mathematics

MDPI

Article

Deep Multi-Semantic Fusion-Based Cross-Modal Hashing

Xinghui Zhu, Liewu Cai, Zhuoyang Zou and Lei Zhu *

College of Information and Intelligence, Hunan Agricultural University, Changsha 410128, China; zhuxh@hunau.edu.cn (X.Z.); liewucai@stu.hunau.edu.cn (L.C.); zzy@stu.hunau.edu.cn (Z.Z.)
* Correspondence: leizhu@hunau.edu.cn

Abstract: Due to the low costs of its storage and search, the cross-modal retrieval hashing method has received much research interest in the big data era. Due to the application of deep learning, the cross-modal representation capabilities have risen markedly. However, the existing deep hashing methods cannot consider multi-label semantic learning and cross-modal similarity learning simultaneously. That means potential semantic correlations among multimedia data are not fully excavated from multi-category labels, which also affects the original similarity preserving of cross-modal hash codes. To this end, this paper proposes deep multi-semantic fusion-based cross-modal hashing (DMSFH), which uses two deep neural networks to extract cross-modal features, and uses a multi-label semantic fusion method to improve cross-modal consistent semantic discrimination learning. Moreover, a graph regularization method is combined with inter-modal and intra-modal pairwise loss to preserve the nearest neighbor relationship between data in Hamming subspace. Thus, DMSFH not only retains semantic similarity between multi-modal data, but integrates multi-label information into modal learning as well. Extensive experimental results on two commonly used benchmark datasets show that our DMSFH is competitive with the state-of-the-art methods.

Keywords: cross-modal hashing; semantic label information; multi-label semantic fusion; graph regularization; deep neural network

Citation: Zhu, X.; Cai, L.; Zou, Z.; Zhu, L. Deep Multi-Semantic Fusion-Based Cross-Modal Hashing. *Mathematics* **2022**, *10*, 430. https://doi.org/10.3390/math10030430

Academic Editor: Alfredo Milani

Received: 21 November 2021
Accepted: 12 January 2022
Published: 29 January 2022

1. Introduction

In recent years, with the rapid development of information technology, massive amounts of multi-modal data (i.e., text [1], image [2], audio [3], video [4], and 3D models [5]) have been collected and stored on the Internet. How to utilize the extensive multi-modal data to improve cross-modal retrieval performance has attracted increasing attention [6,7]. Cross-modal retrieval, a hot issue in the multimedia community, is the use of queries from one modality to retrieve all semantically relevant instances from another modality [8–10]. In general, the structuring of data in different modalities is heterogeneous, but there are strong semantic correlations between these structures. Therefore, the main tasks of cross-modal retrieval are discovering how to narrow the semantic gap and exploring the common representations of multi-modal data, the former being the most challenging problem faced by researchers in this field [11–14].

Most of existing cross-modal retrieval methods, including traditional statistical correlation analysis [15], graph regularization [16], and dictionary learning [17], learn a common subspace [18–21] for multi-modal samples, in which the semantic similarity between different modalities can be measured easily. For example, based on canonical correlation analysis (CCA) [22], several cross-modal retrieval methods [23–25] have been proposed to learn a common subspace in which the correlations between different modalities are easily measured. Besides, graph regularization has been applied in many studies [16,26–28] to preserve the semantic similarity between cross-modal representations in the common subspace. The methods in [17,29,30] draw support from dictionary learning to learn consistent representations for multi-modal data. However, these methods usually have high computational costs and low retrieval efficiency [31]. In order to overcome these shortcomings,

hashing-based cross-modal retrieval techniques are gradually replacing the traditional ones. A practical way to speed up similarity searching is with binary representation learning, referred to as hashing learning, which projects a high-dimensional feature representation from each modality as a compact hash code and preserves similar instances with similar hash codes. In this paper, we focus on the cross-modal binary representation learning task, which can be applied to large-scale multimedia searches in the cloud [32–34].

In general, most of the existing traditional cross-modal hashing methods can be roughly divided into two groups: unsupervised [35–39] and supervised methods [40–44]. Unlike unsupervised methods, supervised methods can excavate similarity relationships between data through semantic labels to achieve better performance. However, these methods rely on shallow features that cannot provide sufficient semantic discrimination information. Recently, deep models [45–48] have been widely adopted to perform feature learning from scratch with very promising performance. This powerful representation learning technique boosts the non-linear correlation learning capabilities of cross-modal hashing models. Thus, lots of deep hashing method [28,49–58] have been developed, which can effectively learn more discriminative semantic representations from multi-modal samples and are gradually replacing the traditional hashing approaches.

Motivation. Although deep hashing algorithms have made remarkable progress in cross-modal retrieval, the semantic gap and heterogeneity gap between different modalities need to be further narrowed. On the one hand, most methods lack mining of ample semantic information from multiple category labels. That means these methods cannot completely retain multi-label semantic information during cross-modal representation learning. Taking [28] as an example, graph regularization is used to support intra-modal and inter-modal similarity learning, but the multi-label semantics are not mined fully during the cross-modal representation learning, which affect the semantic discrimination of hash codes. On the other hand, after the features learned from normal networks are quantized into binary representations, some semantic correlations may be lost in Hamming subspace. For instance, [59] studies the effective distance measurement of cross-modal binary representations in Hamming subspace. However, multi-label semantics learning is ignored, which leads to insufficient semantic discriminability of the hash code. Therefore, to further improve the quality of cross-modal hash codes, two particularly important problems cannot be overlooked during the hashing learning: (1) *how to capture more semantic discriminative features*, and (2) *how to efficiently preserve cross-modal semantic similarity in common Hamming subspaces*. In this work, we consider these two key issues simultaneously during the cross-modal hashing learning to generate more semantically discriminative hash codes.

Our Method. To this end, we propose a novel end-to-end cross-modal hashing learning approach, named deep multi-semantic fusion-based cross-modal hashing (**DMSFH** for short) to efficiently capture multi-label semantics and generate high-quality cross-modal hash codes. Firstly, two deep neural networks are used to learn cross-modal representations. Then, intra-modal loss and inter-modal loss are utilized by generating a semantic similarity matrix to preserve semantic similarity. To further capture the rich semantic information, a multi-label semantic fusion module is used following the feature learning module, which fuses the multiple label semantics into cross-modal representations to preserve the semantic consistency across different modalities. In addition, we introduce a graph regularization method to preserve semantic similarity among cross-modal hash codes in Hamming subspace.

Contributions. The main contributions of this paper are summarized as follows:

- We propose a novel deep learning-based cross-modal hashing method, termed DMSFH, which integrates cross-modal feature learning, multi-label semantic fusion, and hash code learning into an end-to-end architecture.
- We combine the graph regularization method with inter-modal and intra-modal pairwise loss to enhance cross-modal similarity learning in Hamming subspace. Addition-

ally, a multi-label semantic fusion module was developed to enhance the cross-modal consistent semantics learning.
- Extensive experiments conducted on two well-known multimedia datasets demonstrate the outstanding performance of our methods compared to other state-of-the-art cross-modal hashing methods.

Roadmap. The rest of this paper is organized as follows. The related work is summarized in Section 2. The problem definition and the details of the proposed method DMSFH are presented in Section 3. The experimental results and evaluations are reported in Section 4. We discuss the main contributions and characteristics of our research in Section 5. Finally, we conclude this paper in Section 6.

2. Related Work

According to learning manner, the existing cross-modal hashing techniques fall into two categories: unsupervised approaches and supervised approaches. Due to the vigorous development of deep learning, cross-modal deep hashing approaches sprang up in the last decade. This section reviews the works that are related to our paper.

Unsupervised Methods. To learn a hash function, the unsupervised hashing methods aim to mine the unlabeled samples to discover the relationship between multi-modal data. One of the most typical technique is collective matrix factorization hashing (CMFH) [60], which utilizes matrix decomposition to learn two view-specific hash functions, and then different modal data can be mapped into unified hash codes. The latent sematic sparse hashing (LSSH) method [35] uses sparse coding to find the salient structures of images, and matrix factorization to learn the latent concepts from text. Then, the learned latent semantic features are mapped to a joint common subspace. Semantic topic multimodal hashing (STMH) [37], which discovers clustering patterns of texts and factorizes the matrix of images, to acquire multiple semantic of texts and concepts of images in order to learn multimodal semantic features, into a common subspace by their correlations. Multi-modal graph regularized smooth matrix factorization hashing (MSFH) [61] utilizes a multi-modal graph regularization term which includes an intra-modal similarity graph and an inter-modal similarity graph to preserve the topology of the original instances. The latent structure discrete hashing factorization (LSDHF) [62] approach uses the Hadamard matrix to align all eigenvalues of the similarity matrix to generate a hash dictionary, and then straightforwardly distills the shared hash codes from the intrinsic structure of modalities.

Supervised Methods. Supervised cross-modal hashing methods improve the search performance by using supervised information, such as training data labels. Typical supervised approaches include cross-modal similarity sensitive hashing (CMSSH) [40], semantic preserving hashing for cross-view retrieval (SEPH) [41], semantic correlation maximization (SCM) [42], and discrete cross-modal hashing (DCH) [43]. CMSSH applies boosting techniques to preserve the intra-modal similarity. SEPH transforms the semantic similarity of training data into an affinity matrix by using a label as supervised information, and minimizes the Kullback–Leibler divergence to learn hash codes. SCM utilizes all the supervised information for training with linear-time complexity by avoiding explicitly computing the pairwise similarity matrix. DCH learns discriminative binary codes without relaxation, and label information is used to elevate the discriminability of binary codes through linear classifiers. Nevertheless, these cross-modal hashing methods are established on hand-crafted features [43,63]. It is hard to explore the semantic relationships among multi-modal data. Therefore, it is difficult to obtain satisfying retrieval results.

