Multilevel Deep Semantic Feature Asymmetric Network for Cross-Modal Hashing Retrieval

Xiaolong Jiang (iD), Jiabao Fan (iD), Jie Zhang (iD), Ziyong Lin (iD), and Mingyong Li (iD), *Member, IEEE*

Abstract—Cross-modal hash retrieval has been widely applied due to its efficiency and low storage overhead. In the domain of supervised cross-modal hash retrieval, existing methods exhibit limitations in refining data features, leading to insufficiently detailed semantic information extraction and inaccurate reflection of data similarity. The challenge lies in utilizing multi-level deep semantic features of the data to generate more refined hash representations, thereby reducing the semantic gap and heterogeneity caused by different modalities. To address this challenging problem, we propose a multilevel deep semantic feature asymmetric network structure (MDSAN). Firstly, this architecture explores the multilevel deep features of the data, generating more accurate hash representations under richer supervised information guidance. Secondly, we investigate the preservation of asymmetric similarity within and between different modalities, allowing for a more comprehensive utilization of the multilevel deep features to bridge the gap among diverse modal data. Our network architecture effectively enhances model accuracy and robustness. Extensive experiments on three datasets validate the significant improvement advantages of the MDSAN model structure compared to current methods.

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/8718

Index Terms—Cross-modal hashing, cross-modal retrieval, mul-feature method, graph convolutional network.

I. INTRODUCTION

With the continuous development and deepening of the digital society, there has been an explosive growth in multimedia data. The associated challenges in the utilization of multimedia data have become apparent in the public domain, attracting the involvement of numerous researchers dedicated to the study of processing and exploiting large-scale multimedia data [1] [2] [3]. Multimedia data may encompass various modalities, including images, text, audio, video, and more. Faced with the presence of multiple modalities within the same dataset, cross-modal retrieval encounters challenges such as high-dimensional data storage and slow retrieval speed. In recent years, research in the field of cross-modal retrieval has been continuously advancing. Hash retrieval methods, due to their ability to reduce storage space and enhance

retrieval efficiency, have gained significant attention. Hash retrieval methods are favored for their distinct advantages over real-valued retrieval methods, offering improvements in both storage efficiency and retrieval speed [4] [5] [6] [7].

The objective of hash methods is to transform high-dimensional multimedia data from the same category into similar hash codes in Hamming space. Hash methods are categorized into deep hash methods and shallow hash methods. Most shallow hash retrieval methods directly utilize manually extracted features as inputs, greatly impacting the performance of retrieval [8] [9] [10] [11] [12]. In recent years, within the rapidly evolving landscape of deep learning, deep hash methods have garnered significant attention due to their efficiency and accuracy. Deep hash methods have continuously developed and progressed in the exploration of research.

Cross-modal hash retrieval can be classified into two categories: supervised methods and unsupervised methods. Unsupervised hash methods perform retrieval without semantic guidance, utilizing mapping functions between the original feature space and the Hamming space. On the other hand, supervised methods leverage rich semantic information labels to extract more accurate and abundant feature information, leading to higher retrieval performance [13] [14] [15]. This paper focuses on the study of supervised cross-modal deep hash methods.

Challenges persist despite the continuous development in cross-modal hashing and deep hashing retrieval research. These challenges include inconsistent feature extraction within the same modality, as well as the decrease in retrieval accuracy caused by the errors in bridging the semantic gap across modalities [15]. To address these issues, we propose a multi-modal hash retrieval approach based on a multi-level semantic feature network structure. This methodology aims to obtain hash representations of higher quality and greater richness, facilitating improved consistency in feature extraction within modalities, effective enhancement of model robustness through better alignment of semantic features, and efficient integration of local and global feature relationships. Additionally, to mitigate semantic disparities within modalities and heterogeneity across modalities, our proposed model MDSAN employs a cross-modal multi-level semantic asymmetric preservation mechanism. This technique significantly suppresses intra-modal semantic gaps and reduces heterogeneity between different modalities to a considerable degree. Extensive experiments confirm the outstanding performance of the proposed MDSAN model. In conclusion, our contributions can be summarized as:

1. We designed a multilevel deep feature network specifi-

cally for extracting rich semantic features from cross-modal data. This network structure is capable of extracting abundant semantic features to generate accurate hash representations, effectively alleviating the problem of feature loss when compressing high-dimensional data into a low-dimensional space.

2. Our network introduces an asymmetric multilevel deep semantic preservation fusion approach, effectively bridging the semantic gap within and between modalities, as well as mitigating heterogeneity gaps.

3. Our method was extensively validated on three publicly available datasets. The experimental results show that our approach surpasses state-of-the-art methods.

In Section II, we reviewed related research. In Section III, we provided a detailed description of MDSAN. Section IV presented the experimental results. Finally, we summarized this paper in Section V.

II. Related Work

A. Shallow Hashing Method

With the continuous deepening of information digitization, data formats have become increasingly diverse. Faced with a large volume of multimodal data, how to efficiently utilize it has become a noteworthy issue. Traditional shallow hash algorithms primarily employ matrix decomposition and feature mapping techniques to generate a generic representation of multimodal data, which is then binarized to produce hash codes. SCM utilizes label information to construct similarity matrices and further applies spectral relaxation to solve the binary NP-hard problem [7]. GSPH constructs a hashing framework by utilizing label information to build an affinity matrix and applying ridge regression and kernel logistic regression to learn the hash function [16]. SePH uses a probabilistic model to learn and predict unified hash codes from different modalities [17]. SMFH innovatively employs matrix decomposition to explore label semantic information in multimodal data [18]. However, shallow hash methods cannot extract deeper semantic feature information, leaving room for further improvement in retrieval accuracy.

B. Deep Hashing Method

In recent years, with the continuous deepening and application of deep learning, deep learning networks have been introduced into the field of cross-modal hashing to extract deep features. Deep hash methods primarily utilize deep learning to capture correlations within data, thereby expressing relationships between different data modalities more effectively. DCMH proposed an end-to-end learning approach that can directly learn discrete hash codes [19]. SSAH addresses the heterogeneity gap between cross-modal data through the use of adversarial learning [20]. CMHH uses diverse pairwise constraints to learn hash codes between modalities and within modalities, integrating them into an end-to-end framework [21]. MLSPH uses multi-labels as supervised guidance while preserving both inter-modal and intra-modal data for feature learning [22]. MESDCH utilizes multi-label semantic computation to assess the correlation between instances while effectively mitigating the influence of label noise [23]. SDSHL

constructs a dual-semantic-guided method to explore pairwise similarity and builds a connected latent hash space [24]. SDAH utilizes a decoupling approach to separate original features into private and common features. It introduces a variational information bottleneck to preserve more semantic information when compressing high-dimensional information into low-dimensional information [25]. MIAN learns hierarchical semantic features through dual asymmetric matrices and integrates them to address the gap between modalities [26]. MAFH treats labels as a modality and simultaneously designs an extensible weight encoding strategy [27]. LGCNH (Local Graph Convolutional Network Hashing) uses GCN to reconstruct local graphs of multimodal data into different modality features, preserving the underlying data structure [28]. MCGCN addresses the issue of the cross-modal semantic gap by constructing a dedicated cross-modal graph to bridge the semantic gap [29]. Compared to shallow methods, the main advantage of deep hash methods lies in the powerful representation learning capabilities of deep neural networks. However, the above methods have certain shortcomings in the extraction and utilization of semantic features from data. Particularly, when faced with image data with rich semantic features, common feature extraction methods may not effectively capture multi-level features comprehensively. Additionally, classical symmetric metrics in the similarity calculation metric have limitations in preserving features. To tackle these challenges, we put forward a new multilevel semantic feature network solution.

