Expert Systems With Applications 247 (2024) 123333

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Individual mapping and asymmetric dual supervision for discrete cross-modal hashing

Song Wang[a], Huan Zhao[a,*], Zixing Zhang[a], Keqin Li[b]

[a] College of Computer Science and Electronic Engineering of Hunan University, and Key Laboratory for Embedded and Network Computing of Hunan Province, Changsha 410082, China
[b] Department of Computer Science, State University of New York, New Paltz NY, 12561, USA

ARTICLE INFO

Keywords:
Asymmetric dual supervision
Cross-modal hashing
Individual mapping
Optimization algorithm

ABSTRACT

Cross-modal hashing has gained popularity in similarity search due to its excellent query efficiency and economical storage costs. However, current models frequently overlook the distinctive property of each modality, resulting in reduced accuracy due to inadequate utilization of these attributes. Moreover, there is a weak semantic relevance between modality attributes and multiple supervision knowledge (the labels and similarity constraints constructed by labels), accompanied by a cumulative quantization of the models. To address these issues, we propose an Individual Mapping and Asymmetric Dual Supervision method (IMADS). It merges specific and shared information to effectively learn a cross-modal representation space. Furthermore, we present an asymmetric dual supervision learning framework to produce discriminative hash codes. This framework achieves two primary goals: (1) Combing cross-modal representation and multiple supervision information to enhance the consistent relation of distinct modalities, and (2) developing a discrete optimization algorithm to mitigate the information loss caused by the hash code. Comprehensive experimental results illustrate that the introduced IMADS outperforms other stat-of-the-art hashing methods.

1. Introduction

Cross-modal search, a basic yet widely studied topic, aims to explore the semantic correlation between distinct modalities (Li et al., 2022; Seyed, Mohammad, & Müller, 2023; Shen, Sun, Wei, Hu, & Chen, 2022; Wang, Wang, Xu, Cao, & Cai, 2022; Yang, Yao, Liu, & Deng, 2022). Nowadays, as the amount of data increases dramatically on social platforms, cross-modal hashing (CMH) search approaches that convert high-dimensional real features into low-dimensional binary codes, are attracting considerable attention because of their rapid query efficiency and cost-saving storage capability (Qin, Xian, et al., 2022; Tran, Wang, Chen, & Xiao, 2021; Wang, Zhao, & Li, 2022). While CMH is applicable to any combination between modalities, this paper particularly concentrates bidirectional search tasks between images and texts. It intends to bridge the gap between the computer vision and natural language processing communities. Therefore, it is quite significant to conduct an in-depth study over CMH techniques.

Generally, pioneering CMH methods can be broadly grouped into classical unsupervised and supervised categories based on whether labels are utilized. The former (Cheng, Jing, & Ng, 2020; Ding, Guo, & Zhou, 2014; Fang, Jiang, Han, Teng, Zhou et al., 2022; Wang, Gao, Wang, & He, 2015; Wang, Wang, He, Gao, & Tian, 2020; Yao et al., 2023; Zhang, Luo, Huang, Xu, & Song, 2021) conducts the search task by only considering the topological structure of the original data. By contrast, the latter (Chen et al., 2019; Liu, Wang, & Cheung, 2022; Wang & Peng, 2022; Wang, Zareapoor, Yang, & Zheng, 2022; Zhang, Li, Gao, & Chen, 2023; Zhang & Wu, 2022a, 2022b) utilizing the label supervision evidently improves the accuracy of the hashing models in comparison to the unsupervised one. Therefore, leveraging the supervision knowledge proves advantageous for the majority of supervised CMH methods in designing effective hashing models.

Despite the significant breakthroughs, most current supervised CMH search methods still encounter the following challenges. (1) Existing models cannot well address the individual information of each modality. For example, many approaches (Liu, Ji, Wu, & Hua, 2016; Shen et al., 2020; Wang, Gao, Wang, & He, 2019) perform the search tasks by learning the shared feature information of input modalities without considering the inner specific property of each modality. Subsequently, although several supervised CMH methods (Chen, Zhang, Tian, Wang, Zhang et al., 2022; Wang, Zhao, & Nai, 2021b; Wang, Zhao, Wang, Huang, & Li, 2022; Zhang & Wu, 2022b; Zhang, Wu, & Yu, 2021) endeavour to introduce the individual attribute, they give up the original shared attributes. Thus, neither of these can fully

* Corresponding author.
E-mail addresses: swang17@hnu.edu.cn (S. Wang), hzhao@hnu.edu.cn (H. Zhao), zixingzhang@hnu.edu.cn (Z. Zhang), lik@newpaltz.edu (K. Li).

https://doi.org/10.1016/j.eswa.2024.123333
Received 14 August 2023; Received in revised form 10 November 2023; Accepted 23 January 2024
Available online 26 January 2024

explore the specific and shared properties to construct good learning paradigms, which weaken the accuracy of the cutting-edge models. (2) Most state-of-the-art methods cannot correlate feature information and multiple supervision knowledge well. For example, conventional way (Qin, Fei, et al., 2022; Shen et al., 2020; Wang et al., 2019; Zhang, Wu, & Yu, 2021) to accomplish the search tasks is to connect simple shared supervision and common feature representation. However, this strategy neglects the importance of other supervised knowledge and individual feature information. To address this, several models (Fang & Ren, 2020; Ma, Liang, He, & Kong, 2017; Mandal, Chaudhury, & Biswas, 2019) incorporate the common representation and symmetric pairwise semantic similarity matrix builded by the labels to conduct the search tasks. Furthermore, some methods (Chen et al., 2019, 2022; Lin, Ding, Han, & Wang, 2016; Qin, Fei, et al., 2022) have made attempts to merge the common representation and asymmetric multiple supervision knowledge to achieve the hashing framework. Despite promising results, the aforementioned three paradigms have always been unable to construct a good correlation between common representation, individual representation, supervised label, and linear pairwise semantic similarity matrix. Meanwhile, such models often encounter the cumulative quantization of the learned hash code because of the simple treatment over discrete constraints when training optimization.

To cope with above problems, a novel supervised CMH dubbed Individual Mapping and Asymmetric Dual Supervision (IMADS) is proposed. For one thing, this method sufficiently employs the distinctive characteristics of each modality as well as the common property between modalities to acquire the beneficial cross-modal feature representation, going beyond traditional simple shared attribute. For another thing, the IMADS integrates the shared supervise labels, linear asymmetric semantic similarity matrix and the cross-modal feature representation while well reducing the quantization loss by discrete optimization algorithm. The primary advantages of this study are outlined as follows.

- Different from most current supervised methods, we simultaneously capture the specific features of each modality and the shared feature between modalities to maximize the cross-modal representation of the input instance, which well produces discriminative hash code for efficient search.
- IMADS leverages an asymmetric dual supervision learning framework to yield effective hash functions by the labels, linear pairwise semantic similarity, and the specific and shared feature representations while developed discrete optimization algorithm can alleviate cumulative quantization of this framework.
- Abundant experiments managed on three standard datasets substantiate the superiority of our IMADS against several competitive hashing methods and the effectiveness of the proposed learning modules.

The remainder of this study is organized as follows. The related work of supervised hashing models and the framework construction of the IMADS are respectively illustrated in Section 2 and Section 3. Section 4 unveils comparison experiments while the conclusion part is in Section 5.

2. Related work

We roughly survey the literature work of supervised cross-modal hashing methods, namely, common or individual hashing, symmetric or asymmetric hashing.

2.1. Common or individual hashing

Common supervised cross-modal hashing methods utilize supervision knowledge and the shared feature descriptor between modalities to create the hash codes for search tasks. For example, SMFH (Liu et al., 2016) highlights the label-similarity supervision and shared feature of the original instances to build the hashing model. LCMFH (Wang et al., 2019) considers the labels and common feature to directly generate the hash functions and hash codes. SCRATCH (Chen, Li, Luo, Nie, Zhang et al., 2020) adopts the collective matrix factorization technique and the semantic supervision label to discretely obtain the latent hash representations while proposing three learning models including two shallow SCRATCH-o, SCRATCH-t algorithms and a deep SCRATCH-d algorithm. Thereafter, SRLCH (Shen et al., 2020) and SDMSA (Zhang & Wu, 2022c) both devise the learning mapping of labels to the hash codes to design the hashing framework. Differently, individual supervised cross-modal hashing methods exploit the supervision knowledge and the unique geometric distributions to obtain the hash codes. FS-CMFH (Liu, Li, Du, Peng, & Fan, 2018) leverages supervision, the individual information, and the matching relation between individual features to achieve the hashing model. MSLF (Wang et al., 2021b) and LFMH (Zhang, Wu, & Yu, 2021) widely utilize the specifics of each modality and the matching relationship constraint between distinct modalities to produce the hash code so that the obtained code well degrades the quantization loss of the models. EDCAH (Wang, Zhao, Wang, et al., 2022) makes use of the common and unique feature distributions of image–text pairs to construct the efficient unified hashing paradigm. SCLCH (Qin, Fei, et al., 2022) first formulates different feature representations, and then uses the relationship between label-similarity mapping, unique features and hash codes to constitute the hashing paradigm.

2.2. Symmetric or asymmetric hashing

Symmetric supervised cross-modal hashing methods leverage the inner product from two identical matrices to deliver the hash codes during searching. For example, DCMH (Ma et al., 2017) designs the labels and pairwise similarity matrix to gain the unique hash code of each modality. GPSH (Mandal et al., 2019) explores the semantic correlation between data items to develop a generalized hashing framework, making it suitable many applications. SDCH-KDA (Fang & Ren, 2020) integrates multiple supervision information and kernel discriminant to achieve cross-modal hashing retrieval. Despite satisfactory performance, symmetric hashing methods such as DCMH, GPSH, and SDCH-KDA have high complexity in training the model, and meanwhile it is difficult to directly optimize the matrix variables of hashing model. To tackle this, some asymmetric hashing works have been proposed and shown its better search results than symmetric hashing (Da et al., 2017; Meng, Wang, Yu, Chen, & Wu, 2021).

