IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 January 2024, accepted 16 March 2024, date of publication 25 March 2024, date of current version 29 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3380019

RESEARCH ARTICLE

Multi-Granularity Semantic Information Integration Graph for Cross-Modal Hash Retrieval

ZHICHAO HAN, AZREEN BIN AZMAN, (Member, IEEE), FATIMAH BINTI KHALID, (Member, IEEE), AND MAS RINA BINTI MUSTAFFA, (Senior Member, IEEE)
Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang 43400, Malaysia
Corresponding authors: Azreen Bin Azman (azreenazman@upm.edu.my) and Zhichao Han (gs61677@student.upm.edu.my)

ABSTRACT With the development of intelligent collection technology and popularization of intelligent terminals, multi-source heterogeneous data are growing rapidly. The effective utilization of rich semantic information contained in massive amounts of multi-source heterogeneous data to provide users with high-quality cross-modal information retrieval services has become an urgent problem to be solved in the current field of information retrieval. In this paper, we propose a novel cross-modal retrieval method, named MGSGH, which deeply explores the internal correlation between data of different granularities by integrating coarse-grained global semantic information and fine-grained scene graph information to model global semantic concepts and local semantic relationship graphs within a modality respectively. By enforcing cross-modal consistency constraints and intra-modal similarity preservation, we effectively integrate the visual features of image data and semantic information of text data to overcome the heterogeneity between the two types of data. Furthermore, we propose a new method for learning hash codes directly, thereby reducing the impact of quantization loss. Our comprehensive experimental evaluation demonstrated the effectiveness and superiority of the proposed model in achieving accurate and efficient cross-modal retrieval.

INDEX TERMS Cross-modal retrieval, hash, scene graph, multi-granularity.

I. INTRODUCTION

Owing to different data collection methods and data types, multi-source heterogeneous data for the same thing can be described in a variety of different ways, modes, or perspectives. For the description of the same thing, some data are described with a coarse-grained structure, such as labels, while some data are described with a fine-grained structure, such as a scene graph. Therefore, the effective utilization of these massive multi-source heterogeneous data [1] containing rich semantic information, mining the deep internal connections of multi-source heterogeneous data, and providing users with high-quality cross-modal information retrieval services have become key issues in the current field of information retrieval.

Multimodal data contain rich multi-granularity semantic information, including coarse-grained [2] semantic conceptual features and fine-grained semantic features [3]. Mining multi-granularity semantic features can effectively learn the semantic correlations between multimodal data [4], thereby narrowing the ''semantic gap'' [5], which indicates the gap between low-level data features and actual semantics or concepts. For example, a gap exists between the pixel representation of an image and the conceptual or semantic information contained in the image. Resolving the semantic gap requires associating low-level features with higher-level semantics to better understand and retrieve cross-modal data.

Many existing cross-modal related studies have achieved good results on certain publicly available experimental

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

datasets, but they still face some challenges and problems in practical applications, mainly because the representation information of images and text is not fully mined [6]. Traditional cross-modal retrieval models only consider the mining of global semantic information, without considering the objects and their relationships, resulting in some fine-grained semantic information being ignored. If global semantic information and fine-grained semantic information are fully considered and integrated, richer representation information can be obtained. Insufficient cross-modal representation learning [7]. Different modalities of data, such as images and text, may contain different types of information. Preserving inter-modal heterogeneity can ensure a wider range of information coverage and enrich the content and breadth of search results. Additionally, heterogeneity allows for complementarity between modalities, meaning that one modality may contain information that another modality does not possess. For example, in image retrieval, the image itself may convey certain visual features, whereas the relevant text may provide richer semantic information. This complementarity can improve the accuracy and comprehensiveness of the retrieval. Traditional cross-modal retrieval models are simply mapped to a common representation space and then subjected to cross-modal representation learning. This can lead to changes in the heterogeneity between modalities and invariance within modalities, resulting in a loss of correlation and a reduction in the performance of cross-modal retrieval. This requires finding effective representation learning methods, capable of integrating semantic information from different modal data and maintaining data similarity in a shared space. It is difficult to construct a good hash function [8], and many methods in reality cannot consider the nonlinear structure of the data, and instead use continuous solution methods in real space. This method cannot capture nonlinear projections and causes significant quantization losses, particularly when the hash code length is long.

In response to these shortcomings, we propose a scene graph model to perform hierarchical semantic association modeling on multimodal data. For each modality of the data, intra-modal global semantic mapping was used to represent coarse-grained semantic association information, whereas a relational semantic graph based on the scene graph was used to describe fine-grained semantic association information. This can provide sufficient semantic discrimination information for cross-modal quantization coding learning. When cross-modal retrieval is performed between images and text, the visual features of image data and semantic information of text data are heterogeneous, and their expression forms and features are different. Then, by enforcing comprehensive similarity preservation, we narrow the "heterogeneous gap" [9] between different domains and modalities. In addition, a direct learning approach for generating hash codes is introduced, which reduces the impact of quantization loss. This approach not only simplifies the learning process but also enhances the quality of the resultant hash codes. The main contributions of this study are as follows:

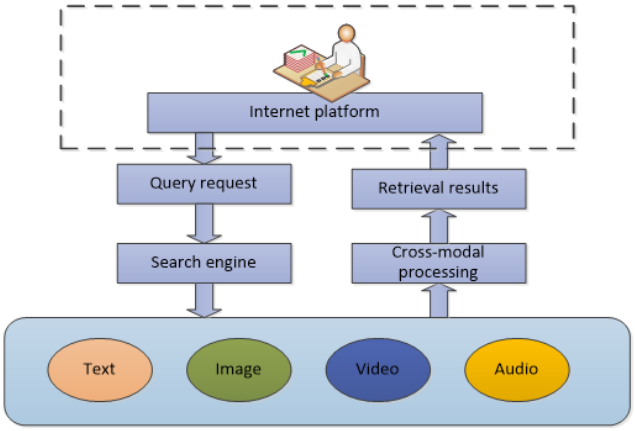

FIGURE 1. The basic schematic diagram of the cross-modal retrieval.

- Multi-granularity semantic feature modeling. We propose a scene graph based multi-granularity semantic information preservation model. This model fully mines more fine-grained and coarse-grained semantic information within images and texts by integrating global semantic information and scene graph semantic relations, which can effectively narrow the low-level features with higher-level semantic gaps and provide a good foundation for hash encoding learning.
- Comprehensive similarity preservation. We propose a triplet loss to reduce the heterogeneity between modalities in cross-modal representation learning and design a reconstruction loss to better capture feature representation and similarity measures within media types.
- Direct Hash Learning. We propose a direct learning method for hash codes that reduces quantization loss and information loss caused by the inability to capture nonlinear semantic information.

II. RELATE WORK

A. CROSS-MODAL RETRIEVAL

The motivation for cross-modal retrieval stems from the inherent heterogeneity of multimedia data and the limitations of traditional unimodal retrieval methods [10]. For example, text-based retrieval methods may not be able to capture the rich visual content in images or videos. By integrating multiple modalities, cross-modal retrieval helps to represent information more comprehensively and accurately, thereby improving the overall retrieval performance [11]. Figure 1 shows a basic schematic diagram of cross-modal retrieval.

The supervised cross-modal hash method is one of the most common types of cross-modal retrieval technique. In the past decade, many important approaches have been proposed and have achieved good retrieval performance. Kumar et al. [12] constructed a set of hash functions for multimodal data and transformed the hash function learning problem into a solvable eigenvalue problem using a new relaxation method. Liu et al. [13] proposed a flexible collaborative matrix

decomposition hash algorithm (FS-CMFH) for efficient cross-modal retrieval that utilizes label consistency between different modalities to preserve semantic information within and between modalities in the common latent semantic representation space. Wang et al. [14] proposed a label consistency matrix decomposition hash method that maps heterogeneous data to a latent semantic representation space, in which multimodal data from the same category share the same feature representation. Jiang et al. [15] initially introduced a cross-modal hash model grounded in deep neural networks, amalgamating multimodal feature and hash learning within a single model. Yang et al. [16] integrated different types of paired constraints using a deep model to enhance the similarity measurement of hash codes from both intra-modal and inter-modal perspectives. Deng et al. [17] formulated a deep hash (TDH) using triplets as supervised information to foster semantic connections among cross-modal instances. The supervised hash method based on the deep learning model not only has strong nonlinear semantic representation ability, but also achieves end-to-end semantic feature learning and hash code generation, substantially enhancing cross-modal hash retrieval accuracy.

