Statistical Model-driven Similarity Hashing: Bridging Modalities for Efficient Unsupervised Retrieval

Mingjin Kuai[1], Jun Long[2], Zhan Yang[2*]

[1]School of Computer Science and Engineering, Central South University, Changsha 410083, China
[2]Big Data Institute, Central South University, Changsha 410083, China
{kuaimingjin, junlong, zyang22}@csu.edu.cn

Abstract

Unsupervised deep cross-modal hash retrieval aims to map multi-modal features into binary hash codes without labels, which is of interest due to its storage efficiency, query speed and convenient applications. However, existing approaches suffer from two main limitations: (1) Slightly insufficient consideration of text instance similarity, along with independent or redundant fusion to learn multi-modal similarity information. (2) They ignore the noisy adjacent correlations between multi-modal instances, leading to a lack of discriminative power in the generated hash codes. To address these challenges, we propose a new approach called Statistical Model-driven Similarity Hashing (SMSH). Specifically, we introduce Jaccard similarity when constructing the text similarity matrix. It reduces the similarity error between text instances while better considering the asymmetry of the elements in the text features. After that, we integrate the original similarity information between various modalities to construct a unified similarity matrix. The gaps between modalities are bridged while reducing the redundant information in them. In addition, we introduce a Statistical Model-driven Similarity Enhancement (SMSE) approach, which reduces the noise of similarity relations between multi-modal instances by using a Gaussian Mixture Model to keep instances with lower semantic similarity as far away from each other as possible. Experiments on three benchmark datasets demonstrate the excellent performance of the SMSH method.

Introduction

With the exponential growth of multi-modal digital content, the single-modal hash retrieval (Shen et al. 2022; Dong et al. 2023) no longer meet the current information retrieval needs. Achieving accurate and efficient cross-modal hash retrieval from huge heterogeneous multi-modal data has become a challenging problem, and the key to this lies in cross-modal feature fusion. The cross-modal feature fusion process aims to map large-scale, high-dimensional multi-modal data into a common space and integrate them into a stable multi-modal representation. Since different modal instances vary in terms of feature representation and distribution, cross-modal hash retrieval requires exploring appropriate methods to bridge the gap between modalities. Cross-modal hash retrieval has been a popular research topic (Luo et al. 2023; Li et al. 2024b; Sun et al. 2024c) aiming to search for semantically similar instances from different modalities, *e.g.*, given one modality (*e.g.*, image) as a query, to retrieve the similar instances in another modality (*e.g.*, text).

Existing cross-modal hashing can be broadly classified into supervised and unsupervised methods based on whether or not semantic labels are used. For supervised methods (Wang et al. 2021; Zhang et al. 2023; Sun et al. 2024a; Chen et al. 2024; Sun et al. 2024b), mainly high quality semantic labels are used to generate hash codes. However, due to the expensive human cost as well as the large amount of time required to obtain correct and reasonable labels (Han et al. 2022; Xu et al. 2022), they cannot be applied on a large scale in the real world. Since deep neural networks (He et al. 2016; Radford et al. 2021) have excellent nonlinear feature extraction capabilities for multi-modal data, many unsupervised deep cross-modal hashing methods have been proposed to avoid relying on large amounts of labelled data. Recent works have focused on constructing unified similarity matrices to guide the learning of hash cods. For example, DJSRH (Su and Zhang 2019), achieves substantial improvement by constructing a novel joint semantic affinity matrix to integrate the original neighbourhood information from different modalities. However, the integration methods are redundant and sub-optimal, involving much unimportant information. Meanwhile, they ignore the noisy adjacent correlations between multi-modal instances, which may result in hash codes that lack discriminative power. Therefore, in order to improve the guidance of similarity signals, similarity enhancement schemes such as DAEH (Shi et al. 2022), CIRH (Zhu et al. 2023), and HEH (Zhong et al. 2023) have been proposed. However, to the best of our knowledge, existing work has not utilized statistical laws to improve it.

Therefore, in order to improve the retrieval performance of unsupervised methods, we need to focus on solving the following two *challenges*: on the one hand, it is to effectively fuse the multi-modal similarity information when considering the asymmetry of the elements in the textual features, so as to preserve the intra- and inter-modal semantic correlations in generated hash codes. On the other hand, it is to improve the discriminative power of hash codes by improving existing similarity enhancement schemes.

*Corresponding author

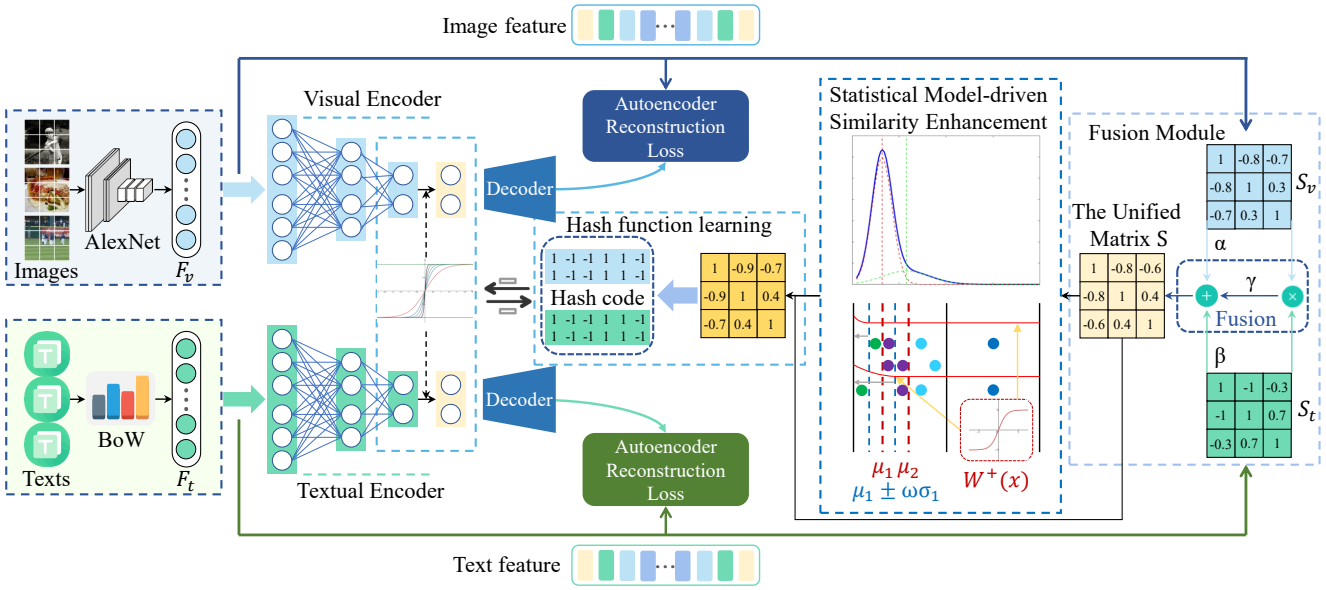


Figure 1: The pipeline of the proposed Statistical Model-driven Similarity Hashing (SMSH).

To overcome the above *challenges*, we propose a novel unsupervised approach called Statistical Model-driven Similarity Hashing (SMSH) for large-scale multi-modal data retrieval. For *Challenge1* , we additionally introduce Jaccard similarity in the construction of text similarity matrix $\mathbf{S}_t$ to solve the problem of asymmetry of text feature elements, so that the similarity between text instances is closer to the real similarity between instances. After that, it is fused with the image similarity matrix $\mathbf{S}_v$ to create a unified similarity matrix $\mathbf{S}$, which solves the redundant information generated during the fusion of cross-modal similarity information. This approach maintains similarity relations between multi-modal instances and effectively bridges the gap between modalities by exploiting the sparse information of textual modalities. For *Challenge2*, we conduct a comprehensive analysis of the statistical distribution laws within $\mathbf{S}_v$, $\mathbf{S}_t$, and $\mathbf{S}$. With the Gaussian Mixture Model, we reveal the underlying statistical properties, enabling us to enhance the similarity between cross-modal instances and effectively separate dissimilar instances. Notably, we identify that the similarity between cross-modal instances is predominantly influenced by the similarity of image instances. Consequently, our primary focus lies in enhancing the similarity between image instances, thereby significantly improving the discriminative power of the generated hash code. This approach, guided by the enhancement of $\mathbf{S}_v$, serves as a key factor in constructing the unified similarity matrix $\mathbf{S}$.