Deep Methods. In recent years, deep learning, as a powerful representation learning technique, has been widely used in cross-modal retrieval tasks. A number of methods integrating deep neural networks and cross-modal hashing have been developed. For example, deep cross-modal hashing (DCMH) [64] firstly applies the end-to-end deep learning architecture for cross-modal hashing retrieval and utilizes the negative logistic likelihood loss to achieve great performance. Pairwise relationship-guided deep hashing (PRDH) [65] uses pairwise label constraints to supervise the similarity learning of inter-modal and intra-

modal data. A correlation hashing network (CHN) [66] adapts the triplet loss measured by cosine distance to find the semantic relationship between pairwise instances. Cross-modal hamming hashing (CMHH) [59] learns high-quality hash representations to significantly penalize similar cross-modal pairs with Hamming distances larger than the Hamming radius threshold. The ranking-based deep cross-modal hashing approach (RDCMH) [49] integrates the semantic ranking information into a deep cross-modal hashing model and jointly optimizes the compatible parameters of deep feature representations and hashing functions. In fusion-supervised deep cross-modal hashing (FDCH) [67], both pair-wise similarity information and classification information are embedded in the hash model, which simultaneously preserves cross-modal similarity and reduces semantic inconsistency. Despite the above-mentioned benefits, most of these methods only use binary similarity to constrain the generation of different instances of hash codes. This causes low correlations between retrieval results and the inputs, as the semantic label information cannot be expressed adequately. Besides, most methods only concentrate on hash code learning, but ignore the deep mining of semantic features. Thus, it is essential to keep sufficient semantic information in the modal structure and generate discriminative hash codes to enhance the cross-modal hashing learning.

To overcome the above challenges, this paper proposes a novel approach to excavate multi-label semantic information to improve the semantic discrimination of cross-modal hash codes. This approach not only uses the negative logistic likelihood loss, but also exploits multiple semantic labels' prediction losses based on cross entropy to enhance semantic information mining. Apart from this, we introduce graph regularization to preserve the semantic similarity of hash codes in Hamming subspace. Therefore, the proposed method is designed to generate high-quality hash codes that better reflect high-level cross-modal semantic correlations.

3. The Proposed Approach

In this section, we propose our method DMSFH, including the model's formulation and the learning algorithm. The framework of the proposed DMSFH is shown in Figure 1, which mainly consists of three parts. The first part is the feature learning module, in which multimedia samples are transformed into high-dimensional feature representations by corresponding deep neural networks. The second part is the multi-label semantic fusion part. This part aims to embed rich multi-label semantic information into feature learning. The third part is the hashing learning module, which retains the semantic similarity of the cross-modal data in the hash codes using a carefully designed loss function. In the following, we introduce the problem definition first, and then discuss DMSFH method in detail.

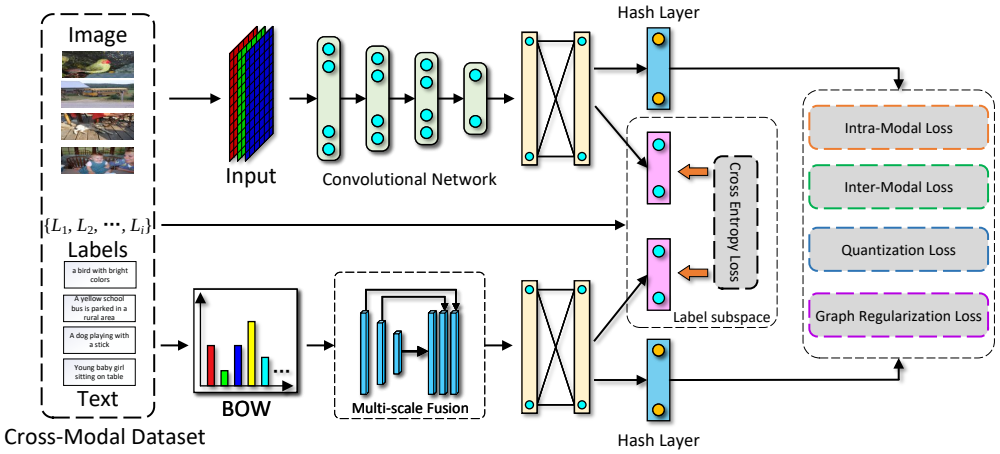


Figure 1. The framework of DMSFH. It contains three main components: (1) the feature learning module, which contains a classical convolutional neural network for image-modality feature learning, and a multi-scale fusion-based convolutional neural network for text-modality feature learning; (2) a multi-label semantic information learning module that is realized by deep neural networks, which is to fuse rich semantic information from multiple labels to generate consistent semantic representations in label subspace; (3) a hash function module that is trained by inter-modal and intra-modal pairwise loss, quantization loss, and graph regularization loss to generate cross-modal hash codes.

3.1. Problem Definition

Without loss of generality, bold uppercase letters, such as $\boldsymbol{W}$, represent matrices. Bold lowercase letters, such as $\boldsymbol{w}$, represent vectors. Moreover, the ij-th element of $\boldsymbol{W}$ is denoted as $\boldsymbol{W}_{ij}$, the i-th row of $\boldsymbol{W}$ is denoted as $\boldsymbol{W}_{i*}$, and the j-th column of $\boldsymbol{W}$ is denoted as $\boldsymbol{W}_{*j}$. $\boldsymbol{W}^T$ is the transpose of $\boldsymbol{W}$. We use $\boldsymbol{I}$ for the identity matrix. $tr(\cdot)$ and $||\cdot||_F$ denote the trace of the matrix and the Frobenius norm of a matrix, respectively. $sign(\cdot)$ is the sign function, shown as follows:

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

To facilitate easier reading, the frequently used mathematical notation is summarized in Table 1.

Table 1. A summary of frequently-used notation.

Notation	Definition
n	the number of training instances
k	length of hash codes
c	the number of categories
$\boldsymbol{v}_i$	the ith image sample
$\boldsymbol{t}_i$	the ith text sample
O	multimedia datasets
$\boldsymbol{L}$	semantic label matrix of instances
$\hat{\boldsymbol{L}}^v$	predicted semantic label matrix of instances in image network
$\hat{\boldsymbol{L}}^t$	predicted semantic label matrix of instances in text network
$\boldsymbol{S}$	binary similarity matrix
$\boldsymbol{F}$	image modality continuous hash code
$\boldsymbol{G}$	textual modality continuous hash code
$\boldsymbol{B}$	the unified binary hash codes

This paper focuses on two common modalities: texts and images. Assume that a cross-modal training dataset consists of n instances, i.e., $O = \{o_1, o_2, \ldots, o_n\}$, where $o_i = (v_i, t_i, \boldsymbol{L}_i)$ denotes the i-th training instances, and v_i and t_i are the i-th image and text,

respectively. $\boldsymbol{L}_i = [\boldsymbol{L}_{i1}, \boldsymbol{L}_{i2}, \ldots, \boldsymbol{L}_{ic}]$ is the multi-label annotation assigned to o_i, where c is the number of categories. If o_i belongs to the jth class, $\boldsymbol{L}_{ij} = 1$; otherwise, $\boldsymbol{L}_{ij} = 0$. In addition, a cross-modal similarity matrix $\boldsymbol{S} = \{\boldsymbol{S}^{vt}, \boldsymbol{S}^{vv}, \boldsymbol{S}^{tt}\}$ is given. If image $\boldsymbol{v}_i$ and text $\boldsymbol{t}_j$ are similar, $\boldsymbol{S}_{ij} = 1$; otherwise, $\boldsymbol{S}_{ij} = 0$.

Given a set of training data O, the goal of cross-modal hashing is to learn two hashing functions, i.e., $h^v(v)$ and $h^t(t)$ for image modality and textual modality, respectively, where $h^v(v) \in \{-1, 1\}^k$, $h^t(t) \in \{-1, 1\}^k$, k is the length of the hash code. In addition, the hash codes preserve the similarities in similarity matrix $\boldsymbol{S}$. If the Hamming distance between the codes $\boldsymbol{b}_i^{(v)} = h^v(v_i)$ and $\boldsymbol{b}_i^{(t)} = h^t(t_i)$ is small, $\boldsymbol{S}_{ij} = 1$; otherwise $\boldsymbol{S}_{ij} = 0$. To easily calculate the similarity between two binary codes b_i and b_j, we use the inner product $\langle b_i, b_j \rangle$ to measure the Hamming distance as follows:

$$dis_H(b_i, b_j) = \frac{1}{2}(K - \langle b_i, b_j \rangle), \tag{2}$$

where K is the length of the hash code.

3.2. Feature Learning Networks

For cross-modal feature learning, deep neural networks are used to extract semantic features from each modality individually. Specifically, for image modality, ResNet34 [46], a well-known deep convolutional network, is used to extract image data features. The original ResNet was pre-trained on imagenet datasets; in addition, excellent results have been achieved on image recognition issues. We replaced the last layer with a network that has $(k + c)$ hidden nodes, which is followed by a hash layer and a tag layer. The hash layer has k hidden nodes for generating binary representations. The label layer has c hidden nodes for generating predictive labels.

For text modality, a deep model named TxtNet is used to generate textual feature representations, which is a three-layer network followed by a multi-scale (MS) fusion model ($T \rightarrow MS \rightarrow 4096 \rightarrow 512 \rightarrow k + c$). The last layer of TxtNet is a fully-connected layer with $(k + c)$ hidden nodes, which outputs deep textual features and prediction labels. The input of TxtNet is the Bag-of-Words (BoW) representation of each text sample. The BoW vector is too sparse, but the features extracted by the multi-scale fusion model are more abundant. Firstly, the BoW vectors are evenly pooled at different scales; then, the semantic information is extracted by nonlinear mapping through a convolution operation and an activation function. Finally, the representations from different scales are fused to obtain richer semantic information. The Ms fusion model contains 5 interpretation blocks. Each block contains a 1×1 convolutional layer and an average pooling layer. The filter sizes of the average pooling layer are set to 50×50, 30×30, 15×15, 10×10 and 5×5, respectively.

