IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 September 2022, accepted 8 September 2022, date of publication 20 September 2022,
date of current version 27 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3208147

RESEARCH ARTICLE

Deep Label Feature Fusion Hashing for Cross-Modal Retrieval

DONGXIAO REN[1], WEIHUA XU[1], ZHONGHUA WANG[2], AND QINXIU SUN[1]
[1]School of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China
[2]Beijing Guodiantong Network Technology Company Ltd., Beijing 100083, China

Corresponding author: Dongxiao Ren (rendx29@163.com)

This work was supported in part by the Youth Fund Project of the National Natural Science Foundation of China under Grant 11801511.

ABSTRACT The rapid growth of multi-modal data in recent years has driven the strong demand for retrieving semantic-related data within different modalities. Therefore, cross-modal hashing has attracted extensive interest and studies due to its fast retrieval speed and good accuracy. Most of the existing cross-modal hashing models simply apply neural networks to extract the features of the original data, ignoring the unique semantic information attached to each data by the labels. In order to better capture the semantic correlation between different modal data, a novel cross-modal hashing model called deep label feature fusion hashing (DLFFH) is proposed in this article. We can effectively embed semantic label information into data features by building label networks in different modal networks for feature fusion. The fused features can more accurately capture the semantic correlation between data and bridge the semantic gap, thus improving the performance of cross-modal retrieval. In addition, we construct feature label branches and the corresponding feature label loss to ensure that the generated hash codes are discriminative. Extensive experiments have been conducted on three general datasets and the results demonstrate the superiority of the proposed DLFFH which performs better than most cross-modal hashing models.

INDEX TERMS Cross-modal retrieval, feature fusion, feature label branch, hashing.

I. INTRODUCTION

The exponential growth of various modal data such as text, image, video and audio has greatly promoted the development of cross-modal retrieval technique, which can retrieve the relevant data of other types when you input one type of data as the query [1], [2]. Generally speaking, the same event or concept can be described with data of different modalities. For example, we can use multi-modal data such as text and photos provided by news media or ordinary audiences to describe the concept *"Beijing Olympic Games"*. Although these different types of data have heterogeneous properties, they are semantically relevant and complement each other, which can be helpful for the users to better understand the target events or topics. However, due to the ''heterogeneous gap'' of various modal data, how to effectively implement cross-modal retrieval is still a challenging task.

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca.

Representational learning [3] is a usual way to eliminate the heterogeneity gap in cross-modal retrieval. It transforms different modal data into value representations in the same semantic space, and semantically similar data have more similar values. Many algorithms of such type are listed in literature [4], such as Canonical Correlation Analysis (CCA) learns the common space by maximizing the pairwise correlation between two sets of heterogeneous data. As the dimension of multi-modal data increases, the storage of large-scale data and the speed of data retrieval are very important. Therefore, the cross-modal hashing method has attracted extensive attention of a large number of researchers due to its lower memory cost and high retrieval speed. We can obtain the Hamming distance by performing a simple bit-wise XOR operations [5] on the two hash codes, and then judge whether the two original data are similar. A small Hamming distance means that the two data are similar and vice versa. The heterogeneity of different modalities makes it difficult to compare the similarity directly. Therefore, the main research

work at present is how to generate efficient hash codes to make cross-modal retrieval more accurate and extensible.

Traditional hashing models [6], [7], [8], [9], [10], [11] generate hash codes based on hand-crafted features, which lack sufficient discrimination ability and cannot represent original data effectively. In addition, a major disadvantage of these models is that the feature learning process and the hash code generation process are separated from each other. For the past few years, with the excellent performance of deep learning in feature extraction and representation, deep cross-modal hashing models [12], [13], [14], [15], [16], [17], [18] have made great progress. On this basis, experiments show that the features based on deep network learning are more representative than the traditional hashing models. However, most of the deep cross-modal hashing models use a single neural network to extract the features of the original data. This ignores the unique semantic information attached to each data, so the generated features cannot accurately represent the original data. The label of each data makes it unique in the dataset, so how to make full use of this effective information in the feature learning process is the key to improve the retrieval efficiency. In order to bridge the semantic gap, we construct label networks and embed semantic label information into data features through feature fusion. The fused features can better capture the semantic correlation between different modal data and improve the accuracy of the model. In addition, in order to make the hash codes generated by the model consistent and distinguishable, we divide the output end of each network into hash code branch and feature label branch to ensure that the hash codes with the same labels are as similar as possible and the hash codes with different labels are discriminative. The main contributions of this work can be summarized as follows:

- The proposed deep label feature fusion hashing (DLFFH) embeds semantic label information into data features through feature fusion between label networks and feature learning networks. In this way, we can better capture the semantic correlation between data and bridge the semantic gap. Our DLFFH integrates the data feature learning process and hash code generation process into a unified deep framework.
- We creatively divide the network output into hash code branch and feature label branch, and guide the generation of more discriminative hash codes according to the proposed feature label loss.
- Numerous experiments on three general datasets prove that this innovative DLFFH performs better than other models.

The rest of this article is organized as follows. Section 2 reviews the related work. Section 3 introduces the innovative DLFFH. The experimental results and corresponding analysis are presented in Section 4. Finally, the conclusion is given in Section 5.

II. RELATED WORK

Depending on whether data labels are applied to model training, cross-modal hashing can be generalized into two categories [19]. One is unsupervised models, the other is supervised models. Unsupervised models refer to the absence of data labels in the process of training retrieval models. To be specific, Collective Matrix Factorization Hashing (CMFH) [20] obtains one different modal common semantic space through collective matrix factorization, and then learns the hash mapping of each modality in this space. Latent Semantic Sparse Hashing (LSSH) [21] applies sparse encoding to process image data and matrix decomposition to process text data, followed by mapping into a common semantic space to learn hashing. Semantic Topic Multimodal Hashing (STMH) [22] applies clustering and matrix factorization to get semantic themes in image and text data respectively, and then learns the relationship between the two modalities data in common subspace through semantic topics. Finally, the mapping of original data to the common subspace is established to obtain the hash code representation.

In contrast, supervised cross-modal hashing can guide the generation of more representative hash codes by applying data labels in the training process. For example, Semantic Correlation Maximization (SCM) [23] guides hash code learning by calculating correlations between data labels of different modalities. Semantic Preserving Hashing (SePH) [24] translates the cross-modal distance of semantic similarity and Hamming space into two probability distributions respectively, and then gets the hash code mapping by decreasing relative entropy of both distributions. Cross-modality Metric Learning using Similarity-Sensitive Hashing (CMSSH) [25] applies boosting strategy to obtain similar hash codes between similar data.