Overall, the contributions of this paper are as follows:

- A novel unsupervised cross-modal training method is proposed, which focuses on constructing a unified similarity enhancement matrix as a supervised signal in generated hash codes.
- A Statistical Model-driven Similarity Enhancement (SMSE) approach is proposed to generate discriminative hash codes using statistical laws for enhancement.
- SMSH is conceptually simple and easily extensible, and is more powerful than many existing unsupervised cross-modal hashing methods.

Related Work

In this section, we introduce unsupervised cross-modal hashing methods, which can be categorized into shallow and deep schemes based on the utilization of deep networks.

Unsupervised shallow Cross-modal Hashing. Inter-Media Hashing (IMH) (Song et al. 2013) is early unsupervised cross-modal hashing methods that extend Spectral Hashing (Weiss and Torralba 2008) to cross-modal hashing. Latent Semantic Sparse Hashing (LSSH) (Zhou, Ding, and Guo 2014) performs cross-modal similarity search via sparse coding and matrix factorization. Joint and Individual Matrix Factorization Hashing (JIMFH) (Wang et al. 2020) learns unified hash codes through joint matrix factorization in addition to learning individual hash codes through individual matrix factorization. Unsupervised Multi-modal Hashing based on Piecewise Learning (UMHPL) (Li et al. 2024a) integrates multi-modal data through adaptive weight factors and nuclear norm minimization. Online Cross-modal Hashing (DPOCH) (Xiao et al. 2024) proposes a dynamic prototype-based online cross-modal hashing method, and gets global adaptive hash codes and hash functions.

Unsupervised deep Cross-modal Hashing. Many unsupervised deep schemes have achieved excellent results due to the powerful nonlinear extraction capability of deep neural networks. To maintain the original data neighborhood structure, Deep Joint-Semantics Reconstructing Hashing (DJSRH) (Su and Zhang 2019) constructs a joint-semantics affinity matrix that captures the latent intrinsic semantic affinities of the input multi-modal instances. Aggregation-based Graph Convolutional Hashing (AGCH) (Zhang et al. 2022) develops an effective aggregation strategy that uses

multiple metrics to construct more accurate affinity matrices for learning. Correlation-Identity Reconstruction Hashing (CIRH) (Zhu et al. 2023) constructs a collaborated graph for modeling to generate a multi-modal complementary representation. Hugging Hashing (HUGH) (Wang et al. 2024) proposes a multi-granularity learning framework called hugging to bridge the modalities.

Notation and Problem Definition

We firstly introduce some notations that are used in this paper. We use $\mathbf{\Phi} = \{v_k, t_k\}_{k=1}^{m}$ to denote the m paired instances in each mini-batch, where v_k and t_k are the image and text raw data of k-th instance. We use $\mathbf{F}_v \in \mathbf{R}^{m \times d_v}$ and $\mathbf{F}_t \in \mathbf{R}^{m \times d_t}$ to denote the feature representations obtained from image and text respectively, where d_v and d_t represent the dimensions of the image and text feature vectors. Similarly, $\mathbf{F}_v^{'}$ and $\mathbf{F}_t^{'}$ represent the reconstructed features of the image and text obtained using the decoders, respectively.

Given $\mathbf{F}_v$ and $\mathbf{F}_t$, this paper aims to use deep hash networks $f_v\left(\mathbf{F}_v, \theta_v\right)$ and $f_t\left(\mathbf{F}_t, \theta_t\right)$, to generate hash codes $\mathbf{B}_v \in \{-1, +1\}^{m \times l}$ and $\mathbf{B}_t \in \{-1, +1\}^{m \times l}$ for image and text respectively, where θ_v and θ_t denote the deep network parameters for image and text respectively, and l denotes the length of the hash code. In addition, $cos\left(\cdot, \cdot\right)$ indicates the cosine similarity function, $\|\cdot\|_F$ denotes Frobenius norm of a matrix, $\|\cdot\|_2$ denotes the ℓ_2 regularization of the vector, $sign(\cdot)$ denotes the $sign$ activation function.

Methodology

Network Architecture

Deep neural networks are pivotal in extracting rich semantic information from raw multi-modal data for cross-modal retrieval. Following (Zhu et al. 2023), we also use AlexNet (Krizhevsky, Sutskever, and Hinton 2012) and Bag of Words (BoW) (Ko 2012) methods to extract raw image and text features $\mathbf{F}_v$ and $\mathbf{F}_t$. In addition, we introduce modality-specific autoencoders (Zhang and Wu 2022; Li et al. 2023) to generate modality-specific hash codes. This approach ensures that the hash codes retain as much feature information of the original modality as possible.

Given $\mathbf{F}_v$ and $\mathbf{F}_t$, they are imported into modality-specific autoencoders to obtain the latent semantic representations $\mathbf{L}_v$ and $\mathbf{L}_t$ of the image and text respectively. After that, we employ the $sign$ activation function to generate the corresponding binary hash codes. However, directly using $sign$ as the activation function in a deep network can lead to the gradient vanishing problem, hindering gradient backpropagation. To address this issue, we use the scaled $tanh$ function to generate the hash codes, where the relation between the $sign$ function and the scaled $tanh$ function can be expressed as:

$$\mathbf{B}_{v/t} = \lim_{\eta \to \infty} tanh(\eta \mathbf{L}_{v/t}) = sign(\mathbf{L}_{v/t}), \tag{1}$$

where $\eta > 0$ is a scaling parameter. We initially set $\eta = 1$ and increase it with the training process. As $\eta \to \infty$, the network will converge to the original hash coding problem.

Further, given $\mathbf{B}_v$ and $\mathbf{B}_t$, we reconstruct the original features of the image and text using a decoder to obtain $\mathbf{F}_v^{'}$ and $\mathbf{F}_t^{'}$. In order to minimize the error between the original modal features and the reconstructed features, as well as to make the hash codes of the different modal instances v_k and t_k of the image-text pairs as consistent as possible, we design the autoencoder reconstruction loss as follows:

$$\begin{aligned} L_{AU} = & \sum \left\| \mathbf{F}_v - \mathbf{F}_v^{'} \right\|_F^2 \\ & + \sum \left\| \mathbf{F}_t - \mathbf{F}_t^{'} \right\|_F^2 + \sum \left\| \mathbf{B}_v - \mathbf{B}_t \right\|_F^2 . \end{aligned} \tag{2}$$