III. Proposed Method

A. Problem Definition and Notation

In this section, we will define and explain the formulas and symbols used in this paper. Our exploration is focused on cross-modal data with images and text. Let us consider a dataset consisting of n image-text pairs denoted as $\Gamma = \{z_i\}_{i=1}^{n}$ where $z_i = (l_i, t_i, p_i)$, $p_i \in Q^{r_1 \times 1}$ and $t_i \in Q^{r_2 \times 1}$ represent the image and text modalities of the i-th object, r_1 and r_2 are the dimensions of the corresponding original data, respectively. $l_i \in \{0,1\}^{a \times 1}$ represents the label vector of the i-th instance z_i, where a denotes the number of classes. If z_i belongs to the j-th class, then $L_{ji} = 1$ otherwise $L_{ji} = 0$. Furthermore, the image feature matrix, text feature matrix, and multi-label annotation matrix for the entire dataset Γ are respectively represented as $P \in Q^{r_1 \times 1}$, $T \in Q^{r_2 \times 1}$, $L \in Q^{a \times n}$.

We define S as the pairwise similarity matrix, representing the semantic connections between each pair of examples. If $S_{ij} = 1$, it signifies z_i and z_j indicating semantic similarity and identical labels. If $S_{ij} = 0$, holds, it is completely opposite. $S_{ij} = dot(P \text{ or } T, L)$, $dot(\cdot)$ represents the dot product similarity calculation function, which expands as follows:

$$S_{Pij} = \sum_{k=1}^{n} p_{ik} \times (l_{jk})^{\mathrm{T}} \tag{1}$$

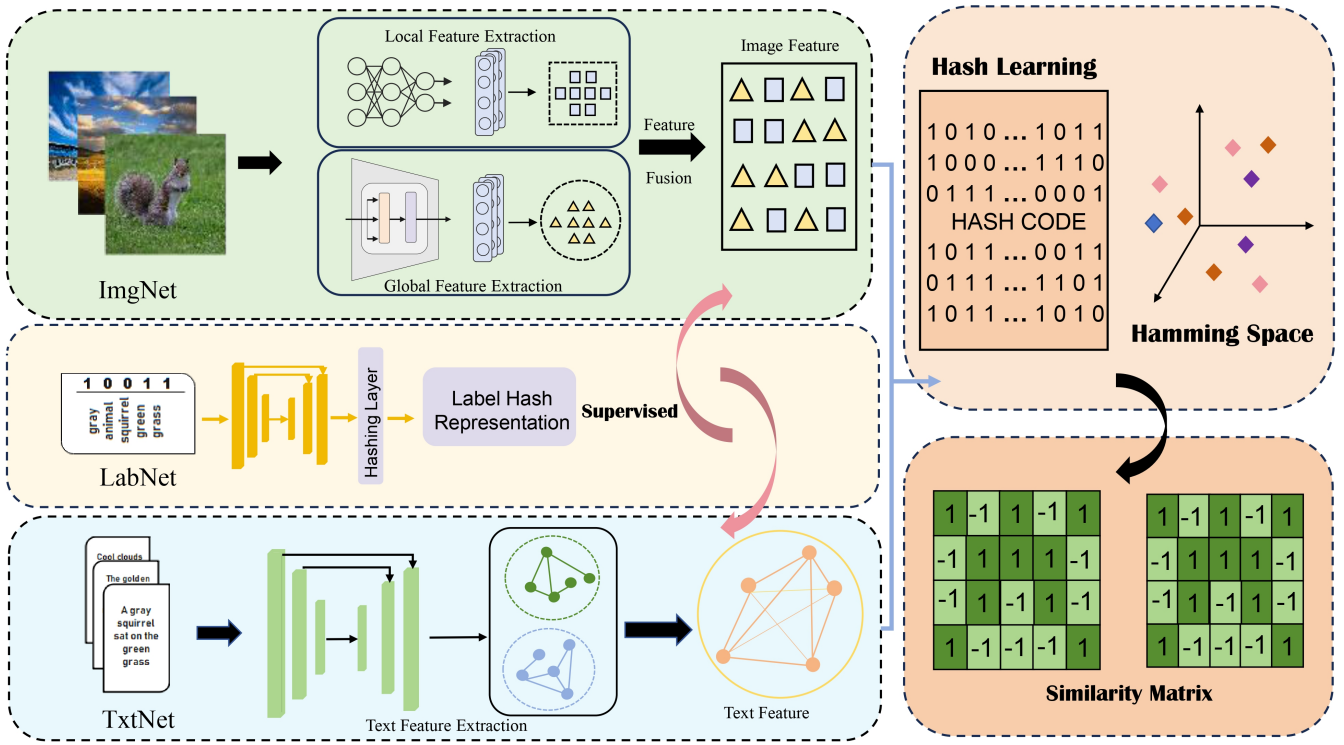$$S_{Tij} = \sum_{k=1}^{n} t_{ik} \times (l_{jk})^{\mathrm{T}} \tag{2}$$

Fig. 1. The framework of MDSAN. Workflow: Initially, data is separately input into the image, text, and label networks for preliminary processing. The deep features learned are embedded into a K-dimensional Hamming space. The final output of the network is a fused hash code from multiple intermediate layer outputs.

where $\boldsymbol{p}$, $\boldsymbol{t}$, and $\boldsymbol{l}$ denote the vector representations of matrices $\boldsymbol{P}$, $\boldsymbol{T}$, and $\boldsymbol{L}$, respectively. In the cross-modal hash retrieval method, we learn a unified hash code representation with two modal semantic features by mapping data from the high-dimensional space of images and text to a common y-byte discrete Hamming space. We use $V^{t,p} = v^{t,p}(\Gamma; \theta^t, \theta^p)$ to represents hash learning function and $O^{t,p} = sign(V^{t,p}) \in \{-1, +1\}^{y \times n}$ to represent the learned hash codes with semantic features, where θ and y are the parameters and length of the hash codes, respectively. The superscripts t and p correspond to the text and image modalities. $sign(\cdot)$ is a sign function defined as:

$$sign(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases} \tag{3}$$

The semantic similarity of data samples can be represented by converting the hash function into binary code. In general, our semantic-preserving hashing function is defined as:

$$\min_{\boldsymbol{O}} \sum_{i=1}^{n} \sum_{j=1}^{n} ||o_i - o_j|| S_{ij} \tag{4}$$