In recent years, the development of deep learning technology provides a new direction for this field. Many cross-modal models have been innovated based on deep learning and achieved good performance. Deep Cross-Modal Hashing (DCMH) [26] builds a network that performs the entire process of converting hash codes from original data. Pairwise Relationship Guided Deep Hashing (PRDH) [27] maintains similarities between and within modalities of data. Deep Multi-Level Semantic Hashing (DMSH) [28] constructs a high-level semantic supervision matrix in the training process, which contains more information than the general similarity matrix. Self-Supervised Adversarial Hashing (SSAH) [29] successfully combines self-supervised networks and adversarial learning into a network. Mask Cross-Modal Hashing (MCMH) [30] applies Mask R-CNN to extract image features. Deep Multiscale Fusion Hashing (DMFH) [31] extracted convolution features at different scales for each image data to represent the image data more accurately. Triplet-Based Deep Cross-Modal Retrieval (TDCMR) [32] applies the improved triplet constraint to generate more accurate hash codes. Semantics-Preserving Hashing based on Multi-Scale Fusion (SPHMF) [33] constructs pairwise loss and inter-modal loss of tag generation network to guide hash code learning. Multi-attention based Semantic Deep Hashing (MSDH) [34] designs a multi attention block to extract more semantic related features from the data.

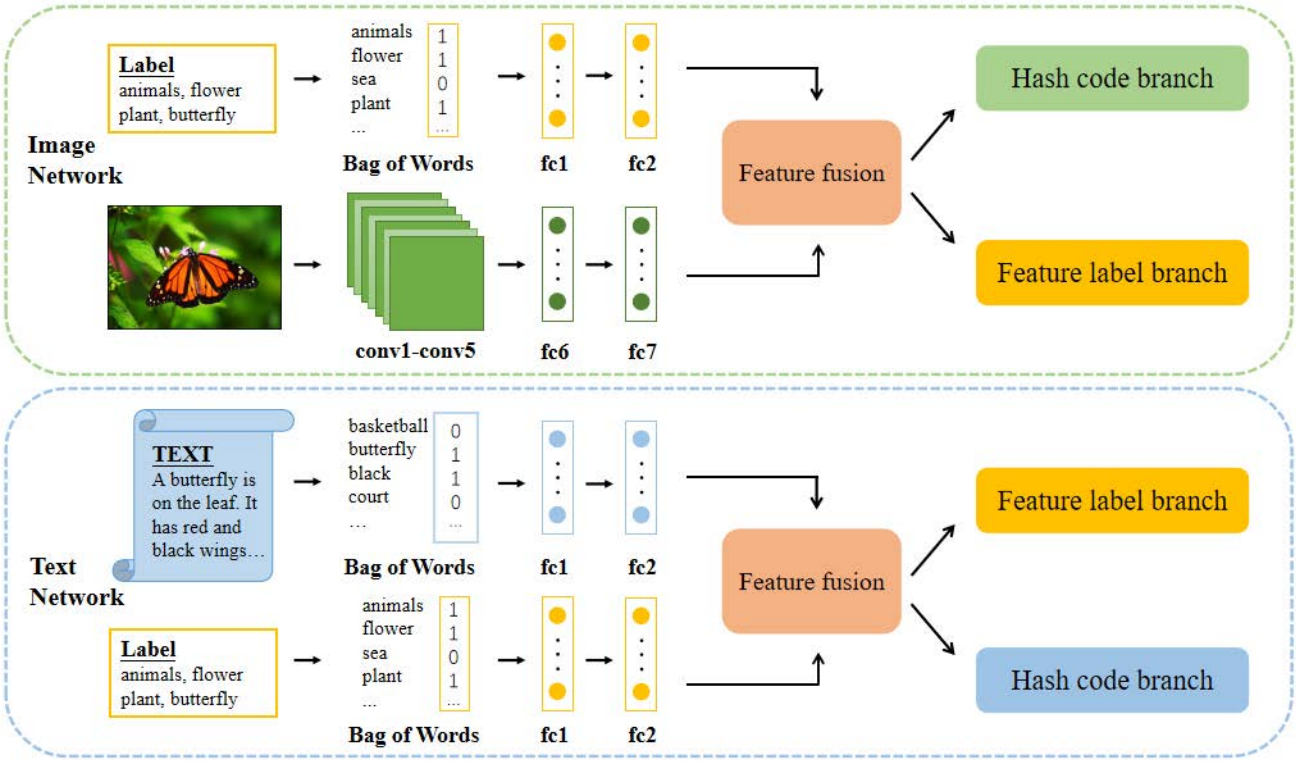


FIGURE 1. Network structure of DLFFH.

Multi-label Modality enhanced Attention based self-supervised deep Cross-modal Hashing (MMACH) [35] first used multi-label cross-modal triplet loss to guide hash code learning, and proposed multi-label modality enhanced attention module to integrate multi-modal data features and label features. Vision and Scene Text Aggregation for Cross-Modal Retrieval (ViSTA) [36] proposed an effective visual and scene text aggregation transformer for cross-modal retrieval. Learning the Best Pooling Strategy for Visual Semantic Embedding [37] learns the best pool strategy to automatically adapt to different data and features through generalized pool operator. Discrete Joint Semantic Alignment Hashing (DJSAH) [38] obtains a distinctive hash code by integrating the high-level semantics of the data.

Although the above models can show good performance, there are still some aspects to be improved. The innovations of our model are as follows: First, DLFFH embeds semantic label information into data features through feature fusion, so that each data feature has its unique semantic label attribute and can more accurately represent original data. Second, we apply hash code branches and feature label branches to generate more discriminative hash codes, where the hash codes of the same labels are more similar, and vice versa.

III. PROPOSED DLFFH

In this section, we will introduce the DLFFH and discuss it in the two most frequently used modalities: image and text. Figure 1 shows the network structure of our DLFFH, which is divided into two segments: image network and text network. We demonstrate the details of the model in the following section.

A. NOTATION

In this article, vectors are represented by lowercase bold letters (e.g., $\mathbf{m}$) and matrices are represented by uppercase bold letters (e.g., $\mathbf{M}$). $\mathbf{M}$ transpose is $\mathbf{M}^T$, the element in *ith* row and *jth* column of matrix $\mathbf{M}$ is represented by M_{ij}. $\|\cdot\|_F$ denotes the Frobenius norm. sign $(\cdot)$ represents the symbolic function, which outputs -1 if its input is negative else outputs 1.

B. PROBLEM DEFINITION

Suppose there exist N data pairs made up of images and text. Let $\mathbf{X} = \{x_1,\ldots,x_N\} \in \mathbb{R}^{N\times D_x}$ refers to the image data, where D_x denotes the dimension of x_i. $\mathbf{Y} = \{y_1,\ldots,y_N\} \in \mathbb{R}^{N\times D_y}$ stands for the text data, where D_y is the dimension of y_j. $\mathbf{L} = \{l_1,\ldots,l_N\} \in \{0,1\}^{N\times C}$ refers to the label matrix, we apply C to stands for the total number of label categories. $\mathbf{S}$ stands for the semantic similarity matrix, $S_{ij} = 1$ means that x_i and y_j have at least one same label. Conversely, they are dissimilar and $S_{ij} = 0$. The Hamming distance between two hash codes reflects their similarity of image data and text data, a small Hamming distance means that two data are similar and vice versa. For different hash codes, we can use the following

formula to calculate their Hamming distance:

$$dis\left\langle \mathbf{h}_i, \mathbf{h}_j \right\rangle = \frac{1}{2}(k - \left\langle \mathbf{h}_i, \mathbf{h}_j \right\rangle) \tag{1}$$

where $\left\langle \mathbf{h}_i, \mathbf{h}_j \right\rangle$ denotes the inner product of two vectors. k denotes hash code length, $\mathbf{h}_i$ and $\mathbf{h}_j$ stands for image x_i hash code and text y_j hash code respectively.