3.3. Hash Function Learning

In the network of image modality, let $f_1^v(v_{i*}; \theta_v, \theta_{vh}) \in \mathbb{R}^{1 \times k}$ denote the learned image feature of the i-th sample v_i, where θ_v is all network parameters before the last layer of the deep neural network, and θ_{vh} is the network parameter of the hash layer. Furthermore, let $f_2^v(v_{i*}; \theta_v, \theta_{vl}) \in \mathbb{R}^{1 \times c}$ denote the output of the label layer for sample v_i, where θ_{vl} is the network parameter of the label layer. In the network of text modality, let $f_1^t(t_{i*}; \theta_t, \theta_{th}) \in \mathbb{R}^{1 \times k}$ denote the learned text feature of the i-th sample t_i, where θ_t is all network parameters before the last layer of deep neural network, and θ_{th} is the network parameter of the hash layer. Furthermore, let $f_2^t(t_{i*}; \theta_t, \theta_{tl}) \in \mathbb{R}^{1 \times c}$ denote the output of the label layer for sample t_i, where θ_{tl} is the network parameter of the label layer.

To capture the semantic consistency between different modalities, the inter-modal negative log likelihood function is used in our approach, which is formulated as:

$$\mathcal{L}_1 = -\sum_{i,j=1}^{n} (\boldsymbol{S}_{ij}^{vt} \boldsymbol{\phi}_{ij}^{vt} - \log(1 + e^{\boldsymbol{\phi}_{ij}^{vt}})), \tag{3}$$

where $\boldsymbol{\phi}_{ij}^{vt} = \frac{1}{2}\boldsymbol{F}_{i*}\boldsymbol{G}_{j*}^{T}$ is the inner product of two instances, $\boldsymbol{F} \in \mathbb{R}^{n\times k}$ with $\boldsymbol{F}_{i*} = f_1^v(v_{i*};\theta_v,\theta_{vh})$, and $\boldsymbol{G} \in \mathbb{R}^{n\times k}$ with $\boldsymbol{G}_{i*} = f_1^t(t_{i*};\theta_t,\theta_{th})$. The likelihood function composed of text feature $\boldsymbol{F}$ and image feature $\boldsymbol{G}$ is as follows:

$$p(S_{ij}|\boldsymbol{F}_{i*},\boldsymbol{G}_{j*}) = \begin{cases} \sigma(\boldsymbol{\phi}_{ij}), & \boldsymbol{S}_{ij} = 1 \\ 1-\sigma(\boldsymbol{\phi}_{ij}), & \boldsymbol{S}_{ij} = 0 \end{cases}, \tag{4}$$

where $\sigma(\boldsymbol{\phi}_{ij}) = \frac{1}{1+e^{-\boldsymbol{\phi}_{ij}}}$ is a sigmoid function, and $\boldsymbol{\phi}_{ij} = \frac{1}{2}\boldsymbol{F}_{i*}\boldsymbol{G}_{j*}^{T}$.

To generate the hash codes with rich semantic discrimination, two essential factors need to be considered: (1) the semantic similarity between different modes should be preserved, and (2) the high-level semantics within each mode should be preserved, which can raise the accuracy of cross-modal retrieval effectively. To realize this strategy, we define the intra-modal pair-wise loss as follows:

$$\mathcal{L}_2 = \mathcal{L}_2^v + \mathcal{L}_2^t, \tag{5}$$

where $\mathcal{L}_2^v$ is the intra-modal pair-wise loss for image-to-image and $\mathcal{L}_2^t$ is the intra-modal pair-wise loss for text-to-text, and $\mathcal{L}_2^v$ and $\mathcal{L}_2^t$ are defined as:

$$\mathcal{L}_2^v = -\sum_{i,j=1}^{n}(\boldsymbol{S}_{ij}^{vv}\boldsymbol{\phi}_{ij}^{vv} - \log(1+e^{\boldsymbol{\phi}_{ij}^{vv}})), \tag{6}$$

$$\mathcal{L}_2^t = -\sum_{i,j=1}^{n}(\boldsymbol{S}_{ij}^{tt}\boldsymbol{\phi}_{ij}^{tt} - \log(1+e^{\boldsymbol{\phi}_{ij}^{tt}})), \tag{7}$$

where $\boldsymbol{\phi}_{ij}^{vv} = \frac{1}{2}\boldsymbol{F}_{i*}\boldsymbol{F}_{j*}^{T}$ is the inner product of image data, and $\boldsymbol{\phi}_{ij}^{tt} = \frac{1}{2}\boldsymbol{G}_{i*}\boldsymbol{G}_{j*}^{T}$ is the inner product of text data.

Based on the negative log likelihood, the loss function can be used to distinguish identical and completely dissimilar instances. However, for more fine-grained hash features, we can extract higher-level semantic information by adding a tag prediction layer, so that the network can learn hash features with deep semantics. The semantic label cross-entropy loss is:

$$\mathcal{L}_3 = \mathcal{L}_3^{v\_label} + \mathcal{L}_3^{t\_label}, \tag{8}$$

where $\mathcal{L}_3^{v\_label}$ is the cross entropy loss for image modalities and $\mathcal{L}_3^{t\_label}$ is the cross entropy loss for text modalities. $\mathcal{L}_3^{v\_label}$ and $\mathcal{L}_3^{t\_label}$ are defined as:

$$\mathcal{L}_3^{v\_label} = \sum_{i}^{n}\sum_{j}^{c}(-\boldsymbol{L}_{ij}\hat{\boldsymbol{L}}_{ij}^{v} + \log(1+e^{\hat{\boldsymbol{L}}_{ij}^{v}})), \tag{9}$$

$$\mathcal{L}_3^{t\_label} = \sum_{i}^{n}\sum_{j}^{c}(-\boldsymbol{L}_{ij}\hat{\boldsymbol{L}}_{ij}^{t} + \log(1+e^{\hat{\boldsymbol{L}}_{ij}^{t}})), \tag{10}$$

where $\boldsymbol{L}_{i*}$ is the original semantic label information, for instance, o_i; and $\hat{\boldsymbol{L}}_{i*}^{v} = f_2^v(v_{i*};\theta_v,\theta_{vl})$ and $\hat{\boldsymbol{L}}_{i*}^{t} = f_2^t(t_{i*};\theta_t,\theta_{tl})$ represent the prediction labels of instance o_i in the image network and text network, respectively.

In order to enhance the correlation between the same hash code in Hamming subspace, we introduce graph regularization to establish the degree of correlation between multi-modal datasets. We formulate a spectral graph learning loss from the label similarity matrix $\boldsymbol{S}$ as follows:

$$\mathcal{L}_4 = \frac{1}{2}\sum_{i,j=1}^{n}||\boldsymbol{b}_i - \boldsymbol{b}_j||_F^2\boldsymbol{S}_{ij}^{vt} = tr(\boldsymbol{B}^T\boldsymbol{L}\boldsymbol{B}), \tag{11}$$

where $\boldsymbol{S}^{vt}$ is the similarity matrix, and $\boldsymbol{B} = \{b_i\}_{i=1}^{n}$ represents the unified hash codes. we define diagonal matrix $\boldsymbol{D} = diag(d_1, \ldots, d_n)$, and $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}^{vt}$ is the graph Laplacian matrix.

We regard $\boldsymbol{F}$ and $\boldsymbol{G}$ as the continuous substitution of the image network hash code $\boldsymbol{B}^v$ and the text network hash code $\boldsymbol{B}^t$ to reduce quantization loss. According to the empirical analysis, the training effect will be better if the same hash code is used for different modes of the same training data, so we set $\boldsymbol{B}^v = \boldsymbol{B}^t = \boldsymbol{B}$. Therefore, quantization loss can be defined as:

$$\mathcal{L}_5 = ||\boldsymbol{B} - \boldsymbol{F}||_F^2 + ||\boldsymbol{B} - \boldsymbol{G}||_F^2. \tag{12}$$

The overall objective function, combining the inter-modality pair-wise loss $\mathcal{L}_1$, the intra-modal pair-wise loss $\mathcal{L}_2$, the cross entropy loss $\mathcal{L}_3$ for the predicted label, graph regularization loss $\mathcal{L}_4$ and quantization loss $\mathcal{L}_5$, is written as below:

$$\begin{aligned} \min_{B,\theta_v,\theta_{vh},\theta_{vl},\theta_t,\theta_{th},\theta_{tl}} & \mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \gamma\mathcal{L}_4 + \beta\mathcal{L}_5 \\ s.t. \quad & \boldsymbol{B} \in \{-1, +1\}^{n \times k}, \end{aligned} \tag{13}$$

where γ and β are hyper-parameters to control the weight of each part.

3.4. Optimization

The objective in Equation (13) can be solved by using an alternative optimization iteratively. We adopt the mini-batch stochastic gradient descent (SGD) method to learn parameter $\vartheta_v = \{\theta_v, \theta_{vh}, \theta_{vl}\}$ in an image network and parameter $\vartheta_t = \{\theta_t, \theta_{th}, \theta_{tl}\}$ in a text network, and $\boldsymbol{B}$. Each time we optimize one network with the other parameters fixed. The whole alternating learning algorithm for DMSFH is briefly outlined in Algorithm 1, and a detailed derivation is described in the following subsections.