Where $o_i = v(x_i)$ is the hash code of sample x_i, $v(\cdot)$ is hash function. To preserve more semantic information, we utilize a $n \times m$ asymmetric computation approach, which, in comparison to the $n \times n$ matrix, possesses more powerful semantic computing capabilities and greater expansibility. The overarching objective formulation for extracting hash codes

imbued with semantic information is defined as follows:

$$\min_{\boldsymbol{O}} \sum_{i=1}^{n} \sum_{j=1}^{m} \|\boldsymbol{o}_i - \boldsymbol{o}_j\| S_{ij}$$
$$= \min_{\boldsymbol{O}} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\boldsymbol{o}_i^\top \boldsymbol{o}_i + \boldsymbol{o}_j^\top \boldsymbol{o}_j - 2\boldsymbol{o}_i^\top \boldsymbol{o}_j S_{ij} \right) \tag{5}$$
$$= \min_{\boldsymbol{O}} \sum_{i=1}^{n} \sum_{j=1}^{m} 2ymn - 2\boldsymbol{o}_i^\top \boldsymbol{o}_j S_{ij}$$
$$s.t. o_i \in \{-1, 1\}^y, S \in \{-1, 1\}^{n \times m}$$

where $m \neq n$, constrained by discrete relationships and the coupling of discrete codes between o_i and o_j, satisfies $o_i^\top o_i = o_j^\top o_j = y$. The objective is to maximize the likelihood of observed paired semantics $\boldsymbol{S}_{ij}$ on the inner product between two binary codes o_i and o_j. Based on the negative log-likelihood similarity matrix $\boldsymbol{S}$ proposed in DCMH, the probability of $\boldsymbol{S}$ under the condition ϖ_{ij} can be expressed as :

$$p(S_{ij}|O_{ij}) = \begin{cases} \sigma(\varpi_{ij}) & S_{ij} = 1 \\ 1 - \sigma(\varpi_{ij}) & S_{ij} = 0 \end{cases} \tag{6}$$

Where $\varpi_{ij} = \frac{1}{2} o_i^\top o_j$, $\sigma(\varpi_{ij}) = \frac{1}{1 + e^{-\varpi_{ij}}}$. Minimizing the negative log-likelihood is equivalent to maximizing the likelihood, which aims to maximize the similarity between image-text pairs at $\boldsymbol{S}_{ij} = 1$ while minimizing it at $\boldsymbol{S}_{ij} = 0$.

B. Model Framework

This section elaborates on the primary network structure of this model. The workflow of this model is illustrated in Fig 1. The model consists primarily of two components: the feature extraction module and the modality fusion module.

1) Feature Extraction Module: Within this module, there are three main networks: image, text, and label network. In the following we uniformly use ImgNet, TxtNet, and LabNet to denote the corresponding networks. We use the negative log-likelihood algorithm to calculate the semantic similarity between cross-modal instances pre-and post-training. The post-training semantic similarity among cross-modal instances is computed as the sum of similarities between the hash representations of intermediate layer features and label hash representations. Traditional image backbone networks commonly utilize widely known coarse-grained convolutional neural networks as their backbone. Following the emergence of Transformer, research has explored the adoption of independently pretrained fine-grained model networks as backbone networks [30] [31].
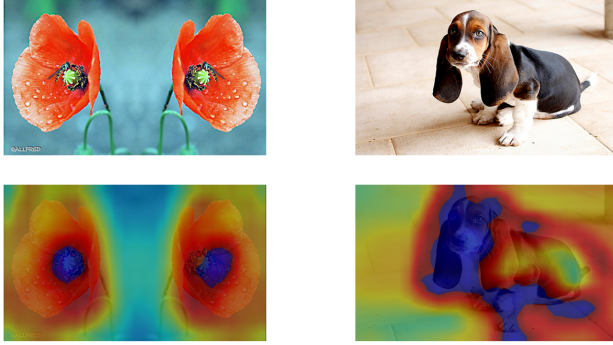

Fig. 2. Examples of features learned by the image network.

However, in the process of research and practical implementation, it is not difficult to identify the following issues: 1. Coarse-grained networks exhibit excellent performance in many tasks, but they also have certain drawbacks. These networks categorize target classes into larger categories, potentially leading to the loss of fine details during the classification process. This may result in the model struggling to differentiate categories with similar features, thereby reducing classification accuracy. Additionally, coarse-grained networks may fail to capture subtle differences between target categories. Due to the merging of categories into larger ones, the model may not learn features that distinguish these subtle differences. When categories are merged into larger ones, the model may encounter information confusion. Feature differences between different categories might be mixed, making it challenging for the model to accurately differentiate between them. 2. In fine-grained classification tasks, there may be an imbalance in the number of samples for different categories. Some categories may have a larger number of training samples, while others may have very few. Due to the small differences between target categories in fine-grained classification tasks, models can easily overfit. To address the aforementioned issues, a carefully designed image network has been developed, incorporating

global and local similarity into the learning of binary codes to generate more effective hash codes. Our backbone network is constructed using a combination of fine-grained and coarse-grained networks. The structure of the image network in this approach adopts a dual-stream feature network pattern for extracting features from images. The coarse-grained network focuses on the coarse extraction of local features, enabling the extraction of richer category information even in datasets with fewer categories. The fine-grained network is employed to extract the global features of our images, allowing us to capture more comprehensive and nuanced deep image features to better understand image semantics. The combination of coarse-grained and fine-grained networks utilizes the coarse-grained network to merge fine-grained categories, obtaining higher-level semantic features. This multi-level semantic feature representation provides more comprehensive semantic knowledge and helps reduce semantic gaps within modalities.

Our ImgNet employs VGG19 [32] and Visual Transformer [33] as backbone networks. These networks seamlessly integrate global and local features into a unified optimization framework. The features learned by our image network are illustrated in Fig. 2. During the learning process, the acquired image features are combined, and the integrated information is input into the hash layer. Subsequently, relative feature hash codes are generated, and similarity matrix calculations are performed. We use $F_{img} = (o_i, v^p)$ to represent the final image features extracted from instances. Additionally, $\boldsymbol{p}_i^{local}$ denotes the local feature vector, and $\boldsymbol{p}_i^{global}$ represents the global feature vector. The objective function for ImgNet is as follows:

$$P = \sum_{i=1}^{n} p_i^{local} \cup p_i^{global} \tag{7}$$

Text information inherently possesses rich semantic content. To better extract text features, based on previous explorations, it is evident that the graph convolutional network (GCN) structure exhibits outstanding performance in cross-modal retrieval. The use of GCN can significantly enhance the accuracy of cross-modal retrieval. Compared to traditional convolutional networks, GCN has the characteristic of sharing information. This property enables it to not only extract information from local features but also effectively utilize global information, greatly improving the model's understanding and abstraction capabilities of the overall text structure. Consequently, it can extract multi-level deep semantic features that are essential for our needs. Therefore, in our TxtNet, we construct a graph convolutional neural network to build a text feature extraction network. This approach thoroughly explores intra-modal semantic similarity. $F_{txt} = (o_i, v^t)$ represents the text features. The propagation process of the Graph Convolutional Network (GCN) is described as follows:

$$H^{(\tau+1)} = \sigma(\tau)(\widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}} H^{(\tau)} W^{(\tau)}) \tag{8}$$

Where, $\widetilde{D}_{ii} = \Sigma_j \widetilde{A}_{ij}$, and $W^{(\tau)}$ represent the layer-specific weight matrix. $\sigma(\tau)$ represents the activation function. $H^{\tau} \in R^{d \times m}$ is the activation matrix of the τ layer. From formula (8), we can observe that Graph Convolutional Networks

(GCNs) excel at capturing spatial relationships within data. Therefore, the generated hash codes can effectively reflect the relationships between instances in the feature space.

