IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 December 2023, accepted 2 January 2024, date of publication 10 January 2024,
date of current version 19 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3352434

RESEARCH ARTICLE

Deep Hashing Similarity Learning for Cross-Modal Retrieval

YING MA[1], MENG WANG[1], GUANGYUN LU[2], AND YAJUN SUN[1]
[1]Key Laboratory of Intelligent Information Processing and Graph Processing, Guangxi University of Science and Technology, Liuzhou 545000, China
[2]College of Information Science and Engineering, Liuzhou Institute of Technology, Liuzhou 545000, China

Corresponding author: Guangyun Lu (170024422@qq.com)

ABSTRACT In the realm of cross-modal retrieval research, hash methods have garnered significant attention from scholars due to their high retrieval efficiency and low storage costs. However, these methods often sacrifice a considerable amount of semantic features when mapping multi-modal characteristics to a low-dimensional space. Moreover, the focus of hash learning has primarily been on inter-modal similarity learning, neglecting the importance of intra-modal similarity learning. To address these issues, this paper proposes a novel cross-modal hash method called Deep Hashing Similarity Learning for Cross-modal Retrieval (DHSL). DHSL incorporates relation networks into the hash method, enabling pairwise matching between images and texts. This approach effectively bridges the heterogeneity gap between images and texts while simultaneously emphasizing the intra-modal similarity information within both modalities. The result is a hash similarity matrix that captures both inter-modal similarity and intra-modal discriminability. Considering that the process of transforming high-dimensional features into hash codes often leads to a loss of important semantic information, we introduce a feature selector to enhance the features. This selector filters out distinctive features from the original feature set and combines them with low-dimensional features to complement the semantic information. Moreover, we introduce weighted cosine triplet loss and quantization loss to constrain the hash representation in the Hamming space, thereby learning high-quality hash codes. Comprehensive experimental results on two benchmark datasets, NUS-WIDE and MIRFlickr25K, demonstrate that DHSL outperforms the state-of-the-art cross-modal hash methods.

INDEX TERMS Cross-modal retrieval, relation network, feature enhancement.

I. INTRODUCTION

With the development of big data, the internet generates a vast amount of multimedia data every day, including texts, images, and videos. Due to heterogeneous differences in data distribution, different modalities manifest disparities, creating an urgent demand for accurate retrieval from multimedia data. As a result, cross-modal retrieval has emerged as an attractive and challenging research field. Among various cross-modal retrieval applications, image-to-text (text-to-image) retrieval is the most widely used, where a text (image) query sample is provided, with the expectation of retrieving images (texts) that contain semantically relevant information from a database.

One commonly used approach for cross-modal retrieval in large-scale data is based on hashing, in which high-dimensional features are stored in binary codes through dimensionality reduction. This allows for similar binary codes for related samples across different modalities, resulting in more accurate cross-modal retrieval. Early cross-modal hashing methods typically relied on manually designed feature extraction methods [1], [2], such as SIFT and HOG, which have limited generalization capability and lower retrieval performance. In recent years, with the widespread application of deep neural networks in cross-modal retrieval, experimental results have shown that neural networks have superior performance in feature extraction compared to shallow methods. Although deep neural networks have achieved significant progress in cross-modal hashing methods [3], [4], [5], [6], [7], [8], [9], there are still several issues that need

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

improvement. Firstly, there are heterogeneity differences among different modalities, and maintaining similarity across modalities is one of the key challenges. Secondly, many cross-modal hashing methods use deep neural networks to convert features into hash codes, potentially resulting in the loss of some semantic information. Thirdly, while most methods focus on constructing cross-modal similarity matrices to preserve similarity between modalities, they often overlook intra-modal similarity instances.

To tackle the mentioned issues, we propose a novel approach for cross-modal retrieval called Deep Hashing Similarity Learning (DHSL). Inspired by relation networks [25], [26], [27], we incorporate relation networks into the cross-modal hashing method. The main contributions of our work can be summarized as follows:

1) We design a feature selector that enhances features using a class residual connection method [28], [29], [30]. Specifically, we employ norm-based feature selection to identify relevant features, and then integrate the selected features with low-dimensional features through a class residual connection. This integration ensures that the hash representation contains more semantic information.
2) In our hash learning approach, we introduce a relational network. Firstly, we directly perform pairwise similarity learning on the enhanced features of images and texts, reducing the loss of semantic information in the hash codes. Secondly, we conduct intra-modality similarity learning on the fused features of low-dimensional and high-dimensional spaces, learning discriminative hash codes.
3) Furthermore, we employ weighted cosine triplet loss and quantization loss to constrain the hash codes. By combining the above loss functions, we optimize the model and enable DHSL to learn high-quality hash codes.

II. RELATED WORK

The essential principle of cross-modal hashing lies in mapping features of different modalities into a shared Hamming space, where the similarity between different modalities can be measured, enabling cross-modal retrieval. The crucial aspects of this process involve the selection of features and the measurement of their similarity, aiming for increased retrieval accuracy.

Cross-modal hashing could be divided into unsupervised hashing and supervised hashing. Unsupervised hashing [10], [11], [12] refers to the process of converting different modal data into hash codes to measure similarity without labeled information. For example, Song et al. [13] sought a common Hamming space that could map multimedia data information, enabling different media data to learn unified hash codes in this space. Ding et al. [2] used collective matrix factorization to learn different view information and mapped them into unified hash codes. Considering the discreteness of hash codes, Wang et al. [14] first learned semantic features (multiple semantic topics or concepts) of multimedia data, mapped these features to a common subspace, and then directly generated hash codes. Supervised hashing [17], compared to unsupervised hashing, utilizes supervised information such as labels or semantic information to further improve cross-modal retrieval performance. Lin et al. [18] added semantic-related supervision into the training data and minimized the Hamming distance of hash codes in the Hamming space by transforming it into a probability distribution, thus obtaining hash codes that preserved semantic structure. Zhang et al. [19] seamlessly integrated semantic labels into the process of hash learning to maximize semantic relevance of modal data. Cao et al. [20] used data category labels as supervised information and learned superior hash codes by maximizing inter-class distance while minimizing intra-class variance. Ma et al. [34] used semantic similarity relationships to learn binary codes and learned hash codes bit by bit through alternating updates, making hash codes more similar.

