Dual Semantic Fusion Hashing for Multi-Label Cross-Modal Retrieval

Kaiming Liu[1] , **Yunhong Gong**[1] , **Yu Cao**[1] , **Zhenwen Ren**[2] , **Dezhong Peng**[1,3] **and Yuan Sun**[1*]
[1]College of Computer Science, Sichuan University
[2]School of National Defense Science and Technology, Southwest University of Science and Technology
[3]National Innovation Center for UHD Video Technology
liukaiming0125@outlook.com, gongyunhong@stu.scu.edu.cn, 371755384@qq.com,
rzw@njust.edu.cn, pengdz@scu.edu.cn, sunyuan_work@163.com

Abstract

Cross-modal hashing (CMH) has been widely used for multi-modal retrieval tasks due to its low storage cost and fast query speed. Although existing CMH methods achieve promising performance, most of them mainly rely on coarse-grained supervision information (*i.e.*, pairwise similarity matrix) to measure the semantic similarities between all instances, ignoring the impact of multi-label distribution. To address this issue, we construct fine-grained semantic similarity to explore the cluster-level semantic relationships between multi-label data, and propose a new dual semantic fusion hashing (DSFH) for multi-label cross-modal retrieval. Specifically, we first learn the modal-specific representation and consensus hash codes, thereby merging the specificity with consistency. Then, we fuse the coarse-grained and fine-grained semantics to mine multiple-level semantic relationships, thereby enhancing hash codes discrimination. Extensive experiments on three benchmarks demonstrate the superior performance of our DSFH compared with 16 state-of-the-art methods.

1 Introduction

In the era of informatization, retrieving relevant instances from vast amounts of multi-modal data [Xu *et al.*, 2024], such as videos, images, and text, has become a great challenge. Cross-modal retrieval (CMR) [Wang *et al.*, 2023; Liu *et al.*, 2024] aims to query semantically related data of other modalities by providing data in one modality, which becomes considerable attention. Early CMR methods [Ranjan *et al.*, 2015; Wang *et al.*, 2016] tend to learn a unified real-valued representation for each instance to eliminate heterogeneity across modalities. Nevertheless, the explosive growth in data volumes brings these methods to the bottleneck of query efficiency. Due to the advantages of retrieval efficiency and storage cost, cross-modal hashing (CMH) [Sun *et al.*, 2023a; Sun *et al.*, 2024a] exhibits significant potential in processing massive-scale data. CMH aims to project high-dimensional features into compact hash codes in the Hamming space, thereby measuring the similarities between all instances through the Hamming distances.

In recent years, numerous supervised CMH methods [Zhu *et al.*, 2024; Sun *et al.*, 2024b] have been proposed to improve retrieval performance. They can be broadly categorized into two types: label matrix supervision [Xu *et al.*, 2017; Shen *et al.*, 2021] and pairwise semantic supervision [Luo *et al.*, 2018; Liu *et al.*, 2019]. Pairwise semantic supervision methods adopt pairwise similarity matrices to estimate instance-level semantic relationships, thereby providing more semantic guidance. Therefore, they obtain more promising performance. However, they still face some limitations. The existing CMH methods always adopt relatively coarse-grained supervision information (*i.e.*, instance-level semantic relationship) to guide hashing learning. In other words, they ignore the intrinsic multi-label semantic correspondences, which makes it difficult to excavate more semantics. For example, as shown in Fig.1, we have three labels A [1,0,0,1], B [0,1,0,1], and C [1,0,1,0], and the number of their corresponding instances are 5, 50, and 5, respectively. According to the instance-level semantic relationship, the similarity between A and B and between A and C is equal. Since B contains a larger number of instances, the more relevant instances could be retrieved if the instances from A are semantically closer to ones from B. However, it is a great challenge how to construct a fine-grained semantic relationship to enhance the expression of semantic information.

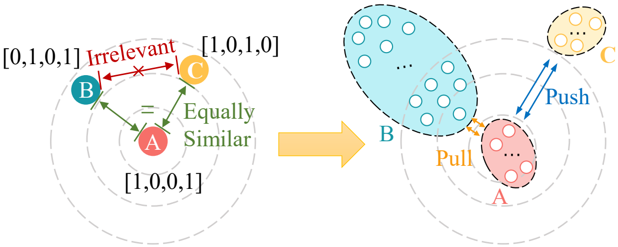


Figure 1: The semantic relationships of multi-label data. From the instance-level semantic view, label A has equivalent similarity to B and C, while B and C are deemed an irrelevant pair. However, since B has more instances, the semantic distance between A and B should be decreased, whereas between A and C should be increased. This fine-grained semantic similarity can more accurately measure the semantic relationships between multi-label data.

[*] Corresponding author.