3.4.1. Optimize ϑ_v

When ϑ_t and $\boldsymbol{B}$ are fixed, we can learn the deep network parameter ϑ_v for the image modality by using SGD with back-propagation(BP). For the i-th image $\boldsymbol{F}_{i*}$, we first calculate the following gradient:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{F}_{i*}} &= \frac{\partial \mathcal{L}_1}{\partial \boldsymbol{F}_{i*}} + \frac{\partial \mathcal{L}_2^v}{\partial \boldsymbol{F}_{i*}} + \frac{\partial \mathcal{L}_5}{\partial \boldsymbol{F}_{i*}} \\ &= \frac{1}{2}\sum_{j=1}^{n}(\sigma(\boldsymbol{\phi}_{ij}^{vt})\boldsymbol{G}_{j*} - \boldsymbol{S}_{ij}^{vt}\boldsymbol{G}_{j*}) + \frac{1}{2}\sum_{j=1}^{n}(\sigma(\boldsymbol{\phi}_{ij}^{vv})\boldsymbol{F}_{j*} - \boldsymbol{S}_{ij}^{vv}\boldsymbol{F}_{j*}) \\ &+ 2\beta(\boldsymbol{F}_{i*} - \boldsymbol{B}_{i*}), \end{aligned} \tag{14}$$

$$\frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{L}}_{ij}^v} = \frac{\partial \mathcal{L}_3^{v\_label}}{\partial \hat{\boldsymbol{L}}_{ij}^v} = (-\boldsymbol{L}_{ij} + \sigma(\hat{\boldsymbol{L}}_{ij}^v)). \tag{15}$$

Then we can compute $\frac{\partial \mathcal{L}}{\partial \theta_v}$, $\frac{\partial \mathcal{L}}{\partial \theta_{vh}}$, and $\frac{\partial \mathcal{L}}{\partial \theta_{vl}}$ by utilizing the chain rule, based on which BP can be used to update the parameters ϑ_v.

3.4.2. Optimize ϑ_t

Similarly, when ϑ_v and $\boldsymbol{B}$ are fixed, we also learn the network parameter ϑ_t of the text modality by using SGD and the BP algorithm. For the i-th text $\boldsymbol{G}_{i*}$, we calculate the following gradient:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{G}_{i*}} &= \frac{\partial \mathcal{L}_1}{\partial \boldsymbol{G}_{i*}} + \frac{\partial \mathcal{L}_2^t}{\partial \boldsymbol{G}_{i*}} + \frac{\partial \mathcal{L}_5}{\partial \boldsymbol{G}_{i*}} \\ &= \frac{1}{2}\sum_{j=1}^{n}(\sigma(\boldsymbol{\phi}_{ij}^{vt})\boldsymbol{F}_{j*} - \boldsymbol{S}_{ij}^{vt}\boldsymbol{F}_{j*}) + \frac{1}{2}\sum_{j=1}^{n}(\sigma(\boldsymbol{\phi}_{ij}^{tt})\boldsymbol{G}_{j*} - \boldsymbol{S}_{ij}^{vv}\boldsymbol{G}_{j*}) \\ &+ 2\beta(\boldsymbol{G}_{i*} - \boldsymbol{B}_{i*}),\end{aligned} \tag{16}$$

$$\frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{L}}_{ij}^t} = \frac{\partial \mathcal{L}_3^{t\_label}}{\partial \hat{\boldsymbol{L}}_{ij}^t} = (-\boldsymbol{L}_{ij} + \sigma(\hat{\boldsymbol{L}}_{ij}^t)). \tag{17}$$

Then we can compute $\frac{\partial \mathcal{L}}{\partial \theta_t}$, $\frac{\partial \mathcal{L}}{\partial \theta_{th}}$, and $\frac{\partial \mathcal{L}}{\partial \theta_{tl}}$ by utilizing the chain rule, based on which BP can be used to update the parameters ϑ_t.

3.4.3. Optimize $\boldsymbol{B}$

When ϑ_v and ϑ_t are fixed, the objective in Equation (13) can be reformulated as follows:

$$\begin{aligned}&\min_{\boldsymbol{B}} \mathcal{L} = \gamma(||\boldsymbol{B} - \boldsymbol{F}||_F^2 + ||\boldsymbol{B} - \boldsymbol{G}||_F^2) + \beta tr(\boldsymbol{B}^T\boldsymbol{L}\boldsymbol{B}) \\ &s.t. \quad \boldsymbol{B} \in \{-1, +1\}^{n\times k}.\end{aligned} \tag{18}$$

We compute the derivation of Equation (18) with respect to $\boldsymbol{B}$ and infer that $\boldsymbol{B}$ should be defined as follows:

$$\boldsymbol{B} = sign((\boldsymbol{F} + \boldsymbol{G})(2\boldsymbol{I} + \frac{\beta}{\gamma}\boldsymbol{L}^{-1})), \tag{19}$$

where γ and β are hyper-parameters, and $\boldsymbol{I}$ denotes the identity matrix.

3.4.4. The Optimization Algorithm

As shown in Algorithm 1, DMSFH's learning algorithm takes raw input training data, including images, text, and labels: $O = \{o_1, o_2, \ldots, o_n\}$, with $o_i = (v_i, t_i, \boldsymbol{L}_i)$. Before the training, parameters ϑ_v and ϑ_t of image network and text network were initialized; mini-batch size $N_v = N_t = 128$; the maximal number of epochs $max\_epoch = 500$; iteration times in each epoch was $iter_v = n/N_v$; $iter_t = n/N_t$, where n is the total number of training data. The training of each epoch consisted of three steps. Step 1: Randomly selecting N_v images from O and setting them as a mini-batch. For each datum in the mini-batch, we calculated $\boldsymbol{F}_{i*} = f_1^v(v_i; \theta_v; \theta_{vh})$ and $\hat{\boldsymbol{L}}_{i*}^v = f_2^v(v_i; \theta_v; \theta_{vl})$ by forward propagation. After the gradient was calculated, the network parameters θ_v, θ_{vh} and θ_{vl} were updated using SGD and back propagation. Step 2: Randomly selecting N_t texts from O and setting them as a mini-batch. For each datum in the mini-batch, we calculated $\boldsymbol{G}_{i*} = f_1^t(t_i; \theta_t; \theta_{th})$ and $\hat{\boldsymbol{L}}_{i*}^t = f_2^t(t_i; \theta_t; \theta_{tl})$ by forward propagation. After the gradient is calculated, the network parameters θ_t, θ_{th} and θ_{tl} were updated using SGD and back propagation. Step 3: Updating $\boldsymbol{B}$ by Equation (19). The above three steps were repeatedly iterated to realize the alternating training of image hash network and text hash network until the maximum epoch number of iterations was reached.

Algorithm 1 The learning algorithm for DMSFH

Require: Training data includes images, text, and labels: $O = \{o_1, o_2, \ldots, o_n\}$, with $o_i = (v_i, t_i, \boldsymbol{L}_i)$.
Ensure: Parameters ϑ_v and ϑ_t of deep neural networks, and binary code matrix $\boldsymbol{B}$.
Initialization
initialize parameters ϑ_v and ϑ_t, mini-batch size $N_v = N_t = 128$, the maximal number of epoches $max\_epoch = 500$, and iteration number $iter_v = n/N_v$, $iter_t = n/N_t$.
repeat
for $iter = 1, 2, \ldots, iter_v$ **do**
Randomly sample N_v images from O to construct a mini-batch of images.
For each instance v_i in the mini-batch, calculate $\boldsymbol{F}_{i*} = f_1^v(v_i; \theta_v; \theta_{vh})$ and $\hat{\boldsymbol{L}}_{i*}^v = f_2^v(v_i; \theta_v; \theta_{vl})$ by forward propagation.
Updata $\boldsymbol{F}$.
Calculate the derivatives according to Equations (14) and (15)
Update the network parameters θ_v, θ_{vh} and θ_{vl} by applying backpropagation.
end for
for $iter = 1, 2, \ldots, iter_t$ **do**
Randomly sample N_t texts from O to construct a mini-batch of texts.
For each instance t_i in the mini-batch, calculate $\boldsymbol{G}_{i*} = f_1^t(t_i; \theta_t; \theta_{th})$ and $\hat{\boldsymbol{L}}_{i*}^t = f_2^t(t_i; \theta_t; \theta_{tl})$ by forward propagation.
Updata $\boldsymbol{G}$.
Calculate the derivatives according to Equations (16) and (17)
Update the network parameters θ_t, θ_{th} and θ_{tl} by applying backpropagation.
end for
Update $\boldsymbol{B}$ using Equation (19)
until the max epoch number max_epoch

4. Experiment

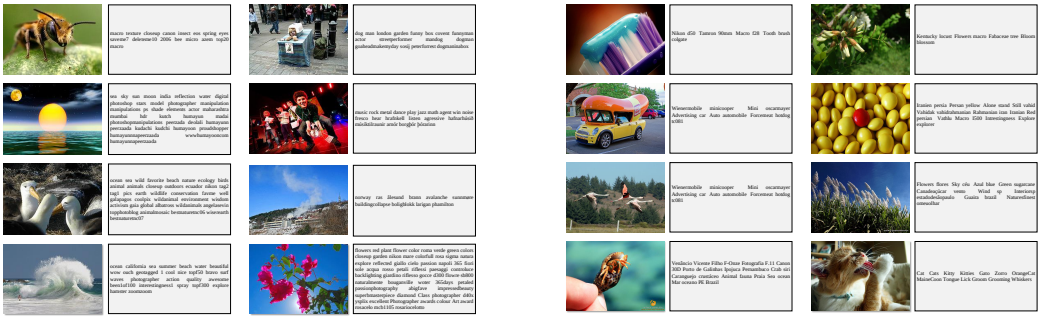
We conducted extensive experiments on two commonly used benchmark datasets, i.e., MIRFLICKR-25K [68] and NUS-WIDE [69], to evaluate the performance of our method, DMSFH. Firstly, we introduce the datasets, evaluation metrics, and implementation details, and then discuss performance comparisons of DMSFH and 6 state-of-the-art methods.