Although supervised cross-modal hashing methods have achieved a good retrieval performance, they rely heavily on annotated data. However, in massive multi-source heterogeneous data, the amount of annotated data is very limited, and annotating a large amount of data consumes huge manpower and resources, which makes the existing supervised hash methods difficult to apply to large-scale multi-source heterogeneous data [18]. Therefore, many researchers have focused on the research and exploration of unsupervised cross-modal hash methods.

To fully explore the potential correlation among multi-modal data, Liu et al. [19] partitioned the feature space into a shared subspace for all hash functions and their complementary subspaces. This shared space captures a common structure across diverse hash functions, encapsulating crucial information for effective hash learning. Hu et al. [20] introduced an iterative multi-view hash algorithm (IMVH) aimed at acquiring optimal alignment within coding schemes to preserve similarity across views. Irie et al. [21] proposed an unsupervised hash method called alternating common quantization, which alternately searches for binary quantizers for each modal space by connecting multimodal data, ensuring the minimum quantization difference while maintaining data similarity. Su et al. [18] introduced an unsupervised deep joint semantic reconstruction hash (DJSRH) method. It amalgamates the distinct modality's original neighborhood details using a semantic similarity matrix to encapsulate the inherent semantic relationships. Huang et al. [22] proposed an unsupervised cross-modal hash framework that utilizes data fusion to capture the underlying manifolds across modalities, avoiding the problem of maximizing mutual constraints between intra-modal and inter modal similarities. Yang et al. [23] investigated a deep semantically aligned hash algorithm (DSAH) that intelligently aligns feature similarity with hash code similarity via a semantic alignment loss function.

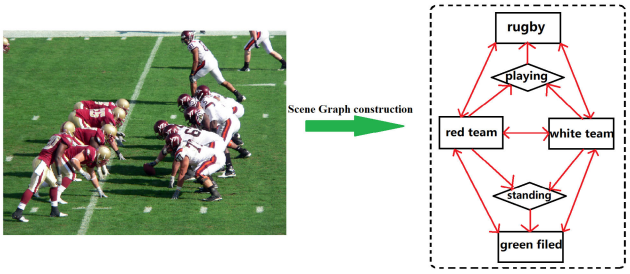

FIGURE 2. The image Scene Graph. This picture depicts the red team and the white team standing on a green field playing rugby, where "red team" "white team" "rugby" and "green field" are objects, and the relationship between them is indicated by two-way arrows; "standing" "playing" are relational words, and the relationship between them and the objects is indicated by single arrows.

To effectively reduce the heterogeneity between modalities, some studies have integrated deep adversarial learning methods into cross-modal hash learning models, thereby improving the quality of hash code generation. For example, He et al. [24] devised an unsupervised cross-modal retrieval technique that leverage adversarial learning. The inclusion of a modal classifier to anticipate the modality of the feature transformation guarantees statistical indistinguishability among these transformed features. Zhang et al. [25] maximized the unsupervised representation learning potential inherent in generative adversarial networks (GANs) to explore the underlying structure of cross-modal data. They formulated an unsupervised cross-modal hash model based on GANs. At present, unsupervised methods lack semantic labels on the data, making it difficult for unsupervised hash models to effectively mine key semantic features and achieve cross-modal semantic alignment. Therefore, fully learning and characterizing the complex correlations between multi-source heterogeneous data and improving the semantic discriminability of multimodal unified hash codes have become key issues for the widespread application of unsupervised methods.

B. SCENE GRAPH

A scene graph is a data structure in the fields of computer vision and graphics and is used to represent the relationships between objects in an image or scene. It represents objects in an image as nodes and uses edges to represent the relationships between objects. The main purpose of scene graphs is to capture the hierarchical and semantic relationships between objects in a structured manner to better understand and process the image content. In computer vision, this structured representation is helpful in tasks such as image understanding, object detection, and scene generation. The scene graph is composed of a triple form of ''object-relationship-object'' where ''object'' is a node and ''relationship'' is an edge connecting two objects. Yu et al. [26] proposed a method called DGCPN. The core idea of this method is to use the

concept of graphs to establish associations between different modal data. The network structure proposed in this study aims to maintain the neighbor relationship of data in the graph, thereby maintaining data consistency when encoding data. This graph neighborhood consistency helps ensure that similar data remain similar in low-dimensional embedding spaces. Johnson et al. [27] explored a scene graph-based image retrieval method to improve the performance and accuracy of image retrieval by utilizing the scene graph information of images. Li et al. [28] introduced a method for the semantic modeling and representation of objects and their relationships in images by generating scene maps from object, phrase, and region descriptions. Figure 2 shows an image of the scene graph.

C. HASH

In the era of big data, cross-modal hashing finds extensive applications in cross-modal retrieval. Its popularity stems from its high efficiency, low dimensionality, minimal storage requirements, and effective representation of high-level feature consistency in diverse multi-source data [29]. Current cross-modal hashing methods have not fully leveraged the abundant semantic details within vast and varied multi-source data. Consequently, the primary challenge in cross-modal hash retrieval research involves effectively exploring these multi-source heterogeneous data to inform the training of deep cross-modal hash models and enhance the quality of the generated hash code. A basic schematic of the cross-modal hash is shown in Figure 3.

In recent years, deep neural networks have substantially enhanced the performance of cross-modal hashing owing to their robust representational capabilities. Zhang et al. [7] introduced a cross-modal hash generation adversarial network model and utilized a reinforcement learning algorithm to guide its training. Shen et al. [30] proposed a semi-supervised graph convolutional hash network (SGCH) method that learns a common cross-modal Hamming space through end-to-end neural networks. Liu et al. [31] proposed a similarity hash (JDSH) method based on a joint modal distribution. This approach employs an unsupervised learning algorithm called Distributed Similarity Decision and Weighting (DSDW) to produce hash codes with increased discriminative capability. Ma et al. [32] introduced DASH in 2016, which is an uncomplicated yet impactful approach designed for cross-modal hashing. The DASH method, which is both non-iterative and devoid of parameters, is remarkably straightforward to implement and requires only three lines of code. They also published a paper called DCMH [33] in 2017, which assessed the efficacy of the proposed discrete optimization by optimizing its objective function using a relax-and-threshold technique. Extensive empirical evaluations conducted on image-text and image-tag datasets show that DCMH is a substantial advancement compared to prior methods, exhibiting notable enhancements in training efficiency and retrieval accuracy. The

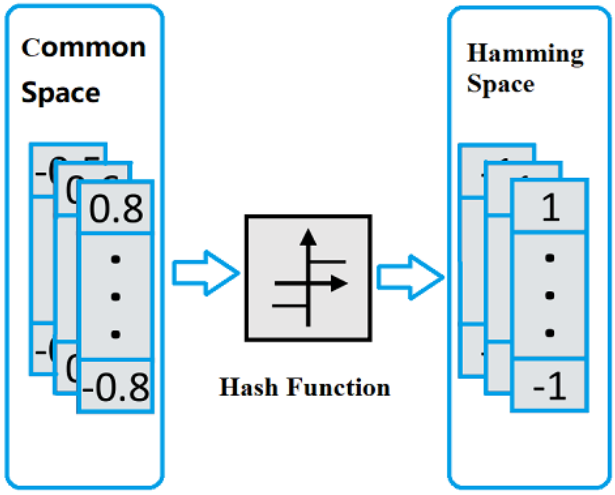

FIGURE 3. The basic schematic diagram of the cross-modal Hash.

performance of semantic information retrieval is enhanced by the incorporation of joint supervision mechanisms, specifically involving instance-pairwise, instance-labeled, and class-wise similarities. Yu et al [34] further augmented the semantic consistency across heterogeneous modalities. The integration of these supervision strategies contributes to an improved and refined retrieval process that ensures a more comprehensive and accurate representation of semantic information in diverse data modalities. These achievements have theoretically verified the feasibility of cross-modal retrieval based on cross-modal hashing, but existing hash methods have not yet fully explored deep internal connections in multi-source heterogeneous data. Therefore, mining deep internal correlations in multi-source heterogeneous data from different granularities and modal data to improve the representation ability of hash codes on multi-source heterogeneous data is of great research value. This study aims to start with deep mining of different granularities data, different modal data, and the inherent correlation relationships between different data sources. Studied the representation of cross-modal hashes for multi-granularity semantic fusion to enhance the representation ability of cross-modal hashes for multi-source heterogeneous data.

