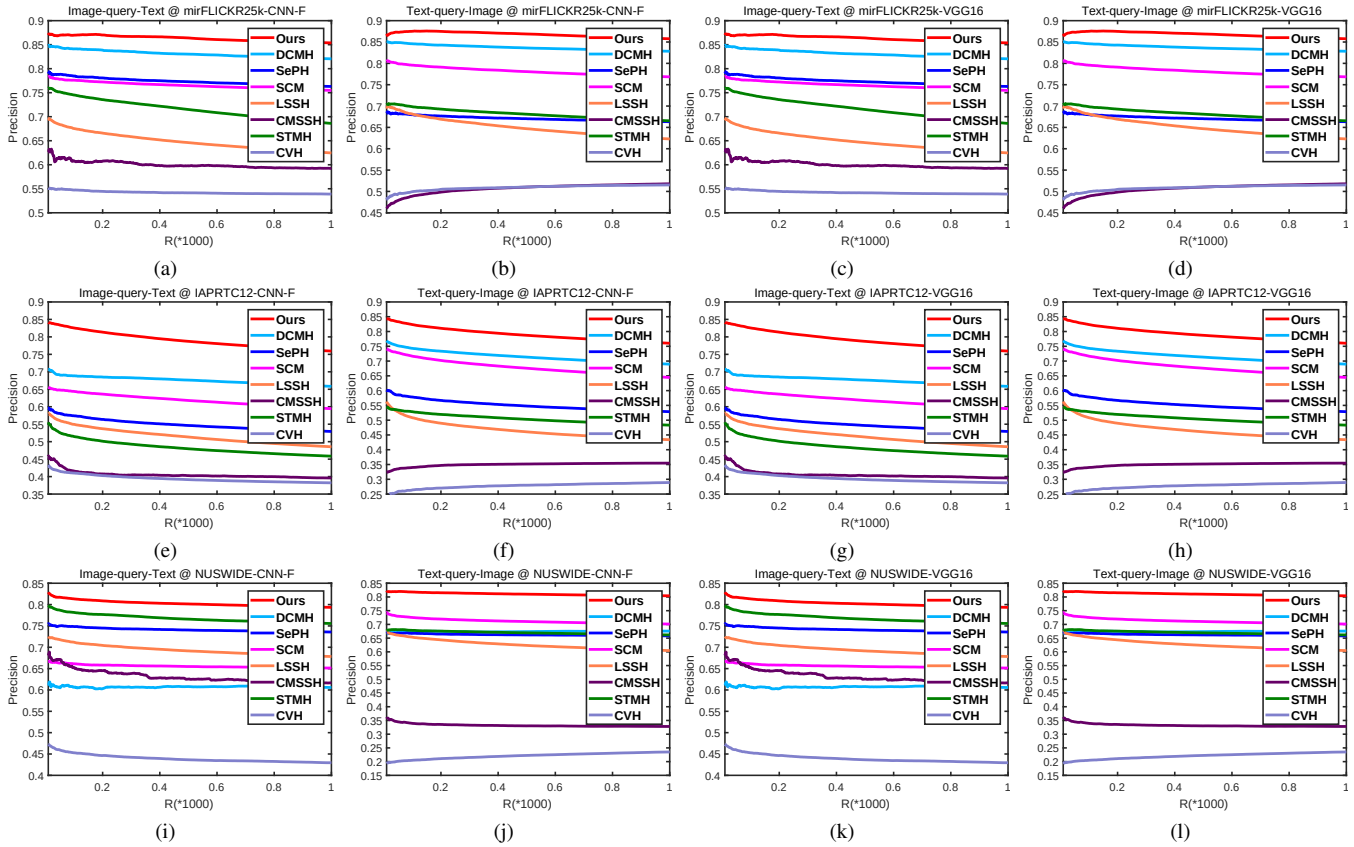


Fig. 3: The Top$N$-precision curves of different methods for image query text and text query image task on the MIRFLICKR-25K, IAPRTC-12 and NUS-WIDE datasets respectively (The code length is 64).