Compared to superficial approaches, deep learning has the ability to extract features with non-linear structures, demonstrating superior performance compared to manually extracting features. Jiang and Li [3] proposed Deep Cross-Modal Hashing (DCMH), which integrated feature extraction and hash codes learning into a framework using deep convolutional neural networks, enabling end-to-end learning. Yang et al. [4] further enhanced DCMH by considering the similarity between instances within each modality and applying pairwise constraints to different types, enhancing the discriminative power of the hash codes. Additionally, Li et al. [6] introduced labels as input information to the neural network, converting labels into binary codes to constrain the hash codes. Yao et al. [35] utilized collaborative filtering to mine the relationship between labels and hash codes, which to some extent reduced memory consumption and enhanced cross-modal alignment through image attributes, thereby improving the quality of hash codes. Wang et al. [21] introduced the concept of adversarial learning to cross-modal retrieval and proposed Adversarial Cross-Modal Retrieval (ACMR). ACMR has prompted many researchers to combine adversarial learning with hashing methods [22], [23], [24] and achieved significant results. AGAH [8] utilized adversarial learning to guide the multi-label attention module in learning feature representations and employed a binary code mapping for multi-label semantic information, leading to improved retrieval performance. DADH [9] employed adversarial learning in both feature learning and hash codes learning, ensuring consistency in cross-modal feature representation through dual adversarial learning. In contrast to adversarial learning, our proposed approach introduces the relational network mechanism in the DRSL [27] method and optimizes it, resulting in improved retrieval performance.

III. METHOD

Without loss of generality, a collection of n instances consisting of images and texts pairs is considered. The image instance is denoted by $X = \{x_i\}_{i=1}^{n}$ and the text instance by $Y = \{y_i\}_{i=1}^{n}$, where x and y represent the image and the text, respectively. $L = \{l_i\}_{i=1}^{n}$ represents the multi-label information, where $l_i = [l_{i1}, l_{i2}, \ldots, l_{ic}]$, and c corresponds to the number of categories. If the i-th instance belongs to the j-th class, then $l_{ij} = 1$; otherwise $l_{ij} = 0$.

Given the training sets X, Y, L, the goal of DHSL is to acquire a hashing function and hash codes for images and texts, as well as a similarity hash matrix R. The DHSL framework is illustrated in Figure 1.

A. FEATURE LEARNING MODULE

1) FEATURE LEARNING

The feature learning part consists of two neural networks, one for learning the image features and another for learning the text features. For the image modality, we utilize the CNN-F network pre-trained on ImageNet to extract 4,096-dimensional deep features from images, with fixed parameters. Since image features contain a lot of redundant information, we set up a 3-layer fully connected network to extract high-level semantic features from images. The last layer of the network serves as the hash layer, mapping image features to a low-dimensional feature space.

In regards to the text characteristics, we begin by utilizing the Bag-of-Words (BoW) model to transform each text into a feature vector, with the specific dimensions determined based on the dataset. Subsequently, we establish a three-layer fully connected neural network to extract high-level semantic information from the text, and employ a hash layer to map the text features to a lower-dimensional feature space.

We shall employ the aforementioned feature learning network as our primary network, with $F^X = f(X; \theta_X)$ and $F^Y = f(Y; \theta_Y)$ denoting the feature projection functions for the images and texts, respectively. Here, F^X and F^Y represent the output image and text features within the primary network. θ_X and θ_Y signify the parameters of the image and the text feature projectors.

B. FEATURE ENHANCEMENT

The process of projecting the original features onto a low-dimensional space through the primary network results in the loss of some semantic information. To address this issue, we propose the utilization of a feature selector, composed of three linear layers, which differentiates from fully connected layers in feature learning. We employ the L_{21}-norm regularization on the feature selector parameters, resulting in a sparse weight matrix. At this stage, the feature selector is able to identify features that contain discriminative edge semantics. Subsequently, these distinctive features are connected in a residual manner to the low-dimensional features, complementing the missing features and achieving feature enhancement. Therefore, the loss function for feature enhancement can be expressed as:

$$L_U = \mu_1 \|U_X\|_{21} + \mu_2 \|U_Y\|_{21} . \tag{1}$$

In this context, the matrices representing the sparsity parameters are denoted as μ_1 and μ_2, and the matrix of sparse weights is denoted as U_*. We can represent the feature-enhanced feature as $H^* = F^* + \sigma M^*, s.t. * \in (X, Y)$. M represents the output of the feature selector, F represents the output of the main network, and σ denotes the weight parameters.

C. HASHING LEARNING MODULE

In the hashing learning phase, a non-linear transformation on image and text features is performed by the hashing layer using the tanh function. It maps the image and text features to a hash code representation ranging from -1 to 1. In the testing phase, the model converts the image and text hash code representations into binary codes using the sign function. The sign function is defined as:

$$sign(x) = \begin{cases} +1, & x > 0 \\ -1, & x \leq 0. \end{cases} \tag{2}$$

The image and text hash codes can be represented as $B^X = sign\left(H^X\right), B^Y = sign\left(H^Y\right)$, respectively. In order to ensure consistency between the hash codes and cross-modal features, we apply quantization loss to enforce balanced constraints on the hash codes.

$$L_q = \left\|B^X - H^X\right\|_F^2 + \left\|B^Y - H^Y\right\|_F^2 . \tag{3}$$

In order to acquire the capability of maintaining the consistency between modalities and the discriminability within each modality, our approach to hash learning consists of two parts. On one hand, we leverage relation networks to conduct similarity learning between image and text hash codes, aiming to learn a similarity hash matrix that captures inter-modality and intra-modality relationships. On the other hand, by employing a weighted cosine triplet loss in the Hamming space, we learn high-quality hash codes that possess similar semantics across different modalities.

1) RELATION NETWORK

In DRSL [27], the author employed a relational network to directly calculate the similarity between images and texts, representing the similarity of each image and text using scalar values, which heavily compromised the modal semantic features. In DHSL, the relational network outputs a similarity matrix that represents the similarity between images and texts, and uses Euclidean distance to measure the distance between the similarity matrix and a priori similarity matrix, thereby encoding more semantic features in the similarity matrix. For the image and text modalities, we utilize a fusion mechanism to directly match and fuse the enhanced features of the image and text, and then calculate the similarity of paired samples using the relational network. The relational network function is represented as $R^{**} = r(G^{**}, \theta_r)$, where G^{**} is the output result of cross-modal feature fusion, with a fusion approach of