III. PROPOSED METHOD

Our model consists of three parts: multi-granularity semantic feature modeling, comprehensive similarity preservice, and cross-modal hash learning. First, feature vectors of images and texts were obtained through convolutional neural networks [35] and bag-of- words [36] models. Then, a global semantic graph was constructed using these feature vectors, which saved coarse-grained semantic information regarding the image and text. Simultaneously, we constructed scene graph semantic information to integrate and connect objects and object relationships in images and texts, to retain more fine-grained semantic information. In this way, we obtained complete coarse-grained and fine-grained

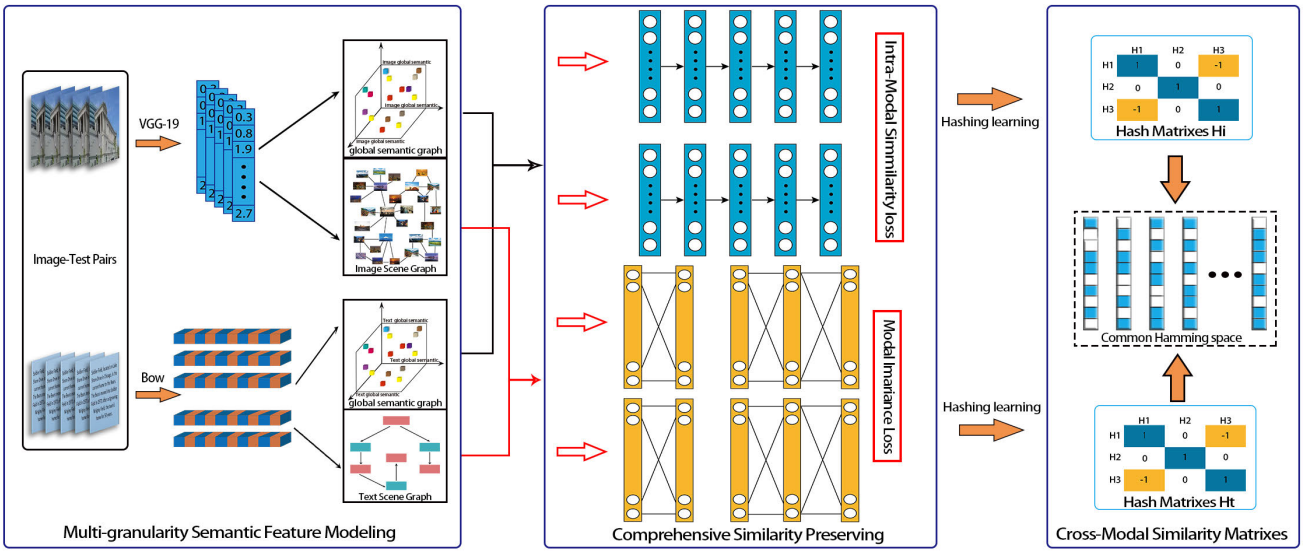

FIGURE 4. Framework of our model.

semantic information of images and texts, providing richer and more accurate semantic descriptions for cross-modal retrieval tasks. Second, when mapping image and text features to a common space [37], we used a triplet constraint [38] method to reduce the heterogeneity between modalities in cross-modal representation learning and designed the reconstruction loss to better capture feature representation and similarity measures within media types. Finally, in the learning process of hash functions, we propose a direct learning method for hash codes that reduces quantization loss and information loss caused by the inability to capture nonlinear semantic information. Figure 4 illustrates the framework of the proposed model.

A. PROBLEM FORMULATION

Let D represent a multimedia dataset containing a set of images, denoted by the capital letter I, containing n images. We can then use sets $I = \{I_1, I_2, I_3, \ldots, I_n\}$ to represent all image examples. In addition, the dataset included a text set, denoted by the capital letter T, containing m texts. Therefore, we can use sets $T = \{T_1, T_2, T_3, \ldots, T_m\}$ to represent all the text examples. We want to calculate the similarity between the images and text. First, features must be extracted from image and text collections. We assume that we use a function $\phi(I)$ to represent the features extracted from image I, with dimension d. Using functions $\phi(T)$ to represent the features extracted from text T, with a dimension of d'. These two feature vectors are then mapped to a shared aligned subspace, where cosine similarity is used to measure the cosine value of the angle between the two vectors, which is used to measure the cross-modal similarity.The formula used is as follows:

$$\text{sim}(I, T) = \cos(\theta) = \frac{\phi(I) \cdot \psi(T)}{\|\phi(I)\| \cdot \|\psi(T)\|}, \tag{1}$$

Here, $\text{sim}(I, T)$ is the similarity score between image I and text T. $\cos(\theta)$ which is the cosine similarity between image features and text features. The upper part of the formula measures the magnitude of the components of the two vectors in the same direction. Representation in the lower half of the formula is the norm of the vectors. Norms are used to normalize vectors such that the cosine similarity is not solely affected by vector length. In the feature space, similar vectors have smaller angles and cosine values close to 1, the angle between dissimilar vectors is larger, and the cosine value is close to 0. To provide more clarity in symbol allocation and sufficient corresponding explanations for readers, Table 1 summarizes the frequently used annotations in this study.

B. GLOBAL SEMANTICS RELATION GRAPH CONSTRUCTION

To realize coarse-grained semantic feature modeling, the Euclidean distance and kNN algorithms were proposed to measure the similarity between images and texts, respectively. where $E_{I:ij}$ is the Euclidean distance to calculate the similarity between nearest neighbors for image and $E_{T:ij}$ for text, and the formula is as follows:

$$E_{I:ij} = \text{EucDst}(\boldsymbol{C}_i, \boldsymbol{C}_j), \tag{2}$$

$$E_{T:ij} = \text{EucDst}(\boldsymbol{W}_i, \boldsymbol{W}_j), \tag{3}$$

where, $\boldsymbol{C}_i \in \mathcal{N}_{knn}(\boldsymbol{C}_j)$, $\boldsymbol{C}_j \in \mathcal{N}_{knn}(\boldsymbol{C}_i)$, $\boldsymbol{W}_i \in \mathcal{N}_{knn}(\boldsymbol{W}_j)$, $\boldsymbol{W}_j \in \mathcal{N}_{knn}(\boldsymbol{W}_i)$, $\mathcal{N}_{knn}(\cdot)$ denotes the set of k-nearest neighbors of a data object. We consider each data object as a node and use the spatial distance between the objects as the relationship weights to form a global semantic relationship graph g_i and g_t for the images and texts.

TABLE 1. Summary of frequently used notations.