TABLE VI: The MAP scores of two retrieval tasks on the NUS-WIDE dataset with different hash code lengths. The baselines are based on CNN-F features and the best accuracy is shown in boldface.

TASK	Method	NUS-WIDE		
		16 bits	32 bits	64 bits
Image Query v.s. Text Database	CVH	0.391	0.384	0.379
	STMH	0.491	0.522	0.535
	CMSSH	0.422	0.399	0.385
	SCM	0.528	0.541	0.548
	LSSH	0.434	0.435	0.480
	SePH	0.592	0.595	0.619
	DCMH	0.526	0.526	0.535
	Ours	**0.607**	**0.627**	**0.634**
Text Query v.s. Image Database	CVH	0.357	0.355	0.353
	STMH	0.484	0.524	0.529
	CMSSH	0.409	0.405	0.394
	SCM	0.544	0.558	0.569
	LSSH	0.399	0.410	0.430
	SePH	0.561	0.559	0.580
	DCMH	0.577	0.584	0.602
	Ours	**0.642**	**0.662**	**0.665**

TABLE VII: The MAP scores of two retrieval tasks on the NUS-WIDE dataset with different hash code lengths. The baselines are based on VGG16 features and the best accuracy is shown in boldface

TASK	Method	NUS-WIDE		
		16 bits	32 bits	64 bits
Image Query v.s. Text Database	CVH	0.407	0.398	0.390
	STMH	0.499	0.520	0.548
	CMSSH	0.507	0.488	0.475
	SCM	0.542	0.555	0.568
	LSSH	0.481	0.495	0.527
	SePH	0.608	0.604	0.628
	DCMH	0.523	0.529	0.552
	Ours	**0.613**	**0.629**	**0.630**
Text Query v.s. Image Database	CVH	0.416	0.404	0.394
	STMH	0.506	0.540	0.550
	CMSSH	0.411	0.408	0.410
	SCM	0.554	0.572	0.584
	LSSH	0.445	0.452	0.459
	SePH	0.536	0.541	0.544
	DCMH	0.566	0.582	0.590
	Ours	**0.649**	**0.669**	**0.668**

retrieval tasks. For example, on IAPRTC-12 dataset, the MAP of our method is, on average, about ten percentage points higher than the second best algorithm. Compared with other methods, our CPAH can effectively capture semantic consistency information through consistency refined module and multi-task adversarial learning module, which significantly boosts the correlation among different modalities.

Fig. 2 and Fig. 3 illustrate the precision-recall curves and top$N$-precision curves on three datasets with 64-bit hash codes. It can be seen that the curves of CPAH are always higher than all other methods. These results are consistent with the MAP evaluation. In short, the proposed method

TABLE VIII: The Top-500 MAP results on IAPRTC-12 dataset under the experimental settings of [19].

TASK	Method	IAPR TC-12		
		16 bits	32 bits	64 bits
Image Query v.s. Text Database	CMDVH	0.7196	0.7727	0.8004
	SSAH	0.7898	0.7970	0.8171
	Ours	**0.8324**	**0.8681**	**0.8690**
Text Query v.s. Image Database	CMDVH	0.7348	0.7744	0.8038
	SSAH	0.7857	0.8185	0.8077
	Ours	**0.8134**	**0.8386**	**0.8598**

TABLE IX: The Top-100 MAP results on NUS-WIDE dataset under the experimental settings of [19].

TASK	Method	NUS-WIDE		
		16 bits	32 bits	64 bits
Image Query v.s. Text Database	CMDVH	0.8503	0.8755	0.8801
	SSAH	0.8507	0.8796	0.8965
	Ours	**0.8978**	**0.8898**	**0.9021**
Text Query v.s. Image Database	CMDVH	0.8270	0.8328	0.8403
	SSAH	0.8317	0.8408	0.8544
	Ours	**0.8940**	**0.9005**	**0.9065**

TABLE X: The Top-500 MAP results on IAPRTC-12 dataset under the experimental settings of [35] and [30].