4.1. Datasets

MIRFLICKR-25K: The original MIRFLICKR-25K [68] dataset contains 25,000 image–text pairs, which were collected from the well-known photo sharing website Flickr. Each of these images has several textual tags. We selected those instances that have at least 20 textual tags for our experiments. The textual tags for each of the selected instances were transformed into a 1386-dimensional BoW vector. In addition, each instance was manually annotated with at least one of the 24 unique labels. We selected 20,015 instances for our experiments.

NUS-WIDE: The NUS-WIDE [69] dataset is a large real-world Web image dataset comprising over 269,000 images with over 5000 user-provided tags, and 81 concepts for the entire dataset. The text of each instance is represented as a 1000-dimensional BoW vector. In our experiment, we removed the instances without labels, and selected instances labeled by the 21 most-frequent categories. This gave 190,421 image–text pairs.

Table 2 presents the statistics of the above two datasets. Figure 2 shows some samples of these two datasets.



NUS-WIDE MIRFlickr-25k

Figure 2. Some examples from the MIRFLICKR-25K and NUS-WIDE datasets.

Table 2. Statistics of the datasets in our experiments.

Dataset	Total	Train/Query/Retrieval	Labels	Text Feature
MIRFLICKR-25K	20,015	10,000/2000/18,015	24	1386d
NUS-WIDE	190,421	10,500/2100/188,321	21	1000d

4.2. Evaluation

Two widely used evaluation methods, i.e., Hamming ranking and hash lookup, were utilized for cross-modal hash retrieval evaluations. Based on the query data and the Hamming distance of the retrieved samples as the sorting criteria, Hamming sorting sorts the retrieved data one by one according to the increasing order of the Hamming distance. In Hamming sorting, mean average precision (MAP) is one of the performance metrics that is commonly used to measure the accuracy of the query results. The larger the MAP value, the better the method retrieval performance. The topN precision curve reflects the changes in precision according to the number of retrieved instances. Besides, a hash search is also based on the criteria of the query data and the Hamming distance of the retrieved samples. However, it only returns the data to be retrieved within the specified Hamming distance as the final result. This can be measured by a precision recall (PR) curve. The larger the area enclosed by the curve and the coordinate axis, the better the retrieval performance of the method.

The value of MAP is defined as:

$$MAP = \frac{1}{M}\sum_{i=1}^{M} AP(q_i), \tag{20}$$

where M is the query dataset and $AP(q_i)$ is the average accuracy of query data q_i. The average value of accuracy is calculated as shown in Equation (21):

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

where N is the number of relevant instances in the retrieved set, and R represents the total amount of data. $p(r)$ denotes the precision of the top r retrieved instances, and $d(r) = 1$ if the r-th retrieved result is relevant to the query instances; otherwise, $d(r) = 0$.

To comprehensively measure the retrieval performance, we utilize another important evaluation metric, i.e., F-score. It is an important evaluation metrics that comprehensively considers precision and recall, which are defined as:

$$F\text{-}score = (1+\beta^2)\frac{Precision * Recall}{\beta^2 * Precision + Recall}; \tag{22}$$

if $\beta = 1$, this measurement is called F1-score. At this time, the accuracy rate and recall rate have the same weight. That means they are same important. In our experiments, we used F1-score to evaluate the cross-modal retrieval performance.

4.3. Baselines and Implementation Detail

Baselines. In this paper, the proposed SFDCH method is compared with several baselines, including SCM [42], SEPH [41], PRDH [65], CMHH [59], CHN [66], and DCMH [64]. SCM and SEPH use manual features, and the other approaches extract features through deep neural networks. Here is a brief introduction to these competitors:

- **SCM** integrates semantic labels into the process of hash learning to conduct large-scale data modeling, which not only maintains the correlation between models, but also achieves good performance in accuracy.
- **SEPH** transforms the semantic similarity of training data into affinity matrix by using a label as supervised information, and minimizes the Kullback–Leibler divergence to learn hash codes.
- **PRDH** integrates two types of pairwise constraints from inter-modality and intra-modality to enhance the similarities of the hash codes.
- **CMHH** learns high-quality hash representations to significantly penalize similar cross-modal pairs with Hamming distances larger than the Hamming radius threshold.
- **CHN** is a hybrid deep architecture that jointly optimizes the new cosine max-margin loss in semantic similarity pairs and the new quantization max-margin loss in compact hash codes.
- **DCMH** integrates features and hash codes learning into a general learning framework. The cross-modal similarities are preserved by using a negative log-likelihood loss.

Implementation Details. Our SFDCH approach was implemented by Pytorch framework. All the experiments were performed on a workstation with Intel(R) Xeon E5-2680_v3 2.5 GHz, 128 GB RAM, 1 TB SSD, and 3TB HDD storage; and 2 NVIDIA GeForce RTX 2080Ti GPUs with Windows 10 64-bit operating system. We set the $max\_epoch = 500$; the learning rate was initialized to $10^{-1.5}$ and gradually lowered to 10^{-6} in 500 epochs. We set the batch size of the mini-batch to 128 and the iteration number of the outer-loop in Algorithm 1 to 500, and the hyper-parameters $\gamma = \beta = 1$. For whole experiment, we used $I \rightarrow T$ to denote using a querying image while returning text, and $T \rightarrow I$ to denote using a querying text while returning an image.

4.4. Performance Comparisons

To evaluate the performance of the proposed method, we compare DMSFH with the six baselines in terms of MAP and PR curves on MIRFLICKR-25K and NUS-WIDE, respectively. Two query tasks, i.e., image-query-text and text-query-image, are considered. Tables 3 and 4 illustrate the MAP results of DMSFH and other methods on different lengths (16, 32, 64 bits) of hash codes on MIRFlickr-25K and NUS-WIDE, respectively. Figures 3–5 demonstrate the PR curves of different coding lengths on MIRFlickr-25K and NUS-WIDE, respectively. Table 5 reports the F1-measure with hash code length 32 bits on the MIRFLICKR-25K dataset.

Hamming Ranking: Tables 3 and 4 report the MAP scores of the proposed method and its competitors for image-query-text and text-query-image on MIRFLICKR-25K and NUS-WIDE, where $I \rightarrow T$ and $T \rightarrow I$ represent image retrieval by text and text retrieval by image, respectively. It is clear from the Tables 3 and 4 that the deep hashing methods perform better than the non-deep methods. Specifically, on MIRFLICKR-25K, we can see in Table 3 that the proposed method DMSFH achieved the highest MAP score for both queries ($I \rightarrow T$: 16 bits MAP = 79.12%, 32 bits MAP = 79.60%, 64 bits MAP = 80.45%; $T \rightarrow I$: 16 bits MAP = 78.22%, 32 bits MAP = 78.62%, 64 bits MAP = 79.50%). It defeated the two most competitive deep learning-based baselines, CNH and DCMH, due to the multiple label semantic fusion. Similarly, we can find from Table 4 that DMSFH won the competition again on NUS-WIDE by $I \rightarrow T$ MAP = 64.08% (16 bits), 65.12% (32 bits), 66.43% (64 bits);

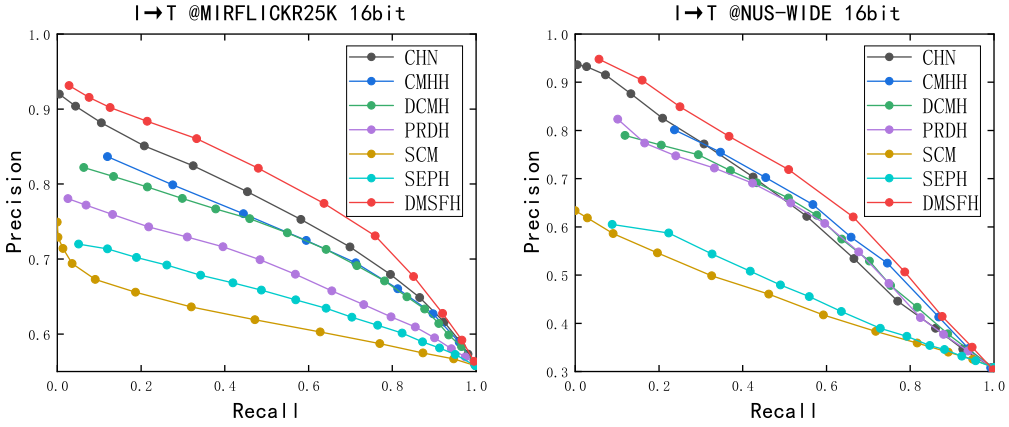
and $T \rightarrow I$ MAP = 63.89% (16 bits), 65.31% (32 bits), 66.08% (64 bits), respectively. This superiority of DMSFH due to the fact that it incorporates richer semantic information than other techniques. In addition, DMSFH leverages graph regularization to measure the semantic correlation of the unified hash codes. That means it can capture more semantic consistent features between different modalities than other deep hashing models, such as CHN and DCMH. Therefore, the above results confirm that the hash codes generated by DMSFH have better semantic discrimination and can better adapt to the task of mutual retrieval of multi-modal data.

Table 3. Mean average precision (MAP) comparison on MIRFLICKR-25K. The best results are in bold font.