Unlike symmetric hashing, asymmetric supervised cross-modal hashing methods utilize asymmetric inner product paradigm to constitute the hashing model. For example, MTFH (Liu, Hu, Ling, & Cheung, 2021) creates a generalized and efficient cross-modal hashing framework by supervised Matrix Tri-Factorization with varying hash lengths to the representation scalability of original paired or unpaired multimodal data, which can be highly applied to many challenging search scenarios. BATCH (Wang et al., 2021) employs the semantic-level information via distance-distance difference optimization, and asymmetric similarity-preserving scheme to build a two-step hashing model. ZS-CMR (Wang, Wang, et al., 2022) achieves the tasks of cross-modal search by efficiently combining the adequate instance-level semantic information and Zero-Shot learning, to address the ignorance problem of intra-modal variance. ADCH (Luo, Zhang, Wu, Chen, Huang et al., 2018) adopts the strategy of substituting a hash code with a real matrix variable to yield discrete hash code, and then also well circumvents quantization errors of the encoding form. Ulteriorly, DJSAH (Wang, Zhao, & Li, 2022), FDCH (Yao et al., 2020), and FADCH (Teng, Ning, Zhang, Wu, & Zeng, 2022) combine the same replacement strategy with multiple supervision knowledge to form a unified hashing framework. TECH (Chen et al., 2019) and EDMH (Chen et al., 2022) individually get binary hash code and hash functions by debasing the model complexity in the pairwise similarity matrix from square terms to linear terms. Despite considerable success, such methods still possess ample room for further upgrading.

3. Proposed approach

This paper takes the boldface uppercase letters to describe the matrices (a.k.a. $\mathbf{M}$) employed in this study. Accordingly, $\mathbf{M}_i$ and $\mathbf{M}_{ij}$ are the ith column matrix and the matrix variable at row i and column j. $\mathbf{X}^{\mathrm{T}}$ is the transpose of matrix $\mathbf{X}$. $\|\cdot\|_F$, $tr(\cdot)$ separately denote Frobenius norm and trace matrix. $\phi(\cdot)$, $exp(\cdot)$, and $sgn(\cdot)$ individually depict the kernel function, the exponential function, and the element-wise sign function.

Assume that a training set $\mathcal{O} = \{o_i\}_{i=1}^n$ produces the image and text data matrices, which $o_i = \{x_1^i, x_2^i, l_i\}$. Thereinto, $x_1^i \in \mathbb{R}^{d_1}$, $x_2^i \in \mathbb{R}^{d_2}$ and $l_i \in \mathbb{R}^n$ express the feature vectors of image, text and shared label data, respectively (generally $d_1 \neq d_2$). Set $\mathbf{X}_1 = \{x_1^i\}_{i=1}^n \in \mathbb{R}^{d_1 \times n}$, $\mathbf{X}_2 = \{x_2^i\}_{i=1}^n \in \mathbb{R}^{d_2 \times n}$ respectively mean image and text feature matrices. To well excavate the high-dimensional nonlinear feature information of cross-modal instances, we introduce Gaussian kernel function to conduct the original image and text features, that is, nucleating $\mathbf{X}_1$ and $\mathbf{X}_2$ to $\phi(\mathbf{X}_1)$ and $\phi(\mathbf{X}_2)$ as the input of our training model, which $\phi\left(\mathbf{X}_1\right) = \{\phi\left(x_1^i\right)\}_{i=1}^n$, $\phi\left(\mathbf{X}_2\right) = \{\phi\left(x_2^i\right)\}_{i=1}^n$. Thereinto, the kernelization features are calculated via

$$\begin{cases} \phi_i\left(x_1\right) = \exp\left(-z\left\|x_1^i - b_1^i\right\|_2^2\right), \\ \phi_i\left(x_2\right) = \exp\left(-z\left\|x_2^i - b_2^i\right\|_2^2\right), \end{cases} \tag{1}$$

where $\left\{x_m^j\right\}_{j=1}^t$ denotes t random anchor points, $z = 1/2\sigma^2$, and the kernel width $\sigma = 1/nt\sum_{i=1}^n \sum_{j=1}^t \left\|x_i - b_j\right\|_2$.

In this work, we utilize two supervision knowledge, namely, the label matrix $\mathbf{L} \in \mathbb{R}^{c\times n}$ and pairwise semantic similarity matrix $\mathbf{S}$, to achieve the cross-modal search tasks. The matrix $\mathbf{L}$ is easily obtained from the original datasets. For matrix $\mathbf{S}$, most methods define that two different samples are similar when $\mathbf{S}_{ij} = 1$; otherwise, $\mathbf{S}_{ij} = -1$. To reduce the computational complexity and well correlate the relation between labels, IMADS gets the formulation by cosine similarity:

$$\tilde{\mathbf{S}}_{ij} = \tilde{\mathbf{L}}_{ki}^{\mathrm{T}}\tilde{\mathbf{L}}_{kj}, \tag{2}$$

where $\tilde{\mathbf{L}}_{ki} = \mathbf{L}_{ki}/\|\mathbf{L}_{ki}\|_2$. We employ matrix $\tilde{\mathbf{L}}$ to save the label characteristics and subsequent the semantic similarity $\tilde{\mathbf{S}} = \tilde{\mathbf{L}}^{\mathrm{T}}\tilde{\mathbf{L}}$. As shown in Eq. (2), the $\tilde{\mathbf{S}}$ value belongs to $[0, 1]$. To ensure that the value falls between -1 and 1, defining the pairwise semantic similarity $\mathbf{S}$ by

$$\mathbf{S} = 2\tilde{\mathbf{S}} - \mathbf{E} = 2\tilde{\mathbf{L}}^{\mathrm{T}}\tilde{\mathbf{L}} - \mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}, \tag{3}$$

where $\mathbf{1}_n \in \{1\}^n$ express the all-one matrix. Benefiting from Eq. (3), there is no need to directly calculate the matrix $\mathbf{S}$ during optimization. Accordingly, the computational cost of the IMADS is transformed from $O(n^2)$ to $O(n)$ while also fully leveraging dual supervision knowledge.

Given the length of binary code l, this work aims to learn the unified hash code $\mathbf{B} \in \{-1,1\}^{l\times n}$ and the hash functions $f_1(\cdot)$, $f_2(\cdot)$ among similarity search tasks. As shown in Fig. 1, the proposed IMADS includes two components: Individual Mapping Learning and Asymmetric Dual Supervision Learning. The following subsections explain the motivational formulations of two learning parts in detail.

3.1. Individual mapping learning

The purpose of individual mapping learning is to yield the optimal common feature representations by considering the individual property of each modality in our hashing model. Collective matrix factorization technique has been proved that it delivers good performance in unsupervised methods. Up to now, many supervised CMH methods (Shen et al., 2020; Wang et al., 2019; Wang, Zhao, Wang, et al., 2022; Zhang & Wu, 2022c) usually adopt collective matrix factorization theory to produce the shared feature representation of image–text pairs. That is to say, $\mathbf{X}_m \approx \mathbf{U}_m\mathbf{V}, m \in \{1, 2\}$, where $\mathbf{X}_m$ denotes cross-modal instances, $\mathbf{U}_m$ is the projection matrix, and $\mathbf{V}$ is the shared feature representation. However, this basic way ignores the specific geometric information from each modality, resulting in the suboptimal accuracy of models. Thus, we design a new learning constraint to better correlate the individual and shared feature representations of the cross-modal data in hashing model by:

$$\min_{\mathbf{U}_m,\mathbf{V}_m,\mathbf{W}_m,\mathbf{V}} \sum_{m=1}^{2} \lambda_m \left\|\phi\left(\mathbf{X}_m\right) - \mathbf{U}_m\mathbf{V}_m\right\|_F^2 \quad + \lambda_3 \sum_{m=1}^{2} \|\mathbf{V} - \mathbf{W}_m\mathbf{V}_m\|_F^2, \tag{4}$$

where $\mathbf{V}_m$ is the individual feature, $\mathbf{U}_m$, $\mathbf{W}_m$ are the projection matrices. To employ the available label knowledge, we build an equivalent relationship between labels and shared feature descriptor as a constraint in Eq. (4), that is $\mathbf{V} = \mathbf{UL}$, in which $\mathbf{U}$ is the projection matrix.

Additionally, we consider to well connect the image and text feature descriptors. Thus, a correlation matrix $\mathbf{Q} \in \mathbb{R}^{l\times l}$ is devised to achieve this goal:

$$\min_{\mathbf{V}_1,\mathbf{V}_2,\mathbf{Q}} \lambda_4 \|\mathbf{V}_2 - \mathbf{Q}\mathbf{V}_1\|_F^2 . \tag{5}$$

Combining Eqs. (4) and (5) by a linear way, we constitute the formulation of individual mapping learning part:

$$\begin{aligned} J_1 = &\min_{\mathbf{U}_m,\mathbf{V}_m,\mathbf{W}_m,\mathbf{U},\mathbf{Q}} \sum_{m=1}^{2} \lambda_m \left\|\phi\left(\mathbf{X}_m\right) - \mathbf{U}_m\mathbf{V}_m\right\|_F^2 + \lambda_3 \sum_{m=1}^{2} \|\mathbf{UL} - \mathbf{W}_m\mathbf{V}_m\|_F^2 \\ &+ \lambda_4 \|\mathbf{V}_2 - \mathbf{Q}\mathbf{V}_1\|_F^2 + \lambda_5 \mathrm{Re}\left(\mathbf{U}_m, \mathbf{V}_m, \mathbf{W}_m, \mathbf{UL}, \mathbf{Q}\right), \end{aligned} \tag{6}$$

where $\mathrm{Re}(\cdot) = \|\cdot\|_F^2$ for avoiding over fitting of Eq. (6).

3.2. Asymmetric dual supervision learning

Asymmetric dual supervision learning aims to produce the discriminative hash code and effective hash functions while improving the semantic correlation the optimal shared feature and multiple supervision. After obtaining the optimal shared feature $\mathbf{V}$ by Eq. (6), most methods (Ding et al., 2014; Wang et al., 2019, 2020; Wang, Zhao, Wang, et al., 2022; Yao et al., 2023) can generate the hash code $\mathbf{B}$ or hash functions $\sum_{m=1}^2 \mathbf{P}_m$. They directly leverage the label $\mathbf{L}$ and transformed linear pairwise semantic similarity formulation $\|l\mathbf{S} - \mathbf{B}^{\mathrm{T}}\mathbf{V}\|_F^2 + \|\mathbf{B} - \mathbf{V}\|_F^2$ to guide the learning of the hash codes. However, this learning scheme suffers from two problems. (1) There is weak connection between the optimal representation $\mathbf{V}$ which enriches the individual and common features, and multiple supervision $\mathbf{L}$ and $\mathbf{S}$. Although possessing some semantic relevance, many studies cannot deliver the high-quality discriminative cross-modal representation and then obtain effective hash codes. (2) Many hashing models learn the hash code $\mathbf{B}$ by minimizing $\min_{\mathbf{B}\in\{-1,1\}^{l\times n}} \alpha \|\mathbf{B}-\mathbf{V}\|_F^2 + \beta \left\|l\mathbf{S}-\mathbf{B}^{\mathrm{T}}\mathbf{V}\right\|_F^2$. For simple optimization, they usually discard the constraint term related to the discrete hash code $\mathbf{B}$ or transform $\mathbf{B} \in \{-1, 1\}^{l\times n}$ as $\mathbf{B} \in [-1, 1]^{l\times n}$. This simple scheme may bring in the available information loss and consequent hash quantization error.