Given dataset $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{L}$ and its semantic similarity matrix $\mathbf{S}$, the DLFFH model can train two hash functions: $f(x_i, l_i)$ and $g(y_j, l_j)$ for image modality and text modality respectively. Therefore, each modal data can generate the corresponding hash code according to its hash function.

C. FEATURES LEARNING PART

For image network, it is composed of image feature learning network, label network, hash code branch and feature label branch. Specifically, we select the first seven layers in the CNN-F [39] model as the image feature learning network, including five convolution layers and two fully connected layers. The initialization parameters of the image feature learning network are trained on ImageNet [40] in advance, and we can obtain the basic image features through this network. The label network is a two-layer fully connected network (4096→4096), which is applied to extract the unique semantic label information of each data. Then feature fusion (concatenating the label features and the image features) is performed to embed the semantic label information into the image features, as shown in Figure 2. Finally, the fused features are connected to two fully connected networks (hash code branch and feature label branch) to generate corresponding hash codes and feature labels, where the number of neurons is hash code length and the label category number respectively.

For text network, it includes text feature learning network, label network, hash code branch and feature label branch. We first apply Bag-of-Words model to convert the text data and label data into vector representations that can be extracted by text network. The text feature learning network is a two-layer fully connected network (8192→4096) for learning the text features of data. The remaining label network, hash code branch and feature label branch are the same as those in the image network. Under the action of feature fusion, the semantic label information in the text data can be effectively embedded into text features to generate more accurate text representations.

D. HASH CODE GENERATION PART

To ensure that the hash code generated by DLFFH can reflect the relationship between original data more accurately, the objective function can be set into three parts: semantic similarity loss, feature label loss and hash code discrete loss. In this article, $\mathbf{U}_{*i} = f\left(x_i, l_i; \theta_x, \theta_{x\_hash}\right)$ denotes the image feature output by hash code branch in image network, where θ_x represents the total parameters of image feature learning network and label network, and θ_{x_hash} represents the parameters of the hash code branch. $\mathbf{L}^x_{*i} = f\left(x_i, l_i; \theta_x, \theta_{x\_label}\right)$ denotes the image feature label output by the feature label branch in image network, where θ_x represents the total parameters of image feature learning network and label network, and θ_{x_label} represents the parameters of the feature label branch. Furthermore, $\mathbf{V}_{*j} = g\left(y_j, l_j; \theta_y, \theta_{y\_hash}\right)$ denotes the text feature output by hash code branch in text network, where θ_y represents the total parameters of text feature learning network and label network, and θ_{y_hash} represents the parameters of the hash code branch. $\mathbf{L}^y_{*j} = g\left(y_j, l_j; \theta_y, \theta_{y\_label}\right)$ refers to the text feature label output by the feature label branch in text network, where θ_y represents the total parameters of text feature learning network and label network, and θ_{y_label} represents the parameters of the feature label branch.

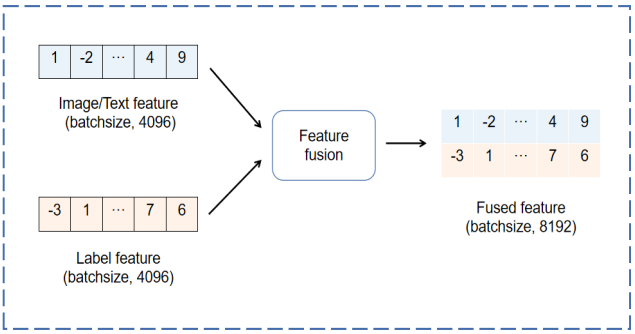


FIGURE 2. Process of feature fusion.

Semantic gaps between different modal data make it impossible to compare directly, so we measure the data similarity by mapping them into a common semantic space. The likelihood function of image data feature and text data feature is shown below:

$$p(S_{ij}|\mathbf{U}_{*i}, \mathbf{V}_{*j}) = \begin{cases} \sigma(\Phi_{ij}), & S_{ij} = 1 \\ 1 - \sigma(\Phi_{ij}), & S_{ij} = 0. \end{cases} \tag{2}$$

where $\Phi_{ij} = \frac{1}{2}\mathbf{U}^T_{*i}\mathbf{V}_{*j}$ and $\sigma(\Phi_{ij}) = \frac{1}{1+e^{-\Phi_{ij}}}$, when $S_{ij} = 1$ denotes the inner product (similarity) between $\mathbf{U}_{*i}$ and $\mathbf{V}_{*j}$ is larger and vice versa. To facilitate the training of model, we apply the negative log likelihood function (semantic similarity loss J_s) of the above equation to get the similarity between image data and text data:

$$J_s = -\sum_{i,j=1}^{N}(S_{ij}\Phi_{ij} - \log(1 + e^{\Phi_{ij}})) \tag{3}$$

where $\Phi_{ij} = \frac{1}{2}\mathbf{U}^T_{*i}\mathbf{V}_{*j}$. Minimizing the negative log likelihood (equivalent to maximizing the likelihood function) above can reduce the Hamming distance between similar image data and text data, thereby improving the accuracy of the model.

Furthermore, we improve the discrimination of hash codes by constraining the distance between the feature label matrix generated by the feature label branch and label matrix. Effectively making hash codes with the same label more similar and vice versa. The feature label loss is defined as follows:

$$J_l = \left\|\mathbf{L}^x - \mathbf{L}\right\|^2_F + \left\|\mathbf{L}^y - \mathbf{L}\right\|^2_F \tag{4}$$

where $\mathbf{L}^x$ refers to image feature label matrix, $\mathbf{L}^y$ refers to text feature label matrix.

Next, there is a certain quantization error when the continuous variables output from the network are converted into hash binary codes. And balancing the -1 and 1 values can effectively maximize the information of hash codes. Therefore, we propose the hash code discrete loss:

$$J_q = \left\|\mathbf{H}^x - \mathbf{U}\right\|_F^2 + \left\|\mathbf{H}^y - \mathbf{V}\right\|_F^2 + \|\mathbf{UE}\|_F^2 + \|\mathbf{VE}\|_F^2 \tag{5}$$

where $\mathbf{H}^x = \text{sign}(\mathbf{U})$, $\mathbf{H}^y = \text{sign}(\mathbf{V})$, $\mathbf{E}$ denotes a vector with all values of 1. Inspired by Jiang and Li [26], we let $\mathbf{H} = \mathbf{H}^x = \mathbf{H}^y$ in the training phase.