To address the above issues, we propose a dual semantic fusion hashing (DSFH) framework for multi-label cross-modal retrieval. Our DSFH explores the multi-level semantic information from multi-label data, and further fuses both the coarse-grained and fine-grained semantic relationships, thereby making the learned hash codes more discriminative. Specifically, as shown in Fig.2, we first learn the modality-specific representations to preserve the consistency and specificity of all modalities to some extent. Then, we adopt K-Means label clustering to capture finer-grained cluster-level relationships between instances. Finally, we fuse the coarse-grained and fine-grained supervision information to guide consensus hashing learning, thereby generating discriminative hash codes. Overall, the main contributions of this paper are summarized as follows:

- We propose a new dual semantic fusion hashing (DFSH) that fully exploits the multi-level semantic information in multi-label multi-modal data. To the best of our knowledge, this is the first work to simultaneously consider the coarse-grained and fine-grained semantic relationships for multi-label cross-modal hashing.
- We establish the cluster-level semantic relationship to capture finer-grained similarities between different instances, thereby promoting that all instances with similar semantics could be clustered together.
- Extensive experiments demonstrate that our DFSH outperforms all state-of-the-art comparison methods on three benchmarks.

2 Related Work

2.1 Label Matrix Supervision Methods

Label matrix supervision Methods [Xu *et al.*, 2017; Wang *et al.*, 2019; Sun *et al.*, 2023b] typically adopt the class label as semantic information to guide hashing learning, thereby improving the discriminative of the learned hash codes. For example, scalable discrete matrix factorization hashing (SCRATCH) [Chen *et al.*, 2020] learns a common latent representation by incorporating original features and labels, thereby generating unified hash codes. To more accurately explore the correlation of heterogeneous data, discrete semantic matrix factorization hashing (DSMFH) [Qin *et al.*, 2021] performs matrix factorization on labels to obtain latent specific representations, then learns hash codes based on these latent representations. To alleviate the interference of noisy labels, robust and discrete matrix factorization hashing (RDMH) [Zhang and Wu, 2022] directly associates hash codes with the label matrix and introduces the $l_{2,1}$-norm to enhance its robustness.

2.2 Pairwise Semantic Supervision Methods

To fully reflect the semantic information between instances, pairwise semantic supervision methods [Zhang and Li, 2014; Lin *et al.*, 2015; Teng *et al.*, 2023] adopt the $n \times n$ similarity matrix to capture semantic similarities between all instances. For example, fast cross-modal hashing (FCMH) [Wang *et al.*, 2022] constructs both global and local similarity matrices to guide hash codes learning. However, the $n \times n$ similarity matrix could lead to high computational cost. To this end, scalable asymmetric discrete cross-modal hashing (BATCH) [Wang *et al.*, 2021] adopts labels to construct pairwise similarity, thereby avoiding the use of the large $n \times n$ similarity matrix. To explore potential semantic correlations of multi-label data, adaptive label correlation based asymmetric discrete hashing (ALECH) [Li *et al.*, 2023] proposes adaptive label correlation to fully utilize the latent label information.

Although the existing CMH methods have obtained promising performance, most of them ignore cluster-level semantic relationships in multi-labels. Traditional coarse-grained semantic similarity makes it difficult to fully mine the correlations of multi-label data.

3 Proposed Method

3.1 Notations

Given a multi-modal dataset $\{\boldsymbol{V}^t\}_{t=1}^M$ with M modalities, we denote $\boldsymbol{V}^t = [\boldsymbol{v}_1^t, \boldsymbol{v}_2^t, \cdots, \boldsymbol{v}_n^t] \in \mathbb{R}^{d^t \times n}$ as the training data of the t-th modality, where n is the number of instances, and d^t is the corresponding feature dimensionality. To capture the nonlinear features from original multi-modal data, we employ the radial basis function (RBF) kernel [Liu *et al.*, 2012] on $\boldsymbol{V}^t$. Therefore, the kernelized features of t modality can be denoted as $\boldsymbol{X}^t = [exp(\frac{\|\boldsymbol{x}_i^t - \boldsymbol{a}_1^t\|_2^2}{-2\sigma^2}), \cdots, exp(\frac{\|\boldsymbol{x}_i^t - \boldsymbol{a}_h^t\|_2^2}{-2\sigma^2})]$. Where $\{\boldsymbol{a}_1^t, \boldsymbol{a}_2^t, \cdots, \boldsymbol{a}_h^t\}$ are h anchors randomly selected from $\boldsymbol{V}^t$, and σ is the kernel width. In addition, $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n] \in \mathbb{R}^{c \times n}$ denotes the ground truth labels, where $\boldsymbol{y}_i \in \{0,1\}^{c \times 1}$ is the label vector of every instance, c is the number of categories. We define $y_{ij} = 1$ if the j-th instance belongs to i-th category, otherwise $y_{ij} = 0$.