Motivated by these, previous works provide the possibility of continuing to study the supervision knowledge in depth. In specific, we construct the constraint relation between the optimal cross-modal representation contained in common and individual features, multiple supervision $\mathbf{L}$ and $\mathbf{S}$. We first thought of dimensionality reduction while avoiding traditional optimization strategies. Then, we convert symmetric form to asymmetric form by replacing one of the hash codes with an auxiliary matrix. Finally, to better correlate the dual supervision knowledge and learned cross-modal feature representation, we design:

$$\min_{\mathbf{B}\in\{-1,1\}^{l\times n}} \alpha \|\mathbf{B}-\mathbf{UL}\|_F^2 + \beta \left\|l\mathbf{S}-\mathbf{B}^{\mathrm{T}}\mathbf{UL}\right\|_F^2 . \tag{7}$$

Meanwhile, the hash functions for querying can be easily computed by the above obtained hash code:

$$\min_{\mathbf{P}_m} \sum_{m=1}^{2} \mu_m \left\|\mathbf{P}_m\phi\left(\mathbf{X}_m\right) - \mathbf{B}\right\|_F^2 + \gamma \|\mathbf{P}_m\|_F^2 . \tag{8}$$

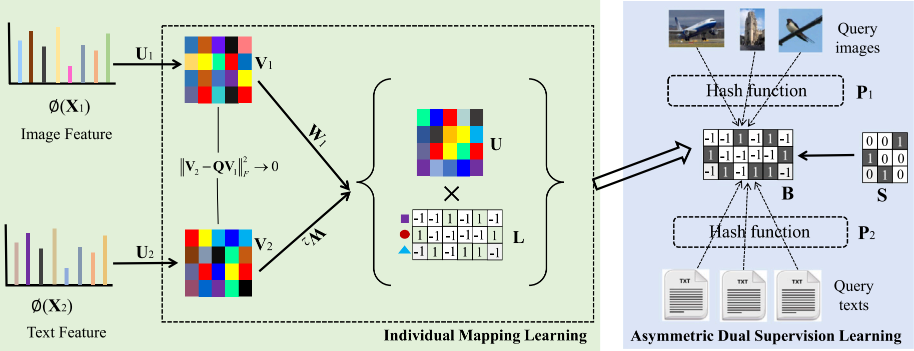

Fig. 1. The primary pipeline of the proposed IMADS. It contains Individual Mapping Learning for mutually obtaining optimal feature representation from image–text instances and Asymmetric Dual Supervision Learning to deliver the effective hash code and hash functions. Meanwhile, this pipeline takes both image and text feature matrices as model input, and generates the hash code and hash functions as model output.

Integrating Eqs. (7) and (8) by a linear way, we define the related form of asymmetric dual supervision learning part:

$$\begin{aligned} J_2 = &\min_{\mathbf{P}_m,\mathbf{B}\in\{-1,1\}^{l\times n}} \alpha\,\|\mathbf{B}-\mathbf{U}\mathbf{L}\|_F^2 + \beta\left\|l\mathbf{S}-\mathbf{B}^{\mathrm{T}}\mathbf{U}\mathbf{L}\right\|_F^2 \\ &+\sum_{m=1}^{2}\mu_m\left\|\mathbf{P}_m\phi\left(\mathbf{X}_m\right)-\mathbf{B}\right\|_F^2+\gamma\,\|\mathbf{P}_m\|_F^2 . \end{aligned} \tag{9}$$

To reduce the cumulative quantization of the asymmetric dual supervision learning during training optimization, we introduce two auxiliary matrices **C** and **D** to measure the difference with **B** and further discretely obtain the closed-form hash code matrix **B**.

To summarize, the final objective function of our IMADS includes two components: individual mapping learning for Eq. (6) and asymmetric dual supervision learning for Eq. (9). Thereinto, the former aims to obtain the optimal common feature descriptor (a.k.a. **V**) via combining the labels, individual and shared feature descriptors. The intention of latter component is to produce the hash code **B** and hash functions $\mathbf{P}_m$ by the generated matrix **V** and multiple supervision knowledge.

3.3. Optimization algorithm

We aim to seek the solution of the matrix variables in Eqs. (6) and (9). Evidently, solving Eq. (6) is a simple optimization problem without any discrete condition. However, directly optimization Eq. (9) is sophisticated because of the discrete constraints attached to the hash code. Thus, we develop a novel optimization algorithm to solve Eq. (9) instead of traditional relaxation scheme.

1. $\mathbf{U}_m$-*Subproblem*. By fixing $\mathbf{V}_1$, $\mathbf{V}_1$, $\mathbf{W}_m$, $\mathbf{V}$ and setting $\partial J_1/\partial \mathbf{U}_m = 0$, the solution of $\mathbf{U}_m$:

$$\mathbf{U}_m = \lambda_m\phi\left(\mathbf{X}_m\right)\mathbf{V}_m{}^{\mathrm{T}}\left(\lambda_m\mathbf{V}_m\mathbf{V}_m{}^{\mathrm{T}}+\lambda_5\mathbf{I}\right)^{-1}, \tag{10}$$

where **I** is defined as identity matrix.

2. $\mathbf{V}_1$-*Subproblem*. Taking $\mathbf{V}_1$ as only argument and letting $\partial J_1/\partial \mathbf{V}_1 = 0$, we get:

$$\begin{aligned} \mathbf{V}_1 = &\left[\lambda_1\mathbf{U}_1{}^{\mathrm{T}}\mathbf{U}_1+\lambda_3\mathbf{W}_1{}^{\mathrm{T}}\mathbf{W}_1+\lambda_4\mathbf{Q}^{\mathrm{T}}\mathbf{Q}+\lambda_5\mathbf{I}\right]^{-1} \\ &\times\left[\lambda_1\mathbf{U}_1{}^{\mathrm{T}}\phi\left(\mathbf{X}_1\right)+\lambda_3\mathbf{W}_1{}^{\mathrm{T}}\mathbf{U}\mathbf{L}+\lambda_4\mathbf{Q}^{\mathrm{T}}\mathbf{V}_2\right]. \end{aligned} \tag{11}$$

3. $\mathbf{V}_2$-*Subproblem*. As $\mathbf{V}_1$-*Subproblem* does, we obtain

$$\begin{aligned} \mathbf{V}_2 = &\left[\lambda_2\mathbf{U}_2{}^{\mathrm{T}}\mathbf{U}_2+\lambda_3\mathbf{W}_2{}^{\mathrm{T}}\mathbf{W}_2+\left(\lambda_4+\lambda_5\right)\mathbf{I}\right]^{-1} \\ &\times\left[\lambda_2\mathbf{U}_2{}^{\mathrm{T}}\phi\left(\mathbf{X}_2\right)+\lambda_3\mathbf{W}_2{}^{\mathrm{T}}\mathbf{U}\mathbf{L}+\lambda_4\mathbf{Q}\mathbf{V}_1\right]. \end{aligned} \tag{12}$$

4. $\mathbf{W}_m$-*Subproblem*. Fixing the other variables and setting $\partial J_1/\partial \mathbf{W}_m = 0$, we possess the solution as

$$\mathbf{W}_m = \lambda_3\mathbf{U}\mathbf{L}\mathbf{V}_m{}^{\mathrm{T}}\left(\lambda_3\mathbf{V}_m\mathbf{V}_m{}^{\mathrm{T}}+\lambda_5\mathbf{I}\right)^{-1}. \tag{13}$$

5. **U**-*Subproblem*. By fixing other variables and setting $\partial J_1/\partial \mathbf{U} = 0$, we have

$$\mathbf{U} = \lambda_3\mathbf{W}_m\mathbf{V}_m\mathbf{L}^{\mathrm{T}}\left[\left(\lambda_3+\lambda_5\right)\mathbf{L}\mathbf{L}^{\mathrm{T}}\right]^{-1}. \tag{14}$$

6. **Q**-*Subproblem*. Analogous to **U**-*Subproblem*, we have

$$\mathbf{Q} = \lambda_4\mathbf{V}_2\mathbf{V}_1{}^{\mathrm{T}}\left(\lambda_4\mathbf{V}_1\mathbf{V}_1{}^{\mathrm{T}}+\lambda_5\mathbf{I}\right)^{-1}. \tag{15}$$

7. **B**-*Subproblem*. With **B** as the only matrix variable, Eq. (9) is reduced as:

$$\min_{\mathbf{B}\in\{-1,1\}^{l\times n}} \alpha\,\|\mathbf{B}-\mathbf{U}\mathbf{L}\|_F^2+\beta\left\|l\mathbf{S}-\mathbf{B}^{\mathrm{T}}\mathbf{U}\mathbf{L}\right\|_F^2+\sum_{m=1}^{2}\mu_m\left\|\mathbf{P}_m\phi\left(\mathbf{X}_m\right)-\mathbf{B}\right\|_F^2. \tag{16}$$