Cross-modal Similarity Matrix Construction

Due to the simplicity and efficiency of the BoW, it is widely used for text feature extraction. However, the text feature vector $\mathbf{F}_i^t \in \{0, 1\}^{1 \times d_t}$ extracted using the BoW contains only two elements. When calculating the text similarity matrix $\mathbf{S}_t$, the asymmetry of the two elements $\mathbf{0}$ and $\mathbf{1}$ inevitably needs to be taken into account. Directly using cosine similarity to compute $\mathbf{S}_t$ does not fully consider the asymmetry of the textual feature elements and may focus too much on directional differences while ignoring the proportion of common elements. To address this, we introduce Jaccard similarity (Zahrotun 2016) in addition to cosine similarity to construct the text similarity matrix $\mathbf{S}_t$. This approach provides a more comprehensive representation of the similarity between textual features. The formula is as follows:

$$\begin{aligned} \mathbf{S}_t &= \zeta \cdot jaccard\left(\mathbf{F}_t, \mathbf{F}_t\right) + (1 - \zeta) \cdot cos\left(\mathbf{F}_t, \mathbf{F}_t\right) \\ &= \left\{ \zeta \cdot jaccard(\mathbf{F}_i^t, \mathbf{F}_j^t) + (1 - \zeta) \cdot cos(\mathbf{F}_i^t, \mathbf{F}_j^t) \right\}_{i,j=1}^{m} \\ &= \{ \zeta \frac{\mathbf{F}_i^t \mathbf{F}_j^{t^\top}}{\|\mathbf{F}_i^t\|_2^2 + \|\mathbf{F}_j^t\|_2^2 - \mathbf{F}_i^t \mathbf{F}_j^{t^\top}} + (1 - \zeta) \frac{\mathbf{F}_i^t \mathbf{F}_j^{t^\top}}{\|\mathbf{F}_i^t\|_2 \|\mathbf{F}_j^t\|_2} \}_{i,j=1}^{m}, \end{aligned} \tag{3}$$

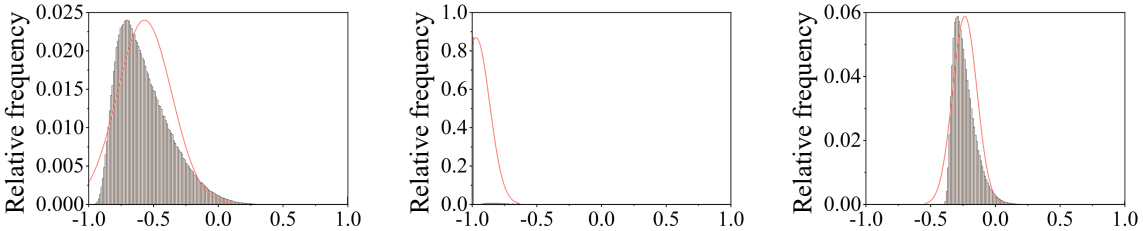
where $jaccard(\mathbf{F}_i^t, \mathbf{F}_j^t)$ and $cos(\mathbf{F}_i^t, \mathbf{F}_j^t)$ refer to the jaccard similarity and cosine similarity relations between text instances t_i and t_j respectively, and ζ is the trade-off parameter of the two types of similarity. By introducing the jaccard similarity, we more comprehensively take into account the asymmetry of text feature elements and the proportion of common elements in text feature vectors, which is more helpful in finding the similarity relation between text instances. Since image features contain rich semantic information, we only use cosine similarity to compute $\mathbf{S}_v$.

After that, we use $\mathbf{S}_v$ and $\mathbf{S}_t$ as follows to construct a cross-modal similarity matrix $\mathbf{S}_{cf}$ to capture the similarity information between multi-modal features. It aims to capture the neighborhood information from different modalities and consider the complementary information between them.

$$\begin{aligned} \mathbf{S}_{cf} &= \frac{1}{2} \left\{ \frac{(\mathbf{S}_{i,*}^t)(\mathbf{S}_{j,*}^v)^\top}{\|\mathbf{S}_{i,*}^t\|_2 \|\mathbf{S}_{j,*}^v\|_2} + \frac{(\mathbf{S}_{i,*}^v)(\mathbf{S}_{j,*}^t)^\top}{\|\mathbf{S}_{i,*}^v\|_2 \|\mathbf{S}_{j,*}^t\|_2} \right\}_{i,j=1}^{m} \\ &= \frac{1}{2} \{cos(\mathbf{S}_{i,*}^t, \mathbf{S}_{j,*}^v)\}_{i,j=1}^{m} + \frac{1}{2} \{cos(\mathbf{S}_{i,*}^v, \mathbf{S}_{j,*}^t)\}_{i,j=1}^{m}, \end{aligned} \tag{4}$$

where $(\mathbf{S}_{i,*}^t)(\mathbf{S}_{j,*}^v)^\top$ denotes the product between the i-th row in $\mathbf{S}_t$ and the j-th row in $\mathbf{S}_v$.

The matrix $\mathbf{S}_{cf}$ integrates the similarity information of different modalities into a single matrix, capturing the neighborhood relations between the modalities. To further enhance the semantic relations and complementarity, we fuse the matrices $\mathbf{S}_v$, $\mathbf{S}_t$, and $\mathbf{S}_{cf}$ into a unified similarity matrix.



(a) $\mathbf{S}_v$ (b) $\mathbf{S}_t$ (c) $\mathbf{S}$

Figure 2: Histogram distribution of NUS-WIDE.

The unified cross-modal similarity matrix $\mathbf{S}$ is as follows:

$$\begin{aligned} \mathbf{S} &= \alpha\mathbf{S}_v + \beta\mathbf{S}_t + \gamma\mathbf{S}_{cf} \\ s.t.\ &\alpha, \beta, \gamma \geq 0,\ \alpha + \beta + \gamma = 1, \end{aligned} \tag{5}$$

where α, β, γ are the balance parameters that regulate the respective independent information in the image and text as well as the fused information. The unified similarity matrix $\mathbf{S}$ not only preserves the co-occurrence information of multi-modalities, but also retains the respective unique neighborhood structure information, which is more effective than previous methods (Su and Zhang 2019; Yu et al. 2021; Mikriukov, Ravanbakhsh, and Demir 2022; Huang et al. 2024).

Statistical Model-driven Similarity Enhancement

Unsupervised hashing methods construct fusion matrices from original features, which may introduce noisy adjacency correlations and negatively impact retrieval performance. Therefore, in this module, we aim to improve the construction process of the unified similarity matrix $\mathbf{S}$ based on statistical distribution laws. This process aims to filter out noisy adjacent correlations between multi-modal instances, thus improving the quality of the similarity matrix.

Using the same sampling method as JDSH (Liu et al. 2020), we randomly select 5000 instance pairs from the NUS-WIDE, MIRFlickr, and MSCOCO datasets and analyze the statistical distributions of the samples in the matrices $\mathbf{S}_v$, $\mathbf{S}_t$ and $\mathbf{S}$. The histogram distributions of $\mathbf{S}_v$, $\mathbf{S}_t$, and $\mathbf{S}$ on NUS-WIDE are shown in Fig. 2, and the distributions for the other two datasets are similar. The left and right sides of $\mathbf{S}_v$ and $\mathbf{S}$ can be approximated as Gaussian distributions with differences. Given that the matrix $\mathbf{S}$ consists of $\mathbf{S}_v$ and $\mathbf{S}_t$, it can be surmised that the statistical distribution of $\mathbf{S}$ is mainly related to $\mathbf{S}_v$. This may be due to the fact that the text feature vector $\mathbf{F}_i^t$ contains more $\mathbf{0}$ elements, resulting in the similarity between two text instances in $\mathbf{S}_t$ mainly focusing on $-\mathbf{1}$, while the matrix $\mathbf{S}_v$ contains more semantically relevant information. Therefore, in order to improve the discriminative power of hash codes by augmenting the matrix $\mathbf{S}$, we focus on utilizing the statistical distribution law of $\mathbf{S}_v$ to filter the noisy adjacent correlations between multi-modal instances and regenerate $\mathbf{S}$.