Notation	Definition
D	Multimedia dataset
I	Image sample
T	Text sample
$E_{I:ij}$	Euclidean distance to calculate similarity between nearest neighbors for image
$E_{T:ij}$	Euclidean distance to calculate similarity between nearest neighbors for text
$\mathcal{N}_{knn}(\cdot)$	knn function
$\boldsymbol{C}_i$and$\boldsymbol{C}_j$	K-nearest neighbors of a data in image
$\boldsymbol{W}_i$and$\boldsymbol{W}_j$	K-nearest neighbors of a data in text
G	Visual scene graph
V	Visual scene graph node-set
E	Visual scene graph edge-set
O	Target region set
R	Object relation set
$\boldsymbol{M}_f^I$	Semantic fusion matrix
$\boldsymbol{\xi}_i$	The semantic representation after fusion
$\boldsymbol{\beta}^I$	Bias term
A_i	Anchor sample
P_i	Positive sample
N_i	Negative sample
X_i	Input data
$\hat{X}_i$	Output data
$[\boldsymbol{H}_I]$	Image feature representation matrix
$[\boldsymbol{H}_T]$	Text feature representation matrix
$\boldsymbol{B}_I$	Image hash code matrices
$\boldsymbol{B}_T$	Text hash code matrices

C. LOCAL SEMANTIC RELATION GRAPH CONSTRUCTION

To achieve fine-grained semantic feature modeling, the proposed approach constructs semantic entities and their semantic relationships in image data based on the local semantic relationship graphs of scene graphs. Specifically, we represent a visual scene graph as $G = \{V, E\}$ where V is the node-set, and E is the edge-set. Target regions and corresponding category labels were extracted from the image using the target detection algorithm Faster R-CNN. Suppose p targets are detected, denoted by O the set of targets containing p targets $O = \{O_1, O_2, O_3, \ldots, O_p\}$, where each target O_i can be represented as $O_i = (o_i, c_i)$, where o_i is the location information of the target and c_i is the category label of the target. To construct a triad of ''object-relationship-object'', we must predict the relationship between the objects. The relationship classification model, the RNN, can be used to predict the relationship between the goals. Suppose we have a collection of relations, denoted by R, containing q relations $R_i = \{R_i1, R_i2, R_i3, \ldots, R_iq\}$. For each pair of targets (O_i, O_j), we compute the relation score $r(i, j) = fi(O_i, O_j)$, where fi is the relation prediction model. Eventually, the scene graph G of an image can be represented as $G = (O, R)$, where O is the set of targets and R is the set of relations. Each relation $r(i, j)$ can be represented by R_k, where k is the category label of the relation.

To achieve fine-grained semantic feature modeling, a hierarchical local semantic graph based on a scene graph was proposed to construct semantic entities and their semantic relations in text data. Specifically, for the text data, the text is first segmented to obtain a word sequence $W = \{W_1, W_2, W_3, \ldots, W_x\}$, where x is the number of words in the text. The vector representation of each word is then obtained by mapping each word to its corresponding vector representation via a bag-of-words. Suppose that the vector representation of the word W_i is v_i. To construct relationships between words, an RNN can be used to predict the relationships between words. Suppose we have a set of relations, denoted by R_t, containing u relations $R_t = \{R_t1, R_t2, R_t3, \ldots, R_tx\}$. For each pair of words (w_i, w_j), we can compute the relationship score $r(i, j) = f_t(w_i, w_j)$, where f_t is the relationship prediction model. Finally, the scene graph G_t of a text can be represented as $G_t = (W, R_t)$, where W is the set of words containing x words, and R_t is the set of relations. Each relation $r(i, j)$ can be represented by R_k, where k is the category label of the relation.

After the global semantic relation graph and local semantic relation graph are constructed, the semantic representations of the network nodes are first learned separately by a graph convolutional neural network, and then the node representations are fused by a semantic fusion network, which can obtain efficient semantic representations ξ_i and ζ_t. The formula for the node representation of semantic fusion network is as follows:

$$\begin{aligned} \xi_i &= \sigma\left(\boldsymbol{M}_f^I\left(\boldsymbol{g}_i^I \bowtie \boldsymbol{G}_i^I\right) + \beta^I\right), \\ \zeta_t &= \sigma\left(\boldsymbol{M}_f^T\left(\boldsymbol{g}_i^T \bowtie \boldsymbol{G}_i^T\right) + \beta^T\right), \end{aligned} \tag{4}$$

where, $\boldsymbol{M}_f^I$ and $\boldsymbol{M}_f^T$ represent the semantic fusion matrix, β^I and β^T represent the bias terms, $\sigma()$ is the nonlinear activation function, and $\bowtie$ represents the vector connection operator.

D. MODALITY INVARIANCE LEARNING

We strive to maintain cross-modal invariance by minimizing the difference between representations of similar items across modalities, and simultaneously maximizing the distinction between dissimilar items within the same modality. To achieve this, we incorporate triplet constraints in embedding via a dedicated triplet loss term. The process of selecting positive and negative sets for triplet loss is as follows. Samples from different modalities but with the same label are composed of sample pairs, for example, let A_I be the anchor image sample, and the positive text sample pairs are paired with $< A_I, P_t >$. Simultaneously, a sample pair, such as $< A_T, P_i >$ was constructed, with text as the anchor and images with the same label as a positive sample. We then selected negative samples from mismatched

image text pairs with different semantic labels to form triplet samples $< A_I, P_t, N_t >$, $< A_T, P_i, N_i >$ where the negative sample N_i is selected from all negative samples in the current batch. The objective function for the modality invariance loss is formulated as follows:

$$L_{\mathrm{i}} = \sum_{i=1}^{N} [d(f(A_i), f(P_t)) - d(f(A_i), f(N_t)) + \alpha]_+ \quad (5)$$

$$L_{\mathrm{t}} = \sum_{t=1}^{N} [d(f(A_t), f(P_i)) - d(f(A_t), f(N_i)) + \alpha]_+ \quad (6)$$

$$L_{\mathrm{triplet}} = L_{\mathrm{i}} + L_{\mathrm{t}}, \quad (7)$$

where, $d()$ is a distance metric, *alpha* is an interval parameter, $[x]_+$ represents the positive part.

E. INTRA-MODAL SIMILARITY LEARNING

In the intra-modal similarity learning step, we focused on data similarity learning within each media type to better capture feature representations and similarity measurements within the media type. An autoencoder enables the input data to be encoded (compressed) and decoded (reconstructed) through a network. We designed a reconstruction loss to calculate the difference between the input of the autoencoder and the decoder output. By optimizing this loss, the autoencoder learns how to encode and decode the input data, thereby capturing their main features of the input data. The reconstruction loss can be expressed as

$$L_{\mathrm{reconstruction}} = \sum_{i=1}^{N} ||X_i - \hat{X}_i||_2^2 \quad (8)$$

Among them, X_i represents the original input sample with index i in the dataset, and $\hat{X}_i$ represents the output reconstructed by the autoencoder on the input sample X_i. By constructing an autoencoder network, the input data were mapped to a low-dimensional encoding space through the encoder and then reconstructed through the decoder. In the process of optimizing reconstruction losses, the network automatically learns how to extract and preserve key features of the input data, thereby obtaining a more compact data representation.

F. HASH CODE LEARNING

After the above learning process, the semantic representations of each modality ξ_i and ζ_i are mapped to the modality-consistent semantic representation space, which can provide a good foundation for hash code learning. For quantized coding of cross-modal feature representation matrices, $[\boldsymbol{H}_I]$ and $[\boldsymbol{H}_T]$, we propose using the following loss function number to train the hash model:

$$\mathcal{L}_{HS} = \gamma \left(\|\boldsymbol{B}_I - [\boldsymbol{H}_I]\|_F^2 + \|\boldsymbol{B}_T - [\boldsymbol{H}_T]\|_F^2 \right) + \eta \left(\|[\boldsymbol{H}_I]\mathbf{1}\|_F^2 + \|[\boldsymbol{H}_T]\mathbf{1}\|_F^2 \right), \quad (9)$$

where $\boldsymbol{B}_I$ and $\boldsymbol{B}_T$ denote the image and text hash code matrices, respectively; 1 denotes all 1 matrix; and γ and η denote the parameters used to balance the two parts. By combining the above loss functions, the cross-modal hash synthesis can be obtained as an objective function.

$$\mathcal{L} = \mathcal{L}_{HS} + L_{\mathrm{reconstruction}} + L_{\mathrm{triplet}}. \quad (10)$$

IV. EXPERIMENTS

A. DATASET

In the experimental section of this study, we provide a broad and in-depth evaluation of the performance of cross-modal retrieval. Experiments were performed on three distinct datasets: Wikipedia [39], NUS-WIDE [40], and MIRFlickr-25k [41]. These datasets cover image and text data from different domains and content types, providing a diverse set of challenges and evaluation environments. To improve the computational efficiency of our experiments, we selected only a part of the data in each dataset as the training and testing samples. A selection of examples from Wikipedia, NUS-WIDE, and MIRFlickr-25k is shown in Figure 5.