Further, Eq. (16) can be derived as

$$\begin{aligned} &\min_{\mathbf{B}} tr\left(\mathbf{B}^{\mathrm{T}}\mathbf{A}+\beta\mathbf{B}^{\mathrm{T}}\mathbf{U}\mathbf{L}(\mathbf{U}\mathbf{L})^{\mathrm{T}}\mathbf{B}\right) \\ \Leftrightarrow &\min_{\mathbf{B}} tr\left(\mathbf{B}^{\mathrm{T}}\mathbf{A}+\beta\mathbf{B}^{\mathrm{T}}\mathbf{U}\mathbf{L}(\mathbf{U}\mathbf{L})^{\mathrm{T}}\mathbf{C}\right)+\frac{\theta}{2}\left\|\mathbf{B}-\mathbf{C}+\frac{\mathbf{D}}{\theta}\right\|_F^2 \\ \Leftrightarrow &\min_{\mathbf{B}} tr\left(\mathbf{B}^{\mathrm{T}}\mathbf{A}+\beta\mathbf{B}^{\mathrm{T}}\mathbf{U}\mathbf{L}(\mathbf{U}\mathbf{L})^{\mathrm{T}}\mathbf{C}-\theta\mathbf{B}^{\mathrm{T}}\mathbf{C}+\mathbf{B}^{\mathrm{T}}\mathbf{D}\right), \end{aligned} \tag{17}$$

where $\mathbf{A} = -2\alpha\mathbf{U}\mathbf{L} - 2\beta l\mathbf{U}\mathbf{L}\mathbf{S}^{\mathrm{T}} - 2\mu_m\mathbf{P}_m\phi\left(\mathbf{X}_m\right)$. According to the formulation of Eq. (17), we obtain

$$\begin{aligned} \mathbf{B} = \operatorname{sgn}\big(&2\alpha\mathbf{U}\mathbf{L}+2\beta l\mathbf{U}\mathbf{L}\mathbf{S}^{\mathrm{T}}+2\mu_m\mathbf{P}_m\phi\left(\mathbf{X}_m\right) \\ &-\beta\mathbf{U}\mathbf{L}(\mathbf{U}\mathbf{L})^{\mathrm{T}}\mathbf{C}+\theta\mathbf{C}-\mathbf{D}\big). \end{aligned} \tag{18}$$

Further, we leverage $\mathbf{S} = 2\tilde{\mathbf{L}}^{\mathrm{T}}\tilde{\mathbf{L}} - \mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ to transform Eq. (18) as Eq. (19), which ensures that the complexity of computing Eq. (19) can be reduced from $O(n^2)$ to $O(n)$.

$$\begin{aligned} \mathbf{B} = \operatorname{sgn}\big(&2\alpha\mathbf{U}\mathbf{L}+2\beta l(2\mathbf{U}\mathbf{L}\tilde{\mathbf{L}}^{\mathrm{T}}\tilde{\mathbf{L}}-\mathbf{U}\mathbf{L}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}})+2\mu_m\mathbf{P}_m\phi\left(\mathbf{X}_m\right) \\ &-\beta\mathbf{U}\mathbf{L}(\mathbf{U}\mathbf{L})^{\mathrm{T}}\mathbf{C}+\theta\mathbf{C}-\mathbf{D}\big). \end{aligned} \tag{19}$$

8. **C**-*Subproblem*. As Eq. (17) does, we deliver the solution:

$$\mathbf{C} = \operatorname{sgn}\left(-\beta\mathbf{U}\mathbf{L}(\mathbf{U}\mathbf{L})^{\mathrm{T}}\mathbf{B}+\theta\mathbf{B}+\mathbf{D}\right). \tag{20}$$

9. **D**-*Subproblem*. Fixing the others and setting $\partial J_2/\partial \mathbf{C} = 0$, matrix **D** present:

$$\mathbf{D} = \mathbf{D}+\theta\left(\mathbf{B}-\mathbf{C}\right). \tag{21}$$

10. $\mathbf{P}_m$-*Subproblem*. Fixing the others and setting $\partial J_2/\partial \mathbf{P} = 0$, we get

$$\mathbf{P}_m = \mu_m\mathbf{B}\mathbf{X}_m(\mu_m\mathbf{X}_m\mathbf{X}_m{}^{\mathrm{T}}+\gamma\mathbf{I})^{-1}. \tag{22}$$

Evidently, we have easily the closed-form solutions of the to-be updated matrix variables over Eqs. (6) and (9) by the proposed optimization algorithm. Moreover, different from the traditional way that expressed in Section 3.3, our work promises to acquire the discrete hash code **B** by mathematical expression of Eq. (19) while effectively weakening the cumulative quantization error of IMADS model. The detailed training stage is depicted in Algorithm 1 and the calculation time of IMADS is resolved via the iteration number ω_1 in lines 4–11 and the iteration number ω_2 in lines 12–17.

Algorithm 1 The IMADS Training Process

```
Input: Features $\sum_{t=1}^{2} \mathbf{X}_t$, label $\mathbf{L}$, parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$, $\lambda_4$, $\alpha$, $\beta$, $\mu_1$, $\mu_2$, $\gamma$, $\theta$, iteration number $w_1$, $w_2$, binary length $l$.
Output: Code matrix $\mathbf{B}$, projection matrices $\mathbf{U}$, $\mathbf{P}_m$.
Random initialize $\mathbf{U}_m$, $\mathbf{W}$, $\mathbf{B}$, $\mathbf{P}_m$.
Acquire the pairwise similarity matrix $\mathbf{S}$ via $\mathbf{L}$.
Convert $\mathbf{X}_t$ into nonlinear kernel features $\phi(\mathbf{X}_t)$.
% Individual Mapping Learning
for $iter1 = 1, \cdots, w_1$ do
    $\mathbf{U}_m \leftarrow$ the value computed by Eq. (10).
    $\mathbf{V}_1 \leftarrow$ the value computed by Eq. (11).
    $\mathbf{V}_2 \leftarrow$ the value computed by Eq. (12).
    $\mathbf{W}_\mathbf{m} \leftarrow$ the value computed by Eq. (13).
    $\mathbf{U} \leftarrow$ the value computed by Eq. (14).
    $\mathbf{Q} \leftarrow$ the value computed by Eq. (15).
end for
% Asymmetric Dual Supervision Learning
for $iter2 = 1, \cdots, w_2$ do
    $\mathbf{B} \leftarrow$ the value computed by Eq. (19).
    $\mathbf{C} \leftarrow$ the value computed by Eq. (20).
    $\mathbf{D} \leftarrow$ the value computed by Eq. (21).
    $\mathbf{P}_m \leftarrow$ the value computed by Eq. (22).
end for
return $\mathbf{B}$, $\mathbf{U}$, $\mathbf{P}_m$.
```

3.4. Out-of-sample learning

As mentioned earlier, IMADS has attained unified hash code during training. When a new query instance x or y appears, the IMADS studies the hash functions $h(x)$ or $h(y)$ to retrieve the homologous samples.

$$\begin{aligned} h(x) &= sgn\left(\mathbf{P}_1 x\right), \\ h(y) &= sgn\left(\mathbf{P}_2 y\right), \end{aligned} \tag{23}$$

where $\mathbf{P}_1$, $\mathbf{P}_2$ are projection matrices, which are learned by using Eq. (22). Afterwards, the querying hash codes are

$$\begin{aligned} b_x &= h(x) = \mathrm{sgn}\left(\mathbf{P}_1 x\right), \\ b_y &= h(y) = \mathrm{sgn}\left(\mathbf{P}_2 y\right). \end{aligned} \tag{24}$$

When searching any a instance, we calculate the distance between querying hash code b_x or b_y and unified hash code $\mathbf{B}$ and then rank the output results.

4. Experiments and results

4.1. Experimental configuration

Datasets. We have conducted experiments on three widely-used datasets, including Wiki (Zhang & Li, 2014) which involves single-label small instance, Flickr25K (Wang, Zhao, & Li, 2022) consisting of multilabel documents, and NUS-WIDE pertaining to large-scale scenarios (Wang et al., 2021b). Therefore, we can comprehensively assess the performance of our proposed IMADS from multiple aspects. The descriptions of all datasets are summarize in Table 1, where d notes the dimension. It is worth noting that based on previous research principles (Wang, Zhao, & Li, 2022; Wang, Zhao, Wang, et al., 2022), we conduct sampling processing about NUS-WIDE to ensure that all comparison methods acquire a stable balance in both performance and efficiency.

Evaluation and Baselines. Our experiments utilize three publicly recognized metrics (Shen et al., 2020; Wang et al., 2020), i.e., mean Average Precision (**mAP**), Precision–Recall curve (**PR**), and topN-Precision curve (**topN**) to verify the effectiveness of the proposed IMADS. In general, higher values of these metrics depict better performance of all methods. This paper compares our IMADS with some competing and state-of-the-art alternatives, including CMFH (Ding et al., 2014), SMFH (Liu et al., 2016), SCMs (Zhang & Li, 2014), SePH (Lin et al., 2016), LCMFH (Wang et al., 2019), JIMFH (Wang et al., 2020), SRLCH (Shen et al., 2020), SCRATCH (Chen et al., 2020), LFMH (Zhang, Wu, & Yu, 2021), RDMH (Zhang & Wu, 2022b), and EDCAH (Wang, Zhao, Wang, et al., 2022). Thereinto, such methods (CMFH, LCMFH, JIMFH, SRLCH, SCRATCH, LFMH, EDCAH) belong to common or individual hashing, and other approaches (SMFH, SCMs, SePH, RDMH, DRMFH) pertain to symmetric or asymmetric hashing. The description of these baselines can be found in the Section of related work. The experimental codes for these baselines are generously provided by the authors, while we independently implemented the SRLCH method.

Table 1
The statistics explanations of datasets employed in experiments.

Datasets	Wiki	Flickr25K	NUS-WIDE
Total	2,866	20,015	186,577
Class	10	24	10
Training Set	2,173	18,015	10,000
Query Set	693	2,000	2,000
Retrieval Set	2,173	18,015	184,577
Image Feature	128-d SIFT	512-d GIST	500-d BOVW
Text Feature	10-d LDA	1,386-d BOW	1,000-d BOW

Implementation Details. We mainly perform the two types of cross-modal search tasks: I2T (query images retrieving similar texts) and T2I (query texts retrieving similar images). In experiments, all comparison alternatives are executed on MATLAB version 2021a version on a workstation equipped with an Intel(R) Core(TM) i9-9820X CPU running at 3.3 GHz, 128 GB of memory. We operate the comparison baselines that recommended by the parameter setups of the original literatures, while recording the best scores of the hashing models. When optimizing the proposed IMADS, this work employs the grid search method to get the experimental parameter values of the objective function. For Eq. (6), IMADS takes $\{\lambda_1 = \lambda_2 = 0.5, \lambda_4 = 10^{-2}\}$ on three datasets. IMADS performs the best when $\{\lambda_3 = 10^4, \lambda_5 = 10^{-3}\}$ for Wiki and $\{\lambda_3 = 10^5, \lambda_5 = 10^{-4}\}$ for Flickr25K and NUS-WIDE. For Eq. (9), IMADS conducts the best when $\{\alpha = 10^3, \beta = 10^{-3}, \mu_1 = \mu_2 = 10^{-3}, \gamma = 10^{-3}, \theta = 10^{-3}\}$ on Wiki and $\{\alpha = 10^5, \beta = 10^{-1}, \mu_1 = \mu_2 = 10^{-3}, \gamma = 10^{-3}, \theta = 10^{-4}\}$ under Flickr25K and NUS-WIDE. In this paper, we set the kernel width as $t = 500$ for Wiki, $t = 1500$ for Flickr25K, and $t = 1000$ for NUS-WIDE. The iteration number $w_1 = 40$, $w_2 = 10$. For fairness, the generation results of all approaches freely run 15 times for reducing the effects of randomness and subsequent average output.