Based on the definition of matrices $\mathbf{S}_v$, $\mathbf{S}_t$, $\mathbf{S}$ and the properties of the statistical distributions, we propose a similarity enhancement scheme called Statistical Model-driven Similarity Enhancement (SMSE) to adjust $\mathbf{S}_v$. According to the statistical distribution law of the matrix $\mathbf{S}_v$, we consider the use of Gaussian mixture distribution for fitting. In this paper, we adopt Gaussian Mixture Model (GMM) (Reynolds 2009) to find the statistical features corresponding to $\mathbf{S}_v$, such as mean and variance. For a sample x in the matrix $\mathbf{S}_v$, the marginal probability given the parameter θ is:

$$\begin{aligned} P(x \mid \theta) &= \sum\nolimits_{k=1}^{K} P\left(x, z = C_k \mid \theta\right) \\ &= \sum\nolimits_{k=1}^{K} P\left(x \mid \theta, z = C_k\right) \cdot P\left(z = C_k \mid \theta\right) \\ &= \sum\nolimits_{k=1}^{K} p_k \cdot \phi\left(x \mid \mu_k, \sigma_k\right) \\ &= \sum\nolimits_{k=1}^{K} p_k \cdot \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right) \\ s.t.\ &p_k \geq 0,\ \sum\nolimits_{k=1}^{K} p_k = 1, \end{aligned} \tag{6}$$

where $z = C_k$ indicates that the sample belongs to a certain class, and $P(z = C_k)$ is the probability distribution of the hidden variable. $P\left(x \mid \theta, z = C_k\right)$ and $\phi(\cdot)$ denote that the conditional probability distribution of the sample obeys a Gaussian distribution, *e.g.*, $(x \mid \theta, z = C_k) \sim N(\mu_k, \sigma_k)$. p_k denotes the weight of the k-th Gaussian distribution in the GMM. μ_k and σ_k denote the mean and the standard deviation of the k-th Gaussian distribution, respectively. In addition, K denotes the number of Gaussian distributions included in the GMM, and $K = 2$ in this paper.

To estimate the mean and standard deviation of the Gaussian mixture distribution in the GMM, we employ the Expectation-Maximization (EM) algorithm. The EM algorithm is an iterative optimization algorithm that aims to estimate the parameters of the GMM. The algorithm consists of two steps: the expectation step (E-step) and the maximization step (M-step). In the E-step, the posterior probability of each sample data belonging to each Gaussian distribution is calculated from the initial estimates of the known parameters. The mathematical formula is as follows:

$$\begin{aligned} \gamma_{ik} &= \frac{p_k \cdot \phi\left(x_i \mid \mu_k, \sigma_k\right)}{\sum_{k=1}^{K} p_k \cdot \phi\left(x_i \mid \mu_k, \sigma_k\right)} \\ i &= 1, 2, \cdots, N;\ k = 1, 2, \cdots, K, \end{aligned} \tag{7}$$

where γ_{ik} denotes the probability that the sample data x_i comes from the k-th Gaussian distribution. N denotes the number of samples in the matrix $\mathbf{S}_v$ and $N = m^2$.

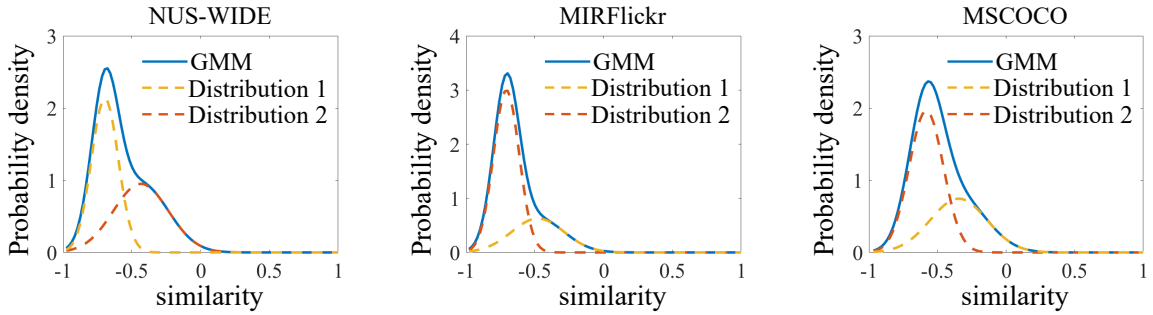


Figure 3: The fitted distribution of elements in $\mathbf{S}_v$.

In the M-step, the model parameters are updated using the posterior probability γ_{ik} from the E-step. The weight, mean, and standard deviation at time $t+1$ are shown below:

$$\begin{aligned} p_k^{(t+1)} &= \frac{1}{N}\sum_{i=1}^{N}\gamma_{ik}, \\ \mu_k^{(t+1)} &= \sum_{i=1}^{N}\left(x_i \cdot \gamma_{ik}\right) / \sum_{i=1}^{N}\gamma_{ik}, \\ \sigma_k^{(t+1)} &= \sqrt{\left(\sum_{i=1}^{N}\left(x_i-\mu_k^{(t+1)}\right)^2\gamma_{ik}\right) / \sum_{i=1}^{N}\gamma_{ik}}, \end{aligned} \tag{8}$$

where $p_k^{(t+1)}, \mu_k^{(t+1)}, \sigma_k^{(t+1)}$ denote the weight, mean and standard deviation of the k-th Gaussian distribution at moment $t+1$, respectively. We use the EM algorithm to iteratively update these parameters until they converge. We randomly select 5000 instances and fit the distribution of the image similarity matrix $\mathbf{S}_v$ using a Gaussian mixture distribution and calculate the parameters using the EM algorithm, the fitting results on three datasets are shown in Fig. 3. In order to verify the effectiveness of the GMM in fitting the similarity between image instances, we adopt the metric of Cross Entropy (CE) as a measure of the difference between the distribution fitted by GMM and the true distribution, and the mathematical formula is defined as follows:

$$CE_{value} = -\sum_{x} Q(x)\log\left(P(x)\right), \tag{9}$$

where x denotes the possible events, $Q(x)$ denotes the true probability density function obtained by using Histogram Method (HMM), which is considered as the target distribution. $P(x)$ denotes the probability density function obtained by using GMM fitting, which is considered as the predictive distribution. The smaller the value of cross entropy, the more similar the two probability distributions are.

Also, since the distribution of samples in the matrices $\mathbf{S}_v$ and $\mathbf{S}$ is right skewed. Except for the high similarity between the image instances themselves, the similarity between most of the instances is low. In order to avoid the effect of noise information on the less similar elements, we consider adjusting the less similar elements in the matrix $\mathbf{S}_v$. To adjust the similarity elements in $\mathbf{S}_v = \{s_{ij}^v\}_{i,j=1}^m$, we define μ_l and σ_l as the mean and standard deviation of the left part of the Gaussian distribution in the GMM of $\mathbf{S}_v$. Since the statistical distribution of the samples in $\mathbf{S}_v$ is a right-skewed distribution with the samples mainly concentrated on the left side, we set a similarity threshold $s_l = \mu_l - \omega\sigma_l$, where ω is a hyper-parameter. If the similarity s_{ij}^v between a pair of image instances is smaller than s_l, indicating a very dissimilar pair, we further decrease s_{ij}^v using a weighted function $W^+(\cdot)$ to increase their dissimilarity. We define $W^+(\cdot)$ for the similarity s_{ij}^v when $s_{ij}^v < s_l$. The mathematical definition is as follows:

$$W^+(x) = \frac{2}{1+e^{-\rho x}} - 1 \quad s.t.\ \rho \in N_+. \tag{10}$$

$W^+(\cdot)$ is the *sigmoid*-like function, that compresses the similarity values between dissimilar pairs to close to -1, which widens the gap between the similarity values of similar pairs and dissimilar pairs. We can use $W^+(\cdot)$ to update $\mathbf{S}_v$ with the following formula:

$$\tilde{s}_{ij}^v = \begin{cases} W^+(s_{ij}^v), & s_{ij}^v < s_l \\ s_{ij}^v, & others, \end{cases} \tag{11}$$

where $\tilde{s}_{ij}^v$ denotes the updated element of the matrix $\mathbf{S}_v$. After that, after obtaining the updated $\mathbf{S}_v$, we are able to use Eq. (5) to obtain the updated $\mathbf{S}$. With the help of the matrix $\mathbf{S}$, which efficiently captures the neighbourhood structure and common information of image and text instances, thus guiding the generation of high-quality hash codes.