To assess the efficacy of the proposed method, we compared it with six representative cross-modal retrieval models. These models cover classical cross-modal retrieval methods. By comparing these models, we aimed to comprehensively evaluate the advantages of our approach under different datasets and model settings. We detail the experimental setup, performance metrics, and comparison results with each model in the following sections to demonstrate the effectiveness and superiority of our approach. The data for most of the comparison methods in the table below are from the original paper or are based on the source code provided in the paper. The following are concise views of these state-of-the-art technologies.

- **CVH** [12] presented at ICJAI in 2011. Proposed a learned hash function method for cross-view similarity search. Maintaining similarity in the hash code space helps achieve efficient similarity matching in multi-view datasets, thus facilitating progress in multimodal data analysis.
- **CMFH** [42] was presented at CVPR in 2014. Their main contribution is the attempt to use the Collective Matrix Decomposition (CMF) method to learn cross-view hash mapping relationships, which supports cross-view searching and enhances search accuracy by integrating information from multiple views.
- **DCMH** [15] presented at the CVPR in 2017. The core idea of this study is to map data from different modalities to a shared low-dimensional binary code space, to achieve similarity matching of the cross-modal data in that space. This method utilizes the ability of deep neural networks to capture the correlations between modalities in high-dimensional feature spaces.
- **DJSRH** [18] Presented at ICCV in 2019. This study proposes a joint reconstruction loss function that combines two key aspects: intra-modal reconstruction and inter-modal reconstruction. Effective retrieval of large-scale

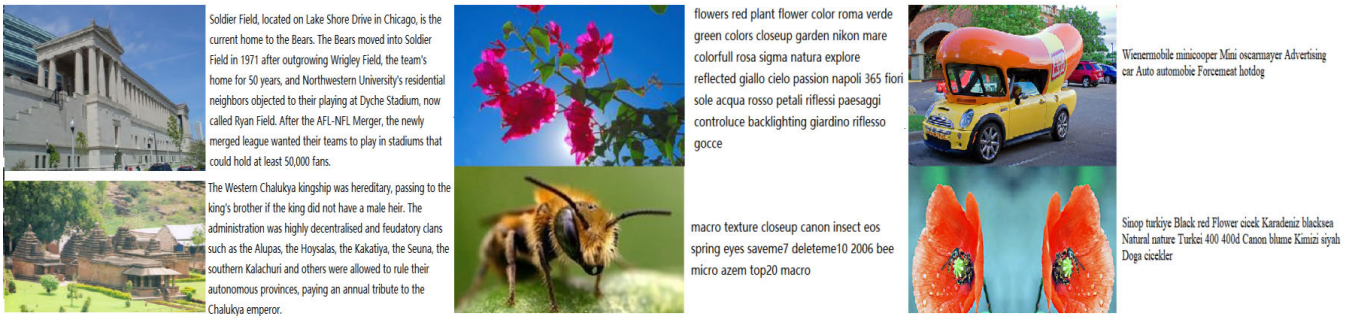

FIGURE 5. A selection of examples from Wikipedia, NUS-WIDE, and MIRFlickr-25k.

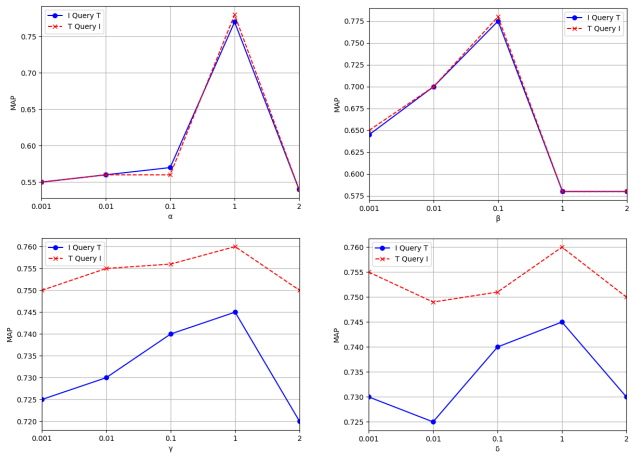

FIGURE 6. The curves of hyper-parameters analysis.

cross-modal datasets was achieved by utilizing the joint semantic information of the cross-modal data.

- **DGCPN** [26] Presented at AAAI in 2021. The method utilizes graph embedding and deep learning to achieve effective hash encoding learning in cross-modal datasets by maintaining data neighborhood consistency. This helps to improve the performance and efficiency of multimodal data retrieval.
- **DFAH** [43] Presented at Knowledge-Based Systems in 2022. A discrete fusion adversarial hashing method for cross-modal retrieval was proposed. It combines the ideas of discrete coding and adversarial training using adversarial training and fusion strategies, as well as the introduction of a discretized quantization loss.

B. SETTING

To ensure a fair comparison, the proposed method is implemented using the DFAH framework. Similarity-preserving subnetwork layers were established with 4096 dimensions for both images and texts. Throughout the training, the parameters of the feature-extracting subnetworks remained fixed, with updates solely applied to similarity-preserving sub-networks. Employing a mini-batch SGD optimizer with a 0.9 momentum and weight decay of 0.0005, we utilized a mini-batch size of 32 with a learning rate set at 0.005. To determine the optimal hyperparameters α, β, γ, and δ. A validation test was conducted using 2000 randomly selected queries from the retrieval database to determine the optimal hyperparameters α, β, γ, and δ. Figure 6 illustrates the sensitivity analysis of each hyperparameter on NUS-WIDE within the range $\{0.001, 0.01, 0.1, 1, 2\}$ for α, β, γ, and δ. The results indicated that the best performance occurred when $\alpha = \delta = \gamma = 1$, and $\beta = 0.1$. Similarly, employing the same method for the other two datasets yielded the optimal hyperparameters $\alpha = \delta = \gamma = 1$, and $\beta = 0.1$, resulting in the best performance. Our experiments were conducted using a workstation equipped with an 11th Gen Intel(R) Core(TM) i7-1165G7 processor at 2.8GHz and an NVIDIA GTX 1080ti GPU with 11GB of memory and CUDA acceleration. The computer ran 16GB of content.

C. PERFORMANCE EVALUATION

To thoroughly assess the performance of our proposed cross-modal retrieval method, we used two main evaluation metrics: mean average precision (mAP) scores and precision-recall curves (PR-curves).

Mean average precision(mAP) is a commonly used performance metric that is particularly applicable to cross-modal retrieval tasks. In our experiments, mAP was used to measure the quality of retrieval results on different datasets. It considers the ranking of retrieved relevant texts/images and the relationships between them, thereby providing a more comprehensive performance evaluation. In calculating mAP, we focus on the average precision of each query (query) and then average the average precision of all queries to obtain the final mAP score. A high mAP score indicates that the retrieval results are of higher quality and ranked higher for relevant items.