4.2. Results and analysis

Search accuracy. Table 2 and Fig. 2 respectively report the mAP values and PR curves of all alternatives using distinct hash bits on three datasets. Obviously, our proposed IMADS outperforms the other baselines in most cases (20 of 24 on mAP metric and 23 of 24 over PR metric). For example, in comparison with the best competitor EDCAH, the mAP values of IMADS have improved by about 2.50% (T2I) on Flickr25K, and 2.21% (I2T) & 2.88% (T2I) on NUS-WIDE. As for the second RDMH method, our IMADS has shown significant improvements in mAP values, with approximately 4.78% (I2T) & 2.00% (T2I) on Wiki, 4.73% (I2T) & 7.35% (T2I) on Flickr25K, and 2.19% (I2T) & 1.46% (T2I) on NUS-WIDE. The main explanation may be that proposed IMADS model enriches distinct feature distributions and reinforcement correlation with multiple supervision. Obviously, the proposed IMADS shows superior performance in most situations against other counterparts, demonstrating the effectiveness of our IMADS and proposed learning modules. Moreover, one finding is that the mAP scores of IMADS are slightly weaker than the SCRATCH model in such cases (16-bit & I2T & Wiki, 16-bit & all tasks & Flickr25K) and than EDCAH model

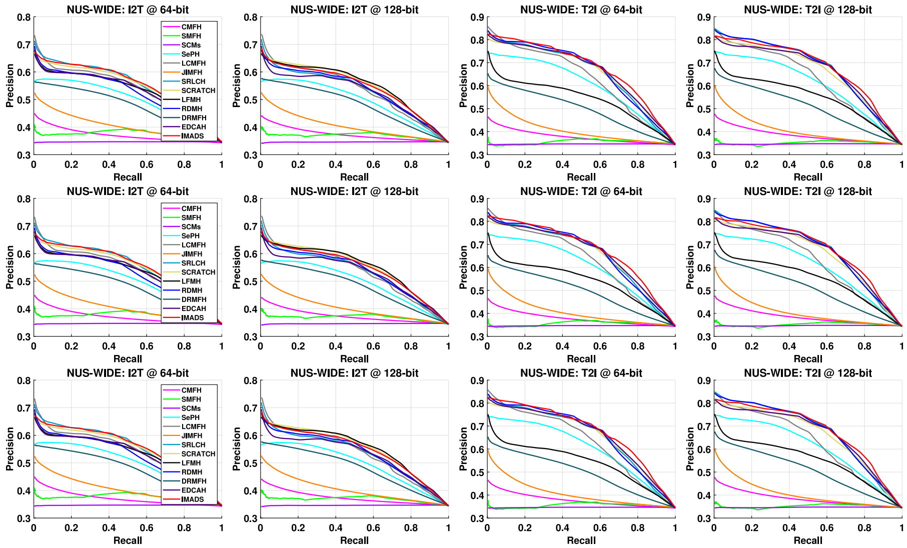

Fig. 2. The PR curves of all alternatives at 64 and 128 bits across three datasets.

Table 2
The mAP comparisons among all baselines with distinct binary lengths.

Task	Method	Wiki				Flickr25K				NUS-WIDE			
		16	32	64	128	16	32	64	128	16	32	64	128
I2T	CMFH (Ding et al., 2014)	0.2411	0.2532	0.2519	0.2567	0.5741	0.5763	0.5772	0.5785	0.3913	0.3925	0.3944	0.3967
	SMFH (Liu et al., 2016)	0.2310	0.2492	0.2653	0.2581	0.6122	0.6254	0.6432	0.6645	0.4614	0.4647	0.4525	0.4686
	SCMs (Zhang & Li, 2014)	0.2474	0.2363	0.2407	0.2601	0.6235	0.6357	0.6451	0.6482	0.5778	0.5859	0.5765	0.6022
	SePH (Lin et al., 2016)	0.2436	0.2568	0.2565	0.2612	0.6532	0.6567	0.6571	0.6595	0.5473	0.5512	0.5624	0.5631
	LCMFH (Wang et al., 2019)	0.3213	0.3382	0.3524	0.3621	0.6812	0.6901	0.7011	0.7042	0.6087	0.6124	0.6222	0.6365
	JIMFH (Wang et al., 2020)	0.2351	0.2425	0.2517	0.2533	0.5914	0.5925	0.5957	0.6022	0.4164	0.4213	0.4176	0.4223
	SRLCH (Shen et al., 2020)	0.3310	0.3424	0.3647	0.3706	0.6571	0.6865	0.6980	0.7026	0.5952	0.6230	0.6326	0.6459
	SCRATCH (Chen et al., 2020)	**0.3601**	0.3722	0.3816	0.3901	**0.7016**	0.7044	0.7052	0.7101	0.6010	0.6123	0.6235	0.6310
	LFMH (Zhang, Wu, & Yu, 2021)	0.3329	0.3335	0.3433	0.3413	0.6607	0.6583	0.6682	0.6735	0.6090	0.6168	0.6190	0.6279
	RDMH (Zhang & Wu, 2022b)	0.3238	0.3311	0.3310	0.3332	0.6525	0.6593	0.6761	0.6872	0.6002	0.6233	0.6261	0.6319
	DRMFH (Yao et al., 2023)	0.2129	0.2193	0.2318	0.2337	0.6281	0.6253	0.6381	0.6410	0.5069	0.5104	0.5180	0.5268
	EDCAH (Wang, Zhao, Wang, et al., 2022)	0.3572	0.3710	0.3802	0.3903	0.6976	0.6940	0.7130	0.7138	**0.6022**	0.6174	0.6321	0.6339
	IMADS	0.3517	**0.3770**	**0.3834**	**0.3981**	0.6956	**0.7082**	**0.7235**	**0.7368**	0.5993	**0.6344**	**0.6543**	**0.6583**
T2I	CMFH (Ding et al., 2014)	0.6073	0.6247	0.6385	0.6436	0.5874	0.5882	0.5893	0.5874	0.3875	0.3890	0.3911	0.3920
	SMFH (Liu et al., 2016)	0.5764	0.6385	0.6568	0.6683	0.6201	0.6375	0.6712	0.7014	0.4295	0.4281	0.4273	0.4269
	SCMs (Zhang & Li, 2014)	0.3819	0.4479	0.4312	0.4328	0.6368	0.6502	0.6577	0.6683	0.5634	0.5924	0.6041	0.6238
	SePH (Lin et al., 2016)	0.6776	0.6841	0.6972	0.6846	0.6944	0.6953	0.7018	0.7039	0.6372	0.6475	0.6683	0.6701
	LCMFH (Wang et al., 2019)	0.6972	0.7128	0.7269	0.7310	0.7361	0.7533	0.7745	0.7759	0.6921	0.7104	0.7197	0.7352
	JIMFH (Wang et al., 2020)	0.5203	0.5424	0.5536	0.5570	0.6024	0.6045	0.6037	0.6107	0.4331	0.4230	0.4225	0.4242
	SRLCH (Shen et al., 2020)	0.7154	0.7222	0.7454	0.7508	0.6969	0.7388	0.7466	0.7558	0.7221	0.7475	0.7628	0.7758
	SCRATCH (Chen et al., 2020)	0.7211	0.7301	0.7534	0.7569	**0.7602**	0.7682	0.7722	0.7829	0.7255	0.7357	0.7572	0.7607
	LFMH (Zhang, Wu, & Yu, 2021)	0.7094	0.7127	0.7228	0.7267	0.6870	0.6930	0.7023	0.7276	0.7143	0.7487	0.7526	0.7662
	RDMH (Zhang & Wu, 2022b)	0.7129	0.7239	0.7316	0.7384	0.6964	0.7027	0.7256	0.7425	0.7374	0.7571	0.7664	0.7763
	DRMFH (Yao et al., 2023)	0.3913	0.4067	0.4504	0.4595	0.6291	0.6308	0.6466	0.6482	0.5302	0.5332	0.5412	0.5535
	EDCAH (Wang, Zhao, Wang, et al., 2022)	0.7224	0.7317	0.7424	0.7516	0.7329	0.7689	0.7722	0.7873	0.7103	0.7232	0.7550	0.7683
	IMADS	**0.7335**	**0.7343**	**0.7581**	**0.7606**	0.7545	**0.7866**	**0.8025**	**0.8174**	**0.7395**	**0.7674**	**0.7878**	**0.7885**

in 16-bit & I2T & NUS-WIDE case. This phenomenon indicates that our IMADS has limited performance improvement under small code lengths, possibly due to the mixture of noise information in encoding. Although our method possesses only a marginal improvement against RDMH and EDCAH on T2I task & NUS-WIDE, it considerably outperforms them in other evaluation metrics. An intriguing observation is that the mAP scores of nearly all methods exhibit a gradual increase as the length grows. We deem that the training model struggles to extract sufficient informative features from a shorter code length, resulting in the suboptimal outcomes. To sum up, the acquired results depict the advantages of IMADS over some competing approaches in mAP metric.