3.2 Model Formulation

Most existing methods emphasize extracting the shared information from different modalities to directly learn consensus hash codes. However, they unconsciously neglect private semantic information from each modality. To this end, we first learn modality-specific representations $\boldsymbol{H}^t$ to maximally preserve the intrinsic semantic correlations of each modality. Then, we adopt the cosine similarity between the corresponding labels of the instances to capture instance-level semantic relationships, *i.e.*, $\boldsymbol{S}_I = \boldsymbol{Y}^\top \boldsymbol{Y}$. Further, to bridge the heterogeneity gaps across different modalities, we propose a consistency learning strategy to learn consensus hash codes $\boldsymbol{B}$ under the guidance of the supervision information. In addition, we add two strong constraints, *i.e.*, $(\boldsymbol{H}^t)^\top \boldsymbol{H}^t = n\boldsymbol{I}$ and $(\boldsymbol{H}^t)^\top \boldsymbol{1} = \boldsymbol{0}$, for the bit decorrelation and bit balance. Mathematically, we can obtain the following objective function:

$$\begin{aligned} \min_{\boldsymbol{H}^t, \boldsymbol{W}^t, \boldsymbol{B}} & \sum_{t=1}^{M} \|\boldsymbol{H}^t - \boldsymbol{W}^t \boldsymbol{X}^t\|^2 + \alpha \|\boldsymbol{B}^\top \boldsymbol{H}^t - r\boldsymbol{Y}^\top \boldsymbol{Y}\|^2 \\ & s.t.\ (\boldsymbol{H}^t)^\top \boldsymbol{1} = \boldsymbol{0}, (\boldsymbol{H}^t)^\top \boldsymbol{H}^t = n\boldsymbol{I}, \\ & \quad (\boldsymbol{W}^t)^\top \boldsymbol{W}^t = \boldsymbol{I}, \boldsymbol{B} \in \{-1,1\}^{r \times n}, \end{aligned} \tag{1}$$

where α is the trade-off parameter, r is the bit length.

Most of the existing methods adopt the pairwise similarity matrix to estimate instance-level semantic relationships.

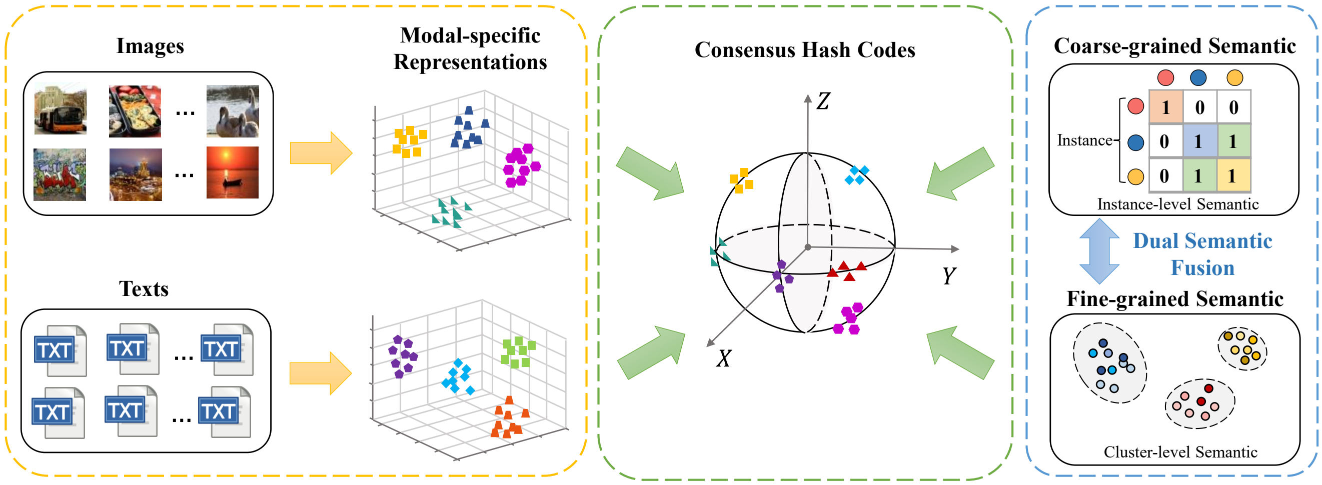


Figure 2: The basic framework of our DSFH. DSFH first learns modality-specific representations to preserve the private semantics of different modalities. Further, DSFH constructs both instance-level and cluster-level semantic relationships for multi-label data. Finally, we fuse both coarse-grained and fine-grained semantics as multi-grained supervision to guide the learning of consensus hash codes.