Precision-recall curves are another method that is commonly used to evaluate retrieval tasks. In our experiments, we plotted PR curves to visualize the trade-off between

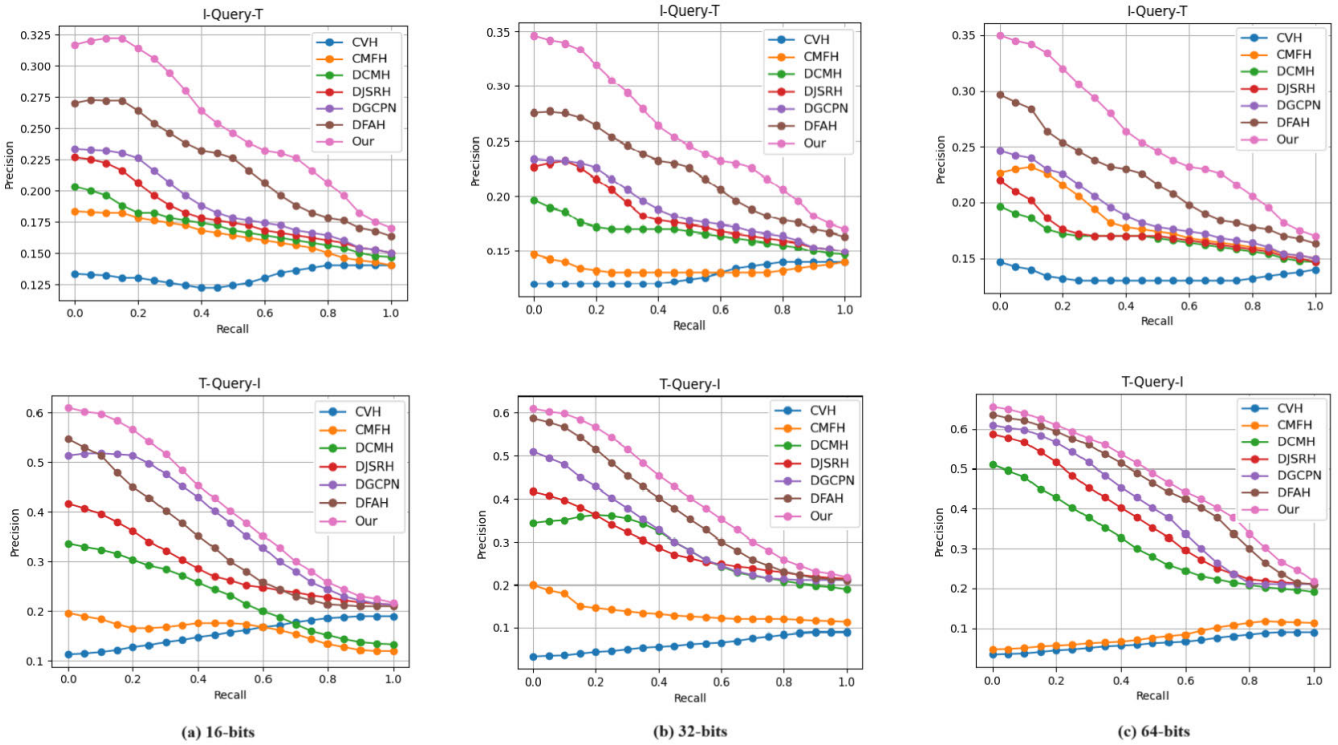

FIGURE 7. PR-curves on Wikipedia.

TABLE 2. Comparison of MAP scores on Wikipedia dataset.

Title 1	Methods	16-bits	32-bits	64-bits
I query T	CVH	0.46	0.149	0.155
	CMFH	0.254	0.258	0.262
	DCMH	0.309	0.318	0.329
	DJSRH	0.388	0.403	0.412
	DGCPN	0.404	0.413	0.420
	DFAH	0.739	0.740	0.751
	MGSGH	**0.751**	**0.757**	**0.759**
T query I	CVH	0.275	0.236	0.168
	CMFH	0.612	0.630	0.630
	DCMH	0.622	0.633	0.645
	DJSRH	0.611	0.635	0.646
	DGCPN	0.539	0.550	0.558
	DFAH	0.761	0.764	0.776
	MGSGH	**0.771**	**0.772**	**0.791**

TABLE 3. Comparison of MAP scores on NUS-WIDE dataset.

Title 1	Methods	16-bits	32-bits	64-bits
I query T	CVH	0.372	0.362	0.406
	CMFH	0.559	0.570	0.578
	DCMH	0.511	0.519	0.524
	DJSRH	0.724	0.773	0.798
	DGCPN	0.625	0.635	0.654
	DFAH	0.639	0.660	0.666
	MGSGH	**0.759**	**0.788**	**0.810**
T query I	CVH	0.474	0.442	0.443
	CMFH	0.661	0.692	0.716
	DCMH	0.637	0.653	0.695
	DJSRH	0.712	0.744	0.771
	DGCPN	0.631	0.648	0.660
	DFAH	0.684	0.698	0.701
	MGSGH	**0.723**	**0.755**	**0.788**

different precision rates and recall rates. The precision rate represents the proportion of retrieved items that are truly relevant and the recall rate represents the proportion of retrieved truly relevant items out of all truly relevant items. By plotting the PR curves, we can see how the precision rate performs under different recall rates, which helps us understand the trade-off between the different performance metrics.

Using these two evaluation metrics, we comprehensively assessed the performance of our method across different datasets and model comparisons. The mAP scores provide an overall performance metric, while the PR-curves help us to gain insight into the performance at different operating points to better evaluate the superiority of our method.

The Wikipedia dataset is a relatively small cross-modal retrieval dataset with only 2866 image text pairs;however, the label classification inside is relatively abstract. Therefore, according to the settings in [44], we selected 2172 examples as training data, 462 examples as test data, and 231 examples as validation examples. From Table 2, we can see that the seven cross-modal retrieval methods, including ours, do not perform very well on the Wikipedia dataset compared with the other datasets. The reason for this lies in the greater abstractness of the categories within Wikipedia compared with the other two datasets, resulting in increased difficulty in acquiring discernible semantic features. However, our method still outperformed the other methods. Compared to the CVH and CMFH methods, our direct learning hash

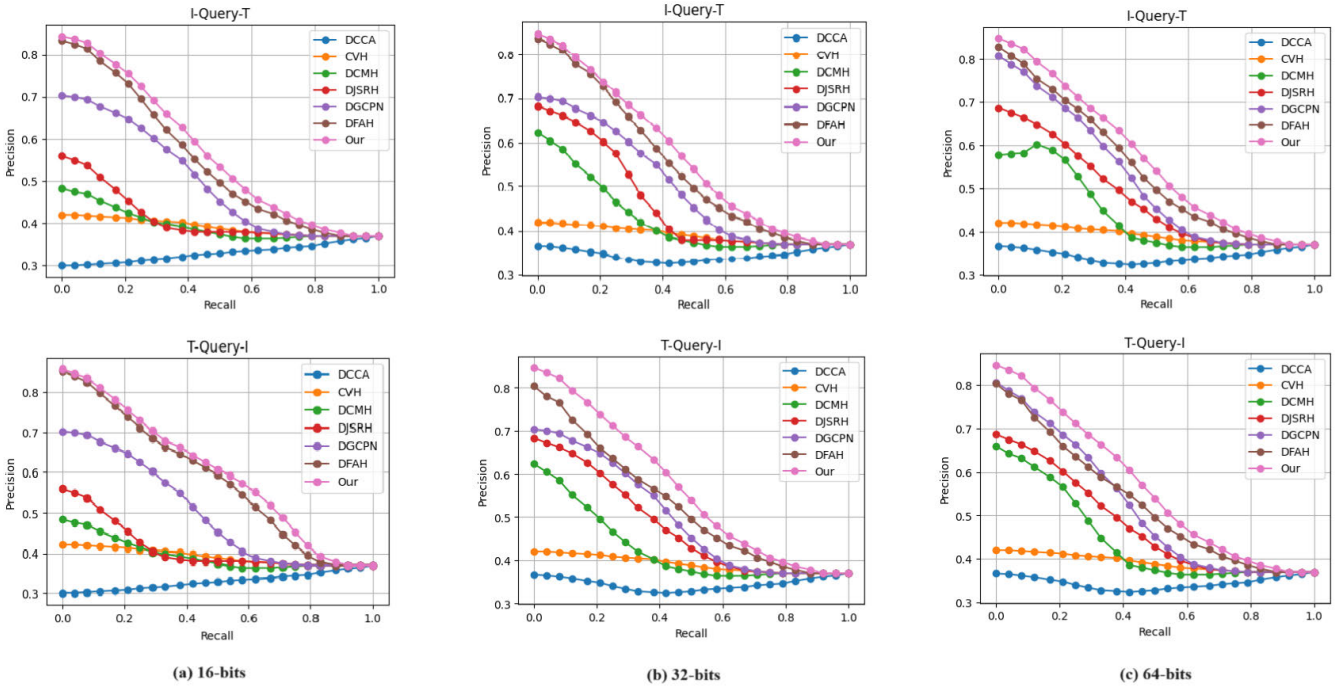

FIGURE 8. PR-curves on NUS-WIDE.