Regarding PR curves, it is evident that the precision of IMADS always remains a higher position over the other baselines as the recall number increases except for the competing LFMH and SCRATCH about the I2T task @ 128-bit on NUS-WIDE. Meanwhile, the trend observed in the PR curves aligns consistently with that of mAP, clearly proving better performance of IMADS in cross-modal search tasks compared to other methods. Fig. 3 reveals the topN curves of all experimental algorithms leveraging 64 and 128 hash lengths. The topN results of our IMADS are better than that of other approaches in the majority of cases (20 of 24), except for slightly weaker precision against the competitor EDCAH (I2T task @ 128-bit on Flickr25K), the SRLCH (I2T tasks @ 64-bit) and LFMH (I2T task @ 128-bit) and RDMH (T2I task @ 128-bit) on NUS-WIDE. In addition, IMADS consumes less training time than two competing SRLCH and RDMH methods and similar to EDCAH method described Fig. 4, indicating that our IMADS possesses fast learning efficiency in the cross-modal search applications. In short,

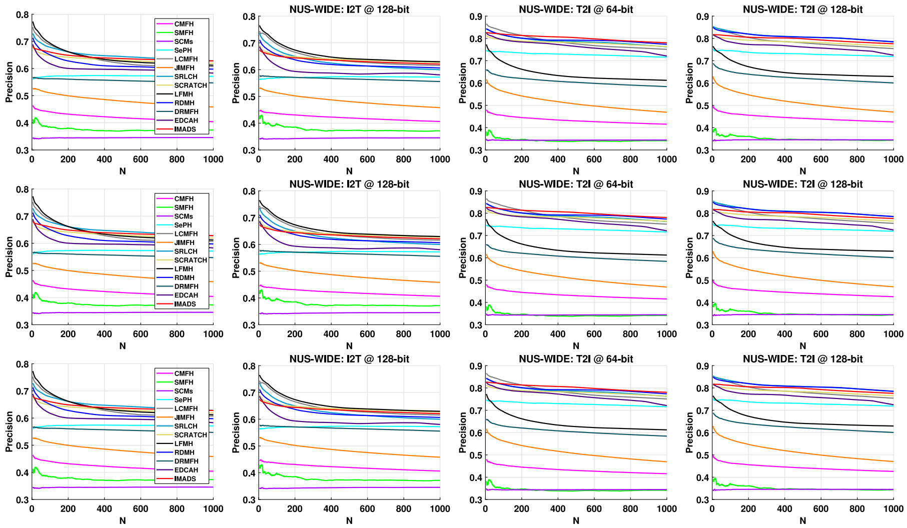

Fig. 3. The topN curves of all alternatives at 64 and 128 bits across three datasets.

Table 3
The mAP comparisons of IMADS and four variants among three datasets.

Task	Method	Wiki				Flickr25K				NUS-WIDE			
		16	32	64	128	16	32	64	128	16	32	64	128
I2T	IMADS-C	0.3441	0.3532	0.3694	0.3896	0.6723	0.6888	0.7126	0.7128	0.5948	0.6213	0.6415	0.6468
	IMADS-L	0.3515	0.3687	0.3741	0.3789	0.6731	0.7028	0.7215	0.7238	0.5892	0.5959	0.6257	0.6390
	IMADS-S	0.1123	0.1121	0.1119	0.1132	0.5594	0.5593	0.5596	0.5560	0.3455	0.3457	0.3459	0.3467
	IMADS-R	0.2830	0.2753	0.3044	0.3137	0.5908	0.5653	0.5809	0.5949	0.4824	0.5107	0.5225	0.4902
	IMADS	**0.3517**	**0.3770**	**0.3834**	**0.3981**	**0.6956**	**0.7082**	**0.7235**	**0.7368**	**0.5993**	**0.6344**	**0.6543**	**0.6583**
T2I	IMADS-C	0.7296	0.7314	0.7509	0.7584	0.7110	0.7281	0.7596	0.7615	0.6868	0.7127	0.7322	0.7413
	IMADS-L	0.7404	0.7472	0.7569	0.7465	0.7316	0.7755	0.7842	0.7961	0.7158	0.7401	0.7684	0.7779
	IMADS-S	0.1164	0.1194	0.1238	0.1345	0.5592	0.5596	0.5600	0.5610	0.3457	0.3463	0.3469	0.3484
	IMADS-R	0.4429	0.5177	0.6443	0.6575	0.5727	0.5876	0.6131	0.6161	0.5546	0.5424	0.6052	0.6203
	IMADS	**0.7335**	**0.7343**	**0.7581**	**0.7606**	**0.7545**	**0.7866**	**0.8025**	**0.8174**	**0.7395**	**0.7674**	**0.7878**	**0.7885**

these observations derived from Table 2, Fig. 2, and Fig. 3 collectively highlight the effectiveness of our IMADS across three assessments.

4.3. Comprehensive analysis

Ablation Study. To confirm the efficacy of each learning component in IMADS, we have formulated four variants for comparative analysis. The **IMADS-C** is the variant that only exploits the shared feature to train the model reflected in $\mathbf{V}_1 = \mathbf{V}_2 = \mathbf{V}$, i.e., $\lambda_2 = \lambda_3 = 0$. The **IMADS-L** is the variant which simply leverages the label supervision knowledge ($\beta = 0$). The **IMADS-S** variant is produced by pairwise semantic similarity matrix, and thus most current studies (Wang, Zhao, & Li, 2022; Yao et al., 2020) only take $\alpha = 0$ to obtain the variant. The IMADS-S is constructed by deleting the constraint $\mathbf{V} = \mathbf{ZL}$. The **IMADS-R** variant utilizes the traditional relaxation scheme to solve the model by $\partial J_2/\partial \mathbf{B} = 0$ on condition of $\mathbf{B} \in [-1, 1]^{l \times n}$.

Table 3 unveils the mAP results of the whole variants with four different hash bits. As anticipated, the IMADS attains the highest mAP values over the other four variants across all comparison datasets. In specific, IMADS outperforms IMADS-C illustrating that merging the shared and individual feature descriptors is conducive to boosting accuracy, which demonstrates the significance of the individual information. Similarly, compared with the variant IMADS-L & IMADS-S, and IMADS-R, the IMADS obtain better performance showing the effectiveness of the asymmetric dual supervision term and discrete optimization algorithm, respectively. Consequently, these findings of Table 3 prove the efficacy of the learning part in our IMADS.

Computational Cost. An excellent model should simultaneously carry good accuracy and fast learning efficiency (a.k.a. computational cost). Thereinto, the computational cost of this paper is mainly determined by the training process of hashing models. Fig. 4 depicts the training time comparisons of four competing alternatives (SCRATCH (Chen et al., 2020), SRLCH (Shen et al., 2020), RDMH (Zhang & Wu, 2022b), EDCAH (Wang, Zhao, Wang, et al., 2022)) and our IMADS on Flickr25K and NUS-WIDE. The training time of all the baselines is very close on small dataset Wiki and is not comparable. We conclude from Fig. 4 that RDMH and SRLCH possess the worst and second worst training efficiency over other methods, respectively because they employ the high-dimensional kernel features as model input and consequent heavy computational burden. The computational cost of our IMADS is similar to that of SCRATCH on Flickr25K, and similar to that of SCRATCH and EDCAH on NUS-WIDE, where the main explanation is that the IMADS adopts the linear hashing computing during training and optimization. Despite analogous results in training time, IMADS shows better mAP and PR values than SCRATCH and EDCAH. Thus, these results indicate the superiority of our IMADS in fast learning efficiency.

Parameter Sensitivity. Analyzing the data parameters in the objective function cannot only help to determine the parameter sensitivity interval, but also contribute to acquiring the optimal parameter settings. Fig. 5 exhibits the parameter analysis results in Eqs. (6) and (9) that influenced our IMADS on three datasets. Thereinto, parameters $\sum_{m=1}^{2} \lambda_m$, λ_3, λ_4, λ_5 over Eq. (6) make effect on the individual mapping learning part and α, β, $\sum_{m=1}^{2} \mu_m$, and θ in Eq. (9) impact the asymmetric dual supervision learning. We can observe from Fig. 5 that: (1) as for individual mapping learning procedure, our IMADS yields stable yet preferable retrieval results when parameters $\sum_{m=1}^{2} \lambda_m \in [0.3, 0.7]$, $\lambda_3 \in \left[10^{-2}, 10\right]$ and $\lambda_4 \in \left[10^{3}, 10^{5}\right]$, and $\lambda_5 \in \left[10^{-4}, 1\right]$. and (2) during asymmetric dual supervision learning, this study also the similar phenomenon in the insensitivity region of $\alpha \in \left[10^{3}, 10^{5}\right]$, $\beta \in \left[10^{-3}, 10^{-1}\right]$, $\sum_{m=1}^{2} \mu_m \in \left[10^{-3}, 10^{-1}\right]$, $\gamma \in \left[10^{-3}, 1\right]$, and $\theta \in \left[10^{-4}, 10^{-3}\right]$ about the IMADS. That

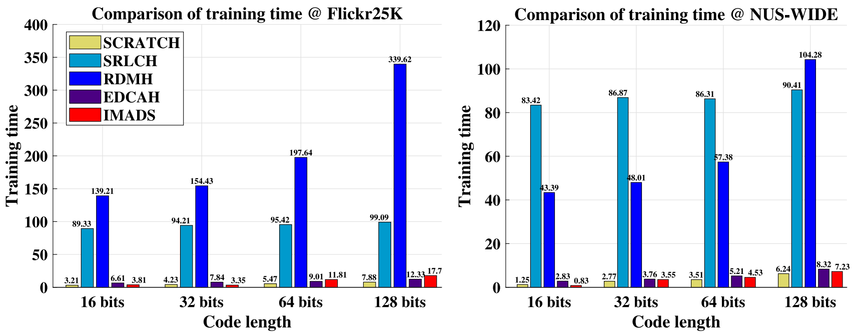

Fig. 4. Training time scores of IMADS and three baselines on Flickr25K and NUS-WIDE.

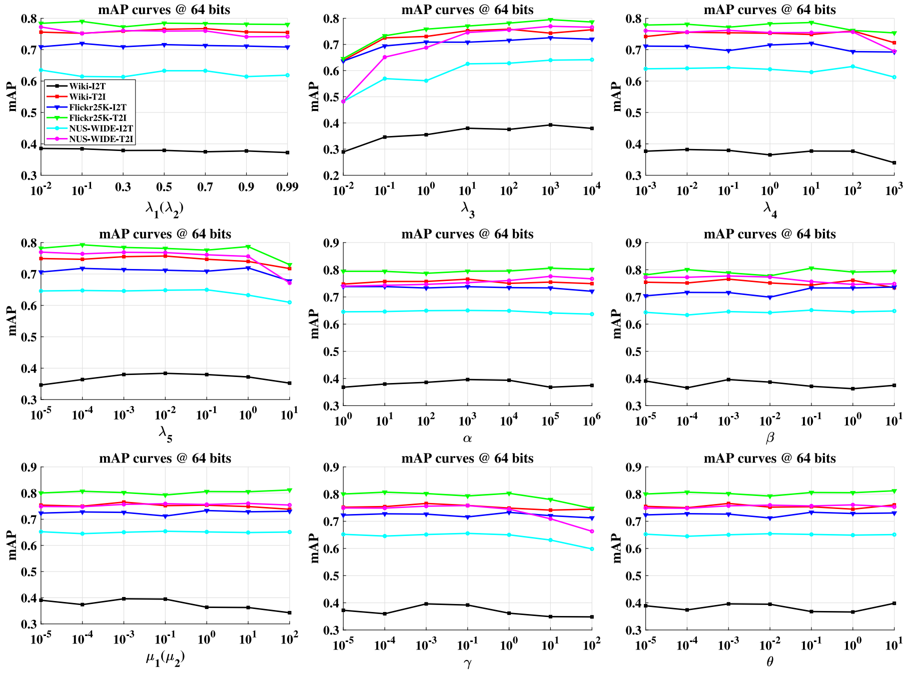

Fig. 5. The mAP scores of our IMADS using various parameters settings on all datasets.