Since multi-label multi-modal data contains more complex semantic relationships, it is difficult to mine deeper semantic information by solely relying on coarse-grained instance-level semantic supervision $\boldsymbol{S}_I$. As shown in Fig.1, when the instance-level semantic similarity between label A and other labels is identical, we hope A to be closer to the label containing more instances. This proximity increases the likelihood of retrieving more relevant instances, leading to more accurate results. To this end, we introduce the fine-grained semantics $\boldsymbol{S}_C$ to capture the cluster-level semantic relationships, thereby improving the retrieval performance. Hence, we adopt label distribution to explore which labels should be in close proximity or clustered together. Specifically, we regard the labels $\boldsymbol{Y}$ as the new features and conduct the K-Means algorithm to obtain cluster distribution $\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_n] \in \mathbb{R}^{k \times n}$. Note k is the number of clusters and $\boldsymbol{l}_i \in \{0,1\}^{k \times 1}$ indicates the cluster assignment of the i-th instance. We further construct the cluster-level semantic similarity relationships by $\boldsymbol{S}_C = \boldsymbol{L}^\top \boldsymbol{L}$. To be specific, we have

$$\mathbf{S}_C^{ij} = \begin{cases} 1, & \text{if } \mathcal{C}(i,j) = 1 \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where $\mathcal{C}(i,j) = 1$ represents that the i-th label and j-th label are from the same cluster.

After constructing the instance-level and cluster-level semantic relationships, we propose a dual semantic fusion learning strategy to ensure the consensus hash codes can encode multi-level semantics. Specifically, we fuse both coarse-grained semantics and fine-grained semantics to simultaneously supervise the learning of hash codes. The joint semantics can be denoted as

$$\boldsymbol{S}_J = \boldsymbol{Y}^\top \boldsymbol{Y} + \gamma \boldsymbol{L}^\top \boldsymbol{L} \quad s.t. \boldsymbol{Y} \in \mathbb{R}^{c \times n}, \boldsymbol{L} \in \mathbb{R}^{k \times n}. \tag{3}$$

We believe the cluster-level and instance-level semantics should be equally important in the joint semantic relationship. Hence, we set the adjusted factor γ to 1. Afterwards, we can obtain the overall objective function as follows:

$$\begin{aligned} \min_{\boldsymbol{H}^t, \boldsymbol{W}^t, \boldsymbol{B}} & \sum_{t=1}^{M} \|\boldsymbol{H}^t - \boldsymbol{W}^t \boldsymbol{X}^t\|^2 \\ & + \alpha \|\boldsymbol{B}^\top \boldsymbol{H}^t - r(\boldsymbol{Y}^\top \boldsymbol{Y} + \boldsymbol{L}^\top \boldsymbol{L})\|^2 \\ s.t.\ & (\boldsymbol{H}^t)^\top \mathbf{1} = \mathbf{0}, (\boldsymbol{H}^t)^\top \boldsymbol{H}^t = n\boldsymbol{I}, \\ & (\boldsymbol{W}^t)^\top \boldsymbol{W}^t = \boldsymbol{I}, \boldsymbol{B} \in \{-1, 1\}^{r \times n}. \end{aligned} \tag{4}$$

In general, we introduce the concept of fine-grained semantics to distinguish which instance semantic relationships need to be emphasized. Fine-grained semantics can provide superior supervision information, especially when the instance-level semantic relationships between labels are the same. As shown in Fig.1, such fine-grained supervision information ensures that more instances can be retrieved. Therefore, our DSFH exhibits the following advantages: 1) It preserves both intra-modal specificity and inter-modal consistency. 2) It reveals the complex semantic relationships in multi-label data, thereby enhancing the discriminative ability of consensus hash codes.

3.3 Optimization

Since the proposed objective function is non-convex and involves three variables, we design an alternating optimization strategy to solve this model. In other words, we update one variable while keeping the others fixed.

▶ $\boldsymbol{W}^t$**-Step:** Fixing the other variables but $\boldsymbol{W}^t$, the $\boldsymbol{W}^t$ step can be simplified as follows:

$$\min_{\boldsymbol{W}^t} \sum_{t=1}^{M} \|\boldsymbol{H}^t - \boldsymbol{W}^t \boldsymbol{X}^t\|^2 \quad s.t.\ (\boldsymbol{W}^t)^\top \boldsymbol{W}^t = \boldsymbol{I}. \tag{5}$$

Afterwards, Eq.5 can be expressed as the trace form, *i.e.*,

$$\max_{(\boldsymbol{W}^t)^\top \boldsymbol{W}^t = \boldsymbol{I}} tr((\boldsymbol{W}^t)^\top \boldsymbol{H}^t (\boldsymbol{X}^t)^\top). \tag{6}$$