Finally, in combination with the semantic similarity loss J_s, feature label loss J_l and hash code discrete loss J_q, the objective function of DLFFH is shown below:

$$\min_{\mathbf{H},\theta_x,\theta_{x\_label},\theta_{x\_hash},\theta_y,\theta_{y\_label},\theta_{y\_hash}} J = J_s + \gamma J_l + \eta J_q \tag{6}$$

where γ, η denote the hyper-parameters.

E. OPTIMIZATION

On account of the hash binary code $\mathbf{H}$ is discrete variable, we apply the alternate learning strategy to settle the problem that $\mathbf{H}$ is not easy to optimize. In each step, only update the parameters in one modality at a time and fix other parameters. The back-propagation (BP) algorithm based on mini-batch stochastic gradient descent (SGD) is applied to update the algorithm. Algorithm 1 summarizes the optimization procedure of DLFFH.

1) OPTIMIZE θ_x, θ_{x_hash} AND θ_{x_label}, WITH OTHER PARAMETERS FIXED

For each image modal data x_i, the derivative of the objective function can be obtained:

$$\frac{\partial J}{\partial \mathbf{U}_{*i}} = \frac{1}{2}\sum_{j=1}^{N}(\sigma(\Phi_{ij})\mathbf{V}_{*j} - S_{ij}\mathbf{V}_{*j}) + 2\eta(\mathbf{U}_{*i} - \mathbf{H}_{*i} + \mathbf{UE}) \tag{7}$$

$$\frac{\partial J}{\partial \mathbf{L}^x_{*i}} = 2\gamma(\mathbf{L}^x_{*i} - \mathbf{L}_{*i}) \tag{8}$$

Then we can apply the chain rule to derive $\frac{\partial J}{\partial \theta_x}$, $\frac{\partial J}{\partial \theta_{x\_hash}}$ and $\frac{\partial J}{\partial \theta_{x\_label}}$.

2) OPTIMIZE θ_y, θ_{y_hash} AND θ_{y_label}, WITH OTHER PARAMETERS FIXED

For each text modal data y_j, the derivative of the objective function can be obtained:

$$\frac{\partial J}{\partial \mathbf{V}_{*j}} = \frac{1}{2}\sum_{i=1}^{N}(\sigma(\Phi_{ij})\mathbf{U}_{*i} - S_{ij}\mathbf{U}_{*i}) + 2\eta(\mathbf{V}_{*j} - \mathbf{H}_{*j} + \mathbf{VE}) \tag{9}$$

Algorithm 1 Optimization Procedure of DLFFH

Input: Image set $\mathbf{X}$, text set $\mathbf{Y}$, label set $\mathbf{L}$ and semantic similarity matrix $\mathbf{S}$.

Output: Parameters θ_x, θ_{x_hash}, θ_{x_label}, θ_y, θ_{y_hash}, θ_{y_label} of two networks, and hash code matrix $\mathbf{H}$.

Initialization

Initialize parameters θ_x, θ_{x_hash}, θ_{x_label}, θ_y, θ_{y_hash}, θ_{y_label}, γ, η, mini-batch size N_x, N_y, maximum iteration number T_{max}, image network iteration number $T_x = \lceil n/N_x \rceil$ and text network iteration number $T_y = \lceil n/N_y \rceil$.

repeat

for *iter* = 1, 2, $\cdots$,T_x **do**

Randomly select N_x samples from $\mathbf{X}$.

Calculate $\mathbf{U}_{*i} = f\left(x_i, l_i; \theta_x, \theta_{x\_hash}\right)$ and $\mathbf{L}^x_{*i} = f\left(x_i, l_i; \theta_x, \theta_{x\_label}\right)$ by forward propagation.

Compute the corresponding derivatives using (7), (8).

Update θ_x, θ_{x_hash} and θ_{x_label} by BP algorithm.

end for

for *iter* = 1, 2, $\cdots$,T_y **do**

Randomly select N_y samples from $\mathbf{Y}$.

Calculate $\mathbf{V}_{*j} = g\left(y_j, l_j; \theta_y, \theta_{y\_hash}\right)$ and $\mathbf{L}^y_{*j} = g\left(y_j, l_j; \theta_y, \theta_{y\_label}\right)$ by forward propagation.

Compute the corresponding derivatives using (9), (10).

Update θ_y, θ_{y_hash} and θ_{y_label} by BP algorithm.

end for

Learn $\mathbf{H}$ using (12).

until a fixed number of iterations

$$\frac{\partial J}{\partial \mathbf{L}^y_{*j}} = 2\gamma(\mathbf{L}^y_{*j} - \mathbf{L}_{*j}) \tag{10}$$

Then we can apply the chain rule to derive $\frac{\partial J}{\partial \theta_y}$, $\frac{\partial J}{\partial \theta_{y\_hash}}$ and $\frac{\partial J}{\partial \theta_{y\_label}}$.

3) OPTIMIZE $\mathbf{H}$, WITH OTHER PARAMETERS FIXED

The objective function is equivalent to the following formula:

$$\max_{\mathbf{H}} tr(\mathbf{H}^T(\eta(\mathbf{U}+\mathbf{V}))) = tr(\mathbf{H}^T\mathbf{P}) = \sum_{i,j} H_{ij}P_{ij}$$
$$s.t. \qquad \mathbf{H} \in \{-1, +1\}^{k\times N} \tag{11}$$

where $\mathbf{P} = \eta(\mathbf{U}+\mathbf{V})$. Therefore, the hash code matrix can be optimized by the following formula:

$$\mathbf{H} = \text{sign}(\eta(\mathbf{U}+\mathbf{V})) \tag{12}$$

F. OUT-OF-SAMPLE EXTENSION

The trained DLFFH can generate hash codes for the data outside the training set. We can take an instance of anyone modality as the input of the network and generate the corresponding hash code through forward propagation. Specifically, when given an instance x_q of image modality and its corresponding label l_q, its hash code can be obtained by the

following formula:

$$\mathbf{h}_q^x = \text{sign}(f(x_q, l_q; \theta_x, \theta_{x\_hash})) \tag{13}$$

Similarly for text modality y_q, we have:

$$\mathbf{h}_q^y = \text{sign}(g(y_q, l_q; \theta_y, \theta_{y\_hash})) \tag{14}$$

IV. EXPERIMENTS

A. DATASETS

The MIRFLICKR-25K dataset [41] contains 25000 image-text pairs collected from the Flickr website, and each data pair is associated with a corresponding label. We eliminate the data with less than 20 text descriptions in the dataset and convert the text descriptions into 1386-dimensional Bag-of-Words (BOW) text vectors. A total of 20015 data pairs of data are used in the experiment after processing, each data pair contains at least one of the 24 labels.

The NUS-WIDE dataset [42] is a collection of 269648 data pairs, each of which contains an image, text description and corresponding labels. Here, 195834 data belonging to the 21 most common labels are selected for the experiment. The text description of each data is converted into a 1000-dimensional Bag-of-Words vector.

The IAPR TC-12 dataset [43] contains 20000 image-text pairs, and each data pair has at least one of 255 labels. Each text description is converted into a 2912-dimensional Bag-of-Words vector. Table 2 summarizes the detailed settings of the above three datasets in the experiment.