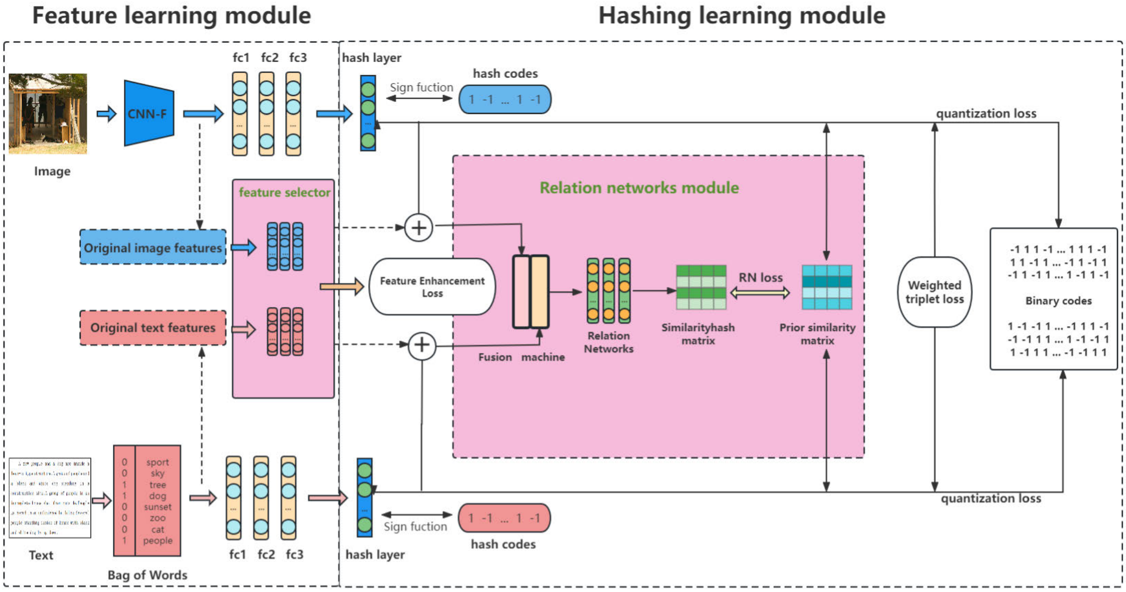


FIGURE 1. The DHSL framework, as depicted in Figure 1, consists of two main components: the feature learning module and the hashing learning module. The feature learning module is further divided into feature learning for the image modality and feature learning for the text modality. The hashing learning module includes hash code learning for relation networks and hash code learning for Hamming space.

concatenation. θ_r represents the parameters of the relational network, and R^{**} denotes the hash similarity matrix calculated by the relational network. For inter-modality fusion, the fused binary code matrix can be represented as $G^{XY} = \left\{G_{pq}^{XY} \mid p = 1, \ldots, n_i; q = 1, \ldots, n_t\right\}$, where G_{pq}^{XY} is the fused binary code of the p-th image and q-th text. On the other hand, the semantic information conveyed by the label matrix reveals that it defines a similarity of 1 for samples of the same class and a similarity of 0 for samples of different classes. S_{pq} represents the similarity between the p-th image and q-th text, defined as follows:

$$S_{pq} = \begin{cases} 1, & l_p^x = l_q^y \\ 0, & otherwise \end{cases} \tag{4}$$

In the module of relational network, we optimize the similarity distance model by minimizing the hash-based similarity matrix R^{XY} and the prior similarity matrix S^{XY}. The corresponding loss function can be represented as:

$$L_1 = \left\|R^{XY} - S^{XY}\right\|_F^2. \tag{5}$$

In the realm of modality, we integrate the low-dimensional spatial features of images and texts together with the enhanced features of the feature enhancement component into binary code representations, G^{XX} and G^{YY}. By utilizing the relation network, we obtain the similarity matrices R^{XX} and R^{YY} for images and texts respectively. Consequently, the modality-intrinsic similarity loss within the relation network can be defined as follows:

$$L_2 = \left\|R^{XX} - S^{YY}\right\|_F^2 + \left\|R^{YY} - S^{YY}\right\|_F^2. \tag{6}$$

In the domain of relational network modules, although the hash similarity matrices may have different forms, they convey similar similarity information. Thus, the similarity hash matrix can be expressed as $R = R^{XX} + R^{YY} + R^{XY}$, and the overall loss of the relational network is:

$$L_{RN} = L_1 + L_2. \tag{7}$$

2) WEIGHTED COSINE TRIPLET LOSS

In the context of multi-label data, where a sample can belong to multiple categories, which results in weak discriminative power between different modalities. Hence, we propose the utilization of a weighted cosine triplet loss [9] to measure the similarity between instances of hash codes, aiming to increase the distance between samples with different semantic meanings while reducing the distance among those with the same semantic meaning.

In the context of the image modality, we construct triplets in the form of $\left(x_i, y_j^+, y_k^-\right)$, where x_i represents an image sample, and y_j^+ and y_k^- represent positive and negative text samples respectively. The positive text samples have similar semantics to the image sample, while the negative text samples have opposite semantics. Typically, cosine distance is used to measure the similarity between instances in the triplet samples. To further accurately describe the association information between data points and multi-class labels, semantic ordering of hash codes in the Hamming space is performed. A weight factor is computed for the ranked hash codes based on the NDGG evaluation criterion, defined as follows:

$$\omega\left(rel_j, rel_k\right) = \frac{2^{rel_j} - 2^{rel_k}}{Z}, \tag{8}$$

TABLE 1. mAP results for cross-modal retrieval task on MIRFlickr25K dataset.

Method	I→T 16bits	 32bits	 64bits	T→I 16bits	 32bits	 64bits
CMSSH [1]	0.5847	0.5842	0.5724	0.5671	0.5688	0.5612
SCM [19]	0.6558	0.6698	0.6732	0.6653	0.6799	0.6862
SePH [18]	0.6707	0.6736	0.6747	0.6907	0.6961	0.6990
DCMH [3]	0.7301	0.7371	0.7413	0.7533	0.7683	0.7737
CMHH [16]	0.7312	0.7226	0.7285	0.7256	0.7335	0.7347
AGAH [8]	0.7969	0.7999	0.8104	0.7952	0.7989	0.8086
SCAHN [15]	0.8138	0.8250	0.8264	0.7971	0.8093	0.8159
DADH [9]	0.8110	0.8165	0.8265	0.7993	0.8051	0.8243
DCMHT [33]	0.8190	0.8229	0.8220	0.8037	0.8137	0.8235
DHSL	**0.8224**	**0.8345**	**0.8361**	**0.8176**	**0.8269**	**0.8355**

where rel_i represents the similarity level of the i-th data point in the sorted list, and Z is the normalization constant, the weighted cosine triplet loss for image modality can be expressed as:

$$L_{X\rightarrow Y} = \sum_{i,j,k} \omega(r_j, r_k)(max(cos(x_i, y_k^-) - cos(x_i, y_j^+) + m, 0)). \tag{9}$$

Similarly, the textual modal weighted cosine triplet loss is defined as follows:

$$L_{Y\rightarrow X} = \sum_{i,j,k} \omega(r_j, r_k)(max(cos(y_i, x_k^-) - cos(y_i, x_j^+) + m, 0)), \tag{10}$$

where ω represents the weight factor, m denotes the marginal parameter, and cos represents the cosine distance. The higher the weight factor, the stronger the semantic relevance between the query sample and the data point from another modality. The overall weighted cosine triplet loss function can be expressed as:

$$L_{tri} = L_{X\rightarrow Y} + L_{Y\rightarrow X}. \tag{11}$$

D. OPTIMIZATION

The overall loss function of the DHSL approach consists of the losses for feature enhancement, relational network, weighted cosine triplet, and quantization. The comprehensive loss function is represented as follows:

$$\min_{\theta_*,\theta_r,R,B^*,U_*} L_{tatal} = \alpha L_{tri} + \beta L_U + \eta L_{RN} + \varphi L_q, \tag{12}$$

where θ_X, θ_Y are the parameters of the main network, θ_r is the parameter of the relationship network, and θ_* represents the parameters of the sparse weight matrix. We optimize the overall objective function using stochastic gradient descent, and a detailed summary of the optimization process is presented in Algorithm 1.