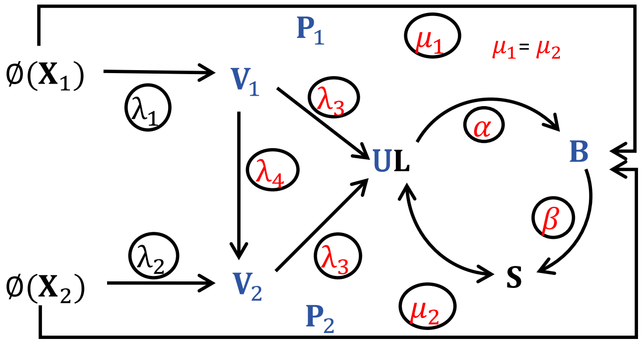

Fig. 6. The graphical network structure of the proposed IMADS during training.

is to say, these wide insensitivity regions produce the optimal scores and meanwhile indicate good generalization performance.

For clear representation, we have shown the graphical network structure of training matrix variables in Fig. 6 while the red circles are main sensitivity parameters and the blue circles denote the to-learned matrices. We conclude from Figs. 5 and 6 that in fact, the five sensitivity parameters (λ_3, λ_4, $\mu_1 = \mu_2$, α, β) greatly impact the performance of IMADS.

Convergence Analysis. To identify the stable convergence of our IMADS, this study conducts the corresponding operations with selected 64 hash length (see Fig. 7), where iter 1 and iter 2 stand for the training stage of the individual mapping learning and the asymmetric dual supervision learning, respectively. For clear and intuitive presentation, the ordinate label in Fig. 7 is normalized by dividing the largest value in the objective function by the other values. Concretely, the first finding is that the two learning procedures of the IMADS both achieve convergence on three datasets, where they hold the fastest rate of convergence on Wiki and the slowest on NUS-WIDE. The fundamental reason is due to the size of these datasets. Besides, the second interesting finding is that iter 1 (around 40) converges much less than iter 2 (within 10) in terms of iteration number. We analyze the possible statements that iter 1 consumes much time to tackle the original features and iter 2 adopts discrete and asymmetric linear optimization.

Cross-modal Visualization. To intuitively display the search results, this study carries out two kinds of cross-modal search cases against the proposed IMADS on Wiki. Fig. 8 reports the top 10 nearest query text and image results on the I2T and T2I tasks, where these green and red boxes respectively denote the right and wrong query samples. Referring to this figure, we gather the following observations: (1) Concerning to the query music class, the IMADS obtains the whole correct top-10 searched results for both two tasks. (2) Regarding the art and biology query categories, there are 1 or 2 incorrect samples in the search cases,

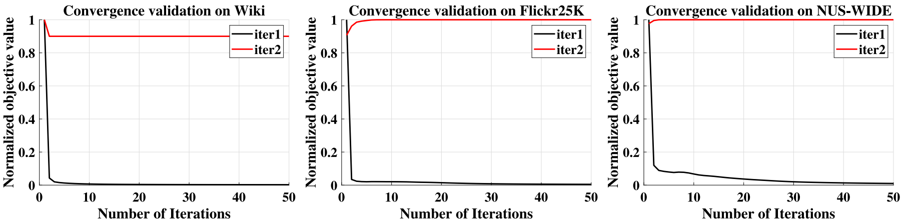

Fig. 7. The convergence curves of our IMADS @ 64-bit on three datasets.

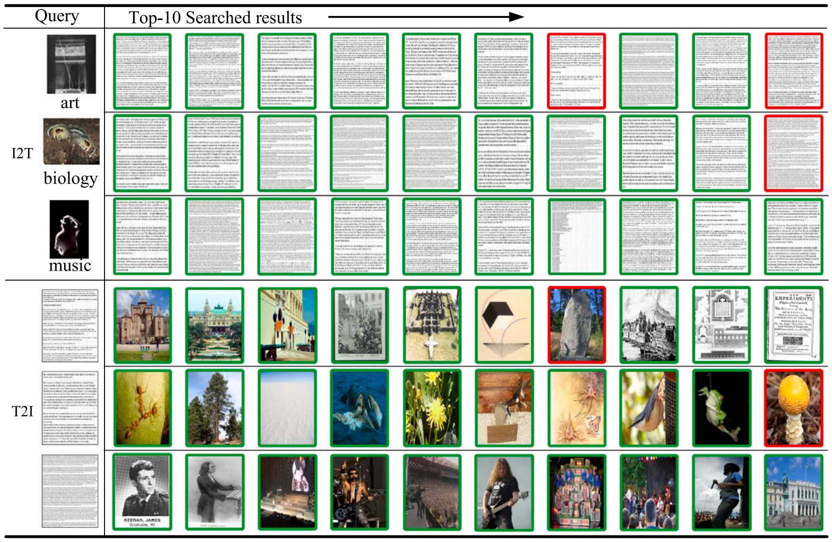

Fig. 8. The visualization display of the search results by our IMADS (Wiki for example).

but they rank lower. We analyze that the original data may be mixed with noise or outliers and subsequent deviations in the search results. Overall, above visual phenomena prove that the IMADS can promote inter-data similarity and semantically correlation between supervision knowledge and data to maintain into the hash codes during querying.

Comparing with deep hashing. To evaluate the functionality of the IMADS, we construct a deep hashing variant (IMADScnn) to compare with some competitive deep cross-modal hashing approaches (PRDH (Yang, Deng, Liu, Liu, Tao et al., 2017), ADAH (Zhang, Lai, & Feng, 2018), EGDH (Shi, You, Zheng, Wang, & Peng, 2019), MLCAH (Ma, Zhang, & Xu, 2020), DADH (Bai, Zeng, Ma, Zhang, & Chen, 2020)) on Flickr25K, which is shown in Fig. 9. The input to the IMADScnn model comprises a 4096-dimensional CNN image feature while retaining the same shallow text feature. From Fig. 9, a significant observation is that IMADScnn performs better than other baselines on I2T task while slightly works worse than EGDH, MLCAH, and DADH w.r.t. 32-bit on T2I task. The primary reason is that the input deep image feature holds a wealth of information, which in turn aids in enhancing accuracy. Besides, it takes three more hours to train all deep models, while the training time of IMADScnn at 32, 64 bits is 3.05, and 3.20 (in seconds) respectively. To sum up, our variant IMADScnn attains the high learning efficiency while delivering comparable accuracy over latest deep hashing alternatives.

Study Limitations and possible explorations. Despite the remarkable experimental performance improvement, there are still two limitations that need to be further promoted in this research work. In specific, the partial failure cases of our IMADS occur in Table 2 and Fig. 3, where four limited mAP scores show in small 16-bit length and the topN scores separately weaker than that of comparative EDCAH (I2T @ 128-bit on Flickr25K), the SRLCH (I2T @ 64-bit) and LFMH (I2T @ 128-bit) and RDMH (T2I @ 128-bit) on NUS-WIDE. Moreover, the visualization of the querying art class (see Fig. 8) shows two incorrect samples on the I2T task (one error on T2I). After deriving the source code, computational optimization, and training process, we find that these two limitations share a common cause, namely the sorting learning problem, which in turn leads to the similarity problem, with the source resorting to data noise and class label noise problems.

According to the aforementioned analysis, we plan to carry out the possible solutions from two aspects. One is to reduce the data noise by eliminating redundant information except for the common and individual features, and the second one is to remove the erroneous supervision tag to well maintain the beneficial knowledge of the corresponding data in the original labels. Thereinto, the key techniques involve discrete optimal transport theory and Bayesian personalized ranking matrix factorization in the follow-up explorations.

5. Conclusion

This work presents a novel supervised hashing method called IMADS, specifically designed for cross-modal search applications. Extensive experiments show that our IMADS works more efficient than other state-of-the-art alternatives. Specifically, the IMADS highlights three main advantages: (1) it better excavates the individual features of each modality while fusing the shared feature between modalities, and (2) the asymmetric dual supervision framework can well strengthen the semantic correlation between multiple supervision knowledge, the common and individual features, and (3) different from current relaxation paradigm, this work develops a novel optimization scheme to optimization the model IMADS. The upcoming plan contains building a knowledge graph modality for cross-modal instances, aiming to effectively maximize the similarity correlation between distinct modalities and the robust semantic relationship between modality and supervision.

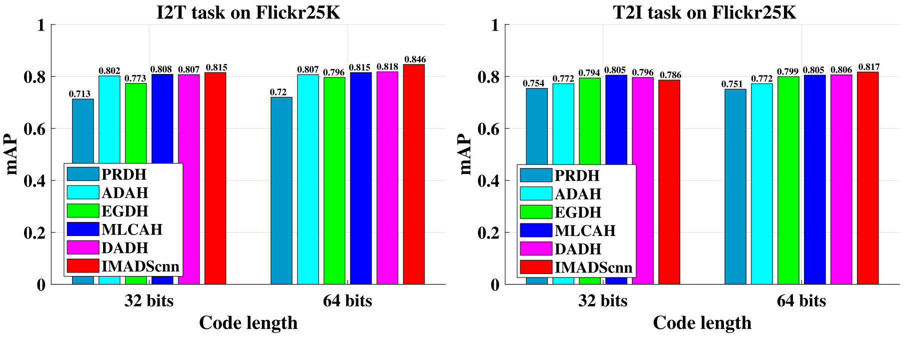

Fig. 9. The mAP scores of IMADScnn and several deep hashing methods.