Since the orthogonal constrain, we perform the singular value decomposition (SVD) on $\boldsymbol{H}^t(\boldsymbol{X}^t)^\top$, *i.e.*,

$$\boldsymbol{H}^t(\boldsymbol{X}^t)^\top = \boldsymbol{P^t}\boldsymbol{\Sigma}(\boldsymbol{Q^t})^\top, \tag{7}$$

where $\boldsymbol{P^t}$ and $\boldsymbol{Q^t}$ are the left and right singular values, respectively. Therefore, we can obtain $\boldsymbol{W}^t$ as follows:

$$\boldsymbol{W}^t = \boldsymbol{P^t}(\boldsymbol{Q^t})^\top. \tag{8}$$

► $\boldsymbol{H}^t$**-Step:** To optimize $\boldsymbol{H}^t$, we fix the other two variables and obtain the following sub-problem,

$$\begin{aligned}\min_{\boldsymbol{H}^t} \sum_{t=1}^{M} &\|\boldsymbol{H}^t - \boldsymbol{W}^t\boldsymbol{X}^t\|^2 \\ &+ \alpha\|\boldsymbol{B}^\top\boldsymbol{H}^t - r(\boldsymbol{Y}^\top\boldsymbol{Y} + \boldsymbol{L}^\top\boldsymbol{L})\|^2 \\ s.t.\ &(\boldsymbol{H}^t)^\top\boldsymbol{1} = \boldsymbol{0}, (\boldsymbol{H}^t)^\top\boldsymbol{H}^t = n\boldsymbol{I}.\end{aligned} \tag{9}$$

Then, we can simplify Eq.9 as

$$\begin{aligned}&\max_{\boldsymbol{H}^t} tr(\boldsymbol{G}^t(\boldsymbol{H}^t)^\top) \\ &s.t.\ (\boldsymbol{H}^t)^\top\boldsymbol{1} = \boldsymbol{0},\ (\boldsymbol{H}^t)^\top\boldsymbol{H}^t = n\boldsymbol{I},\end{aligned} \tag{10}$$

where $\boldsymbol{G}^t = \boldsymbol{W}^t\boldsymbol{X}^t + \alpha r\boldsymbol{H}(\boldsymbol{Y}^\top\boldsymbol{Y} + \boldsymbol{L}^\top\boldsymbol{L})$. To avoid the large $n \times n$ matrices (*i.e.*, $\boldsymbol{Y}^\top\boldsymbol{Y}, \boldsymbol{L}^\top\boldsymbol{L}$) to reduce both the time and space costs, we can obtain

$$\boldsymbol{G}^t = \boldsymbol{W}^t\boldsymbol{X}^t + \alpha r(\boldsymbol{H}\boldsymbol{Y}^\top\boldsymbol{Y} + \boldsymbol{H}\boldsymbol{L}^\top\boldsymbol{L}). \tag{11}$$

Then, we define $\boldsymbol{U} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n^\top$ and compute eigendecomposition on $\boldsymbol{G}^t\boldsymbol{U}(\boldsymbol{G}^t)^\top$ as follows:

$$\boldsymbol{G}^t\boldsymbol{U}(\boldsymbol{G}^t)^\top = \begin{bmatrix}\boldsymbol{Z}^t & \hat{\boldsymbol{Z}}^t\end{bmatrix}\begin{bmatrix}\Theta^t & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{Z}^t & \hat{\boldsymbol{Z}}^t\end{bmatrix}^\top, \tag{12}$$

where $\boldsymbol{\Theta}^t \in \mathbb{R}^{l\times l}$, $\boldsymbol{Z}^t \in \mathbb{R}^{r\times l}$, and l denotes the rank of $\boldsymbol{G}^t\boldsymbol{U}(\boldsymbol{G}^t)^\top$. By adopting the Gram-Schmidt process on $\hat{\boldsymbol{Z}}^t$, we can obtain $\tilde{\boldsymbol{Z}}^t \in \mathbb{R}^{r\times(r-l)}$.To solve $\boldsymbol{H}^t$, we further define $\boldsymbol{J}^t = \boldsymbol{U}(\boldsymbol{G}^t)^\top\boldsymbol{Z}^t(\Theta^t)^{-1/2} \in \mathbb{R}^{n\times l}$ and a random orthogonal matrix $\tilde{\boldsymbol{J}}^t \in \mathbb{R}^{n\times(r-l)}$. Therefore, the final optimization equation of $\boldsymbol{H}^t$ can be written as:

$$\boldsymbol{H}^t = \sqrt{n}\begin{bmatrix}\boldsymbol{Z}^t & \tilde{\boldsymbol{Z}}^t\end{bmatrix}\begin{bmatrix}\boldsymbol{J}^t & \tilde{\boldsymbol{J}}^t\end{bmatrix}^\top. \tag{13}$$

Notably, $\hat{\boldsymbol{Z}}^t$, $\tilde{\boldsymbol{Z}}^t$ and $\tilde{\boldsymbol{J}}^t$ are empty when $r = l$.