Algorithm 1 Learning of DHSL

```
Input: Training set $X, Y, L$, learning rate $\xi$;
Output: Optimized parameters $\theta_X, \theta_Y$ and $\theta_r$, sparse matrix
    $U_*$, hash similarity matrix $R$ and hash codes $B^*$;
Initialize the parameter $\theta_X, \theta_Y$ and $\theta_r$ hyper-parameters,
    hash similarity matrix $R$ and hash codes $B^*$;
repeat
    for iteration=1,2,...,$\frac{n}{m}$ do
        Update parameters $\theta_*$ by using back propagation:
        $\theta_* \leftarrow \theta_* - \xi \frac{\partial L_{tatal}}{\partial \theta_*}$ $s.t. * \in (X, Y)$;
        Update parameters $U_*$ by using back propagation:
        $U_* \leftarrow U_* - \xi \frac{\partial L_{tatal}}{\partial U_*}$ $s.t. * \in (X, Y)$;
        Update parameters $\theta_r$ by using back propagation:
        $\theta_r \leftarrow \theta_r - \xi \frac{\partial L_{tatal}}{\partial \theta_r}$;
    end for
    calculate $R$;
    calculate $B^*$;
until convergence;
```

TABLE 2. mAP results for cross-modal retrieval task on NUS-WIDE dataset.

Method	I→T 16bits	 32bits	 64bits	T→I 16bits	 32bits	 64bits
CMSSH [1]	0.5114	0.5060	0.4932	0.4489	0.3890	0.3804
SCM [19]	0.5393	0.5528	0.5577	0.4747	0.4813	0.4895
SePH [18]	0.5065	0.5140	0.5189	0.5334	0.5437	0.5499
DCMH [3]	0.5666	0.5710	0.5874	0.5703	0.5869	0.6041
CMHH [16]	0.5605	0.5751	0.5916	0.5677	0.5756	0.5890
AGAH [8]	0.6487	0.6609	0.6579	0.6345	0.6469	0.6426
SCAHN [15]	0.6535	0.6653	0.6686	0.6654	0.6712	0.6716
DADH [9]	0.6594	0.6685	0.6768	0.6694	0.6733	0.6851
DCMHT [33]	0.6770	**0.6911**	0.7030	0.6922	0.7055	0.7201
DHSL	**0.6795**	0.6890	**0.7059**	**0.6971**	**0.7163**	**0.7292**

TABLE 3. The experimental result in the ablation configuration on datasets.

Task	Method	MIRFlickr25K 16bits	 32bits	 64bits	NUS-WIDE 16bits	 32bits	 64bits
I2T	DHSL1	0.8122	0.8136	0.8301	0.6746	0.6789	0.6971
	DHSL2	0.8014	0.8099	0.8134	0.6669	0.6718	0.6855
	DHSL3	0.8101	0.8211	0.8252	0.6610	0.6728	0.6997
	DHSL4	0.7963	0.8017	0.8105	0.6513	0.6635	0.6752
	DHSL	**0.8224**	**0.8345**	**0.8361**	**0.6795**	**0.6874**	**0.7059**
T2I	DHSL1	0.8130	0.8116	0.8274	0.6906	0.7101	0.7219
	DHSL2	0.8008	0.8071	0.8125	0.6776	0.6908	0.7037
	DHSL3	0.8089	0.8107	0.8212	0.6803	0.7007	0.7102
	DHSL4	0.7901	0.7993	0.8104	0.6684	0.6821	0.6965
	DHSL	**0.8172**	**0.8275**	**0.8327**	**0.6975**	**0.7180**	**0.7268**

TABLE 4. Computational overhead of different models.

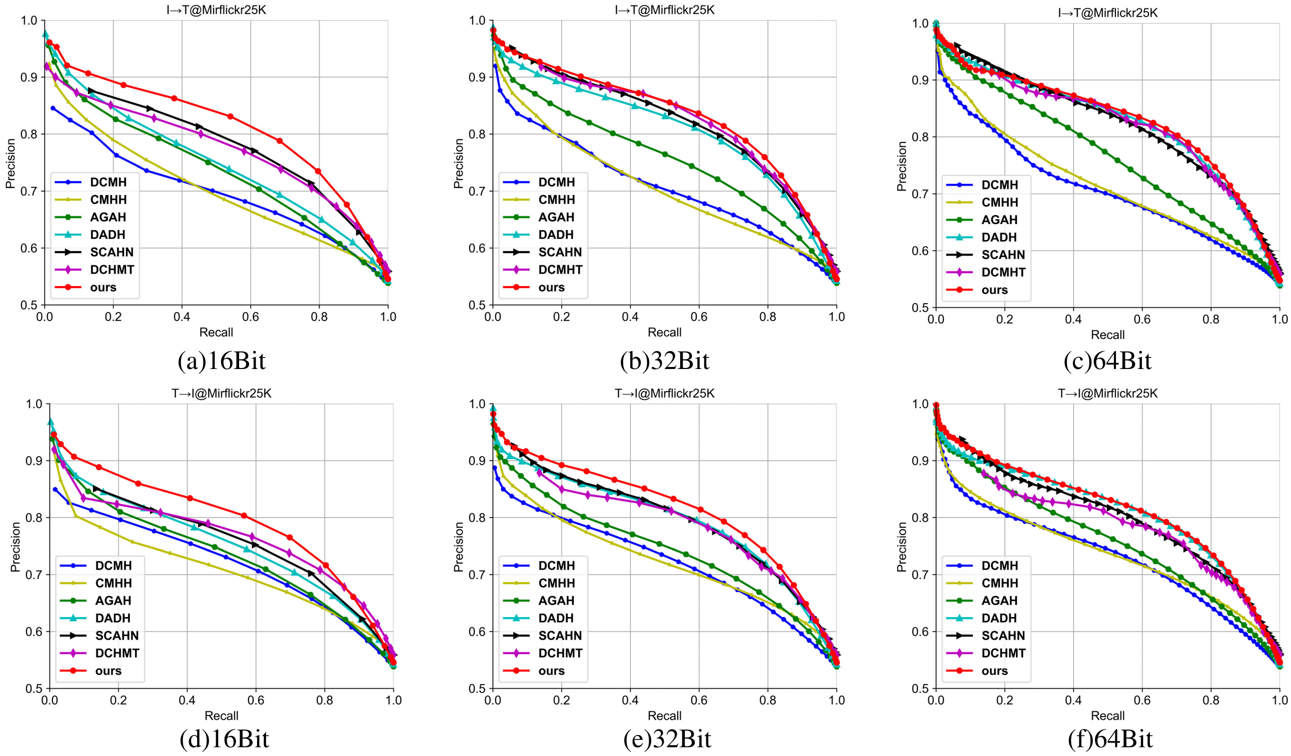
Method	The time used for training one epoch(s)	Parameter quantity(e6)
DADH	101.55	257.97
DHSL	89.61	168.87

IV. EXPERIMENTS

A. DATABASES AND EVALUATION CRITERIA

In this piece of writing, the experiment with two commonly used datasets in cross-modal retrieval will be illustrated.