Methods	MIRFLICKR-25K					
	Image-Query-Text			Text-Quary-Image		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
SCM [42]	0.6410	0.6478	0.6608	0.6450	0.6532	0.6623
SEPH [41]	0.6785	0.6853	0.6884	0.7168	0.7298	0.7325
PRDH [65]	0.7016	0.7101	0.7184	0.7663	0.7764	0.7811
CMHH [59]	0.7374	0.7328	0.7510	0.7388	0.7241	0.7326
CHN [66]	0.7543	0.7533	0.7512	0.7724	0.7782	0.7810
DCMH [64]	0.7406	0.7415	0.7434	0.7617	0.7716	0.7748
DMSFH	**0.7912**	**0.7960**	**0.8045**	**0.7822**	**0.7862**	**0.7950**

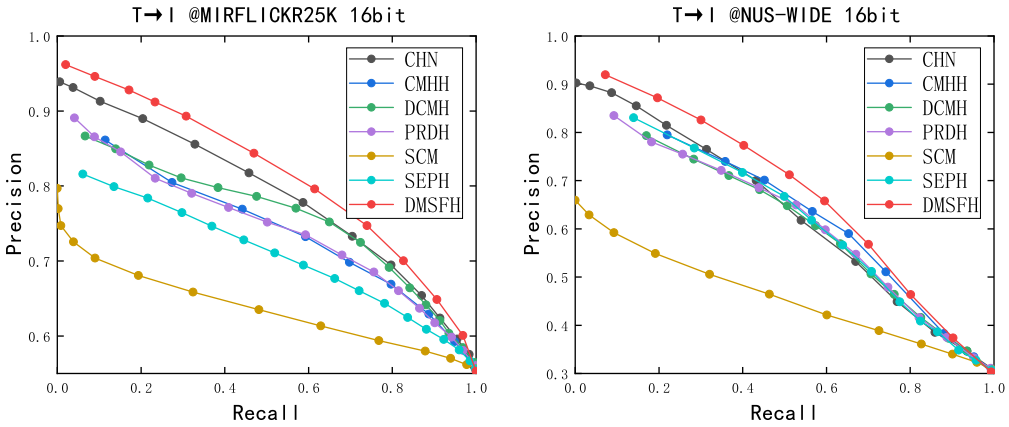
Table 4. Mean average precision (MAP) comparison on NUS-WIDE. The best results are in bold font.

Methods	NUS-WIDE					
	Image-Query-Text			Text-Quary-Image		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
SCM [42]	0.4642	0.4825	0.4910	0.4308	0.4414	0.4536
SEPH [41]	0.4831	0.4898	0.4953	0.6117	0.6322	0.6342
PRDH [65]	0.6002	0.6118	0.6180	0.6214	0.6302	0.6357
CMHH [59]	0.5574	0.5720	0.6021	0.5798	0.5834	0.5935
CHN [66]	0.5802	0.6024	0.6086	0.5878	0.6034	0.6045
DCMH [64]	0.5512	0.5638	0.5940	0.5878	0.6011	0.6106
DMSFH	**0.6408**	**0.6512**	**0.6643**	**0.6389**	**0.6531**	**0.6608**



(**a**) MIRFLICKR-25K (**b**) NUS-WIDE

Figure 3. *Cont.*

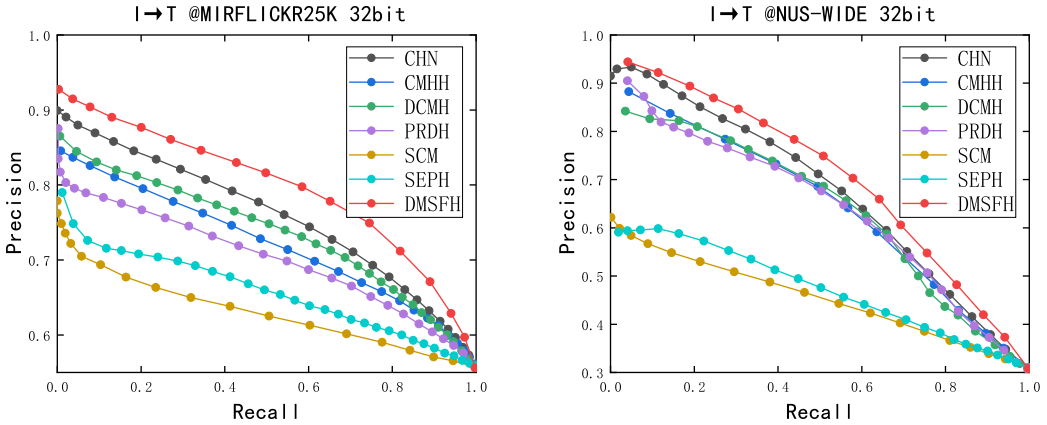


(**c**) MIRFLICKR-25K (**d**) NUS-WIDE

Figure 3. Precision–recall curves on MIRFLICKR-25K and NUS-WIDE. I→T query: (**a**,**b**); T→I query: (**c**,**d**). The code length was 16 bits.

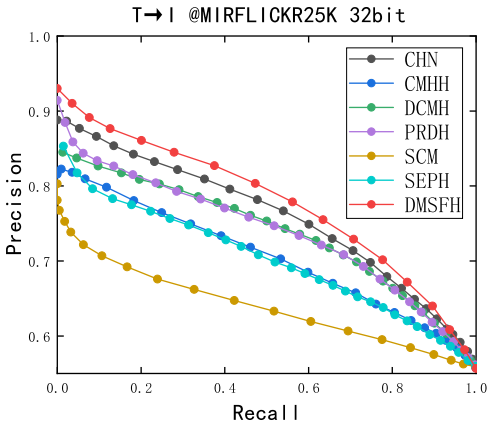
Table 5. F1-measures of our method and the competitors on MIRFLICKR-25K. The code length was 32 bit. The best results are in bold font.

Methods	MIRFLICKR-25K					
	Image-Query-Text			Text-Quary-Image		
	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure
DMSFH	0.9135	0.0616	**0.1154**	0.9046	0.0562	**0.1058**
CHN	0.8852	0.0376	0.0721	0.8741	0.0321	0.0619
DCMH	0.8682	0.0525	0.0990	0.8556	0.0428	0.0815
CMHH	0.8373	0.0412	0.0785	0.8216	0.0308	0.0694
PRDH	0.8586	0.0206	0.0402	0.8742	0.0326	0.0628

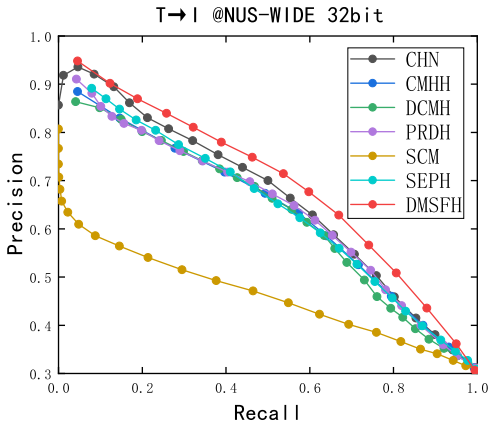


(**a**) MIRFLICKR-25K (**b**) NUS-WIDE

Figure 4. *Cont*.



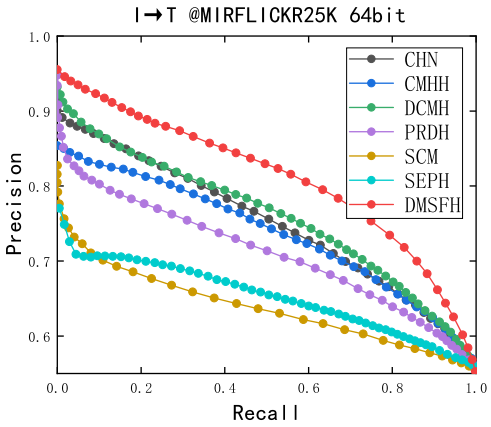
(**c**) MIRFLICKR-25K



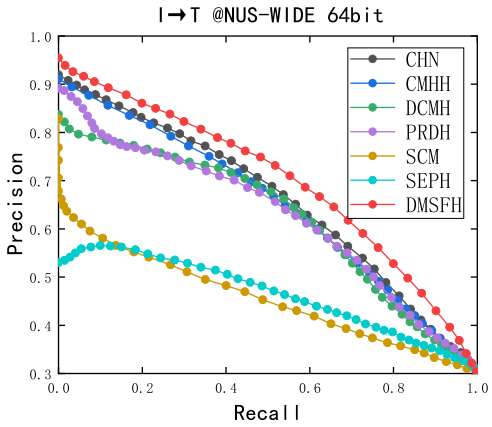
(**d**) NUS-WIDE

Figure 4. Precision-Recall curves on MIRFLICKR-25K and NUS-WIDE dataset. I→T query: (**a**,**b**); T→I query: (**c**,**d**). The code length was 32 bits.

Hash Lookup: To further demonstrate the comparison of the proposed model with these baselines, we used PR curves to evaluate their retrieval performances. Figures 3–5 show the PR curves with different coding lengths (16 bits, 32 bits, and 64 bits) on MIRFLICKR-25K and NUS-WIDE datasets, respectively. As expected, the deep learning-based models had better performances than the manual features-based models, mainly due to the powerful representation capabilities of deep neural networks. Besides, no matter what the length of the hash code was, our method performed better, obviously, on the PR curve than the other deep based competitors. That happened mainly because DMSFH has stronger cross-modal consistent semantic learning capabilities by not only considering both the intra-modal and inter-modal semantic discriminative information, but integrating graph regularization into hashing learning as well. Besides, we selected the best five methods, and report their average precision, average recall, and average F1-measure with Hamming radius $r = 0, 1, 2$ in Table 5 on MIRFLICKR-25K for when the code length was 32. We found that in all cases our DMSFH can achieve the best F1-measure.