B. EVALUATION PROTOCOL AND BASELINE

1) EVALUATION PROTOCOL

In this article, we apply two classical cross-modal hashing evaluation protocols: Hamming ranking and hash lookup to verify the validity of DLFFH.

Hamming ranking refers to the ascending order of the Hamming distance between the query data and the retrieval dataset. The accuracy of Hamming ranking can be calculated by applied the Mean Average Precision (MAP) [44], which can be obtained by averaging the average accuracy. The MAP calculation equation is as follow:

$$\text{MAP} = \frac{1}{n}\sum_{i=1}^{n}\text{AP}(q_i) \tag{15}$$

$$\text{AP} = \frac{1}{K}\sum_{s=1}^{N}\text{M}(s)\text{R}(s) \tag{16}$$

where n denotes the number of query data, q_i refers to the *ith* query data, N denotes the number of retrieved data. K refers to the number of retrieved data related to query data. $M(s)$ is the accuracy of the first s retrieved data. If the *sth* retrieved data is similar to the query data, $R(s) = 1$, otherwise $R(s) = 0$.

The hash lookup protocol returns retrieved results within the specified Hamming radius, and its performance is measured by the Precision-Recall curve. We can obtain the corresponding precision and recall by changing the Hamming radius, and draw the Precision-Recall curve on this basis.

2) BASELINE

In order to prove the performance of DLFFH, we compare it with seven currently representative models. According to the model structure, these can be divided into hand-crafted models (CMFH [20], SCM [23], SePH [24]) and deep network models (DCMH [26], SSAH [29], MSDH [34], MMACH [35]).

C. IMPLEMENTATION DETAILS

In this article, we build the DLFFH based on the TensorFlow framework. Except that the image feature learning network adopts the trained parameters, other network parameters are randomized. The hyper-parameters of the objective function is set to: $\gamma = 1$, $\eta = 0.1$, the detailed hyper-parameter analysis will be explained in the following sections. The mini-batch size is 128 and the number of model training iterations is 300. The learning rate decreases from 10^{-2} to 10^{-6} with the increase of iterations. For all models, we run five times in turn to get the average.

For activation functions applied in DLFFH, we apply identity function in hash code branches. Sigmoid function is adopted in feature label branches, and the remaining neural networks all apply the Rectified Linear Unit (ReLU) [45].

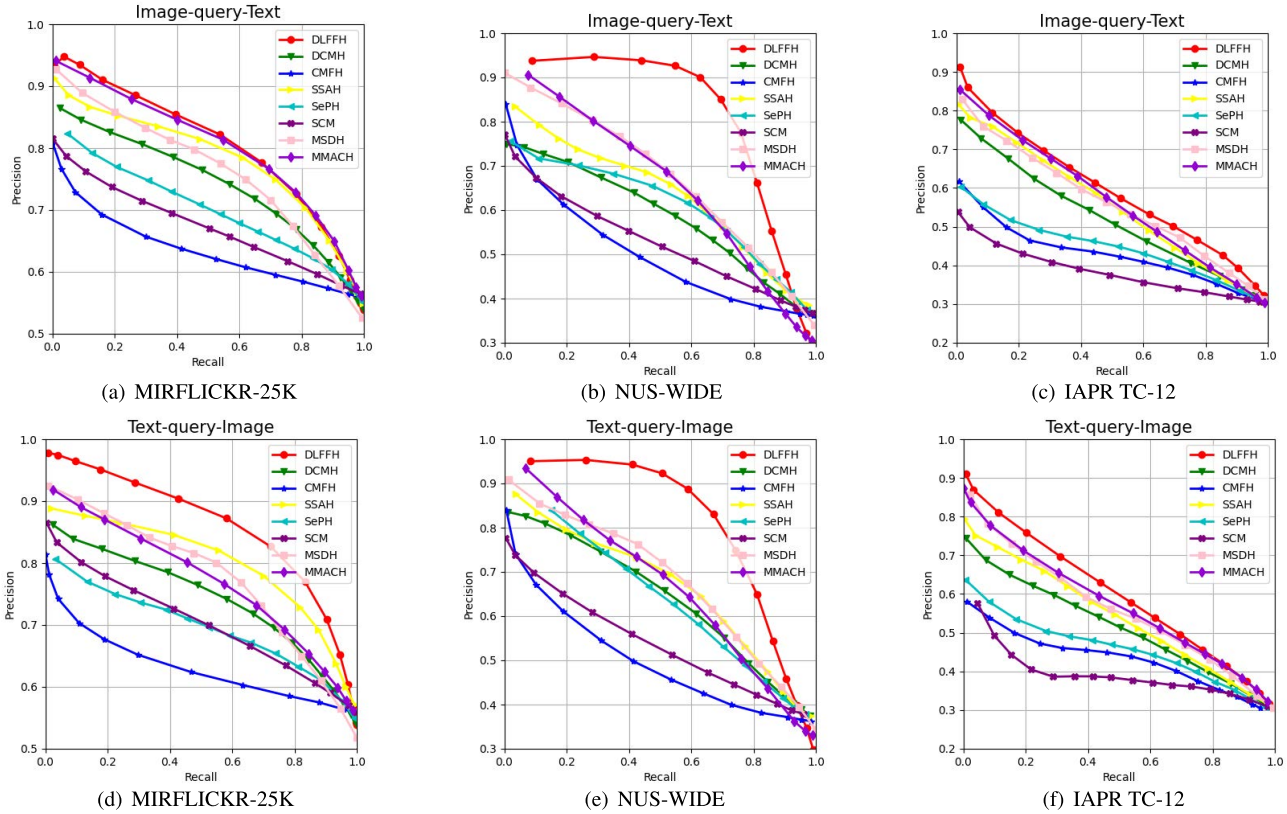
D. PERFORMANCE

1) HAMMING RANKING

Table 1 records the MAP values (16 bits, 32 bits, 64 bits) of DLFFH and seven baselines in two cross-modal retrieval tasks on MIRFLICKR-25K, NUS-WIDE and IAPR TC-12 datasets. "I→T" refers to apply images to retrieve the corresponding text, and "T→I" refers to apply text to retrieve the corresponding images. It can be seen from the table that the MAP values of DLFFH on the three datasets are greater than those of other baselines, achieving excellent performance. Compared with hand-crafted models, deep network models perform better because of their excellent performance in feature learning process. Specifically, on MIRFLICKR-25K, compared with the most representative deep network model DCMH, the MAP values of DLFFH on the two retrieval tasks increased by 9.25%/11.02% on average, and increased by 1%/8.62% on average compared with the most advanced MMACH. On NUS-WIDE, the MAP for "I→T"/"T→I" achieves an average increase of 34.64%/23.50% and 23.13%/21.28% compared with DCMH and MMACH. Similarly, there is an average increase in 18.10%/16.77% (DCMH) and 2.55%/7.34% (MMACH) on IAPR TC-12, demonstrating the effectiveness of DLFFH. Although the deep network models can achieve good performance, they lack the unique label features of each data in the feature learning process. On the contrary, we embed the semantic label information into the hash code through feature fusion and set feature label branches to further increase the discrimination of hash codes. Therefore, the performance of DLFFH can be effectively improved. In addition, the performance of most models is positively correlated with the length

TABLE 1. Performance comparison of MAP values.