MIRFlickr25K consists of 24 categories with a total of 25,000 image-text pairs. Each pair has at least one label. For our experimental database, we select 20,015 instance pairs, with each pair having a minimum of 20 text tags. Each instance in the text modality is represented as a 1,386-dimensional bag-of-words vector. For the experimental dataset, we randomly choose 2,000 instances as the query set, while the remaining instances serve as the retrieval database.



(a)16Bit (b)32Bit (c)64Bit

(d)16Bit (e)32Bit (f)64Bit

FIGURE 2. The Precision-recall curves on the MIRFLICKR25K dataset.

Additionally, we extract 10,000 image-text pairs from the database for training.

NUS-WIDE contains 81 categories and a total of 269,648 images, each accompanied by relevant text descriptions. We select 195,834 image-text pairs from this dataset as our experimental database, covering the most common 21 categories. Each instance in the text modality is represented as a 1,000-dimensional bag-of-words vector. For the experimental dataset, we randomly choose 2,100 pairs as the query dataset and extract 10,000 image-text pairs from the database for training.

As evaluation metrics, we adopt three commonly used indicators in cross-modal retrieval: mAP, Precision-recall curves, and top-N precision. We compute these three metrics for two different tasks: image retrieval text (I→T) and text retrieval image (T→I).

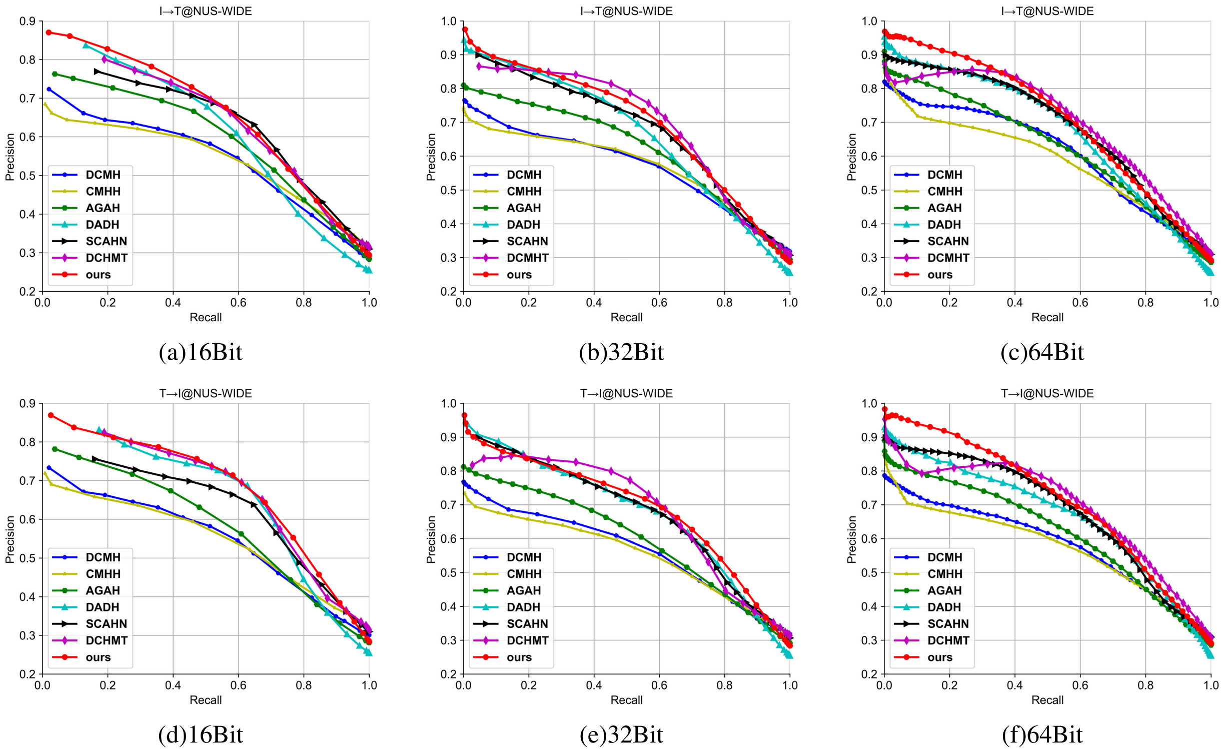
B. EXPERIMENTAL SETUP

After conducting experimental analysis, we tailor different hyperparameters for various networks and datasets. For the image and text main networks, each is a 3-layer fully connected network. The dropout rate is set to 0.2 for the first two layers, and the activation function used is ReLU. Batch normalization is applied to each layer, and the number of neurons in the three-layer network is set to 8,192-2,048-k. Taking into account the disparities between the relational network and the main networks in learning features, we have assigned distinct initial learning rates: 0.0001 for the main network and 0.0005 for the relational network. The batch size remains uniform at 128. The relational network consists of three linear layers, with each layer containing 2k-256-c neurons. For the MIRFlickr25K dataset, the hyperparameters are as follows: $\alpha = 1$, $\beta = 1$, $\eta = 0.1$, $\varphi = 0.01$, $\sigma = 0.1$, $\mu_1 = 1$, $\mu_2 = 5$, $m = 0.00001$, with an epoch of 120. As for the NUS-WIDE dataset, the hyperparameters are set as $\alpha = 1$, $\beta = 1$, $\eta = 0.1$, $\varphi = 0.15$, $\sigma = 0.01$, $\mu_1 = 3$, $\mu_2 = 3$, $m = 0.00001$, with an epoch of 120. Notably, on both datasets, adjusting the hyperparameters to $\mu_1 = 4$ and $\mu_2 = 5$ improves mAP results when the hash codes length is set to 16 bits.

C. COMPARISON WITH EXISTING METHODS

To demonstrate the superiority of the DHSL approach, we compare it with eight advanced cross-modal hashing methods, namely CMSSH [1], SCM [19], SePH [18], DCMH [3], CMHH [16], AGAH [8], SCAHN [15], DADH [9] and DCMHT [33].