Due to the importance of label information as a significant supervisory signal in supervised hashing methods, and the richness of category information within labels, we introduce labels as a modality into the training process of our image and text modalities. However, due to the specific characteristics of label information, which is low-dimensional and simple, differing from the rich features of image and text information, we only use features from the last fully connected layer in the label network. These extracted features are input into the hash layer to generate feature hash codes for supervising and guiding the training of the image and text networks. Employing a neural network to convert original labels into low-dimensional binary codes is an effective strategy for reducing the impact of noise present in the original labels. This concise and versatile representation helps prevent the model from relying too heavily on high-dimensional label features during training, thereby reducing overfitting and improving the model's ability to generalize and handle new data. This process enhances the training quality for ImgNet and TxtNet, ultimately strengthening the robustness of the trained models. Therefore, the binary code generation representation for our ImgNet and TxtNet is as follows:

$$O_{img,txt} = sign(F_{img}, F_{txt}) \tag{9}$$

Where F_{img} and F_{txt} represent semantic embedding of features for images and text.

2) Fusion Module for Multi-level Semantic Preservation: Generally, symmetry computation can be optimized through manifold graph learning strategies, including spectral hashing [34] and anchor graph hashing [35]. However, due to the computational constraints of a $n \times n$ matrix, it is challenging to optimize such a large manifold graph. Compared to most supervised methods that predefine a similarity matrix to preserve pairwise similarities between all instances, this Symmetry approach may not fully leverage label information and rich semantic sample information. Binary similarity may not accurately reflect the semantic relationships between samples. Moreover, it is infeasible to scale due to substantial memory and computational costs. The primary challenges in current cross-modal retrieval methods remain the intra-modal semantic gap and inter-modal heterogeneity. Multi-level deep semantic representation can help minimize intra-modal distances within the same category and maximize distances between different categories within the same modality. Simultaneously, it effectively reduces the heterogeneity gap between modalities caused by insufficient fine-grained semantic information, addressing the issue of semantic mismatch between modalities. As the similarity metric between cross-modal instances can accurately represent multi-level semantic relationships, we choose to adopt an asymmetric similarity metric. Compared to a symmetric metric n×n matrix, an asymmetric metric n×m matrix can effectively preserve multi-level deep semantics. Due to its characteristics, the asymmetric metric is more flexible in describing complex relationships compared to the

symmetric metric, which is well-suited for expressing the multi-level deep semantic information we need.

According to formula (5), we can redefine the objective function for intra-modal similarity calculation as follows:

$$\min_{O} \sum_{i=1}^{n} \sum_{j=1}^{m} - \{ [(f_i^{p,t})^{\mathrm{T}} f_j^{p,t} S_{ij}] + [(v_i^{p,t})^{\mathrm{T}} v_j^{p,t} S_{ij}] \} \tag{10}$$

Where $\boldsymbol{f}_i$ and $\boldsymbol{v}_i$ representing i-th column vector of matrix $\boldsymbol{F}$ and $\boldsymbol{V}$. p and t representing image and text modality. The objective function for inter-modal similarity calculation is as follows:

$$\min_{O} \sum_{i=1}^{n} \sum_{j=1}^{m} -(\phi_i^{\mathrm{T}} v_j^p S_{ij} + \phi_i^{\mathrm{T}} v_j^t S_{ij}) \tag{11}$$

By optimizing Eq. 11 as follows:

$$\min_{O} \sum_{i=1}^{n} \sum_{j=1}^{m} -Tr \left\{ sign(\Psi) S[v^p (P_s)^{\mathrm{T}} + v^t (T_s)^{\mathrm{T}}] \right\} \tag{12}$$

Where $Tr(\cdot)$ represents the trace of the matrix. P_S and T_S are samples selected from images and text samples, respectively. $\Psi = [\phi_1, \cdots, \phi_n]$ is our semantic encoding matrix.

C. Hash Learning

The quality of hash codes determines the accuracy of retrieval to some extent. In the process of transforming multi-level deep semantic information into corresponding hash representations, how to compress and retain information to the maximum extent becomes a crucial issue. Meanwhile, considering the reduction of losses between semantic feature hash code representations within modalities and the handling of semantic heterogeneity gaps between different modalities, for similar situations, we use intra-modal loss and inter-modal loss to generate distinctive hash codes. Intra-modal loss consists of two parts: image-image loss and text-text loss defined as L_{intra}^{Img} and L_{intra}^{Txt}. Labels, as crucial information for supervised hash code generation, are defined as L_{intra}^{Lab} our label learning loss. We aim to maintain as much inter-modal similarity as possible in the cross-modal hash field. If the dot product of the feature vectors of two instances is large, the probability that the two instances are more similar to each other is higher. Therefore, determining the similarity of various hash codes can be achieved by calculating the dot product of individual codes in the Hamming space. Through the learning in formulas (6) and (10), we can optimize the asymmetric pairwise loss function through negative log-likelihood, where the optimization objective is:

$$L_{intra}^{Img,Txt} = -\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{2} (f_i^{p,t})^{\mathrm{T}} f_j^{p,t} S_{ij} - log(1 + e^{\frac{1}{2}(f_i^{p,t})^{\mathrm{T}} f_j^{p,t}}) \tag{13}$$

$$L_{intra}^{Lab} = -\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{2} (v_i^{p,t})^{\mathrm{T}} v_j^{p,t} S_{ij} - log(1 + e^{\frac{1}{2}(v_i^{p,t})^{\mathrm{T}} v_j^{p,t}}) \tag{14}$$

To minimize the error between the predicted hash codes and the target binary codes, we define $L_{intra}^{Q} = \|V^{p,t} - O^{p,t}\|_F^2$

TABLE I
MDSAN MEAN AVERAGE PRECISION (MAP) PERFORMANCE COMPARISON OF 16, 32, AND 64-BIT HASH CODE
LENGTHS ON THREE DATASETS