Task	Model	MIRFLICKR-25K			NUS-WIDE			IAPR TC-12		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
I→T	CMFH [20]	0.5526	0.5865	0.5907	0.4427	0.4527	0.4623	0.4042	0.4168	0.4198
	SCM [23]	0.6225	0.6379	0.6508	0.4807	0.4845	0.4882	0.3641	0.3655	0.3713
	SePH [24]	0.6571	0.6652	0.6717	0.5752	0.5838	0.5902	0.4365	0.4472	0.4548
	DCMH [26]	0.7413	0.7462	0.7549	0.5903	0.6031	0.6093	0.4837	0.4926	0.5218
	SSAH [29]	0.7801	0.7837	0.7879	0.6322	0.6368	0.6397	0.5318	0.5419	0.5681
	MSDH [34]	0.7532	0.7635	0.7813	0.6363	0.6585	0.6832	0.5315	0.5483	0.5728
	MMACH [35]	0.7981	0.8107	0.8169	0.6459	0.6567	0.6689	0.5412	0.5783	0.6057
	OURS	**0.8023**	**0.8212**	**0.8263**	**0.7941**	**0.8117**	**0.8216**	**0.5531**	**0.5919**	**0.6243**
T→I	CMFH [20]	0.5638	0.5949	0.5972	0.4515	0.4548	0.4614	0.4193	0.4251	0.4267
	SCM [23]	0.6801	0.6889	0.6941	0.4895	0.4917	0.5073	0.3657	0.3723	0.3749
	SePH [24]	0.7183	0.7247	0.7278	0.5883	0.5943	0.6124	0.4489	0.4544	0.4603
	DCMH [26]	0.7792	0.7783	0.7805	0.6389	0.6511	0.6571	0.4970	0.5153	0.5379
	SSAH [29]	0.7912	0.7927	0.7948	0.6691	0.6732	0.6793	0.5339	0.5494	0.5701
	MSDH [34]	0.7812	0.7849	0.7980	0.6589	0.6822	0.7082	0.5338	0.5541	0.5736
	MMACH [35]	0.7886	0.7940	0.8069	0.6448	0.6679	0.6703	0.5343	0.5629	0.5892
	OURS	**0.8487**	**0.8695**	**0.8775**	**0.7816**	**0.8031**	**0.8205**	**0.5705**	**0.6068**	**0.6329**



(a) MIRFLICKR-25K (b) NUS-WIDE (c) IAPR TC-12

(d) MIRFLICKR-25K (e) NUS-WIDE (f) IAPR TC-12

FIGURE 3. Precision-recall curves with 16 bits hash codes.

of hash codes, which indicates that longer hash codes can contain more discrimination information.

2) HASH LOOKUP

Figure 3 plots the Precision-Recall curves with 16 bits hash codes on three general datasets. The area under the Precision-Recall curve is positively correlated with the performance of the model. We can see that the Precision-Recall curve of DLFFH is higher than other curves, which fully demonstrates that our innovative model is superior to other baselines and further verifies the results of MAP comparison.

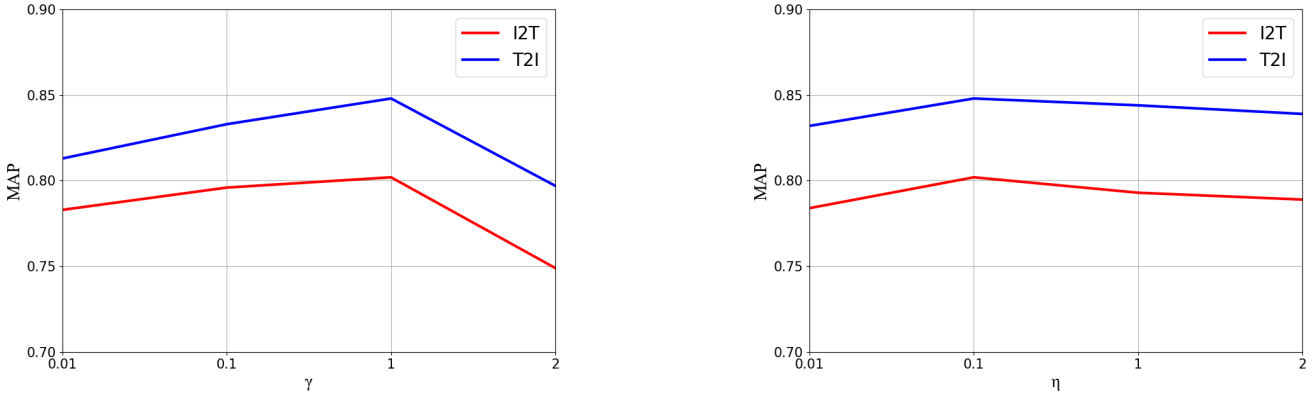


FIGURE 4. Sensitivity analysis of hyper-parameters with 16 bits hash codes.

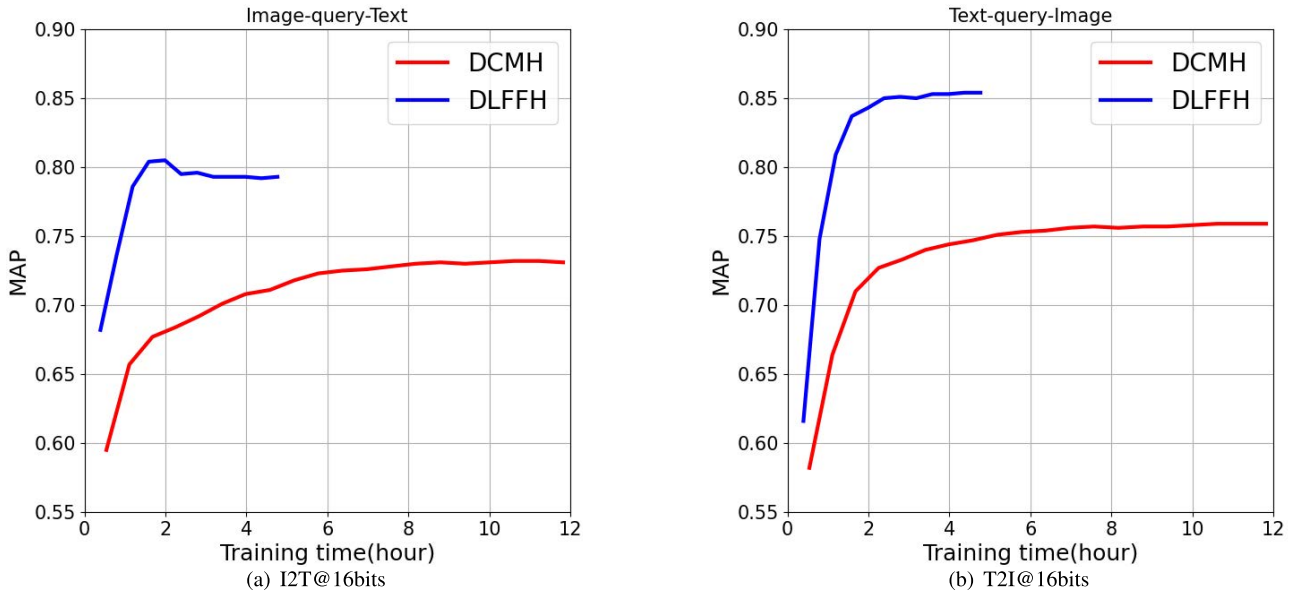


FIGURE 5. Training efficiency of DLFFH and DCMH. The code length is 16 bits.