The results, as shown in Table 1 and Table 2, reveal that DHSL achieved the best performance on both datasets. Overall, on both datasets, the DHSL method shows improved performance on both tasks as the number of hash bits increases.This improvement is attributed to the feature enhancement part, which allows longer hash codes to contain more discriminative semantic features, thus aiding in the learning of sample similarities across modalities. Additionally, compared to the DADH baseline, DHSL achieves a higher performance improvement on the NUS-WIDE dataset. The reason for this is that each data point in NUS-WIDE has more labels, allowing a better expression of inter-modality relationships in the reduced-dimensional hash representation.



(a)16Bit (b)32Bit (c)64Bit

(d)16Bit (e)32Bit (f)64Bit

FIGURE 3. **The Precision-recall curves on NUS-WIDE.**

Moreover, the relational network can directly calculate the similarity between cross-modal features, enhancing the alignment of cross-modal data and subsequently improving accuracy. In comparison to the DCMHT baseline, DHSL shows a slight performance improvement. This is mainly due to the use of the CLIP pre-trained model [31], [32] in the main network of DCMHT, which extracts cross-modal high-level semantic representations and enhances its potential for cross-modal alignment. However, this also significantly increases the complexity of the model. Therefore, DHSL improves retrieval performance while reducing computational space requirements.

This text illustrates the Precision-recall curves and top-N precision curves of the DHSL method and the top six state-of-the-art baseline methods on two public datasets, with hash codes of 16 bits, 32 bits, and 64 bits. The curves are shown in Figures 2,3, 4 and 5 respectively.

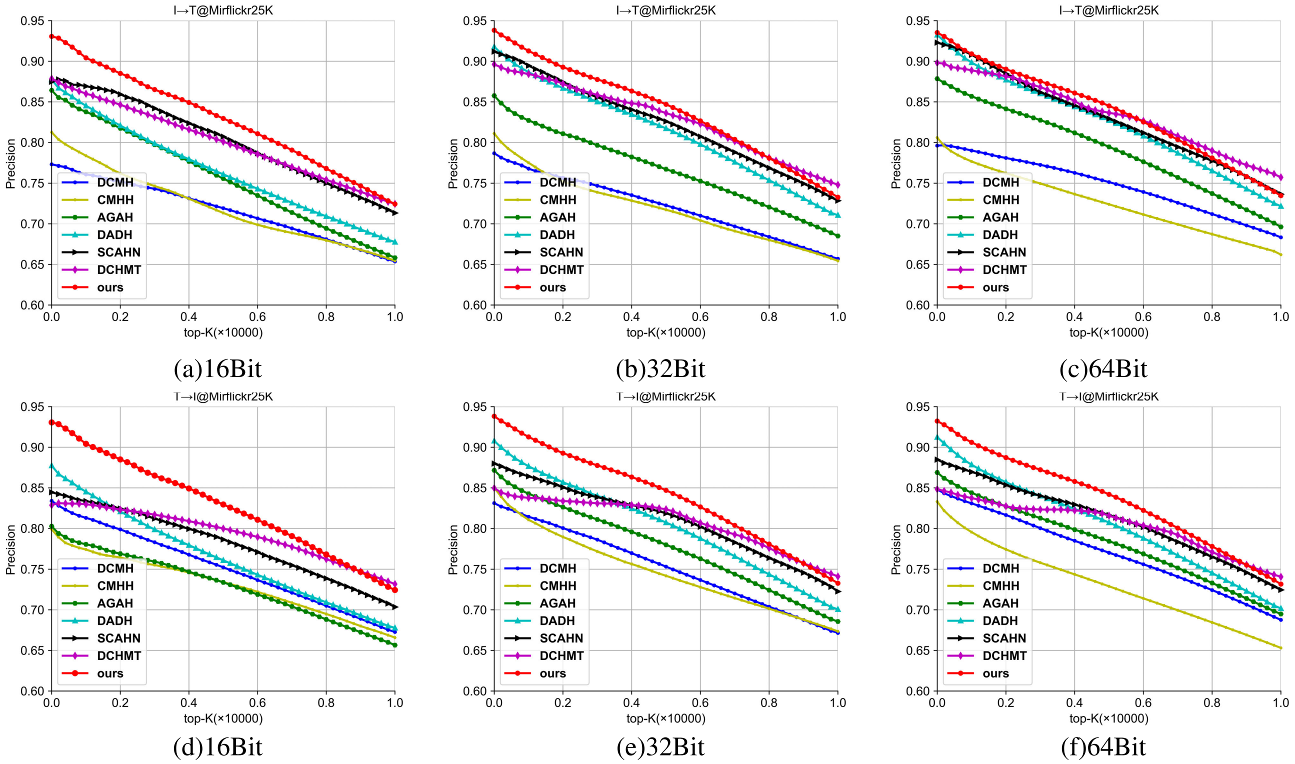
D. ANALYSIS ON ABLATION EXPERIMENTS

In order to verify the effectiveness of the DHSL modules, four ablation experiments are designed for validation. (1) DHSL1 abandons the selection of original features and directly combines them with the hashed output features at the fusion layer. (2) DHSL2 removes the class residual connections and deletes the feature enhancement part. (3) DHSL3 removes the intra-modal similarity hash metric from the relational network module. (4) DHSL4 removes the relational network module altogether.

Table 3 presents the results of four ablation experiments. For DHSL1, low-dimensional features are directly fused with original features without using L_{21} norm for constraint, leading to the inclusion of many irrelevant features. This results in feature repetition and redundancy, causing a decrease in retrieval accuracy. The results of DHSL2 demonstrate that the feature enhancement component can improve the model performance. The fusion of filtered high-dimensional features in the model partially compensates for the shortcomings of feature dimension reduction, adding semantic features from multiple modalities in cross-modal hashing. DHSL3 demonstrates the effectiveness of incorporating intra-modality similarity hashing learning in the relational network. By combining inter-modality and intra-modality similarity hashing learning in the relational network, not only the similarity between modalities can be learned, but also the discriminative power within each modality can be increased, thereby improving retrieval accuracy. DHSL4 demonstrates the positive effect of the relational network module on cross-modal hashing learning. The relation network directly performs similarity learning on the hash representation, reducing matching errors for cross-modal data points. It simultaneously conducts similarity learning on inter-modality and intra-modality features, allowing the similarity hash matrix to contain more semantic features and generating more similar and discriminative hash codes.