Task	Method	IAPR TC-12			NUS-WIDE			MIRFlickr-25k		
		16bit	32bit	64bit	16bit	32bit	64bit	16bit	32bit	64bit
	SCM [7]	0.3887	0.3945	0.4068	0.4626	0.4792	0.4886	0.6354	0.6407	0.6556
	SePH [17]	0.4186	0.4298	0.4315	0.4797	0.4859	0.4906	0.6740	0.6813	0.6803
	CMSSH [36]	0.3049	0.3074	0.3010	0.3092	0.3099	0.3396	0.5600	0.5709	0.5836
	GSPH [16]	0.3716	0.3921	0.4015	0.4015	0.4151	0.4214	0.6068	0.6191	0.6230
	DCMH [19]	0.4530	0.4727	0.4919	0.5495	0.5820	0.5833	0.6526	0.6630	0.6658
	PRDH [37]	0.4761	0.4883	0.4925	0.5480	0.5865	0.5782	0.6513	0.6620	0.6712
I2T	CMHH [21]	0.4903	0.5074	0.5152	0.5541	0.5764	0.5733	0.6325	0.6524	0.6373
	SSAH [20]	0.5348	0.5619	0.5781	0.6163	0.6278	0.6140	0.7745	0.7882	0.7990
	SCAHN [38]	0.5206	0.5438	0.5551	0.6649	0.6631	0.6698	0.8155	0.8224	0.8243
	MLSPH [22]	0.5342	0.5727	0.5772	0.6375	0.6599	0.6796	0.7776	0.8041	0.8180
	MESDCH [23]	0.5426	0.5735	0.5764	0.6475	0.6687	0.6836	0.8034	0.8170	0.8320
	MIAN [26]	0.5014	0.5472	0.5705	0.6742	0.6913	0.7473	0.8261	0.8312	0.8427
	MAFH [27]	0.5584	**0.5967**	0.6056	0.6679	0.6962	0.7562	0.8157	0.8365	0.8462
	OURS	**0.5641**	0.5943	**0.6157**	**0.7724**	**0.7978**	**0.7976**	**0.8998**	**0.8953**	**0.9083**
	SCM [7]	0.3824	0.3897	0.4002	0.4261	0.4372	0.4478	0.6340	0.6458	0.6541
	SePH [17]	0.4667	0.4857	0.4936	0.6072	0.6280	0.6291	0.7139	0.7258	0.7294
	CMSSH [36]	0.3189	0.3282	0.3229	0.3167	0.3171	0.3179	0.5726	0.5776	0.5753
	GSPH [16]	0.4177	0.4452	0.4641	0.4995	0.5233	0.5351	0.6282	0.6458	0.6503
	DCMH [19]	0.4851	0.4976	0.5171	0.5480	0.5824	0.5857	0.6527	0.7072	0.7091
	PRDH [37]	0.5112	0.5283	0.5403	0.5081	0.5773	0.5797	0.6527	0.6713	0.6960
T2I	CMHH [21]	0.4790	0.4951	0.4963	0.5582	0.5714	0.5773	0.6865	0.7042	0.6834
	SSAH [20]	0.5265	0.5594	0.5726	0.6204	0.6251	0.6215	0.7860	0.7974	0.7910
	SCAHN [38]	0.5194	0.5379	0.5481	0.6665	0.6720	0.6740	0.8033	0.8091	0.8128
	MLSPH [22]	0.5061	0.5258	0.5518	0.6403	0.6560	0.6777	0.7615	0.7807	0.7994
	MESDCH [23]	0.5293	0.5622	0.5657	0.6368	0.6578	0.6697	0.7798	0.7971	0.8098
	MIAN [26]	0.5366	0.5592	0.5736	0.6755	0.7037	0.7295	0.7803	0.7964	0.7983
	MAFH [27]	0.5566	0.5833	0.5944	0.6745	0.7244	0.7323	0.7984	0.8198	0.8210
	OURS	**0.6262**	**0.6731**	**0.7008**	**0.7530**	**0.7686**	**0.7814**	**0.9148**	**0.9200**	**0.9327**

as the loss for the distance between the quantized feature representation and the hash representation. $L_{intra}^{C} = \left\| \widetilde{L}^{p,t} - L \right\|_{F}^{2}$ serves as the loss to measure the difference between the predicted class labels $\widetilde{L}$ and the original labels L. Through formulas (13) and (14), we intra-modal loss function is defined as:

$$L_{intra} = \alpha(L_{intra}^{Img,Txt} + L_{intra}^{Lab}) + \mu L_{intra}^{C} + \vartheta L_{intra}^{Q} \quad (15)$$

Where α, μ and ϑ are hyperparameters. The inter-modality loss is defined as: L_{inter}^{Img} and L_{inter}^{Txt}. Simultaneously, to better reduce the loss of conversion error between semantic features and target binary codes, we have defined $L_{inter}^{Q} = \|\Psi - O^{p,t}\|_{F}^{2}$, where ρ is hyperparameter. L_{inter}^{Img} and L_{inter}^{Txt} loss objective function is defined as follows:

$$L_{inter}^{Img} = -\sum_{i=1}^{n}\sum_{j=1}^{m} \left\{ \left[\frac{1}{2}(\phi_i^p)^{\mathrm{T}} v_j^p S_{ij}\right] - [log(1 + e^{\frac{1}{2}(\phi_i^p)^{\mathrm{T}} v_j^p})]\right\}$$
$$(16)$$

$$L_{inter}^{Txt} = -\sum_{i=1}^{n}\sum_{j=1}^{m} \left\{ \left[\frac{1}{2}(\phi_i^t)^{\mathrm{T}} v_j^t S_{ij}\right] - [log(1 + e^{\frac{1}{2}(\phi_i^t)^{\mathrm{T}} v_j^t})]\right\}$$
$$(17)$$

According to formulas (16) and (17), we inter-modal loss function is defined as:

$$L_{inter} = \alpha(L_{inter}^{Img} + L_{inter}^{Txt}) + \rho L_{inter}^{Q} \quad (18)$$

Based on formulas (15) and (18), we can derive our overall objective function as follows:

$$\min_{O} L_{ours} = L_{intra} + L_{inter} \quad (19)$$

Where L_{intra} helps us learn more accurate hash representations within each modality, L_{inter} enables us to reduce the gap between modalities, thereby enhancing the reliability of the generated hash codes.

IV. EXPERIMENTS

In this section, we will describe our experimental setup and outline the evaluation details. We conducted extensive experiments on different datasets to validate the effectiveness of our approach. By comparing with some currently advanced
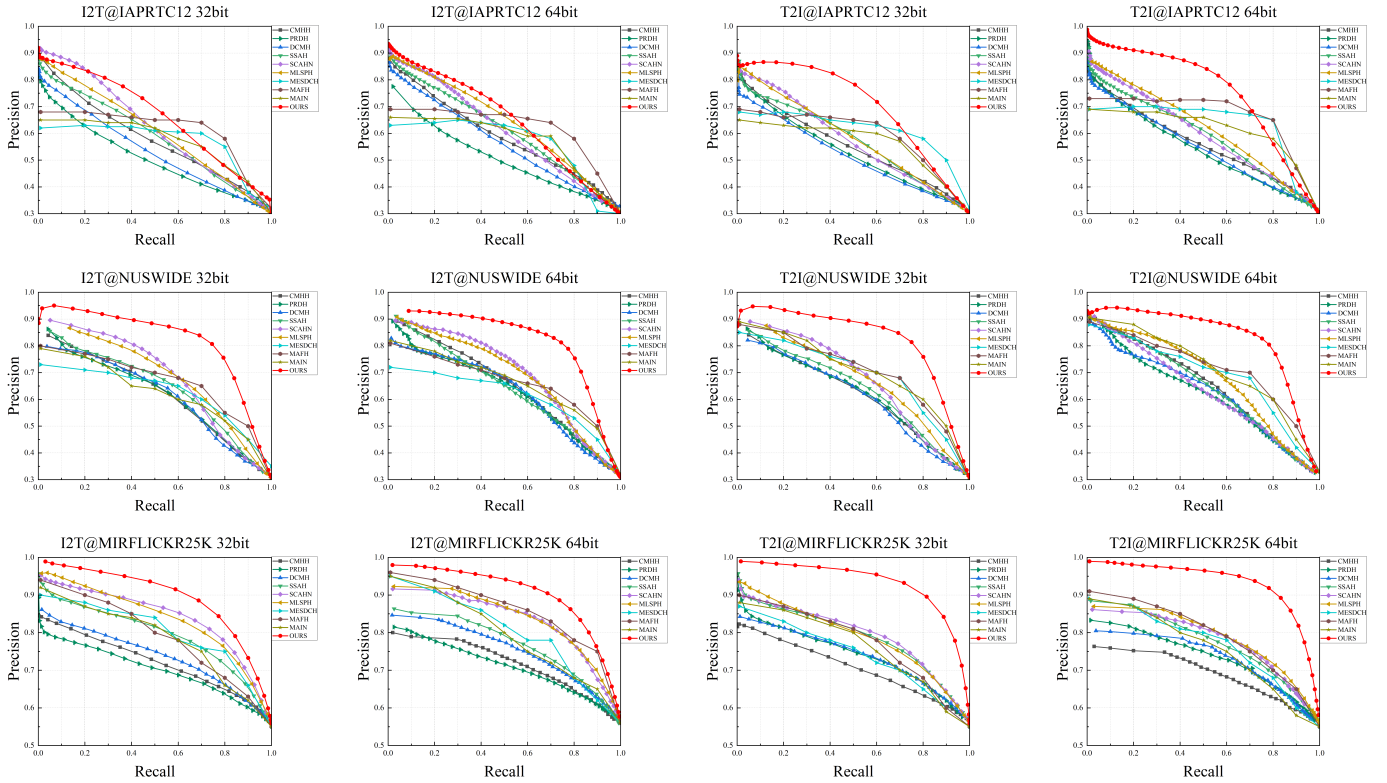
Fig. 3. The Precision-Recall (PR) curves of MDSAN on three datasets at hash code lengths of 32 and 64 bits. I2T denotes Image-to-Text retrieval and T2I denotes Text-to-Image retrieval.