Algorithm 1: Statistical Model-driven Similarity Hashing

Input: Training image-text pairs n; training set $\{v_k, t_k\}_{k=1}^m$; network parameters θ_v and θ_t; hyper-parameters $\zeta, \omega, \rho, \varphi_1, \varphi_2, \xi$; Estimated parameters of the statistical distribution μ_l, σ_l; number of epochs t;
Output: Deep hash networks $f_v(\mathbf{F}_v, \theta_v)$ and $f_t(\mathbf{F}_t, \theta_t)$; hash codes $\mathbf{B}_v$ and $\mathbf{B}_t$ for image and text modalities respectively;

```
Initialize epoch t = 0;
repeat
    t = t + 1; η = √t;
    for all i ∈ [1, ⌈n/m⌉] do
        Randomly select m training instance pairs;
        Obtain image feature F_v and text feature F_t;
        Calculate S_t in Eq. (3); S_v = cos(F_t, F_t);
        Calculate S in Eq. (5); Update S in Eq. (11);
        Hashing codes B_{v/t} with tanh function in Eq. (1);
        Calculate the objective function Eq. (13), and update network parameters by backpropagation;
    end for
until convergence;
```

Task	Methods	Reference	NUS-WIDE				MIRFlickr				MSCOCO			
			16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
$I \to T$	CVH	IJCAI11	0.372	0.363	0.404	0.390	0.606	0.599	0.596	0.589	0.505	0.509	0.519	0.510
	IMH	SIGMOD13	0.470	0.473	0.476	0.459	0.612	0.601	0.592	0.579	0.570	0.615	0.613	0.587
	CMFH	CVPR14	0.455	0.459	0.465	0.467	0.621	0.624	0.625	0.627	0.621	0.669	0.525	0.562
	DBRC	TMM18	0.424	0.459	0.447	0.447	0.617	0.619	0.620	0.621	0.567	0.591	0.617	0.627
	UDCMH	IJCAI18	0.511	0.519	0.524	0.558	0.689	0.698	0.714	0.717	—	—	—	—
	DJSRH	ICCV19	0.724	0.773	0.798	0.817	0.810	0.843	0.862	0.876	0.678	0.724	0.743	0.768
	JDSH	SIGIR20	0.736	0.793	0.832	0.835	0.832	0.853	0.882	0.892	0.694	0.738	0.769	0.788
	AGCH	TMM21	0.809	0.830	0.831	0.852	0.865	0.887	0.892	0.912	0.741	0.772	0.789	0.806
	DAEH	TCSVT22	0.788	0.824	0.833	0.849	0.890	0.906	0.918	0.920	0.735	0.770	0.791	0.809
	CIRH	TKDE23	0.815	0.836	0.854	0.862	0.901	0.913	0.929	0.937	0.797	0.819	0.830	0.849
	UCCH	TPAMI23	0.809	0.837	0.845	0.859	0.885	0.906	0.919	0.920	0.794	0.839	0.866	0.887
	HUGH	IJCV24	0.799	0.831	0.839	0.841	0.896	0.912	0.920	0.936	0.808	0.852	0.874	0.899
	SMSH	Ours	**0.816**	**0.840**	**0.858**	**0.865**	**0.904**	**0.919**	**0.932**	**0.942**	**0.821**	**0.873**	**0.905**	**0.916**
$T \to I$	CVH	IJCAI11	0.401	0.384	0.442	0.432	0.591	0.583	0.576	0.576	0.543	0.553	0.560	0.542
	IMH	SIGMOD13	0.478	0.483	0.472	0.462	0.603	0.595	0.589	0.580	0.641	0.709	0.705	0.652
	CMFH	TIP16	0.529	0.577	0.614	0.645	0.642	0.662	0.676	0.685	0.627	0.667	0.554	0.595
	DBRC	TMM18	0.455	0.459	0.468	0.473	0.618	0.622	0.626	0.628	0.635	0.671	0.697	0.735
	UDCMH	IJCAI18	0.637	0.653	0.695	0.716	0.692	0.704	0.718	0.733	—	—	—	—
	DJSRH	ICCV19	0.712	0.744	0.771	0.789	0.786	0.822	0.835	0.847	0.650	0.753	0.805	0.823
	JDSH	SIGIR20	0.721	0.785	0.794	0.804	0.825	0.864	0.878	0.880	0.703	0.759	0.793	0.825
	AGCH	TMM21	0.769	0.780	0.798	0.802	0.829	0.849	0.852	0.880	0.746	0.774	0.797	0.817
	DAEH	TCSVT22	0.759	0.800	0.808	0.811	0.842	0.866	0.883	0.891	0.741	0.780	0.811	0.832
	CIRH	TKDE23	0.774	0.803	0.810	0.817	0.867	0.885	0.900	0.901	0.811	0.847	0.872	0.895
	UCCH	TPAMI23	0.790	0.809	0.821	0.825	0.876	0.899	0.905	0.911	0.806	0.860	0.885	0.913
	HUGH	IJCV24	0.778	0.806	0.812	0.819	0.881	0.900	0.908	0.912	0.819	0.874	0.896	0.919
	SMSH	Ours	**0.794**	**0.813**	**0.830**	**0.837**	**0.890**	**0.908**	**0.914**	**0.917**	**0.831**	**0.894**	**0.915**	**0.931**

Table 1: The mAP@50 results at various code lengths and datasets and the best result in each column is marked in bold.

Hash Code Reconstruction and Generation

We learn hash codes by minimizing the reconstruction error between the unified similarity matrix $\mathbf{S}$ and the inter- and intra-modal cosine similarity matrices of the hash codes $\mathbf{B}_v$ and $\mathbf{B}_t$. The mathematical formulation of this is as follows:

$$L_C = \|\xi\mathbf{S} - cos(\mathbf{B}_v, \mathbf{B}_t)\|_F^2 + \varphi_1 \|\xi\mathbf{S} - cos(\mathbf{B}_v, \mathbf{B}_v)\|_F^2 + \varphi_2 \|\xi\mathbf{S} - cos(\mathbf{B}_t, \mathbf{B}_t)\|_F^2, \tag{12}$$

where φ_1 and φ_2 are trade-off parameters for hash code reconstruction, ξ is a scale parameter that makes the reconstruction more flexible.

Overall Objective Function and Optimization

By integrating Eq. (2) and Eq. (12) into a unified framework, the SMSH is optimized by minimising the loss function:

$$\min_{\mathbf{B}_v, \mathbf{B}_t} L = L_{AU} + L_C \quad s.t.\ \mathbf{B}_{v/t} \in \{-1, +1\}^{m \times l}, \tag{13}$$

where L_{AU}, L_C are the reconstruction losses of the autoencoder and hash code, respectively. We summarize the basic process of SMSH in Algorithm 1.

Experiments

Datasets

NUS-WIDE (Chua et al. 2009) consists of 269,648 image-text pairs in 81 concepts. Following (Zhu et al. 2023), we select 186,577 image-text pairs corresponding to the 10 most common concepts, 2,000 image-text pairs are randomly selected as the query, and the rest as a database retrieval set (containing 10,000 training pairs).

MIRFlickr (Huiskes and Lew 2008) consists of 25,000 image-text pairs containing 24 unique concepts. We remove those pairs with less than 20 labels and end up with 20,015 image-text pairs. For a fair comparison, we follow the experimental settings of (Zhu et al. 2023) to randomly select 2,000 pairs as the query set and the rest as the database retrieval set (containing 10,000 training pairs).