► $\boldsymbol{B}$**-Step:** Removing irrelevant variables, we can optimize $\boldsymbol{B}$ as follows:

$$\begin{aligned}&\min_{\boldsymbol{B}} \sum_{t=1}^{M}\alpha\|\boldsymbol{B}^\top\boldsymbol{H}^t - r(\boldsymbol{Y}^\top\boldsymbol{Y} + \boldsymbol{L}^\top\boldsymbol{L})\|^2 \\ &\quad s.t.\ \boldsymbol{B} \in \{-1, 1\}^{r\times n}.\end{aligned} \tag{14}$$

Then, Eq.14 can be represented as the trace form

$$\max_{\boldsymbol{B}}\ tr(\alpha r(\sum_{t=1}^{M}\boldsymbol{H}^t(\boldsymbol{Y}^\top\boldsymbol{Y} + \boldsymbol{L}^\top\boldsymbol{L}))\boldsymbol{B}^\top). \tag{15}$$

Finally, we can obtain the optimal solution $\boldsymbol{B}$, *i.e.*,

$$\boldsymbol{B} = sgn(\alpha r\sum_{t=1}^{M}(\boldsymbol{H}^t\boldsymbol{Y}^\top\boldsymbol{Y} + \boldsymbol{H}^t\boldsymbol{L}^\top\boldsymbol{L})). \tag{16}$$

3.4 Out-of-Sample Extension

To achieve out-of-sample extension, we further learn modality-specific hash function $\boldsymbol{F^t}$ by the learned hash codes $\boldsymbol{B}$. Afterwards, we adopt the following linear regression, *i.e.*,

$$\min_{\boldsymbol{F}^t}\sum_{t=1}^{M}\|\boldsymbol{B} - \boldsymbol{F}^t\boldsymbol{X}^t\|^2 + \lambda\|\boldsymbol{F}^t\|^2. \tag{17}$$

where λ is a hyper-parameter to avoid the trivial solution. Then, the hash function $\boldsymbol{F^t}$ can be solved as follows:

$$\boldsymbol{F}^t = \boldsymbol{B}(\boldsymbol{X}^t)^\top(\boldsymbol{X}^t(\boldsymbol{X}^t)^\top + \lambda\boldsymbol{I})^{-1}. \tag{18}$$

Given a new query $\boldsymbol{X}_q^t$, we can directly obtain the corresponding hash codes $\boldsymbol{B}_q^t$ of each modality, *i.e.*,

$$\boldsymbol{B}_q^t = sgn(\boldsymbol{F}^t\hat{\boldsymbol{X}}_q^t). \tag{19}$$

where $\hat{\boldsymbol{X}}_q^t$ is the RBF kernelized features of $\boldsymbol{X}_q^t$.

3.5 Time Complexity Analysis

The time complexity of our proposed DSFH mainly depends on three steps, *i.e.*, costs of updating $\boldsymbol{W}^t$, $\boldsymbol{H}^t$, and $\boldsymbol{B}$. In every iteration step, the time complexity for calculating the three variables is about $\mathcal{O}(rhn + r^2h)$, $\mathcal{O}(rhn + rcn + r^3)$, and $\mathcal{O}(rn)$, respectively. Therefore, the total time complexity is roughly $\mathcal{O}(rhn + r^2h + rcn + r^3 + rn)$. Since $r, h, c \ll n$, the time complexity is approximately $\mathcal{O}(n)$, which is linear to the size of training data.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of the proposed DSFH, we conduct numerous experiments on three benchmarks. **MIRFlickr-25K** [Huiskes and Lew, 2008] has 25,000 image-text pairs sourced from the Flickr website. Each image and text are characterized by a 512-dimensional GIST vector and a 1386-dimensional BOW vector, respectively. In our experiments, we select the instances associated with a minimum of 20 textual labels, resulting in 20,015 instances. Further, we randomly choose 2,000 instances as the query set, while the remaining image-text pairs constitute the training set. **IAPR-TC12** [Escalante *et al.*, 2010] consists of 20,000 geographical images belonging to 255 categories. The features for the images and textual descriptions are 512-dimensional GIST vectors and 2912-dimensional BOW vectors, respectively. In our experiments, we randomly select 2,000 instances from the entire dataset to form the query set, and the remaining instances are used for the training set. **NUS-WIDE** [Chua *et al.*, 2009] contains 269,648 image-text pairs, where each image is semantically associated with one or more of the 81 textual labels. The image instances are represented as 500-dimensional SIFT vectors, while the textual descriptions are represented as 1000-dimensional binary tagging vectors. In our experiments, we deliberately choose 184,710 data pairs with the top 10 most frequent class labels. We randomly select 1,867 data pairs as the query set and the remains as the training set.