supervised deep hash retrieval methods, we demonstrate the outstanding performance of our method.

A. Experimental Dataset

To rigorously evaluate the efficacy of our model, we opted for the inclusion of three widely recognized datasets: MIRFlickr-25k, NUS-WIDE, and IAPR TC-12. A comprehensive series of effectiveness experiments was conducted on these datasets to validate the robustness and performance of our approach.

MIRFlickr-25k: The original MIRFlickr-25k dataset comprises 25,000 instances of image-text pairs. Given its nature as a multi-label dataset, we conducted experiments exclusively on samples possessing a minimum of 20 textual labels, resulting in a total of 20,015 instances forming our experimental dataset. For the image, we resized it to $224*224*3$ and the text converted to a 1,386-dimensional bag-of-words vector. NUS-WIDE: As one of the most widely applied and extensive datasets in the realm of cross-modal datasets, the NUS-WIDE dataset encompasses 269,648 instances of image-text pairs, associated with 21 semantic class labels. IAPR TC-12: The IAPR TC-12 dataset comprises 20,000 instances of image-text pairs, spanning a total of 275 categories.

B. Results Evaluation Method

To evaluate the performance of the cross-modal retrieval model, assessments will be conducted for both image-to-text (image information retrieval text) and text-to-image (text

information retrieve images) retrieval capabilities. To comprehensively gauge the model's proficiency, we employ three widely recognized evaluation metrics: Mean Average Precision (MAP), Precision-Recall (PR) curves, and top-N accuracy curves. Calculating the MAP allows for an assessment of the model's holistic performance, with retrieval capabilities being directly correlated with the MAP. The Precision-Recall (PR) curve illustrates the precision performance of the model at different recall rates, aiding in the analysis of the model's balanced performance. The top-N accuracy curve depicts the accuracy of the model in the top-N retrieval results.

In the experiments, for the hyperparameters α, μ, ϑ and ρ, five experimental values were selected for each hyperparameter to be tested separately according to the selection range of the classical method [19] [26] about hyperparameters. Specifically, the experimental values of α, μ, and ϑ were set to [0.01, 0.1, 1, 2, 5], while the experimental values of ρ were [10, 50, 100, 500, 1000]. Based on the analysis of the experimental results for five different experimental values of each hyperparameter, we finally chose the respective optimal values as the setting values of these hyperparameters, i.e., α was set to 0.1, μ to 0.1, ϑ to 1, and ρ to 500. For the network configuration, we fixed the batch size at 128 and set the maximum training epoch at 50 and Adam was used. We chose the learning rate for the image network in the range of 10^{-6} to 10^{-4}, and for the text and label networks, it ranged from 10^{-6} to 10^{-2}. All experiments were conducted under the aforementioned settings. Our experiments were conducted using the PyTorch open-source environment and an NVIDIA 3090 GPU.
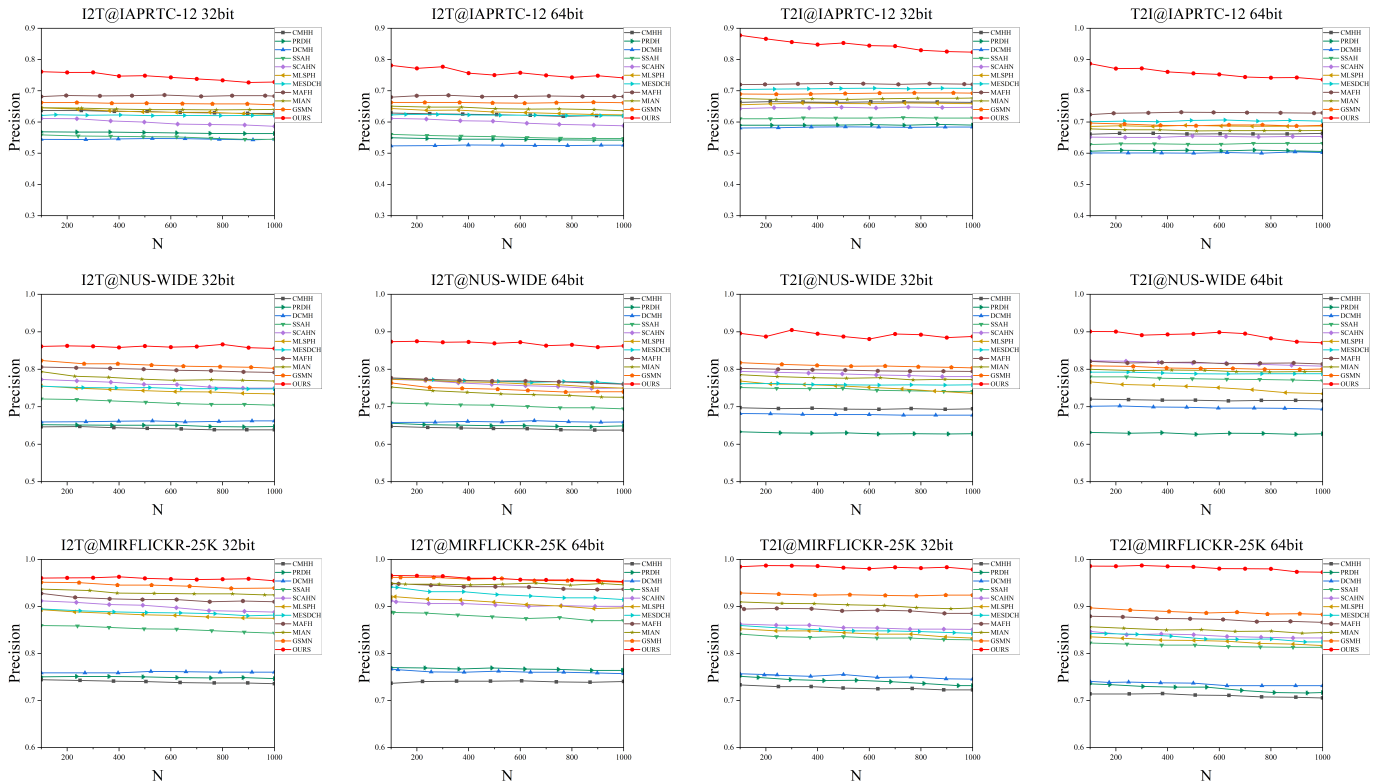
Fig. 4. The top-N accuracy curves of MDSAN on three datasets with hash code lengths of 32 and 64 bits. I2T represents Image-to-Text retrieval, and T2I represents Text-to-Image retrieval.