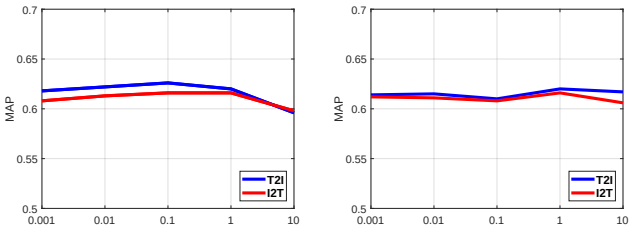
TASK	Method	IAPR TC-12		
		16 bits	32 bits	64 bits
Image Query v.s. Text Database	DVSH	0.6037	0.6395	0.6806
	DCMH	0.6594	0.6744	0.6905
	ADAH	0.7018	0.6893	0.6941
	Ours	**0.7613**	**0.7889**	**0.8056**
Text Query v.s. Image Database	DVSH	0.5696	0.6321	0.6964
	DCMH	0.5780	0.6061	0.6310
	ADAH	0.6464	0.6373	0.6668
	Ours	**0.7496**	**0.7727**	**0.7760**

TABLE XI: Ablation Study of consistency refined module and multi-task adversarial learning module on MIRFLICKR-25K dataset.

TASK	Method	MIRFLICKR-25K		
		16 bits	32 bits	64 bits
Image Query v.s. Text Database	No	0.733	0.745	0.751
	MA	0.772	0.783	0.785
	MA+CR	**0.775**	**0.791**	**0.787**
Text Query v.s. Image Database	No	0.749	0.764	0.766
	MA	0.773	0.781	0.786
	MA+CR	**0.777**	**0.787**	**0.789**

outperforms all of the comparisons.

We also make a comparison between our CPAH and other state-of-the-art methods: DVSH [35], CMDVH [19], SSAH [22] and ADAH [30]. Since the codes of these methods is not publicly available, we compare these methods by utilizing the same experimental settings. For CMDVH and SSAH, we follow the same experimental settings in [19] and directly cite the results from [22] for a fair comparison. The top-500 MAP results on IAPRTC-12 are shown in Table VIII and the Top-100 MAP results on IAPRTC-12 are shown in Table IX. It can be seen that our proposed method outperforms CMDVH as and as SSAH. In additional, The SSAH is also a adversarial learning based method. These results demonstrate that our method has more powerful capability to explore the semantic correlation and consistency.



(a) α (b) β

Fig. 4: The influence of hyper-parameters. (a) The MAP on the IAPRTC-12 dataset with 16 bits under different values of α. (b) The MAP on the IAPRTC-12 dataset with 16 bits under different values of β.

In order to compare with DVSH and ADAH, we implement our method under the same experimental settings in [35] and [30]. The top-500 MAP results of DVSH, DCMH, ADAH and our CPAH on IAPRTC-12 are illustrate in Table X. Notably, the DVSH adopts the LSTM recurrent neural network to extract sentences features for text representation, while DCMH, ADAH and our method only use bag-of-words vectors. The results in Table X can prove that our method can achieve better performance than DVSH, DCMH and ADAH in all cases, even we use the one-hot bag-of-words vectors.

D. Hyper-parameter Sensitivity Analysis

We studied the influence of the hyper-parameters α and β. Fig. 4 shows the MAP results on IAPRTC-12 dataset with different values of α and β, where the code length is 16 bits. When one parameter is analyzed, the other parameter is fixed. According to the results in Fig. 4, we finally choose $\alpha = 0.1$ and $\beta = 1$ for our hyper-parameters.

E. Ablation study

In order to justify the effectiveness of the proposed consistency refined module and multi-task adversarial learning module, we conduct an ablation study for them, which are performed on MIRFLICKR-25K dataset. The MAP results are shown in Table XI. The "No" means the original networks without consistency refined module and multi-task adversarial learning module. The "MA" means we only use multi-task adversarial learning module. It has about 3%-4% improvement in MAP compared to "No", which can prove the effectiveness of the proposed multi-task adversarial learning module. The "MA+CR" means we use both multi-task adversarial learning module and consistency refined module. It has about 1% improvement in performance compared to "MA", that can prove the effectiveness of the proposed consistency refined module.