Task	Method	MIRFlickr-25K				IAPR-TC12				NUS-WIDE			
		8-bit	16-bit	32-bit	64-bit	8-bit	16-bit	32-bit	64-bit	8-bit	16-bit	32-bit	64-bit
Img2Txt	RFDH	58.25	58.46	58.28	58.22	35.32	44.85	45.53	45.83	34.54	47.33	57.76	58.32
	LCMFH	67.44	69.86	70.88	69.49	32.69	42.73	44.70	45.69	55.52	63.18	64.21	64.87
	MTFH	65.82	72.45	73.12	73.64	47.14	48.32	50.45	51.98	/	/	/	/
	FCMH	72.59	73.69	75.20	75.15	46.25	49.48	51.70	53.20	64.66	65.82	66.40	67.20
	FDDH	70.38	72.97	73.33	75.81	44.37	48.04	52.29	53.89	59.75	62.07	65.79	68.60
	BATCH	71.63	73.31	73.69	74.10	44.91	48.05	50.40	52.62	63.17	65.72	66.49	67.47
	EDMH	71.18	73.22	73.87	74.00	46.39	49.86	50.85	52.43	64.56	65.83	67.16	67.44
	DAH	70.32	72.33	72.46	72.63	43.40	44.72	48.15	52.01	62.63	63.58	66.29	66.31
	ALECH	71.95	73.54	74.00	74.28	45.68	48.30	50.35	52.07	65.02	66.08	67.85	68.22
	WASH	71.18	72.53	72.72	73.03	46.75	48.25	51.00	53.45	62.45	64.04	64.18	63.34
	AMSH	72.56	73.78	74.29	74.89	46.81	49.05	51.62	53.66	64.63	65.37	67.60	67.34
	DSFH	**74.62**	**75.41**	**75.13**	**76.66**	**50.53**	**52.81**	**54.64**	**56.73**	**67.92**	**68.07**	**67.94**	**69.34**
Txt2Img	RFDH	58.64	57.66	57.97	57.85	34.83	45.52	46.40	57.54	35.48	53.66	58.22	62.73
	LCMFH	70.93	74.48	74.65	74.15	34.69	49.86	53.68	56.42	58.43	67.08	72.23	73.64
	MTFH	69.42	79.44	81.73	80.24	52.27	57.36	60.92	62.33	/	/	/	/
	FCMH	79.76	81.75	**83.57**	83.69	53.47	58.50	61.92	65.13	75.57	77.64	78.84	80.76
	FDDH	74.53	78.09	79.45	82.54	49.33	55.16	61.14	65.00	70.20	74.79	77.98	81.58
	BATCH	79.01	80.65	81.35	82.05	52.75	57.77	61.85	64.88	76.57	77.58	79.41	80.20
	EDMH	79.59	81.53	82.61	83.20	53.61	58.70	60.53	63.53	73.12	78.50	79.61	79.64
	DAH	77.47	79.20	81.03	81.55	49.80	54.75	58.17	61.17	73.82	77.45	78.05	79.09
	ALECH	78.06	80.75	81.73	82.15	52.55	57.74	61.44	64.61	76.26	77.64	78.89	79.77
	WASH	76.69	78.53	79.59	79.77	50.89	54.25	61.50	65.02	73.31	77.70	80.39	81.09
	AMSH	80.12	81.69	82.90	82.86	53.89	58.87	62.98	66.32	77.05	78.46	80.12	80.83
	DSFH	**80.89**	**82.60**	82.58	**84.25**	**56.46**	**61.03**	**64.56**	**67.55**	**78.56**	**80.69**	**81.19**	**82.58**

Table 1: Performance comparison (mAP) of DSFH and baselines on three datasets.

Method	Img2Txt			Txt2Img		
	16-bit	32-bit	64-bit	16-bit	32-bit	64-bit
DADH	80.20	80.72	81.79	79.20	79.59	80.64
MDCH	80.63	81.66	82.32	80.48	82.15	83.37
DMFH	78.02	79.19	79.46	79.78	80.97	81.01
DCMHT	82.63	82.72	84.41	81.16	81.37	82.81
CMGCAH	79.01	80.30	81.23	78.23	79.32	80.45
ALECH$_{cnn}$	83.23	84.55	85.12	80.93	82.16	82.55
WASH$_{cnn}$	81.60	82.21	82.70	79.47	79.94	80.51
AMSH$_{cnn}$	84.52	85.66	86.21	81.80	83.21	83.51
DSFH$_{cnn}$	**84.93**	**85.97**	**86.35**	**82.45**	**83.71**	**84.01**

Table 2: Performance comparison (mAP) of DSFH and other deep methods on MIRFlickr-25K.

4.2 Baselines and Evaluation Metrics

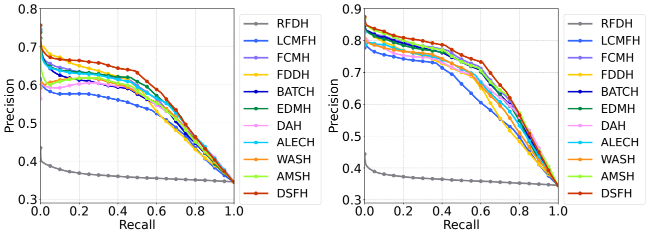
To comprehensively evaluate our proposed DSFH, we compare DSFH with 16 competitive baselines. They are 11 shallow CMH methods (*i.e.*, RFDH [Wang *et al.*, 2017], LCMFH [Wang *et al.*, 2018], MTFH [Liu *et al.*, 2019], FCMH [Wang *et al.*, 2022], FDDH [Liu *et al.*, 2022], BATCH [Wang *et al.*, 2021], EDMH [Chen *et al.*, 2022], DAH [Zhang *et al.*, 2023b], ALECH [Li *et al.*, 2023], WASH [Zhang *et al.*, 2023a] and AMSH [Luo *et al.*, 2023]) and 5 deep CMH methods (DADH [Bai *et al.*, 2020], MDCH [Lin *et al.*, 2021], DMFH [Nie *et al.*, 2021], DCMHT [Tu *et al.*, 2022], and CMGCAH [Ou *et al.*, 2023]). Following the common evaluation metrics in the CMR field, we adopt the Mean Average Precision (mAP), the Precision-Recall (PR) curve, and the Precision at Top-N (Precision@TopN) curve to show the performance of the proposed DSFH.