E. PARAMETER ANALYSIS

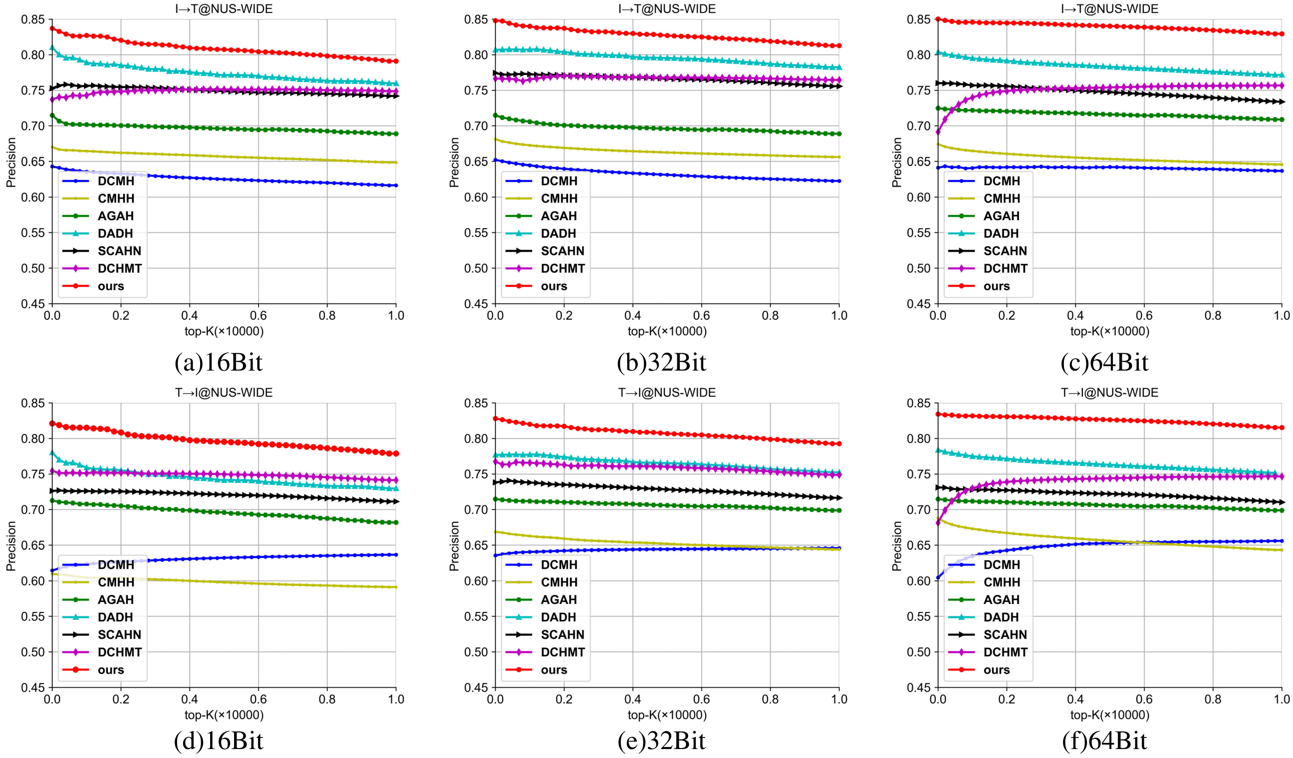
Experiments are conducted on the MIRFlickr25K dataset to analyze the impact of parameter variations η, σ, μ_1 and μ_2 on the mAP results. While testing each parameter, the other parameters are kept constant using the optimal values from



(a)16Bit (b)32Bit (c)64Bit

(d)16Bit (e)32Bit (f)64Bit

FIGURE 4. **The Precision-recall curves on the MIRFLICKR25K dataset.**

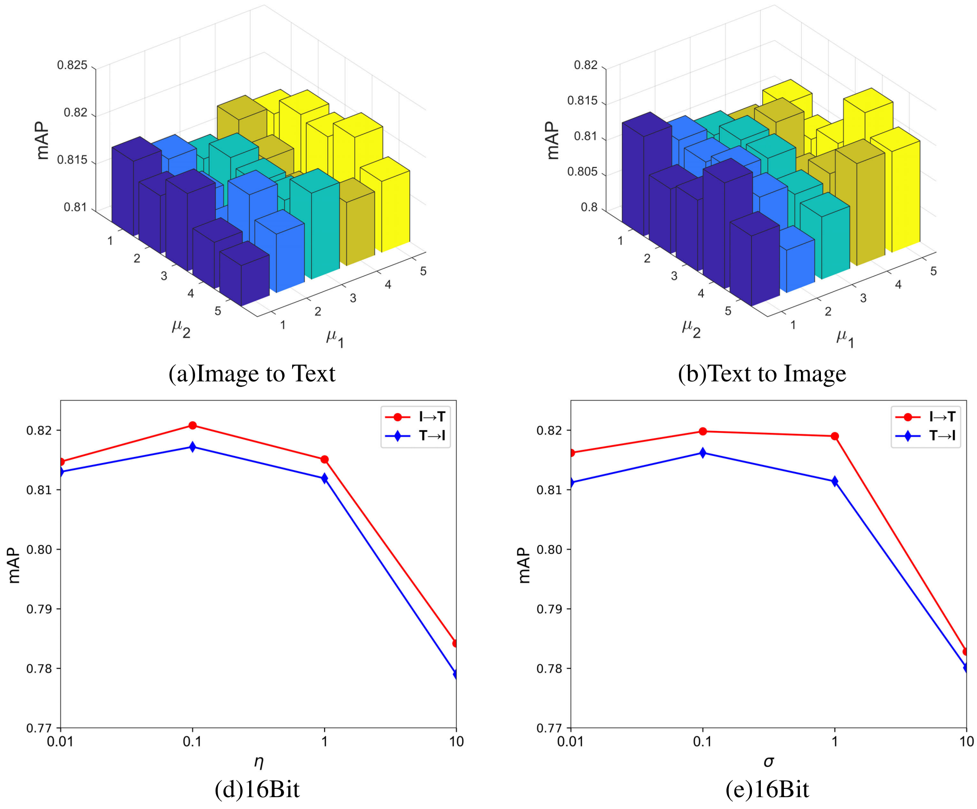


(a)16Bit (b)32Bit (c)64Bit

(d)16Bit (e)32Bit (f)64Bit

FIGURE 5. **The Precision-recall curves on NUS-WIDE.**

section IV-B. The parameters μ_1 and μ_2 control the selection of image and text features in feature enhancement, while η and σ influence the overall objective loss with different losses. We optimize parameters μ_1 and μ_2 using grid search, and parameters η and σ using traditional manual search. We evaluate the model's performance with a hash code length



(a)Image to Text

(b)Text to Image

(d)16Bit

(e)16Bit

FIGURE 6. Parameter analysis on MIRFLICKR25K.

of 16 bits, and the results are shown in Figure 6. The model achieves the best performance when $\mu_1 = 4$ and $\mu_2 = 5$, as well as when $\eta = 0.1$ and $\sigma = 0.1$.