V. Conclusion

In this paper, we present a novel deep hashing method, namely multi-task consistency-preserving adversarial Hashing (CPAH) for cross-modal retrieval. Compact hash codes of

image and text are generated in an end-to-end deep learning architecture. Specifically, a consistency refined module (CR) is designed to divide the representation of different modalities into uncorrelated parts: modality-common and modality-private representation. Furthermore, we design a multi-task adversarial learning module (MA), which not only make the modality-common representation of different modalities close to each other but also preserve the semantic consistency effectively. Finally, we employ each modality-common representation to generate compact and discriminative hash codes. Extensive experimental results on three datasets, MIRFLICKR-25K, IAPRTC-12 and NUS-WIDE, show that the proposed CPAH yields state-of-the-art performance in cross-modal retrieval task.

REFERENCES

[1] D. Tao, Y. Guo, M. Song, Y. Li, Z. Yu, and Y. Y. Tang, "Person re-identification by dual-regularized kiss metric learning," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2726–2738, 2016.

[2] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, 2015.

[3] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *CVPR*, 2012, pp. 2160–2167.

[4] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, 2016.

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

[6] 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, 2016.

[7] 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, 2017.

[8] X. Shi, F. Xing, K. Xu, M. Sapkota, and L. Yang, "Asymmetric discrete graph hashing," in *AAAI*, 2017.

[9] X. Shi, M. Sapkota, F. Xing, F. Liu, L. Cui, and L. Yang, "Pairwise based deep ranking hashing for histopathology image classification and retrieval," *Pattern Recogn.*, vol. 81, pp. 14–22, 2018.

[10] M. Sapkota, X. Shi, F. Xing, and L. Yang, "Deep convolutional hashing for low-dimensional binary embedding of histopathological images," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 805–816, 2018.

[11] W. Liu, C. Mu, S. Kumar, and S.-F. Chang, "Discrete graph hashing," in *NeurIPS*, 2014, pp. 3419–3427.

[12] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *ACM MM*. ACM, 2014, pp. 7–16.

[13] H. Liu, R. Ji, Y. Wu, F. Huang, and B. Zhang, "Cross-modality binary code learning via fusion similarity hashing," in *CVPR*, 2017.

[14] X. Liu, C. Deng, B. Lang, D. Tao, and X. Li, "Query-adaptive reciprocal hash tables for nearest neighbor search," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 907–919, 2016.

[15] X. Liu, Z. Li, C. Deng, and D. Tao, "Distributed adaptive binary quantization for fast nearest neighbor search," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5324–5336, 2017.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.

[17] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[18] C. Deng, Y. Xue, X. Liu, C. Li, and D. Tao, "Active transfer learning network: A unified deep joint spectral–spatial feature learning model for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1741–1754, 2019.

[19] V. Erin Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Cross-modal deep variational hashing," in *ICCV*, 2017, pp. 4077–4085.

[20] Q. Jiang and W. Li, "Deep cross-modal hashing," in *CVPR*, 2017.

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

[22] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *CVPR*, 2018, pp. 4242–4251.

[23] Y. Cao, B. Liu, M. Long, and J. Wang, "Cross-modal hamming hashing," in *ECCV*, 2018, pp. 202–218.

[24] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3893–3903, 2018.

[25] C. Deng, E. Yang, T. Liu, W. Liu, J. Li, and D. Tao, "Unsupervised semantic-preserving adversarial hashing for image search," *IEEE Trans. Image Process.*, 2019.

[26] E. Yang, T. Liu, C. Deng, and D. Tao, "Adversarial examples for hamming space search," *IEEE Trans. Cybern.*, 2018.

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

[28] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *NeurIPS*, 2016, pp. 343–351.

[29] 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.*, no. 99, pp. 1–12, 2018.