As shown in Table I. We selected the most representative cross-modal hashing retrieval methods for comparison, including shallow hashing methods CMSSH [36], SePH [17], SCM [7], and GSPH [16], as well as deep hashing methods DCMH [19], PRDH [37], CMHH [21], MIAN [26], SSAH [20], SCAHN [38], MLSPH [22], MESDCH [23] and MFAH [27] were chosen as the comparative methods in the proposed experiments. The results from all methods are from either recurrence experiments or original paper data.

Through extensive comparisons across three datasets, our approach exhibits notable advantages over state-of-the-art methods like MFAH, MIAN, and MESDCH. This underscores the efficacy of our multi-level deep feature extraction network in generating precise hash representations, consequently elevating retrieval performance. It further substantiates the profound practical significance of multi-level deep feature representation in addressing the complexities of multimodal data.

By contrasting SCAHN and SSAH, as well as comparing with PRDH, our method demonstrates significant advantages in adversarial learning, label-constrained learning, and constrained structural learning of hash codes. This indicates that our multi-layered semantic structure combined with an asymmetric metric can learn more accurate hash representations. In comparison to alternative learning approaches, our model structure exhibits robust advantages.

As can be seen by comparing Table I, compared to the MAFH method, our method has a significant improvement. Specifically, on the NUS-WIDE dataset, it achieves improvements of 11.67% in image-to-text and 8.06% in text-to-image,

respectively. On the MIRFlickr-25k dataset, the improvements are 8.20% in image-to-text and 13.45% in text-to-image. However, regrettably, our method fails to outperform MAFH (Multi-Label Method) in the IAPR TC-12 dataset in the image-to-text retrieval task (32-bit).

By analyzing this we can find that the IAPR TC-12 dataset has 275 label categories, which is more than ten times the number of label categories in the NUS-WIDE(21 label categories) and MIRFlickr-25k(24 label categories) datasets. As the number of label categories increases, the inter-category relationships and feature combinations that the model needs to handle become more complex [22] [27]. The excessive number of label categories in the dataset makes the feature information-rich leading to the fact that our extraction of multilevel features does not open a big gap with other methods. Moreover, our multi-level similarity matrix calculation method with a large number of label categories lacks targeting compared to multi-label methods, resulting in little improvement in the image-to-text retrieval task. This is an area where our method deserves improvement in the future, and the processing of multi-category labeled data needs to be more targeted.

As the Fig. 3 Precision-Recall (PR) curves, it is evident that our method surpasses others across different code lengths (32 and 64) on the three datasets. As the Fig. 4 top-N accuracy curve, it is evident that our recall top-N accuracy curves outperform those of other baseline methods. The outcomes of the two evaluations indicate that MDSAN significantly surpasses all baseline methods, consistent with the MAP analyses.

C. Ablation Experiment

In this section, we will conduct ablation experiments on our modules to examine their effectiveness. In Table II, T indicates the utilization of the network structure under that module, while F signifies the removal of the network structure under that module, as shown in Table II and Table III.

TABLE II
ABLATION EXPERIMENTS

Version	TxtNet	ImgNet	Multi-level Semantic Metric
V1	F	T	T
V2	T	F	T
V3	T	T	F
V4-ALL	T	T	T

In the V1 version, we replaced our text network with BottleNeck. In the V2 version, we substituted our image network with AlexNet. In V3, we replaced the multi-level semantic deep feature similarity Metric with a conventional similarity Metric. The V4 version represents our model's complete structure. In the ablation experiments, we conducted multiple experiments on the MIRFlickr-25k dataset with different hash lengths (16-128).

TABLE III
ABLATION EXPERIMENTS RESULT OF MEAN AVERAGE
PRECISION (MAP)

Task	Version	16bit	32bit	64bit	128bit
I2T	V1	0.8095	0.8139	0.8195	0.8286
	V2	0.8481	0.8460	0.8608	0.8664
	V3	0.8475	0.8707	0.8815	0.8911
	V4-All	**0.8998**	**0.8953**	**0.9083**	**0.9109**
T2I	V1	0.7940	0.8037	0.8075	0.8118
	V2	0.8822	0.8871	0.9028	0.9100
	V3	0.8418	0.8702	0.8915	0.9064
	V4-All	**0.9148**	**0.9200**	**0.9327**	**0.9385**

The experimental results are shown in Table III. Comparing V1-3 versions with V4 versions in the table, we can conclude that in V1, our text feature extraction network significantly enhances our ability to retrieve text from images. This strongly supports the notion that our text network effectively extracts multi-level semantic deep features, a capability not present in traditional local feature extraction networks. In the V2, we observe a 5.64% average improvement in image-to-text retrieval capability and a 3.46% average improvement in text-to-image retrieval capability. This robustly demonstrates the effectiveness of the image network we adopted and further validates the crucial role of multi-level semantic deep features in the cross-modal data processing domain. In the V3, we see a 3.53% average improvement in image-to-text retrieval capability and a 5.58% average improvement in text-to-image retrieval capability. This further validates that the asymmetrical similarity metric we employed can preserve more comprehensive semantic features, presenting a significant advantage over the symmetric similarity metric.

D. Limitations and Future Perspectives

While our model performs well in terms of retrieval performance, it is currently limited to focusing on the retrieval capabilities of both modalities. Notably, on datasets with a large number of labeled categories, our model falls short in some aspects of performance, which highlights the limitations of the model in handling these richly labeled data. In order to further improve the comprehensiveness and adaptability of the model, future research will focus on addressing the problem of a large number of label categories processing and exploring how to extend it to more modal applications so that our model can be more optimized and all-purpose.

V. CONCLUSION

In this study, we introduced a cross-modal hashing retrieval model based on a novel multi-level semantic deep feature network structure. Our approach integrates a deep feature network that can extract multi-level semantic information with an asymmetric similarity measure that retains a greater amount of semantic features within a unified framework. The proposed multi-level semantic feature extraction network proves effective in capturing nuanced feature information, showcasing superior capabilities compared to traditional feature extraction networks in balancing both local and global features of the data. Furthermore, the asymmetrical similarity metric demonstrates enhanced expandability and semantic calculation prowess when compared to the symmetric metric. Experimental evaluations conducted on three benchmark datasets substantiate the exceptional performance of our model in cross-modal hashing retrieval tasks.

REFERENCES

[1] X. Li, J. Yu, S. Jiang, H. Lu, and Z. Li, "Msvit: training multiscale vision transformers for image retrieval," *IEEE Transactions on Multimedia*, 2023. doi:10.1109/TMM.2023.3304021.