MSCOCO (Lin et al. 2014) contains 123,287 image-text pairs in 80 separate categories. We remove pairs without any label information and end up with 122,218 image-text pairs. Similar to the setup of the previous two datasets, we randomly selected 2,000 image-text pairs as the query and the rest as the database retrieval set (containing 10,000 training pairs).

Baseline and Evaluation Metrics

We compare the proposed method with 12 state-of-the-art unsupervised baselines: CVH (Kumar and Udupa 2011), IMH (Song et al. 2013), CMFH (Ding, Guo, and Zhou 2014), DBRC (Hu, Nie, and Li 2019), UDCMH (Wu et al. 2018), DJSRH, JDSH, AGCH, DAEH, CIRH, UCCH (Hu et al. 2023), HUGH. Among them, CVH, IMH and CMFH are shallow methods, while DBRC, UDCMH, DJSRH, JDSH, AGCH, DAEH, CIRH, UCCH and HUGH are deep methods.

We employ two widely-used standard metrics for retrieval evaluation: mean Average Precision (mAP) and precision@top-N curves (P@N) to evaluate the retrieval performance of all methods.

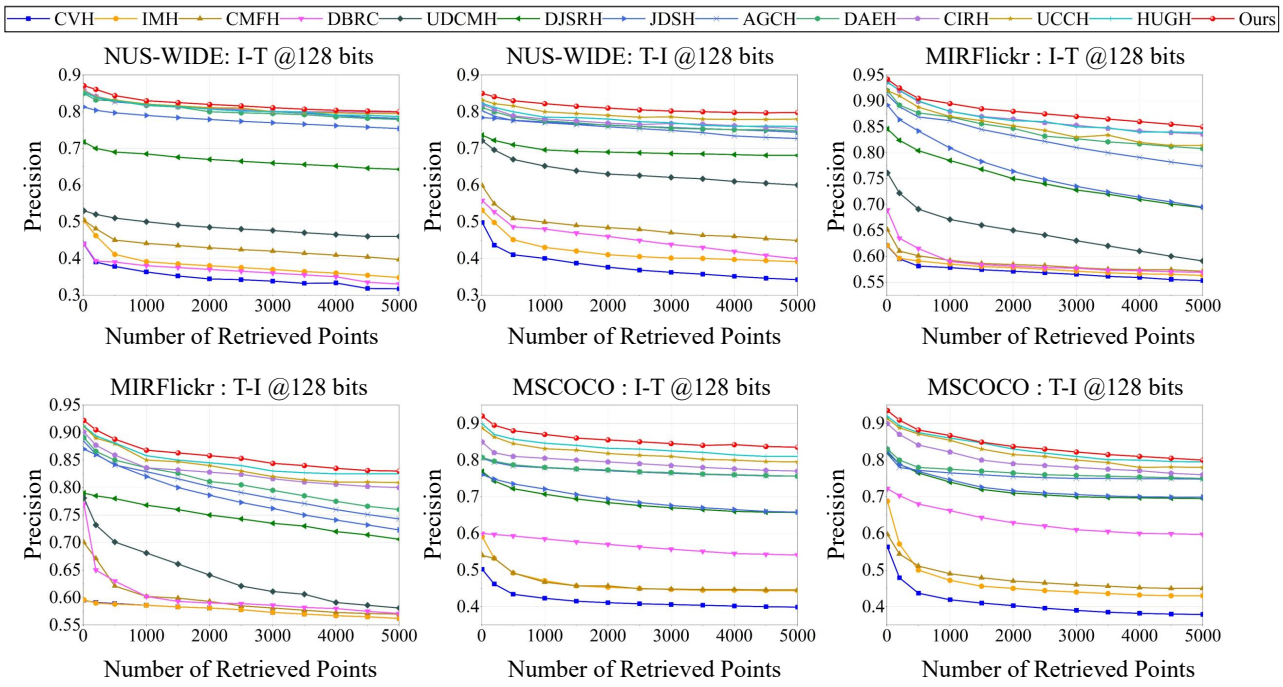


Figure 4: The P@N curves at 128 bits.

Implementation Details

As shown in Fig. 1, our model focuses on extracting features from images and text via AlexNet and BoW, respectively. The visual encoder is a multi-layer perceptron structure($d_v \rightarrow 4096 \rightarrow ReLU \rightarrow l$) and the decoder is also a multi-layer perceptron structure($l \rightarrow 4096 \rightarrow ReLU \rightarrow d_v$), where $ReLU$ is the activation function. The structure of the textual encoder and decoder is similar to the vision. We adopt the Adam optimization algorithm (Kingma and Ba 2015) with learning rates set to 0.0001 and 0.0001 on the three datasets. The mini-batch size m is 64.

The selection of hyper-parameters. For simplicity, we set $\varphi_1 = \varphi_2 = 3$ on three datasets. we cross-validate the hyper-parameters $\zeta, \alpha, \beta, \gamma, \xi$, and set $\zeta = 0.8, \alpha = 0.4, \beta = 0.2, \gamma = 0.4, \xi = 3$ for NUS-WIDE, $\zeta = 0.6, \alpha = 0.3, \beta = 0.2, \gamma = 0.5, \xi = 3$ for MIRFlikcr, and $\zeta = 0.6, \alpha = 0.3, \beta = 0.3, \gamma = 0.4, \xi = 1.5$ for MSCOCO. Meanwhile, we cross-validate the hyper-parameters ρ and ω, and set $\rho = 6, \omega = -2$ for NUS-WIDE, $\rho = 6, \omega = -0.5$ for MIRFlikcr, and $\rho = 4, \omega = -2$ for MSCOCO.

Results and Analysis

Table 1 shows the results of mAP@50 for SMSH, where two cross-modal retrieval tasks are compared on three datasets. We can derive the following analyses: (1) Our proposed method, SMSH, usually achieves the best retrieval accuracy for different hash code lengths (from 16 bits to 128 bits), especially on the MSCOCO. Taking the hash code length of 128 bits as an example, our method achieves retrieval accuracy that are 0.6% and 1.2% higher than the optimal deep learning baseline UCCH on the NUS-WIDE, 0.6% and 0.5% higher than HUGH on the MIRFlickr, and 1.7% and 1.2% higher than HUGH on MSCOCO. (2) Furthermore, the P@N curves shown in Fig. 4 demonstrate that our method consistently outperforms all the compared baselines. This observation aligns with the mAP@50 results and provides further evidence of the excellent performance of our method.

Statistical Indicator	NUS-WIDE		MIRFlickr		MSCOCO	
	left	right	left	right	left	right
Mean	-0.6959	-0.4340	-0.7079	-0.4821	-0.5819	-0.3460
Standard deviation	0.0980	0.2000	0.0924	0.1906	0.1249	0.2060
Weight	0.5242	0.4758	0.6920	0.3080	0.6151	0.3849
CE_{value}	0.26036		0.48452		0.25003	

Table 2: The statistical indicators in the fitted GMM.