TABLE 4. Comparison of MAP scores on MIRFlickr-25K dataset.

Title 1	Methods	16-bits	32-bits	64-bits
I query T	CVH	0.606	0.599	0.596
	CMFH	0.648	0.660	0.669
	DCMH	0.689	0.698	0.714
	DJSRH	0.810	0.843	0.862
	DGCPN	0.732	0.742	0.751
	DFAH	0.739	0.740	0.751
	MGSGH	**0.832**	**0.855**	**0.873**
T query I	CVH	0.591	0.583	0.576
	CMFH	0.617	0.624	0.631
	DCMH	0.692	0.704	0.718
	DJSRH	0.786	0.822	0.835
	DGCPN	0.729	0.741	0.749
	DFAH	0.761	0.764	0.776
	MGSGH	**0.806**	**0.835**	**0.893**

function strategy reflects advantages, indicating that our direct learning hash function strategy not only avoids the quantization loss but also maintains the similarity of similar objects in the binary space in a very good way. In addition, as shown Figure 7, the longer the hash code length, the better is the cross-modal retrieval performance, indicating that longer hash codes can reduce the impact of hash conflicts.

The NUS-WIDE dataset is an open-source cross-modal retrieval dataset that is very large, and some text descriptions are not only in English but also in other languages. Therefore, based on the ACMR settings [38], we selected 8000 samples as training data, 1000 samples as test data, and 1000 samples as validation examples. From Table 3, we can see that our method has better cross-modal retrieval results because we build a graph structure for tags and use graph strategies to explore the semantic dependencies. This is mainly because semantic knowledge is comprehensively and accurately represented by a multigraph so that more fine-grained semantic relationships can be mined to improve retrieval accuracy. This validates the importance of considering semantic connections among tags and highlights the value of investigating multiple tags as a promising avenue for research. Compared to the DCMH method, our method is more advantageous, because our intra-modal invariance strategy plays an important role in the common space mapping process. This strategy effectively reduced the heterogeneity between modalities in cross-modal representation learning, demonstrating the effectiveness of our method. However, based on Figure 8, our observation indicates that retrieving text from images yields superior results compared to retrieving images from text. This disparity stems from the fact that textual features capture the semantic essence of an object more effectively, whereas image features solely delineate the rudimentary visual attributes.

The MIRFLICKR-25K dataset contains many variations and a large amount of data. To ensure fairness in comparison, we selected 20015 samples as training data, 2000 samples as test data, and 2000 samples as validation examples based on the DFAH setting. As shown in Table 4, our method obtains better results than all other methods, compared to the CMFH method, which only uses coarse-grained features and ignores fine-grained complementary cues in cross-modal similarity learning. Our proposed method has a more powerful semantic representation capability, which means that it can capture more abstract conceptual information, and fully considered a combination of global coarse-grained semantic information

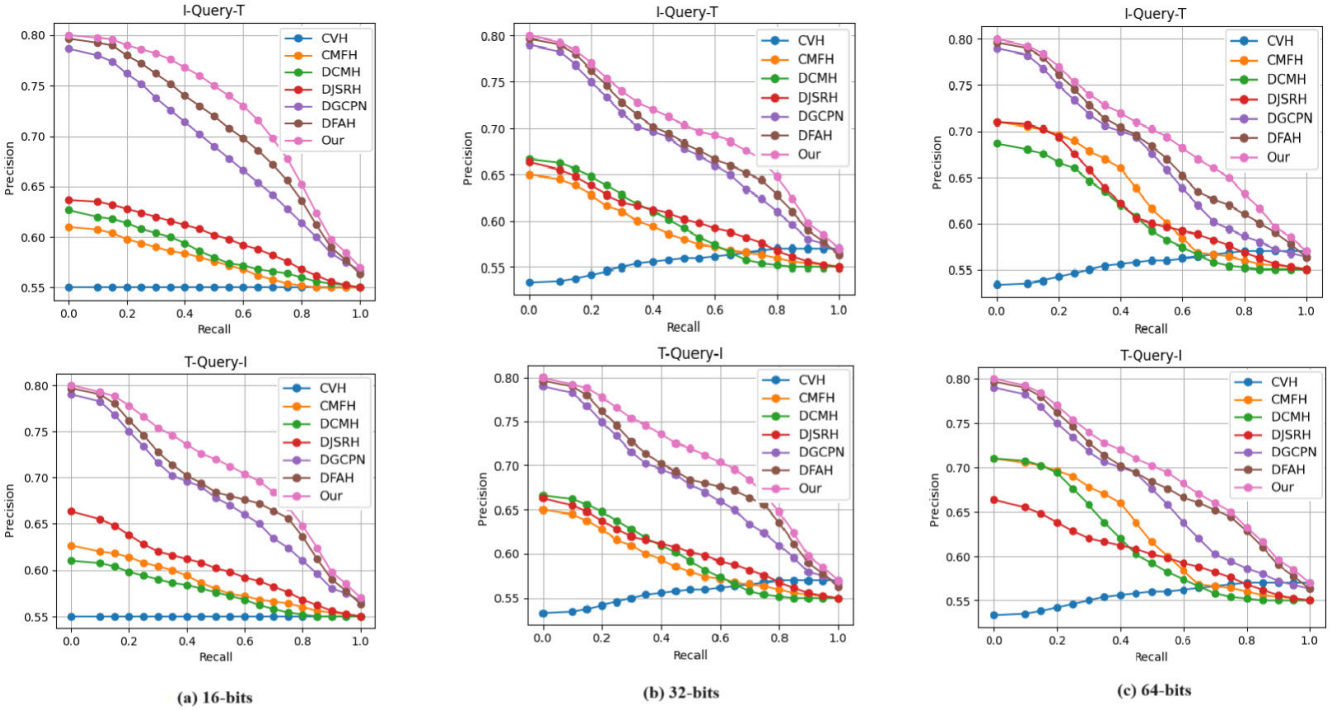

FIGURE 9. PR curves on MIRFLICKR-25K.

TABLE 5. The mAP@50 results on MIRFlickr-25k to evaluate the effectiveness of each part in our method.

Title 1	Methods	16-bits	32-bits	64-bits
I query T	MGSGH	0.832	0.855	0.873
	MGSGH-Nothing	0.499	0.518	0.518
	MGSGH-Only Graph	0.612	0.621	0.624
T query I	MGSGH	0.806	0.835	0.893
	MGSGH-Nothing	0.507	0.524	0.531
	MGSGH-Only Graph	0.661	0.694	0.716

and fine-grained local semantic information. In addition, from Figure 9, we can see that the DGCPN method also achieves relatively good results because the DGCPN method also considers fine-grained linguistic information, whereas the DGCPN method employs comprehensive similarity to prevent losses to pay attention to fine-grained linguistic information. DFAH, on the other hand, introduces the adversarial idea, which makes full use of semantic concepts to expand the interval distance between different semantic concepts, but does not consider inter-modal heterogeneity and intra-modal invariance.

D. FURTHER ANALYSIS

To further demonstrate the advantages of our proposed method in some aspects, we made some variations in the feature extraction preservation of image text and the learning of hash functions while keeping the other settings the same. MGSGH was the method proposed in this study. MGSGH-Nothing simply uses deep neural networks to extract features of images and texts, and then maps them to a binary common subspace for cross-modal retrieval. MGSGH-Only Graph fully explores the inherent correlation between data of different granularities by constructing a scene graph neural network, fusing coarse-grained and fine-grained representations, and then simply mapping them to a binary common subspace for cross-modal retrieval.