[2] C. Deng, E. Yang, T. Liu, and D. Tao, "Two-stream deep hashing with class-specific centers for supervised image search," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 6, pp. 2189–2201, 2019. doi:10.1109/tip.2018.2821921.

[3] X. Yao, D. She, S. Zhao, J. Liang, Y.-K. Lai, and J. Yang, "Attention-aware polarity sensitive embedding for affective image retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1140–1150, 2019. doi:10.1109/ICCV.2019.00123.

[4] Q.-Y. Jiang and W.-J. Li, "Discrete latent factor model for cross-modal hashing," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3490–3501, 2019. doi:10.1109/TIP.2019.2897944.

[5] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 102–112, 2018. doi:10.1109/TIP.2018.2863040.

[6] X. Nie, B. Wang, J. Li, F. Hao, M. Jian, and Y. Yin, "Deep multiscale fusion hashing for cross-modal retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 401–410, 2020. doi:10.1109/TCSVT.2020.2974877.

[7] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 28, 2014. doi:10.1609/aaai.v28i1.8995.

[8] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pp. 415–424, 2014. doi: 10.1145/2600428.2609610.

[9] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2075–2082, 2014. doi:10.1109/CVPR.2014.267.

[10] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016. doi:10.1016/j.neucom.2015.09.116.

[11] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3594–3601, IEEE, 2010. doi:10.1109/cvpr.2010.5539928.

[12] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3893–3903, 2018. doi:10.1109/tip.2018.2821921.

[13] P.-F. Zhang, Y. Li, Z. Huang, and X.-S. Xu, "Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 24, pp. 466–479, 2021. doi:10.1109/tmm.2021.3053766.

[14] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018. doi:10.1109/tpami.2018.2798607.

[15] L. Zhu, T. Wang, F. Li, J. Li, Z. Zhang, and H. T. Shen, "Cross-modal retrieval: A systematic review of methods and future directions," *arXiv preprint arXiv:2308.14263*, 2023. doi:10.1145/3539618.3594245.

[16] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for n-label cross-modal retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4076–4084, 2017. doi:10.1109/TIP.2018.2863040.

[17] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," pp. 3864–3872, 2015. doi:10.1109/cvpr.2015.7299011.

[18] H. Liu, R. Ji, Y. Wu, and G. Hua, "Supervised matrix factorization for cross-modality hashing," *arXiv preprint arXiv:1603.05572*, 2016. doi:10.48550/arXiv.1603.05572.

[19] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3232–3240, 2017. doi:10.1109/cvpr.2017.348.

[20] S. Jin, S. Zhou, Y. Liu, C. Chen, X. Sun, H. Yao, and X.-S. Hua, "Ssah: Semi-supervised adversarial deep hashing with self-paced hard sample generation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 11157–11164, 2020. doi:10.1609/aaai.v34i07.6773.

[21] Y. Cao, B. Liu, M. Long, and J. Wang, "Cross-modal hamming hashing," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 202–218, 2018. doi:10.1007/978-3-030-01246-5-13.

[22] X. Zou, X. Wang, E. M. Bakker, and S. Wu, "Multi-label semantics preserving based deep cross-modal hashing," *Signal Processing: Image Communication*, vol. 93, p. 116131, 2021. doi:10.1016/j.image.2020.116131.

[23] X. Zou, S. Wu, N. Zhang, and E. M. Bakker, "Multi-label modality enhanced attention based self-supervised deep cross-modal hashing," *Knowledge-Based Systems*, vol. 239, p. 107927, 2022. doi:10.1016/j.knosys.2021.107927.

[24] S. Teng, S. Lin, L. Teng, N. Wu, Z. Zheng, L. Fei, and W. Zhang, "Joint specifics and dual-semantic hashing learning for cross-modal retrieval," *Neurocomputing*, vol. 565, p. 126993, 2024. doi:10.1016/j.neucom.2023.126993.

[25] M. Meng, J. Sun, J. Liu, J. Yu, and J. Wu, "Semantic disentanglement adversarial hashing for cross-modal retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. doi:10.1109/tcsvt.2023.3293104.

[26] Z. Zhang, H. Luo, L. Zhu, G. Lu, and H. T. Shen, "Modality-invariant asymmetric networks for cross-modal hashing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 5, pp. 5091–5104, 2022. doi:10.1109/tkde.2022.3144352.

[27] X. Li, J. Yu, H. Lu, S. Jiang, Z. Li, and P. Yao, "Mafh: Multilabel aware framework for bit-scalable cross-modal hashing," *Knowledge-Based Systems*, vol. 279, p. 110922, 2023. doi:10.1016/j.knosys.2023.110922.

[28] Y. Chen, S. Wang, J. Lu, Z. Chen, Z. Zhang, and Z. Huang, "Local graph convolutional networks for cross-modal hashing," in *Proceedings of the 29th ACM international conference on multimedia*, pp. 1921–1928, 2021. doi:10.1145/3474085.3475346.

[29] F. Wu, S. Li, G. Gao, Y. Ji, X.-Y. Jing, and Z. Wan, "Semi-supervised cross-modal hashing via modality-specific and cross-modal graph convolutional networks," *Pattern Recognition*, vol. 136, p. 109211, 2023. doi:10.1016/j.patcog.2022.109211.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. doi:10.48550/arXiv.1706.03762.

[31] S. R. Dubey, S. K. Singh, and W.-T. Chu, "Vision transformer hashing for image retrieval," in *2022 IEEE international conference on multimedia and expo (ICME)*, pp. 1–6, IEEE, 2022. doi:10.1109/ICME52920.2022.9859900.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. doi:10.48550/arXiv.1409.1556.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. doi:10.48550/arXiv.2010.11929.

[34] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," *Advances in neural information processing systems*, vol. 21, 2008. doi:10.5555/2981780.2981999.

[35] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 1–8, 2011. doi:10.5555/3104482.3104483.

[36] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3594–3601, IEEE, 2010. doi:10.1109/cvpr.2010.5539928.

[37] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017. doi:10.1609/aaai.v31i1.10719.

[38] H. Cui, L. Zhu, J. Li, Y. Yang, and L. Nie, "Scalable deep hashing for large-scale social image retrieval," *IEEE Transactions on image processing*, vol. 29, pp. 1271–1284, 2019. doi:10.1109/tip.2019.2940693.

Xiaolong Jiang is a master student at the College of Computer and Information Science of Chongqing Normal University. His current research interests include Cross-modal retrieval and deep learning.

Jiabao Fan is a master student at the College of Computer and Information Science of Chongqing Normal University. His current research interests include Cross-modal retrieval and deep learning.

Jie Zhang is a master student at the College of Computer and Information Science of Chongqing Normal University. His current research interests include Cross-modal retrieval and deep learning.

Ziyong Lin is a master student at the College of Computer and Information Science of Chongqing Normal University. His current research interests include Cross-modal retrieval and deep learning.

Mingyong Li (Member, IEEE) received the B.S. degree from Central China Normal University, in 2003, and the Ph.D. degree from the Department of Computer Science and Technology, Donghua University, in 2021. He is currently a Professor with the School of Computer and Information Science, Chongqing Normal University. His current research interests include Cross-modal big data processing, large-scale data retrieval and deep learning.