TABLE 2. Detailed setup of datasets.

Dataset	Total	Training	Query	Labels
MIRFLICKR-25K	20015	10000	2000	24
NUS-WIDE	195834	10000	2000	21
IAPR TC-12	20000	10000	2000	255

E. SENSITIVITY TO PARAMETERS

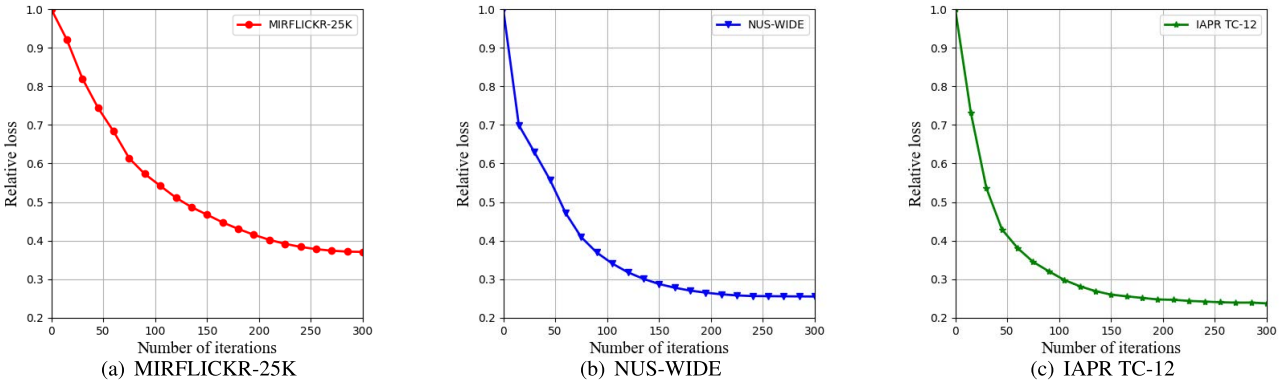
In this section, we perform sensitivity analysis on the hyper-parameters γ and η. The experimental process is divided into two stages, and only one of the hyper-parameters is changed in each stage. The variation range of hyper-parameters is set to 0.01, 0.1, 1, 2. The experimental results are shown in Figure 4, where we take MIRFLICKR-25K as the training set and the hash code length is 16 bits. We can see that DLFFH is not sensitive to parameters, which relatively proves the stability and validity of our model. Therefore, we set the hyper-parameters to $\gamma = 1$ and $\eta = 0.1$.

TABLE 3. MAP comparison of DLFFH and its variants.

Task	Method	MIRFLICKR-25K		
		16bits	32bits	64bits
I→T	DLFFH	0.8023	0.8212	0.8263
	DLFFH-1	0.6768	0.7103	0.7064
	DLFFH-2	0.7925	0.8167	0.8247
T→I	DLFFH	0.8487	0.8695	0.8775
	DLFFH-1	0.7132	0.7148	0.7238
	DLFFH-2	0.8401	0.8551	0.8641

F. ABLATION STUDY

To verify the effect of feature fusion and feature label branches in DLFFH, we design two variants for comparison: (a) DLFFH-1 removes the label network and feature fusion in each modality; (b) DLFFH-2 is constructed by deleting the feature label branch and only retaining the hash code branch.



(a) MIRFLICKR-25K (b) NUS-WIDE (c) IAPR TC-12

FIGURE 6. Convergence curve of DLFFH.

Table 3 shows the MAP values of DLFFH and its variants in different retrieval tasks on MIRFLICKR-25K dataset, from which we can find: (1) the MAP performance of DLFFH is better than that of DLFFH-1. We can conclude that feature fusion can effectively embed the semantic label information learned by the label network into the data features and make up the semantic gap. Therefore, the fused features with semantic label information can more accurately represent the original data and improve the efficiency of the model. (2) DLFFH performs better than DLFFH-2. The reason is that more discriminative hash codes can be generated under the action of feature label branches, which improves the performance to a certain extent.

G. TRAINING EFFICIENCY

Figure 5 shows the changes of MAP with training time between DLFFH and DCMH on MIRFLICKR-25K. It can be seen from the figure that DLFFH can train a model with higher accuracy in a shorter time and achieve convergence faster. Compared with DCMH, DLFFH can better capture the semantic correlation between different modal data and bridge the modal gap through feature fusion. Therefore, the training efficiency of our model can be effectively improved.

H. CONVERGENCE ANALYSIS

To verify the convergence of DLFFH, we perform experiments on three datasets with 16 bits hash codes. The experimental results are shown in Figure 6. In order to better intuitively show the convergence of the model, we apply relative loss to record. The relative loss is to divide all the objective function values by the first iteration value. It can be seen from the figure that the relative loss of DLFFH decreases rapidly and converges gradually with the increase of iteration times, which also verifies that the model has excellent training efficiency.

V. CONCLUSION

In this article, we propose an innovative model called DLFFH. Compared with other models, DLFFH embeds semantic label information into the feature learning process through label network and feature fusion, which can make the data features generated by the network more representative. This can more effectively capture the semantic correlation and make up the semantic gap between multi-modal data. In addition, the feature label branch makes the generated hash codes more discriminative. Numerous experimental results on three general datasets prove that the deep label feature fusion hashing achieves satisfactory performance.