Statistical Evaluation

We follow the setup of the previous section and randomly select 5000 instances to be fitted using the GMM, and the statistical indicators obtained are shown in Table 2. We find that the means of the two Gaussian distributions in the GMM are negative, which coincides with the fact that the histogram of the similarity distribution between image instances is a right-skewed distribution. Also in addition to visualisation, the model parameters themselves verify the feasibility of keeping dissimilar instances as far apart as possible. The small standard deviation indicates that the data distribution is tight, due to the fact that the similarity is restricted to the interval $[-1, 1]$. Meanwhile, we choose the CE_{value} to measure the degree of the model in fitting similarity, and find that CE_{value} is very small and close to $\mathbf{0}$. This shows that the fitted model is very close to the true dis-

Methods	NUS-WIDE		MIRFlickr		MSCOCO	
	$I \to T$	$T \to I$	$I \to T$	$T \to I$	$I \to T$	$T \to I$
SMSH-1@128	0.856	0.832	0.936	0.914	0.909	0.924
SMSH-2@128	0.855	0.829	0.939	0.914	0.909	0.928
SMSH-3@128	0.839	0.826	0.930	0.913	0.891	0.908
SMSH@128	**0.865**	**0.837**	**0.942**	**0.917**	**0.916**	**0.931**
SMSH-1@16	0.809	0.790	0.898	0.882	0.808	0.821
SMSH-2@16	0.808	0.789	0.899	0.885	0.809	0.828
SMSH-3@16	0.800	0.789	0.883	0.877	0.779	0.825
SMSH@16	**0.816**	**0.794**	**0.904**	**0.890**	**0.821**	**0.831**

Table 3: The mAP@50 of ablation experiments.

tribution, validating the reasonableness of our method.

Ablation Study

In this section, in order to demonstrate the effectiveness of each module in the proposed SMSH, three model variants are designed to evaluate the contribution of each proposed module. The variants of SMSH are designed as follows:

- **SMSH-1** abandons the autoencoder reconstruction module, *i.e.*, removes L_{AU} from the loss function Eq. (13).
- **SMSH-2** abandons the jaccard similarity and adopts the cosine similarity instead of Eq. (3) in calculating $\mathbf{S}_t$.
- **SMSH-3** abandons the SMSE method and no longer utilize Eq. (11) to update $\mathbf{S}$, *i.e.*, no longer use the similarity enhancement scheme to generate hash codes, and directly using Eq. (5) to generate $\mathbf{S}$.

Table 3 shows the results of the ablation study on three datasets. The results show that all three modules play an important role, especially when the SMSE method is not used, the performance of the SMSH-3 module decreases most significantly. Specific analyses are as follows:

(1) Compared with **SMSH-1**, SMSH shows that minimizing the error between original modal features and reconstructed features can significantly improve the retrieval performance, especially for low-bit hash codes.

(2) Compared with **SMSH-2**, SMSH shows that fully considering textual features can improve the performance. Effectively capturing the asymmetry of text vector elements helps to construct text similarity relations.

(3) Compared with **SMSH-3**, SMSH demonstrates a significant performance improvement of the SMSE method. This method generates more discriminative hash codes through a statistical model-driven scheme.

Parameter Sensitivity

In this section, we analyze the sensitivity of three hyper-parameters in SMSH: ζ, ρ, and ω. We test the sensitivity of ζ in the range 0 to 1 with a step size of 0.2, ρ in the range 1 to 10 with a step size of 2, and ω in the range -3 to 2 with a step size of 1. We evaluate the retrieval performance of these hyper-parameters and report the results in Fig. 5. Based on these experimental results, our observations are as follows:(1) The construction of the text similarity matrix $\mathbf{S}_t$ is insensitive to ζ. (2) ρ determines the effect of the

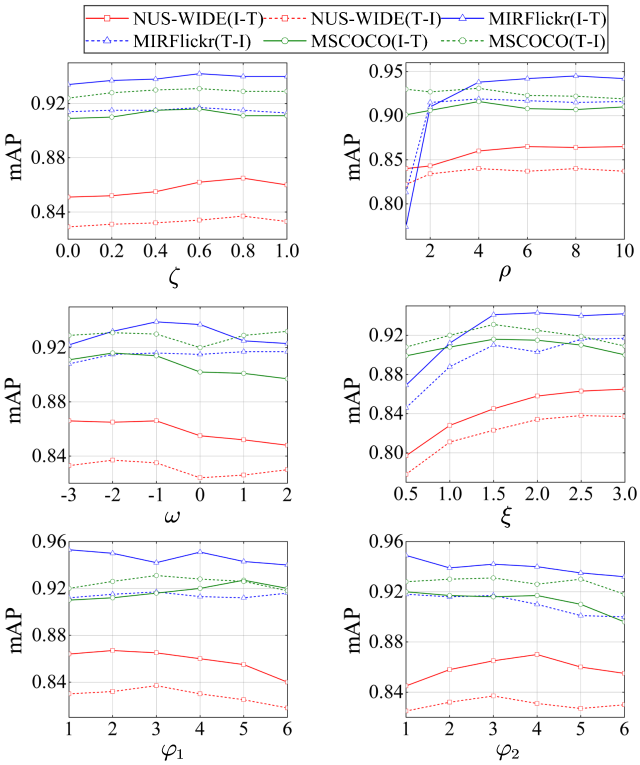


Figure 5: Parameter Sensitivity for parameters at 128 bits.

enhancement of the similarity relation, and the wide range of variation after starting from 2 is stable in our method, suggesting that the SMSE method is insensitive to ρ. (3) ω determines the range of similarity relation enhancement and is more related to the statistical distribution of the data. We find that the SMSE method is insensitive to ω, and therefore, our method can be well applied in practical applications.

At the same time, we also analyze the sensitivity of these three hyper-parameters of ξ, φ_1 and φ_2 in the Hash Code Reconstruction and Generation module. We evaluate the retrieval performance of these three hyper-parameters and the results are shown in Fig. 5. We find that our model is not sensitive to changes in ξ except when the parameter ξ takes on a small value. We find that SMSH is also not sensitive to φ_1 and φ_2, this shows the robustness of our method.

Conclusion

In this paper, we propose a Statistical Model-driven Similarity Hashing (SMSH) for large-scale cross-modal retrieval. SMSH takes into account both the asymmetry of elements in text features and the similarity relation between multi-modal instances to better bridge the gap between modalities. In order to improve the existing similarity enhancement schemes for multi-modal instances, a Statistical Model-driven Similarity Enhancement (SMSE) method is proposed based on the statistical distribution laws of $\mathbf{S}_v$, $\mathbf{S}_t$ and $\mathbf{S}$. It generates more discriminative hash codes by enhancing the matrix $\mathbf{S}_v$ to guide the reconstruction of $\mathbf{S}$. Extensive experiments demonstrate the excellent performance of the proposed SMSH method.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grant 62202501, in part by the Science and Technology Plan of Hunan Province under grant 2023GK2013 and in part by the National Key R&D Program of China under grant 2021YFB3900902.

References

Chen, Y.; Long, J.; Guo, L.; and Yang, Z. 2024. Supervised Semantic-Embedded Hashing for Multimedia Retrieval. *Knowl. Based Syst.*, 299: 112023.

Chua, T.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval*. ACM.

Ding, G.; Guo, Y.; and Zhou, J. 2014. Collective Matrix Factorization Hashing for Multimodal Data. In *CVPR*, 2083–2090. IEEE Computer Society.

Dong, G.; Zhang, X.; Shen, X.; Lan, L.; Luo, Z.; and Ying, X. 2023. Discriminative Geometric-Structure-Based Deep Hashing for Large-Scale Image Retrieval. *IEEE Trans. Cybern.*, 53(10): 6236–6247.

Han, L.; Li, P.; Plaza, A.; and Ren, P. 2022. Hashing for Localization (HfL): A Baseline for Fast Localizing Objects in a Large-Scale Scene. *IEEE Trans. Geosci. Remote. Sens.*, 60: 1–16.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. IEEE Computer Society.

Hu, D.; Nie, F.; and Li, X. 2019. Deep Binary Reconstruction for Cross-Modal Hashing. *IEEE Trans. Multim.*, 21(4): 973–985.

Hu, P.; Zhu, H.; Lin, J.; Peng, D.; Zhao, Y.; and Peng, X. 2023. Unsupervised Contrastive Cross-Modal Hashing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3): 3877–3889.