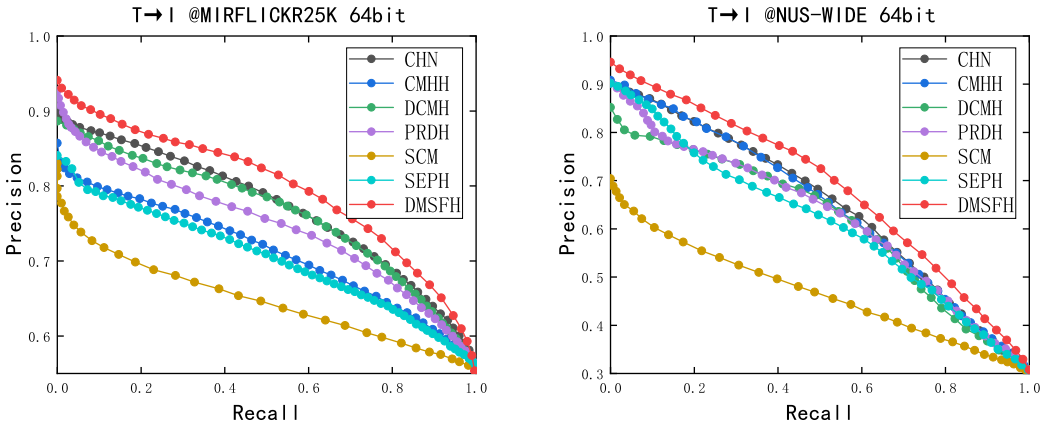


(**a**) MIRFLICKR-25K



(**b**) NUS-WIDE

Figure 5. *Cont*.



(**c**) MIRFLICKR-25K (**d**) NUS-WIDE

Figure 5. Precision–recall curves on MIRFLICKR-25K and NUS-WIDE datasets. I→T query: (**a**,**b**); T→I query: (**c**,**d**). The code length was 64 bits.

4.5. Ablation Experiments of DMSFH

To verify the validity of the DMSFH components, we conducted ablation experiments on the MIRFLICKR-25K dataset, and the experimental results are shown in Table 6. We define DMSFH-P as employing only intra-modal pairwise loss and inter-modal pairwise loss, and DMSFH-S removed the graph regularization loss. From Table 6, we can see that both the semantic prediction discriminant loss and graph regularization loss employed by DMSFH can effectively improve the retrieval accuracy. From the results, it can be seen that DMSFH can obtain better performance when using the designed modules.

Table 6. Ablation experiments of DMSFH on the MIRFLICKR-25K dataset. The best results are in bold font.

Methods	**MIRFLICKR-25K**					
	Image-Query-Text			**Text-Quary-Image**		
	16 bits	**32 bits**	**64 bits**	**16 bits**	**32 bits**	**64 bits**
DMSFH-P	0.7708	0.7812	0.7904	0.7690	0.7728	0.7802
DMSFH-S	0.7876	0.7926	0.8012	0.7786	0.7836	0.7918
DMSFH	**0.7912**	**0.7960**	**0.8045**	**0.7822**	**0.7862**	**0.7950**

5. Discussion

This paper proposes deep multi-semantic fusion-based cross-modal hashing (DMSFH) for cross-modal retrieval tasks. Firstly, it preserves the semantic similarity between data through intra-modal loss and inter-modal loss, and then introduces a multi-label semantic fusion module to further capture more semantic discriminative features. In addition, semantic similarity in Hamming space is preserved by graph regularization loss.

We compared DMSFH with other methods. We used the cross-modal multi-label datasets MIRFLICKR-25K and NUS-WIDE, which have 24 and 21 label attributes, respectively. According to Tables 3 and 4, it can be seen that the map scores of DMSFH are better than those of the other methods. As for DCMH and PRDH, DMSFH outperformed these two deep learning methods based on the same inter-modal and intra-modal pairwise loss, precisely because it captured more semantic information with the addition of new losses. Therefore, DMSFH is able to optimize the semantic heterogeneity problem to a certain extent and improve the accuracy. In addition, the computational cost of the model is measured using floating point operations (FLOPs), with an approximate number of FLOPs of 3.67 billion for DMSFH. Compared with real-valued cross-modal retrieval methods, the computational and retrieval cost of our method is quite low due to the shorter binary

cross-modal representations (i.e., 64 bits hash codes) and Hamming distance measurements. As they generate higher dimensional feature representations (i.e., 1000 dimensional feature map), the real-valued cross-modal representation learning models always have higher complexity.

Although our study achieved some degree of performance improvement, there are limitations. First, when constructing the sample similarity matrix, our method in this paper does not fully extract the fine-grained labeling information between data, and there is still a higher performance improvement in fine-grained semantic information extraction. Second, our method mainly focuses on the construction and optimization of the loss function, but how to improve the cross-modal semantic feature representation learning is also an important issue. Therefore, deeper semantic mining in the semantic feature learning part is also a direction for our future research. Third, our method was tested on a specific dataset, and common cross-modal hash retrieval methods use data of known categories, but in practical applications, the rapid emergence of new unlabeled things often affects the accuracy of cross-modal data retrieval. How to achieve high precision cross-modal retrieval in the absence of annotation information is also an important research problem.

6. Conclusions

In this paper, we proposed an effective hashing approach dubbed deep multi-semantic fusion-based cross-modal hashing (DMSFH) to improve semantic discriminative feature learning and similarity preserving of hash codes in common Hamming subspace. This method learns an end-to-end framework to integrate feature learning and hash code learning. A multi-label semantic fusion method is used to realize cross-modal consistent semantic learning to enhance the semantic discriminability of hash codes. Moreover, we designed the loss function with graph regularization from inter-modal and intra-modal perspectives to enhance the similarity learning of hash codes in Hamming subspace. Extensive experiments on two cross-modal datasets demonstrated that our proposed approach can effectively improve cross-modal retrieval performance, which is significantly superior to other baselines.

In future work, we will consider the heterogeneous semantic correlations between multi-modal samples in both aspects of high-level semantics and fine-grained semantics, which can be formulated as heterogeneous information networks (HIN) to capture more semantic information and realize cross-modal semantic alignment in a more effective manner. In addition, how to measure the distance of the relation distribution of semantic details between different modalities will be studied. An essential problem will be enhancing the cross-modal semantic representation learning.

Author Contributions: Conceptualization, X.Z. and L.C.; methodology, L.C. and L.Z.; software, L.C.; validation, L.C., L.Z. and Z.Z.; formal analysis, L.Z.; investigation, X.Z. and L.C.; resources, X.Z. and L.Z.; data curation, Z.Z.; writing—original draft preparation, L.C.; writing—review and editing, L.Z.; visualization, Z.Z.; supervision, X.Z. and L.Z.; project administration, X.Z.; funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Research and Development Program of Hunan Province (2020NK2033), and the National Natural Science Foundation of China (62072166).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: This work was supported in part by the Key Research and Development Program of Hunan Province (2020NK2033) and the National Natural Science Foundation of China (62072166).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, Y. Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2021**, *17*, 1–25. [CrossRef]
2. Zhu, L.; Zhang, C.; Song, J.; Liu, L.; Zhang, S.; Li, Y. Multi-Graph Based Hierarchical Semantic Fusion for Cross-Modal Representation. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Virtual, 5–9 July 2021; pp. 1–6.
3. Morgado, P.; Vasconcelos, N.; Misra, I. Audio-visual instance discrimination with cross-modal agreement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 12475–12486.
4. Zhang, B.; Hu, H.; Sha, F. Cross-modal and hierarchical modeling of video and text. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 374–390.
5. Jing, L.; Vahdani, E.; Tan, J.; Tian, Y. Cross-Modal Center Loss for 3D Cross-Modal Retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 3142–3151.
6. Liu, R.; Wei, S.; Zhao, Y.; Zhu, Z.; Wang, J. Multiview Cross-Media Hashing with Semantic Consistency. *IEEE Multimed.* **2018**, *25*, 71–86. [CrossRef]
7. Zhang, D.; Wu, X.J.; Yu, J. Label consistent flexible matrix factorization hashing for efficient cross-modal retrieval. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2021**, *17*, 1–18. [CrossRef]
8. Zhu, L.; Song, J.; Zhu, X.; Zhang, C.; Zhang, S.; Yuan, X. Adversarial Learning-Based Semantic Correlation Representation for Cross-Modal Retrieval. *IEEE Multimed.* **2020**, *27*, 79–90. [CrossRef]
9. Wu, L.; Wang, Y.; Shao, L. Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Trans. Image Process.* **2018**, *28*, 1602–1612. [CrossRef] [PubMed]
10. Wang, Y.; Zhang, W.; Wu, L.; Lin, X.; Fang, M.; Pan, S. Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering. *arXiv* **2016**, arXiv:1608.05560.
11. Zhao, M.; Liu, Y.; Li, X.; Zhang, Z.; Zhang, Y. An end-to-end framework for clothing collocation based on semantic feature fusion. *IEEE Multimed.* **2020**, *27*, 122–132. [CrossRef]
12. Shen, H.T.; Liu, L.; Yang, Y.; Xu, X.; Huang, Z.; Shen, F.; Hong, R. Exploiting subspace relation in semantic labels for cross-modal hashing. *IEEE Trans. Knowl. Data Eng.* **2020**, *33*, 3351–3365. [CrossRef]
13. Zhang, C.; Song, J.; Zhu, X.; Zhu, L.; Zhang, S. HCMSL: Hybrid Cross-modal Similarity Learning for Cross-modal Retrieval. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2021**, *17*, 1–22. [CrossRef]
14. Zhao, S.; Xu, M.; Huang, Q.; Schuller, B.W. Introduction to the special issue on MMAC: Multimodal affective computing of large-scale multimedia data. *IEEE Multimed.* **2021**, *28*, 8–10. [CrossRef]
15. Sharma, A.; Kumar, A.; Daume, H.; Jacobs, D.W. Generalized multiview analysis: A discriminative latent space. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2160–2167.
16. Wang, K.; He, R.; Wang, L.; Wang, W.; Tan, T. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2010–2023. [CrossRef] [PubMed]
17. Deng, C.; Tang, X.; Yan, J.; Liu, W.; Gao, X. Discriminative dictionary learning with common label alignment for cross-modal retrieval. *IEEE Trans. Multimed.* **2015**, *18*, 208–218. [CrossRef]
18. Li, K.; Qi, G.J.; Ye, J.; Hua, K.A. Linear subspace ranking hashing for cross-modal retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1825–1838. [CrossRef] [PubMed]
19. Zhu, L.; Song, J.; Wei, X.; Yu, H.; Long, J. CAESAR: Concept augmentation based semantic representation for cross-modal retrieval. *Multimed. Tools Appl.* **2020**, 1–31. [CrossRef]
20. Chen, Y.; Wang, Y.; Ren, P.; Wang, M.; de Rijke, M. Bayesian feature interaction selection for factorization machines. *Artif. Intell.* **2022**, *302*, 103589. [CrossRef]
21. Wei, Y.; Zhao, Y.; Lu, C.; Wei, S.; Liu, L.; Zhu, Z.; Yan, S. Cross-modal retrieval with CNN visual features: A new baseline. *IEEE Trans. Cybern.* **2016**, *47*, 449–460. [CrossRef]
22. Hotelling, H. Relations between Two Sets of Variates. In *Breakthroughs in Statistics*; Springer: New York, NY, USA, 1992; pp. 162–190.
23. Gong, Y.; Ke, Q.; Isard, M.; Lazebnik, S. A multi-view embedding space for modeling internet images, tags, and their semantics. *Int. J. Comput. Vis.* **2014**, *106*, 210–233. [CrossRef]
24. Zu, C.; Zhang, D. Canonical sparse cross-view correlation analysis. *Neurocomputing* **2016**, *191*, 263–272. [CrossRef]
25. Ballan, L.; Uricchio, T.; Seidenari, L.; Del Bimbo, A. A cross-media model for automatic image annotation. In Proceedings of the International Conference on Multimedia Retrieval, Glasgow, UK, 1–4 April 2014; pp. 73–80.
26. Wang, L.; Zhu, L.; Dong, X.; Liu, L.; Sun, J.; Zhang, H. Joint feature selection and graph regularization for modality-dependent cross-modal retrieval. *J. Vis. Commun. Image Represent.* **2018**, *54*, 213–222. [CrossRef]
27. Zhang, C.; Liu, M.; Liu, Z.; Yang, C.; Zhang, L.; Han, J. Spatiotemporal activity modeling under data scarcity: A graph-regularized cross-modal embedding approach. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 531–538.
28. Deng, C.; Chen, Z.; Liu, X.; Gao, X.; Tao, D. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Trans. Image Process.* **2018**, *27*, 3893–3903. [CrossRef]