Table 5 lists the data with hash code lengths of 16, 32, and 64 bits obtained by applying the above three methods to the MIRFlickr-25k dataset. From this table, we can see that MGSGH achieves the best results, indicating that the construction of scene graph neural networks and direct learning of hash codes play important key roles. Scene graphs can provide rich correlation modeling for different object and relationship combinations, and help achieve more accurate and comprehensive cross-modal retrieval. According to the experimental data from the MGSGH and MGSGH-Only Graph, directly learning the hash code can effectively reduce quantization loss, retain more semantic information, and improve the efficiency of cross-modal hash retrieval.

V. CONCLUSION

This study introduces a scene graph model to perform hierarchical semantic association modeling on multimodal data. For each modality of data, coarse-grained semantic association information is represented using intra-modal global semantic maps, whereas fine-grained semantic association information is described using hierarchical local semantic maps based on scene graphs, providing sufficient semantic discrimination information for cross-modal quantization

coding learning. By enforcing comprehensive similarity preservation, we establish effective cross-modal representations to bridge the differences in the feature space or data structure between image and text modal data. In addition, a direct learning approach for generating hash codes is introduced, which reduces the impact of quantization loss. This approach not only simplifies the learning process but also enhances the quality of the resultant hash codes. The proposed framework was rigorously validated on various benchmark datasets, thereby demonstrating its superior performance in retrieval tasks.

REFERENCES

[1] C. Zhang, J. Song, X. Zhu, L. Zhu, and S. Zhang, "HCMSL: Hybrid cross-modal similarity learning for cross-modal retrieval," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 1, pp. 1–22, Jan. 2021.

[2] L. Zhu, J. Song, X. Zhu, C. Zhang, S. Zhang, and X. Yuan, "Adversarial learning-based semantic correlation representation for cross-modal retrieval," *IEEE MultimediaMag.*, vol. 27, no. 4, pp. 79–90, Oct. 2020.

[3] Z. Ji, H. Wang, J. Han, and Y. Pang, "SMAN: Stacked multimodal attention network for cross-modal image–text retrieval," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 1086–1097, Feb. 2022.

[4] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

[5] S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen, "Cross-modal scene graph matching for relationship-aware image-text retrieval," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1497–1506.

[6] C. Wang, H. Yang, and C. Meinel, "Deep semantic mapping for cross-modal retrieval," in *Proc. IEEE 27th Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2015, pp. 234–241.

[7] J. Zhang, Y. Peng, and M. Yuan, "SCH-GAN: Semi-supervised cross-modal hashing by generative adversarial network," *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 489–502, Feb. 2020.

[8] L. Wu, Y. Wang, and L. Shao, "Cycle-consistent deep generative hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1602–1612, Apr. 2019.

[9] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[10] L. Zhen, P. Hu, X. Peng, R. S. M. Goh, and J. T. Zhou, "Deep multimodal transfer learning for cross-modal retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 798–810, Feb. 2022.

[11] L. Zhang and X. Wu, "Multi-task framework based on feature separation and reconstruction for cross-modal retrieval," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108217.

[12] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1–6.

[13] X. Liu, A. Li, J.-X. Du, S.-J. Peng, and W. Fan, "Efficient cross-modal retrieval via flexible supervised collective matrix factorization hashing," *Multimedia Tools Appl.*, vol. 77, no. 21, pp. 28665–28683, Nov. 2018.

[14] D. Wang, X. Gao, X. Wang, and L. He, "Label consistent matrix factorization hashing for large-scale cross-modal similarity search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2466–2479, Oct. 2019.

[15] 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.

[16] 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. Artif. Intell.*, Feb. 2017, vol. 31, no. 1, pp. 1–8.

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

[18] S. Su, Z. Zhong, and C. Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3027–3035.

[19] X. Liu, Y. Mu, D. Zhang, B. Lang, and X. Li, "Large-scale unsupervised hashing with shared structure learning," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1811–1822, Sep. 2015.

[20] Y. Hu, Z. Jin, H. Ren, D. Cai, and X. He, "Iterative multi-view hashing for cross media indexing," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 527–536.

[21] G. Irie, H. Arai, and Y. Taniguchi, "Alternating co-quantization for cross-modal hashing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1886–1894.

[22] J. Huang, C. Min, and L. Jing, "Unsupervised deep fusion cross-modal hashing," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 358–366.

[23] D. Yang, D. Wu, W. Zhang, H. Zhang, B. Li, and W. Wang, "Deep semantic-alignment hashing for unsupervised cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 44–52.

[24] L. He, X. Xu, H. Lu, Y. Yang, F. Shen, and H. T. Shen, "Unsupervised cross-modal retrieval through adversarial learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1153–1158.

[25] J. Zhang, Y. Peng, and M. Yuan, "Unsupervised generative adversarial cross-modal hashing," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–8.

[26] J. Yu, H. Zhou, Y. Zhan, and D. Tao, "Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 5, pp. 4626–4634.

[27] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3668–3678.

[28] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1270–1279.

[29] D. Zhang and X.-J. Wu, "Robust and discrete matrix factorization hashing for cross-modal retrieval," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108343.

[30] Z. Shen, D. Zhai, X. Liu, and J. Jiang, "Semi-supervised graph convolutional hashing network for large-scale cross-modal retrieval," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2366–2370.

[31] S. Liu, S. Qian, Y. Guan, J. Zhan, and L. Ying, "Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1379–1388.

[32] D. Ma, J. Liang, X. Kong, and R. He, "Frustratingly easy cross-modal hashing," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 237–241.

[33] 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.

[34] E. Yu, J. Ma, J. Sun, X. Chang, H. Zhang, and A. G. Hauptmann, "Deep discrete cross-modal hashing with multiple supervision," *Neurocomputing*, vol. 486, pp. 215–224, May 2022.

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

[36] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[37] F. Zhang, M. Xu, and C. Xu, "Geometry sensitive cross-modal reasoning for composed query based image retrieval," *IEEE Trans. Image Process.*, vol. 31, pp. 1000–1011, 2022.

[38] 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.

[39] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 251–260.

[40] 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.*, Jul. 2009, pp. 1–9.

[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] 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.

[43] J. Li, E. Yu, J. Ma, X. Chang, H. Zhang, and J. Sun, "Discrete fusion adversarial hashing for cross-modal retrieval," *Knowl.-Based Syst.*, vol. 253, Oct. 2022, Art. no. 109503.
[44] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 53–3846.

ZHICHAO HAN was born in Hebei, China. He received the M.Sc. degree from Guangxi Normal University, China, in 2018. He is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. His research interests include machine learning, deep learning, and cross-modal retrieval.

AZREEN BIN AZMAN (Member, IEEE) received the Diploma degree in software engineering from the Institute of Telecommunication and Information Technology, in 1997, the Bachelor of Information Technology degree in information systems engineering from Multimedia University, Malaysia, in 1999, and the Ph.D. degree in computing science specializing in information retrieval from the University of Glasgow, Scotland, in September 2007. Before joining the Ph.D. degree, he was with the industry, for a few years. He is currently an Associate Professor with University Putra Malaysia. His current research interests include information retrieval, text mining, natural language processing, and intelligent systems. He serves as a Committee Member for Malaysian Society of Information Retrieval and Knowledge Management (PECAMP) and Malaysian Information Technology Society (MITS).

FATIMAH BINTI KHALID (Member, IEEE) received the B.Sc. degree in computer science from University Technology Malaysia (UTM), in 1992, the master's degree from University Kebangsaan Malaysia (UKM), in 1997, and the Ph.D. degree in system science and management from UKM, in 2008. From 1993 to 1995, she was a System Analyst with UKM. After the master's degree, she started involved in teaching with the Sal College, until 1999, and continued with University Putra Malaysia, in June 1999. In January 2016, she was a Secondment with the Computer Science Department, Tabuk University, Saudi Arabia, for one and half years. She is currently with the Faculty of Computer Science and Information Technology, as a Lecturer and an Associate Professor.

MAS RINA BINTI MUSTAFFA (Senior Member, IEEE) received the Ph.D. degree in multimedia systems from Universiti Putra Malaysia, Malaysia. She is currently an Associate Professor with University Putra Malaysia. Her current research interests include multimedia information retrieval, computer vision, pattern recognition, image processing, multimedia computing, multimedia systems, and applications.