Huang, J.; Kang, P.; Han, N.; Chen, Y.; Fang, X.; Gao, H.; and Zhou, G. 2024. Two-Stage Asymmetric Similarity Preserving Hashing for Cross-Modal Retrieval. *IEEE Trans. Knowl. Data Eng.*, 36(1): 429–444.

Huiskes, M. J.; and Lew, M. S. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval*, 39–43. ACM.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations*.

Ko, Y. 2012. A study of term weighting schemes using class information for text classification. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval*, 1029–1030. ACM.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, 1106–1114.

Kumar, S.; and Udupa, R. 2011. Learning Hash Functions for Cross-View Similarity Search. In *IJCAI*, 1360–1365. IJCAI/AAAI.

Li, J.; Zheng, K.; Li, Z.; Gao, L.; and Jia, X. 2023. X-Shaped Interactive Autoencoders With Cross-Modality Mutual Learning for Unsupervised Hyperspectral Image Super-Resolution. *IEEE Trans. Geosci. Remote. Sens.*, 61: 1–17.

Li, Y.; Long, J.; Tu, Z.; and Yang, Z. 2024a. Fast unsupervised multi-modal hashing based on piecewise learning. *Knowl. Based Syst.*, 299: 112111.

Li, Z.; Yao, T.; Wang, L.; Li, Y.; and Wang, G. 2024b. Supervised Contrastive Discrete Hashing for cross-modal retrieval. *Knowl. Based Syst.*, 295: 111837.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference*, volume 8693 of *Lecture Notes in Computer Science*, 740–755. Springer.

Liu, S.; Qian, S.; Guan, Y.; Zhan, J.; and Ying, L. 2020. Joint-modal Distribution-based Similarity Hashing for Large-scale Unsupervised Deep Cross-modal Retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 1379–1388. ACM.

Luo, X.; Wang, H.; Wu, D.; Chen, C.; Deng, M.; Huang, J.; and Hua, X. 2023. A Survey on Deep Hashing Methods. *ACM Trans. Knowl. Discov. Data*, 17(1): 15:1–15:50.

Mikriukov, G.; Ravanbakhsh, M.; and Demir, B. 2022. Deep Unsupervised Contrastive Hashing for Large-Scale Cross-Modal Text-Image Retrieval in Remote Sensing. *CoRR*, abs/2201.08125.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 8748–8763. PMLR.

Reynolds, D. A. 2009. Gaussian Mixture Models. In *Encyclopedia of Biometrics*, 659–663. Springer US.

Shen, X.; Dong, G.; Zheng, Y.; Lan, L.; Tsang, I. W.; and Sun, Q. 2022. Deep Co-Image-Label Hashing for Multi-Label Image Retrieval. *IEEE Trans. Multim.*, 24: 1116–1126.

Shi, Y.; Zhao, Y.; Liu, X.; Zheng, F.; Ou, W.; You, X.; and Peng, Q. 2022. Deep Adaptively-Enhanced Hashing With Discriminative Similarity Guidance for Unsupervised Cross-Modal Retrieval. *IEEE Trans. Circuits Syst. Video Technol.*, 32(10): 7255–7268.

Song, J.; Yang, Y.; Yang, Y.; Huang, Z.; and Shen, H. T. 2013. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 785–796. ACM.

Su, S.; and Zhang, C. 2019. Deep Joint-Semantics Reconstructing Hashing for Large-Scale Unsupervised Cross-Modal Retrieval. In *2019 IEEE/CVF International Conference on Computer Vision*, 3027–3035. IEEE.

Sun, Y.; Dai, J.; Ren, Z.; Chen, Y.; Peng, D.; and Hu, P. 2024a. Dual Self-Paced Cross-Modal Hashing. In *Thirty-Eighth AAAI Conference on Artificial Intelligence*, 15184–15192. AAAI Press.

Sun, Y.; Liu, K.; Li, Y.; Ren, Z.; Dai, J.; and Peng, D. 2024b. Distribution Consistency Guided Hashing for Cross-Modal Retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5623–5632. ACM.

Sun, Y.; Ren, Z.; Hu, P.; Peng, D.; and Wang, X. 2024c. Hierarchical Consensus Hashing for Cross-Modal Retrieval. *IEEE Trans. Multim.*, 26: 824–836.

Wang, D.; Wang, Q.; He, L.; Gao, X.; and Tian, Y. 2020. Joint and individual matrix factorization hashing for large-scale cross-modal retrieval. *Pattern Recognit.*, 107: 107479.

Wang, J.; Zeng, Z.; Chen, B.; Wang, Y.; Liao, D.; Li, G.; Wang, Y.; and Xia, S. 2024. Hugs Bring Double Benefits: Unsupervised Cross-Modal Hashing with Multi-granularity Aligned Transformers. *Int. J. Comput. Vis.*, 132(5): 1–33.

Wang, Y.; Luo, X.; Nie, L.; Song, J.; Zhang, W.; and Xu, X. 2021. BATCH: A Scalable Asymmetric Discrete Cross-Modal Hashing. *IEEE Trans. Knowl. Data Eng.*, 33(11): 3507–3519.

Weiss, Y.; and Torralba, A. 2008. Spectral Hashing. In *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems*, 1753–1760. Curran Associates, Inc.

Wu, G.; Lin, Z.; Han, J.; Liu, L.; Ding, G.; Zhang, B.; and Shen, J. 2018. Unsupervised Deep Hashing via Binary Latent Factor Models for Large-scale Cross-modal Retrieval. In *IJCAI*, 2854–2860. ijcai.org.

Xiao, K.; Xingbo, L.; Wen, X.; Xiushan, N.; and Yilong, Y. 2024. Online Cross-modal Hashing With Dynamic Prototype. *ACM Transactions on Multimedia Computing Communications and Applications*, 20(8): 1–18.

Xu, L.; Zeng, X.; Zheng, B.; and Li, W. 2022. Multi-Manifold Deep Discriminative Cross-Modal Hashing for Medical Image Retrieval. *IEEE Trans. Image Process.*, 31: 3371–3385.

Yu, J.; Zhou, H.; Zhan, Y.; and Tao, D. 2021. Deep Graph-neighbor Coherence Preserving Network for Unsupervised Cross-modal Hashing. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 4626–4634. AAAI Press.

Zahrotun, L. 2016. Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method. *Computer Engineering and Applications*, 5: 11–18.

Zhang, D.; and Wu, X. 2022. Scalable Discrete Matrix Factorization and Semantic Autoencoder for Cross-Media Retrieval. *IEEE Trans. Cybern.*, 52(7): 5947–5960.

Zhang, D.; Wu, X.; Xu, T.; and Kittler, J. 2023. WATCH: Two-Stage Discrete Cross-Media Hashing. *IEEE Trans. Knowl. Data Eng.*, 35(6): 6461–6474.

Zhang, P.; Li, Y.; Huang, Z.; and Xu, X. 2022. Aggregation-Based Graph Convolutional Hashing for Unsupervised Cross-Modal Retrieval. *IEEE Trans. Multim.*, 24: 466–479.

Zhong, F.; Chu, C.; Zhu, Z.; and Chen, Z. 2023. Hypergraph-Enhanced Hashing for Unsupervised Cross-Modal Retrieval via Robust Similarity Guidance. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3517–3527. ACM.

Zhou, J.; Ding, G.; and Guo, Y. 2014. Latent semantic sparse hashing for cross-modal similarity search. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 415–424. ACM.

Zhu, L.; Wu, X.; Li, J.; Zhang, Z.; Guan, W.; and Shen, H. T. 2023. Work Together: Correlation-Identity Reconstruction Hashing for Unsupervised Cross-Modal Retrieval. *IEEE Trans. Knowl. Data Eng.*, 35(9): 8838–8851.