[30] X. Zhang, H. Lai, and J. Feng, "Attention-aware deep adversarial hashing for cross-modal retrieval," in *ECCV*, 2018, pp. 591–606.

[31] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *IJCAI*, vol. 22, no. 1, 2011, p. 1360.

[32] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *ACM SIGIR*, 2014, pp. 415–424.

[33] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *CVPR*, 2010, pp. 3594–3601.

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

[35] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *ACM SIGKDD*, 2016, pp. 1445–1454.

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

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

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

[39] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *CVPR*, 2018, pp. 5542–5551.

[40] D. Xie, C. Deng, H. Wang, C. Li, and D. Tao, "Semantic adversarial network with multi-scale pyramid attention for video classification," *arXiv preprint arXiv:1903.02155*, 2019.

[41] Y. Cao, B. Liu, M. Long, and J. Wang, "Hashgan: Deep learning to hash with pair conditional wasserstein gan," in *CVPR*, 2018, pp. 1287–1296.

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

[43] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[44] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *ACM CIVR*, 2008, pp. 39–43.

[45] H. J. Escalante, C. A. Hernández, J. A. Gonzalez, A. López-López, M. Montes, E. F. Morales, L. E. Sucar, L. Villaseñor, and M. Grubinger, "The segmented and annotated iapr tc-12 benchmark," *Comput. Vis. Image Understand.*, vol. 114, no. 4, pp. 419–428, 2010.

[46] 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 *ACM CIVR*, 2009, p. 48.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[49] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *CVPR*, 2015, pp. 3864–3872.

De Xie received the B.E. degree in Xi'an University of Architecture & Technology, Xi'an, China, in 2015. He is currently pursuing his Ph.D. degree at School of Electronic Engineering, Xidian University. His research interests focus on computer vision, natural language processing and multi-modal analysis.

Cheng Deng (M'11) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China. He is currently a Full Professor with the School of Electronic Engineering, Xidian University. He has authored or coauthored more than 80 scientific articles at top venues, including IEEE TNNLS, TIP, TCYB, TMM, TSMC, ICCV, CVPR, ICML, NIPS, IJCAI, and AAAI. His research interests include computer vision, pattern recognition, and information hiding.

Chao Li received the B.E. degree in Electronic and Information Engineering from Inner Mongolia University of Science & Technology, China, in 2014. He is currently pursuing his Ph.D. degree at School of Electronic Engineering, Xidian University. His main research interests include computer vision and machine learning.

Xianglong Liu received the B.S. and Ph.D. degrees in computer science from Beihang University, Beijing, in 2008 and 2014. From 2011 to 2012, he visited the Digital Video and Multimedia Laboratory, Columbia University as a joint Ph.D. student. He is currently an Assistant Professor with Beihang University. His research interests include machine learning, computer vision, and multimedia information retrieval.

Dacheng Tao (F'15) is currently a Professor of computer science and an ARC Laureate Fellow with the School of Computer Science, Faculty of Engineering and Information Technologies, and the Inaugural Director of the UBTECH Sydney Artificial Intelligence Centre, The University of Sydney. He mainly applies statistics and mathematics to artificial intelligence and data science. His research results have expounded in one monograph and more than 200 publications in prestigious journals and at prominent conferences, such as IEEE T-PAMI, T-IP, T-NNLS,T-CYB, IJCV, JMLR, NIPS, ICML, CVPR, ICCV, ECCV, ICDM, and ACM SIGKDD, with several best paper awards, such as the best theory/algorithm paper runner up award in IEEE ICDM 07, the Best Student Paper Award in IEEE ICDM 13, the 2014 ICDM 10-year highest-impact paper award, the 2017 IEEE Signal Processing Society Best Paper Award, and the Distinguished Paper Award in the 2018 IJCAI. He received the 2015 Australian Scopus-Eureka Prize and the 2018 IEEE ICDM Research Contributions Award. He is a Fellow of the Australian Academy of Science, AAAS, IAPR, OSA, and SPIE.