REFERENCES

[1] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," 2016, *arXiv:1607.06215*.
[2] L. Jin, X. Shu, K. Li, Z. Li, G.-J. Qi, and J. Tang, "Deep ordinal hashing with spatial attention," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2173–2186, May 2019.
[3] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10386–10395.
[4] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.
[5] G. Mikriukov, M. Ravanbakhsh, and B. Demir, "Unsupervised contrastive hashing for cross-modal retrieval in remote sensing," 2022, *arXiv:2204.08707*.
[6] F. Shen, C. Shen, W. Liu, and H. T. Shen, "Supervised discrete hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 37–45.
[7] 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. Int. Conf. Manage. Data (SIGMOD)*, 2013, pp. 785–796.
[8] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. 22nd Int. Joint Conf. Artif. Intell. (IJCAI)*, Barcelona, Spain, Jul. 2011, pp. 1360–1365.
[9] K. Li, G.-J. Qi, J. Ye, and K. A. Hua, "Linear subspace ranking hashing for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1825–1838, Sep. 2017.
[10] M. Rastegari, J. Choi, S. Fakhraei, D. Hal, and L. Davis, "Predictable dual-view hashing," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, Atlanta, GA, USA, 2013, pp. 1328–1336.
[11] Y. Zhen and D.-Y. Yeung, "Co-regularized hashing for multimodal data," in *Proc. 25th Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1376–1384.
[12] X. Zou, X. Wang, E. M. Bakker, and S. Wu, "Multi-label semantics preserving based deep cross-modal hashing," *Signal Process., Image Commun.*, vol. 93, Apr. 2021, Art. no. 116131.
[13] H. Qiang, Y. Wan, Z. Liu, L. Xiang, and X. Meng, "Discriminative deep asymmetric supervised hashing for cross-modal retrieval," *Knowl.-Based Syst.*, vol. 204, Sep. 2020, Art. no. 106188.
[14] H. Peng, J. He, S. Chen, Y. Wang, and Y. Qiao, "Dual-supervised attention network for deep cross-modal hashing," *Pattern Recognit. Lett.*, vol. 128, pp. 333–339, Dec. 2019.

[15] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1445–1454.
[16] Y. Shi, X. You, F. Zheng, S. Wang, and Q. Peng, "Equally-guided discriminative hashing for cross-modal retrieval," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4767–4773.
[17] C. Wang, H. Yang, and C. Meinel, "Deep semantic mapping for cross-modal retrieval," in *Proc. IEEE 27th Int. Conf. Tools Artif. Intell. (ICTAI)*, Washington, DC, USA, Nov. 2015, pp. 234–241.
[18] X. Wu, T. Wang, and S. Wang, "Cross-modal learning based on semantic correlation and multi-task learning for text-video retrieval," *Electronics*, vol. 9, no. 12, p. 2125, Dec. 2020.
[19] Q.-Y. Jiang and W.-J. Li, "Discrete latent factor model for cross-modal hashing," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3490–3501, Jul. 2019.
[20] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2083–2090.
[21] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2014, pp. 415–424.
[22] D. Wang, X. Gao, X. Wang, and L. He, "Semantic topic multimodal hashing for cross-media retrieval," in *Proc. 24th Int. Joint Conf. Artif. Intell. (IJCAI)*, Buenos Aires, AR, USA, Jul. 2015, pp. 3890–3896.
[23] D. Zhang and W. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2177–2183.
[24] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3864–3872.
[25] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3594–3601.
[26] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3270–3278.
[27] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *Proc. AAAI Conf. Art. Intell.*, Feb. 2017, pp. 1618–1625.
[28] Z. Ji, W. Yao, W. Wei, H. Song, and H. Pi, "Deep multi-level semantic hashing for cross-modal retrieval," *IEEE Access*, vol. 7, pp. 23667–23674, 2019.
[29] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4242–4251.
[30] Q. Lin, W. Cao, Z. He, and Z. He, "Mask cross-modal hashing networks," *IEEE Trans. Multimedia*, vol. 23, pp. 550–558, 2020.
[31] X. Nie, B. Wang, J. Li, F. Hao, M. Jian, and Y. Yin, "Deep multiscale fusion hashing for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 401–410, Jan. 2021.
[32] J. Song, Y. Lin, J. Song, W. Yu, and L. Zhang, "TDCMR: Triplet-based deep cross-modal retrieval for geo-multimedia data," *Appl. Sci.*, vol. 11, no. 22, p. 10803, Nov. 2021.
[33] H. Zhang and M. Pan, "Semantics-preserving hashing based on multi-scale fusion for cross-modal retrieval," *Multimedia Tools Appl.*, vol. 80, no. 11, pp. 17299–17314, May 2020.
[34] L. Zhu, G. Tian, B. Wang, W. Wang, D. Zhang, and C. Li, "Multi-attention based semantic deep hashing for cross-modal retrieval," *Int. J. Speech Technol.*, vol. 51, no. 8, pp. 5927–5939, Aug. 2021.
[35] X. Zou, S. Wu, N. Zhang, and E. M. Bakker, "Multi-label modality enhanced attention based self-supervised deep cross-modal hashing," *Neurocomputing*, vol. 467, pp. 138–162, Jan. 2022.
[36] M. Cheng, Y. Sun, L. Wang, X. Zhu, K. Yao, J. Chen, G. Song, J. Han, J. Liu, E. Ding, and J. Wang, "ViSTA: Vision and scene text aggregation for cross-modal retrieval," 2022, *arXiv:2203.16778*.
[37] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang, "Learning the best pooling strategy for visual semantic embedding," 2020, *arXiv:2011.04305*.
[38] T. Yao, X. Kong, H. Fu, and Q. Tian, "Discrete joint semantic alignment hashing for cross-modal image-text search," *IEEE Trans. Cybern.*, vol. 50, no. 12, pp. 4896–4907, Dec. 2020.
[39] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," 2014, *arXiv:1405.3531*.
[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and A. C. Berg, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[41] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retr.*, 2008, pp. 39–43.
[42] 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 Retr.*, 2009, pp. 1–9.
[43] 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.
[44] W. Liu, C. Mu, S. Kumar, and S.-F. Chang, "Discrete graph hashing," in *Proc. 27th Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 3419–3427.
[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1106–1114.

DONGXIAO REN was born in Nanyang, Henan, China, in 1982. She received the B.S. degree in computer science and technology from Henan University, in 2005, the M.S. degree in computer software and theory from Southwest Jiaotong University, in 2008, and the Ph.D. degree in information and communication engineering from the University of Electronic Science and Technology, in 2012. From 2012 to 2017, she was a Senior Engineer with State Grid Ningxia Electric Power Company and successfully completed dozens of projects. Since 2017, she has been an Assistant Professor with the School of Science, Zhejiang University Science and Technology, Hangzhou, China. She is the author of more than ten papers. Her research interests include image processing, pattern recognition, and big data processing technology.

WEIHUA XU was born in Quanzhou, Fujian, China, in 1998. He received the B.S. degree in industrial engineering from the Xiamen University of Technology, China, in 2020. He is currently pursuing the master's degree in applied statistics with the Zhejiang University of Science and Technology, Hangzhou, China. His major research interests include computer vision, deep learning, and cross-modal retrieval.

ZHONGHUA WANG was born in Zhoukou, Henan, China, in 1980. He received the B.S. degree in computer science and technology from Henan University, in 2005, and the M.S. degree in business administration from Ningxia University, in 2015. He has worked in industry for decades and accumulated rich project experience. He is currently a Project Manager with Beijing Guodiantong Network Technology Company Ltd. He has finished more than ten projects. His research interest includes information processing and machine learning.

QINXIU SUN was born in Shandong, China, in 1980. She received the Ph.D. degree from Zhejiang University, in 2011. Since 2012, she has been working with the Zhejiang University of Science and Technology, Hangzhou, China, where she has been an Associated Professor, since 2013. She is the author of more than 20 papers. She has hosted two Natural Science Foundations of China and two Natural Science Foundations of Zhejiang Province. Her research interests include quantum groups and Lie algebra.