F. COMPLEXITY ANALYSIS

This section compares the parameter size and training time of the proposed model with the baseline DADH on MIRFlickr25K. From table 4, it can be observed that the proposed model achieves a reduction of 89.10e6 in parameter size compared to DADH, with a decrease in training time of approximately 10 seconds per epoch. The introduction of adversarial networks in DADH contributes to a significant increase in parameter complexity, with the discriminator alone having 67.15e6 parameters, while DHSL's relation network only has 1e5 paramter, which can be considered negligible. Overall, DHSL substantially reduces the model's complexity when compared to DADH.

V. CONCLUSION

The paper introduces a novel cross-modal hashing technique called Deep Hashing Similarity Learning for Cross modal Retrieval (DHSL). DHSL aims to address the heterogeneity between modalities and achieve efficient hashing. In the feature learning process, DHSL incorporates class residual connections to enhance features and integrates them with low-dimensional features to supplement missing information. Regarding the hashing learning process, DHSL incorporates relation networks to learn the similarity between modalities and within modalities. Our goal is to learn a hash similarity matrix that approximates the prior similarity matrix containing label information. This process effectively bridges the heterogeneity between images and texts, preserving the semantic relevance between cross-modal features and the discriminative semantics within each modality. The hash codes in Hamming space are constrained by introducing the weighted cosine triplet loss and quantization loss, preserving the cross-modal semantic data structure and ultimately learning efficient hash codes. Through multiple experiments on two widely used datasets, the results demonstrate that DHSL outperforms several state-of-the-art methods. However, there are still some limitations in this study. In future work, we will further improve the utilization of label information in relation networks and explore superior feature extraction methods.

CONFLICT OF INTEREST

There is no conflict of interest with any individual/ organization for the present work.

REFERENCES

[1] 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.

[2] 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. 2075–2082.

[3] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3232–3240.

[4] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *Proc. 31st AAAI Conf. Artif. Intell.*, vol. 31, no. 1, Feb. 2017, pp. 1618–1625.

[5] J. Lu, V. E. Liong, and Y.-P. Tan, "Adversarial multi-label variational hashing," *IEEE Trans. Image Process.*, vol. 30, pp. 332–344, 2021.

[6] Y. Li, X. Wang, S. Qi, C. Huang, Z. L. Jiang, Q. Liao, J. Guan, and J. Zhang, "Self-supervised learning-based weight adaptive hashing for fast cross-modal retrieval," *Signal, Image Video Process.*, vol. 15, no. 4, pp. 673–680, Jun. 2021.

[7] L. Shi, J. Du, G. Cheng, X. Liu, Z. Xiong, and J. Luo, "Cross-media search method based on complementary attention and generative adversarial network for social networks," *Int. J. Intell. Syst.*, vol. 37, no. 8, pp. 4393–4416, Aug. 2022.

[8] W. Gu, X. Gu, J. Gu, B. Li, Z. Xiong, and W. Wang, "Adversary guided asymmetric hashing for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2019, pp. 159–167.

[9] C. Bai, C. Zeng, Q. Ma, J. Zhang, and S. Chen, "Deep adversarial discrete hashing for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 525–531.

[10] W. Liu, C. Mu, S. Kumar, and S. F. Chang, "Discrete graph hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 3419–3427.

[11] D. Wang, X. Gao, X. Wang, and L. He, "Semantic topic multimodal hashing for cross-media retrieval," in *Proc. Int. Conf. Artif. Intell.*, 2015, pp. 3890–3896.

[12] 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, doi: 10.1145/2600428.2609610.

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

[14] D. Wang, P. Cui, M. Ou, and W. Zhu, "Learning compact hash codes for multimodal representations using orthogonal deep structure," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1404–1416, Sep. 2015.

[15] X. Wang, X. Zou, E. M. Bakker, and S. Wu, "Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval," *Neurocomputing*, vol. 400, pp. 255–271, Aug. 2020.

[16] Y. Cao, B. Liu, M. Long, and J. Wang, "Cross-modal Hamming hashing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 202–218.

[17] M. Meng, J. Sun, J. Liu, J. Yu, and J. Wu, "Semantic disentanglement adversarial hashing for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, early access, 2023, doi: 10.1109/TCSVT.2023.3293104.

[18] 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.

[19] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 28, no. 1, Jun. 2014, pp. 2177–2183.

[20] Y. Cao, M. Long, J. Wang, and H. Zhu, "Correlation autoencoder hashing for supervised cross-modal search," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2016, pp. 197–204.

[21] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 154–162.

[22] J. Zhang, Y. Peng, and M. Yuan, "Unsupervised generative adversarial cross-modal hashing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Apr. 2018, pp. 539–546.

[23] X. Zhang, H. Lai, and J. Feng, "Attention-aware deep adversarial hashing for cross-modal retrieval," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 591–606.

[24] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[25] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4967–4976.

[26] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3588–3597.

[27] X. Wang, P. Hu, L. Zhen, and D. Peng, "DRSL: Deep relational similarity learning for cross-modal retrieval," *Inf. Sci.*, vol. 546, pp. 298–311, Feb. 2021.

[28] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "ExFuse: Enhancing feature fusion for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 269–284.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, 2016, pp. 630–645.

[30] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, Feb. 2017, pp. 4278–4284.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2021, pp. 8748–8763.

[33] J. Tu, X. Liu, Z. Lin, R. Hong, and M. Wang, "Differentiable cross-modal hashing via multimodal transformers," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 453–461.

[34] D. Ma, J. Liang, R. He, and X. Kong, "Nonlinear discrete cross-modal hashing for visual-textual data," *IEEE MultimediaMag.*, vol. 24, no. 2, pp. 56–65, Apr. 2017.

[35] T. Yao, X. Kong, H. Fu, and Q. Tian, "Discrete semantic alignment hashing for cross-media retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 12, pp. 4896–4907, Dec. 2020.

YING MA received the Bachelor of Science degree from Yantai University, in 2019. She is currently pursuing the master's degree with the School of Science, Guangxi University of Science and Technology. Her research interests include machine learning and multimodal learning.

MENG WANG received the master's degree from Huazhong Normal University in 2005. He is a Professor at the School of TUS-Digit, Guangxi University of Science and Technology. He has published more than 20 papers in domestic and international academic journals as well as conferences. His research focuses on natural language understanding, cross-modal retrieval, and graph neural networks.

GUANGYUN LU was born in Yulin, Guangxi, in 1983. He is currently a Senior Engineer/Information System Project Manager. His research interests include natural language processing, image recognition and processing, and multimodal retrieval.

YAJUN SUN received the Bachelor of Engineering degree in data science and big data technology from the Hubei University of Economics, in 2021. He is currently pursuing the master's degree in applied statistics with the Guangxi University of Science and Technology. His main research interests include computer vision and multimodal learning.