CRediT authorship contribution statement

Song Wang: Methodology, Conceptualization, Formal analysis, Data curation, Software, Writing – original draft, Investigation, Writing – review & editing, Validation. **Huan Zhao:** Funding acquisition, Writing–original draft, Validation, Supervision, Writing – review & editing. **Zixing Zhang:** Investigation, Conceptualization, Writing – original draft, Supervision, Writing – review & editing. **Keqin Li:** Resources, Writing – Original Draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgments

This work received support from the National Natural Science Foundation of China (No. 62076092, 61772188).

References

Bai, C., Zeng, C., Ma, Q., Zhang, J., & Chen, S. (2020). Deep adversarial discrete hashing for cross-modal retrieval. In *International conference on multimedia retrieval* (pp. 525–531).

Chen, Z., Li, C., Luo, X., Nie, L., Zhang, W., & Xu, X. (2020). SCRATCH: A scalable discrete matrix factorization hashing framework for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(7), 2262–2275.

Chen, Z.-D., Wang, Y., Li, H.-Q., Luo, X., Nie, L., & Xu, X.-S. (2019). A two-step cross-modal hashing by exploiting label correlations and preserving similarity in both steps. In *ACM international conference on multimedia* (pp. 1694–1702).

Chen, Y., Zhang, H., Tian, Z., Wang, J., Zhang, D., & Li, X. (2022). Enhanced discrete multi-modal hashing: More constraints yet less time to learn. *IEEE Transactions on Knowledge and Data Engineering*, *34*(3), 1177–1190.

Cheng, M., Jing, L., & Ng, M. K. (2020). Robust unsupervised cross-modal hashing for multimedia retrieval. *ACM Transactions on Information Systems*, *38*(3), 30:1–30:25.

Da, C., Xu, S., Ding, K., Meng, G., Xiang, S., & Pan, C. (2017). AMVH: Asymmetric multi-valued hashing. In *IEEE conference on computer vision and pattern recognition* (pp. 898–906).

Ding, G., Guo, Y., & Zhou, J. (2014). Collective matrix factorization hashing for multimodal data. In *IEEE conference on computer vision and pattern recognition* (pp. 2083–2090).

Fang, X., Jiang, K., Han, N., Teng, S., Zhou, G., & Xie, S. (2022). Average approximate hashing-based double projections learning for cross-modal retrieval. *IEEE Transactions on Cybernetics*, *52*(11), 11780–11793.

Fang, Y., & Ren, Y. (2020). Supervised discrete cross-modal hashing based on kernel discriminant analysis. *Pattern Recognition*, *98*, Article 107062.

Li, H., Wang, W., Liu, Z., Niu, Y., Wang, H., Zhao, S., Liao, Y., Yang, W., & Liu, X. (2022). A novel locality-sensitive hashing relational graph matching network for semantic textual similarity measurement. *Expert Systems with Applications*, *207*, Article 117832.

Lin, Z., Ding, G., Han, J., & Wang, J. (2016). Cross-view retrieval via probability-based semantics-preserving hashing. *IEEE Transactions on Cybernetics*, *47*(12), 4342–4355.

Liu, X., Hu, Z., Ling, H., & Cheung, Y.-M. (2021). MTFH: A matrix tri-factorization hashing framework for efficient cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(3), 964—981.

Liu, H., Ji, R., Wu, Y., & Hua, G. (2016). Supervised matrix factorization for cross-modality hashing. In *International joint conferences on artificial intelligence* (pp. 1767–1773).

Liu, X., Li, A., Du, J., Peng, S., & Fan, W. (2018). Efficient cross-modal retrieval via flexible supervised collective matrix factorization hashing. *Multimedia Tools and Applications*, *77*(21), 28665–28683.

Liu, X., Wang, X., & Cheung, Y. (2022). FDDH: fast discriminative discrete hashing for large-scale cross-modal retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(11), 6306–6320.

Luo, X., Zhang, P., Wu, Y., Chen, Z., Huang, H., & Xu, X. (2018). Asymmetric discrete cross-modal hashing. In *ACM on international conference on multimedia retrieval* (pp. 204–212).

Ma, D., Liang, J., He, R., & Kong, X. (2017). Nonlinear discrete cross-modal hashing for visual-textual data. *IEEE Multimedia*, *24*(2), 56–65.

Ma, X., Zhang, T., & Xu, C. (2020). Multi-level correlation adversarial hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, *22*(12), 3101–3114.

Mandal, D., Chaudhury, K. N., & Biswas, S. (2019). Generalized semantic preserving hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, *28*(1), 102–112.

Meng, M., Wang, H., Yu, J., Chen, H., & Wu, J. (2021). Asymmetric supervised consistent and specific hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, *30*, 986–1000.

Qin, J., Fei, L., Zhang, Z., Wen, J., Xu, Y., & Zhang, D. (2022). Joint specifics and consistency hash learning for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing*, *31*, 5343–5358.

Qin, Q., Xian, L., Xie, K., Zhang, W., Liu, Y., Dai, J., & Wang, C. (2022). Deep Multi-Similarity Hashing with semantic-aware preservation for multi-label image retrieval. *Expert Systems with Applications*, *205*, Article 117674.

Seyed, M. A., Mohammad, S. H., & Müller, H. (2023). A novel Siamese deep hashing model for histopathology image retrieval. *Expert Systems with Applications*, *225*, Article 120169.

Shen, H. T., Liu, L., Yang, Y., Xu, X., Huang, Z., Shen, F., & Hong, R. (2020). Exploiting subspace relation in semantic labels for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, *33*(10), 3351–3365.

Shen, Y., Sun, X., Wei, X.-S., Hu, H., & Chen, Z. (2022). A channel mix method for fine-grained cross-modal retrieval. In *IEEE international conference on multimedia and expo* (pp. 1–6).

Shi, Y., You, X., Zheng, F., Wang, S., & Peng, Q. (2019). Equally-guided discriminative hashing for cross-modal retrieval. In *International joint conference on artificial intelligence* (pp. 4767–4773).

Teng, S., Ning, C., Zhang, W., Wu, N., & Zeng, Y. (2022). Fast asymmetric and discrete cross-modal hashing with semantic consistency. *IEEE Transactions on Computational Social Systems*, *10*(2), 577–589.

Tran, V., Wang, L., Chen, H., & Xiao, Q. (2021). MCHT: A maximal clique and hash table-based maximal prevalent co-location pattern mining algorithm. *Expert Systems with Applications*, *175*, Article 114830.

Wang, D., Gao, X., Wang, X., & He, L. (2015). Semantic topic multimodal hashing for cross-media retrieval. In *International joint conference on artificial intelligence* (pp. 3890–3896).

Wang, D., Gao, X.-B., Wang, X., & He, L. (2019). Label consistent matrix factorization hashing for large-scale cross-modal similarity search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(10), 2466–2479.

Wang, Y., Luo, X., Nie, L., Song, J., Zhang, W., & Xu, X.-S. (2021). BATCH: A scalable asymmetric discrete cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, *33*(11), 3507–3519.

Wang, Y., & Peng, Y. (2022). MARS: Learning modality-agnostic representation for scalable cross-media retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(7), 4765–4777.

Wang, D., Wang, Q., He, L., Gao, X., & Tian, Y. (2020). Joint and individual matrix factorization hashing for large-scale cross-modal retrieval. *Pattern Recognition*, *107*, Article 107479.

Wang, K., Wang, Y., Xu, X., Cao, Z., & Cai, X. (2022). Instance-level semantic alignment for zero-shot cross-modal retrieval. In *IEEE international conference on multimedia and expo* (pp. 1–6).

Wang, L., Zareapoor, M., Yang, J., & Zheng, Z. (2022). Asymmetric correlation quantization hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, *24*, 3665–3678.

Wang, S., Zhao, H., & Li, K. (2022). Discrete joint semantic alignment hashing for cross-modal image-text search. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(11), 8022–8036.

Wang, S., Zhao, H., & Nai, K. (2021). Learning a maximized shared latent factor for cross-modal hashing. *Knowledge-Based Systems*, *228*, Article 107252.

Wang, S., Zhao, H., Wang, Y., Huang, J., & Li, K. (2022). Cross-modal image-text search via efficient discrete class alignment hashing. *Information Processing & Management*, *59*(3), Article 102886.

Yang, E., Deng, C., Liu, W., Liu, X., Tao, D., & Gao, X. (2017). Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI conference on artificial intelligence* (pp. 1618–1625).

Yang, E., Yao, D., Liu, T., & Deng, C. (2022). Mutual quantization for cross-modal search with noisy labels. In *IEEE conference on computer vision and pattern recognition* (pp. 7541–7550).

Yao, T., Li, Y., Guan, W., Wang, G., Li, Y., Yan, L., & Tian, Q. (2023). Discrete robust matrix factorization hashing for large-scale cross-media retrieval. *IEEE Transactions on Knowledge and Data Engineering*, *35*(2), 1391–1401.

Yao, T., Yan, L., Ma, Y., Yu, H., Su, Q., Wang, G., & Tian, Q. (2020). Fast discrete cross-modal hashing with semantic consistency. *Neural Networks*, *125*, 142–152.

Zhang, X., Lai, H., & Feng, J. (2018). Attention-aware deep adversarial hashing for cross-modal retrieval. In *European conference on computer vision* (pp. 614–629).

Zhang, D., & Li, W.-J. (2014). Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI conference on artificial intelligence* (pp. 2177–2183).

Zhang, C., Li, H., Gao, Y., & Chen, C. (2023). Weakly-supervised enhanced semantic-aware hashing for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, *35*(6), 6475–6488.

Zhang, P., Luo, Y., Huang, Z., Xu, X., & Song, J. (2021). High-order nonlocal hashing for unsupervised cross-modal retrieval. *World Wide Web*, *24*(2), 563–583.

Zhang, L., & Wu, X. (2022a). Latent space semantic supervision based on knowledge distillation for cross-modal retrieval. *IEEE Transactions on Image Processing*, *31*, 7154–7164.

Zhang, D., & Wu, X. (2022b). Robust and discrete matrix factorization hashing for cross-modal retrieval. *Pattern Recognition*, *122*, Article 108343.

Zhang, D., & Wu, X. (2022c). Scalable discrete matrix factorization and semantic autoencoder for cross-media retrieval. *IEEE Transactions on Cybernetics*, *52*(7), 5947–5960.

Zhang, D., Wu, X.-J., & Yu, J. (2021). Label consistent flexible matrix factorization hashing for efficient cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *17*(3), 1–18.