29. Xu, X.; Shimada, A.; Taniguchi, R.I.; He, L. Coupled dictionary learning and feature mapping for cross-modal retrieval. In Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME), Turin, Italy, 29 June–3 July 2015; pp. 1–6.
30. Xu, X.; Yang, Y.; Shimada, A.; Taniguchi, R.I.; He, L. Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts. In Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 847–850.
31. Zhang, C.; Zhong, Z.; Zhu, L.; Zhang, S.; Cao, D.; Zhang, J. M2GUDA: Multi-Metrics Graph-Based Unsupervised Domain Adaptation for Cross-Modal Hashing. In Proceedings of the 2021 International Conference on Multimedia Retrieval, Taipei, Taiwan, 16–19 November 2021; pp. 674–681.
32. Zhu, L.; Song, J.; Yang, Z.; Huang, W.; Zhang, C.; Yu, W. DAP ^{2}2 CMH: Deep Adversarial Privacy-Preserving Cross-Modal Hashing. *Neural Process. Lett.* **2021**, 1–21. [CrossRef]
33. Mithun, N.C.; Sikka, K.; Chiu, H.P.; Samarasekera, S.; Kumar, R. Rgb2lidar: Towards solving large-scale cross-modal visual localization. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 934–954.
34. Zhan, Y.W.; Wang, Y.; Sun, Y.; Wu, X.M.; Luo, X.; Xu, X.S. Discrete online cross-modal hashing. *Pattern Recognit.* **2022**, *122*, 108262. [CrossRef]
35. Zhou, J.; Ding, G.; Guo, Y. Latent semantic sparse hashing for cross-modal similarity search. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, Gold Coast, QLD, Australia, 6–11 July 2014; pp. 415–424.
36. Song, J.; Yang, Y.; Yang, Y.; Huang, Z.; Shen, H.T. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 22–27 June 2013; pp. 785–796.
37. Wang, D.; Gao, X.; Wang, X.; He, L. Semantic topic multimodal hashing for cross-media retrieval. In Proceedings of the Twenty-fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015; pp. 3890–3896.
38. Hu, M.; Yang, Y.; Shen, F.; Xie, N.; Hong, R.; Shen, H.T. Collective reconstructive embeddings for cross-modal hashing. *IEEE Trans. Image Process.* **2018**, *28*, 2770–2784. [CrossRef]
39. Zhang, J.; Peng, Y.; Yuan, M. Unsupervised generative adversarial cross-modal hashing. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 539–546.
40. Bronstein, M.M.; Bronstein, A.M.; Michel, F.; Paragios, N. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3594–3601.
41. Lin, Z.; Ding, G.; Hu, M.; Wang, J. Semantics-preserving hashing for cross-view retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3864–3872.
42. Zhang, D.; Li, W.J. Large-scale supervised multimodal hashing with semantic correlation maximization. In Proceedings of the AAAI Conference on Artificial Intelligence, Québec City, QC, Canada, 27–31 July 2014; Volume 28, pp. 2177–2183.
43. Xu, X.; Shen, F.; Yang, Y.; Shen, H.T.; Li, X. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Trans. Image Process.* **2017**, *26*, 2494–2507. [CrossRef] [PubMed]
44. Mandal, D.; Chaudhury, K.N.; Biswas, S. Generalized semantic preserving hashing for cross-modal retrieval. *IEEE Trans. Image Process.* **2018**, *28*, 102–112. [CrossRef] [PubMed]
45. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
47. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv* **2014**, arXiv:1405.3531.
48. Yang, W.; Peng, J.; Wang, H.; Wang, M. Progressive Learning with Multi-scale Attention Network for Cross-domain Vehicle Re-identification. *Sci. China Inf. Sci.* **2021**. [CrossRef]
49. Liu, X.; Yu, G.; Domeniconi, C.; Wang, J.; Ren, Y.; Guo, M. Ranking-based deep cross-modal hashing. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 4400–4407.
50. Zhen, L.; Hu, P.; Wang, X.; Peng, D. Deep supervised cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10394–10403.
51. Jiang, Q.Y.; Li, W.J. Discrete latent factor model for cross-modal hashing. *IEEE Trans. Image Process.* **2019**, *28*, 3490–3501. [CrossRef]
52. Wei, J.; Xu, X.; Yang, Y.; Ji, Y.; Wang, Z.; Shen, H.T. Universal weighting metric learning for cross-modal matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13005–13014.
53. Zhang, M.; Li, J.; Zhang, H.; Liu, L. Deep semantic cross modal hashing with correlation alignment. *Neurocomputing* **2020**, *381*, 240–251. [CrossRef]
54. Li, C.; Deng, C.; Li, N.; Liu, W.; Gao, X.; Tao, D. Self-supervised adversarial hashing networks for cross-modal retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4242–4251.

55. Xie, D.; Deng, C.; Li, C.; Liu, X.; Tao, D. Multi-task consistency-preserving adversarial hashing for cross-modal retrieval. *IEEE Trans. Image Process.* **2020**, *29*, 3626–3637. [CrossRef] [PubMed]
56. Wang, X.; Shi, Y.; Kitani, K.M. Deep Supervised Hashing with Triplet Labels. In *Asian Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 70–84.
57. Chen, S.; Wu, S.; Wang, L. Hierarchical semantic interaction-based deep hashing network for cross-modal retrieval. *PeerJ Comput. Sci.* **2021**, *7*, e552. [CrossRef] [PubMed]
58. Zou, X.; Wang, X.; Bakker, E.M.; Wu, S. Multi-label semantics preserving based deep cross-modal hashing. *Signal Process. Image Commun.* **2021**, *93*, 116131. [CrossRef]
59. Cao, Y.; Liu, B.; Long, M.; Wang, J. Cross-modal hamming hashing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 202–218.
60. Ding, G.; Guo, Y.; Zhou, J. Collective matrix factorization hashing for multimodal data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2075–2082.
61. Fang, Y.; Zhang, H.; Ren, Y. Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing. *Knowl.-Based Syst.* **2019**, *171*, 69–80. [CrossRef]
62. Fang, Y.; Li, B.; Li, X.; Ren, Y. Unsupervised cross-modal similarity via Latent Structure Discrete Hashing Factorization. *Knowl.-Based Syst.* **2021**, *218*, 106857. [CrossRef]
63. Kumar, S.; Udupa, R. Learning hash functions for cross-view similarity search. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011; pp. 1360–1365.
64. Jiang, Q.Y.; Li, W.J. Deep cross-modal hashing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3232–3240.
65. Yang, E.; Deng, C.; Liu, W.; Liu, X.; Tao, D.; Gao, X. Pairwise relationship guided deep hashing for cross-modal retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31, pp. 1618–1625.
66. Cao, Y.; Long, M.; Wang, J.; Yu, P.S. Correlation hashing network for efficient cross-modal retrieval. *arXiv* **2016**, arXiv:1602.06697.
67. Wang, L.; Zhu, L.; Yu, E.; Sun, J.; Zhang, H. Fusion-supervised deep cross-modal hashing. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 37–42.
68. Huiskes, M.J.; Lew, M.S. The mir flickr retrieval evaluation. In Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, Vancouver, BC, Canada, 30–31 October 2008; pp. 39–43.
69. Chua, T.S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; Zheng, Y. Nus-wide: A real-world web image database from national university of singapore. In Proceedings of the ACM International Conference on Image and Video Retrieval, Santorini Island, Greece, 8–10 July 2009; pp. 1–9.