4.3 Experimental Setup

In our experiments, we adopt two common CMR tasks (*i.e.*, Img2Txt and Txt2Img) to validate the performance of all competition methods. Img2Txt and Txt2Img represent using images to retrieve relevant texts and using texts to retrieve relevant images, respectively. Experimentally, we set hash lengths from 8 to 64 bits. The number of anchors for RBF is set to 1500, and the maximum iteration step is set to 10. The hyper-parameters α and λ are set to $\{10^{-3}, 10^{-3}\}$, $\{10^{-4}, 10^{-3}\}$, and $\{10^{-4}, 10^{-4}\}$ for MIRFlickr-25K, IAPR-TC12, and NUS-WIDE, respectively. In addition, the number of clusters k is 400, 300, and 400 for three datasets, respectively. For fair comparisons, the experimental settings of all baselines are consistent with those reported in the original papers. To comprehensively evaluate the retrieval performance, we conduct extensive experiments on a Windows server equipped with 64GB of RAM.

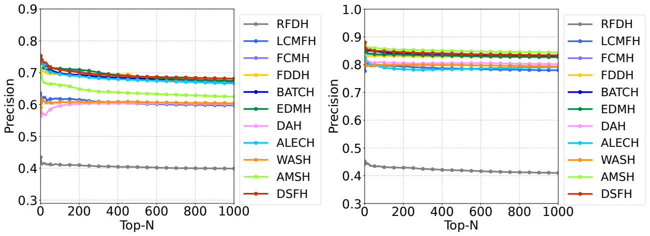
4.4 Comparison with Shallow Baselines

To show the effectiveness of the proposed DSFH, we compare our method with 11 shallow baselines on three benchmark datasets. The mAP scores of all methods with differ-



(a) Img2Txt (b) Txt2Img

Figure 3: PR curves with 8 bits on NUS-WIDE.



(a) Img2Txt (b) Txt2Img

Figure 4: Precision@TopN curves with 8 bits on NUS-WIDE.

ent hash lengths are reported in Table 1. Additionally, Fig.3 and Fig.4 depict the corresponding PR curves and the Precision@TopN curves on the NUS-WIDE dataset, respectively. Note here that we cannot perform MTFH on the NUS-WIDE dataset due to being out of memory. According to the experimental results, we can obtain the following observations: 1) Clearly, our DSFH outperforms nearly all baselines under various experimental conditions. Specifically, For Img2Txt tasks with different hash lengths, the improvement of the retrieval performance ranges from a minimum of 0.34% to a maximum of 3.76%. For Txt2Img tasks with different hash lengths, the gains of the retrieval performance are between 0.77% and 2.65%. Such improvements could be attributed to the dual semantic fusion scheme in DSFH, which can enhance the discriminative ability of hash codes.

2) DSFH boasts the largest area under the PR curves, which indicates that DSFH achieves optimal performance in all cases. Moreover, the Precision@TopN curves show that DSFH maintains high precision with different N values. Note here that DSFH appears to be inferior to AMSH in Fig.4(b). This could be because AMSH introduces the adaptive margin matrix factors on semantic labels, thereby alleviating the rigid zero-one linear regression.

3) In general, the performance of all methods on Txt2Img tasks surpasses that of Img2Txt tasks. This discrepancy could be attributed to the text features having more discriminative semantics, whereas image features are relatively more abstract. In addition, the mAP scores for all methods increase with the length of hash codes because longer hash codes can encode more information.

4.5 Comparison with Deep Baselines

To further show the performance of our DSFH, we compare it with several shallow methods with deep features and deep methods on the MIRFlickr-25K dataset. The mAP results are shown in Table 2. Specifically, the first five rows represent the latest deep methods, and the latter four rows correspond to shallow methods fed with deep features. The deep image features are 4096-dimensional vectors extracted by the pre-trained CNN-F model [Chatfield *et al.*, 2014]. In general, although DSFH$_{cnn}$ is not an end-to-end deep method, it still outperforms these comparison methods. This could be because DSFH$_{cnn}$ can encode the more semantic information of multi-label data into the discriminative hash codes.

Method	MIRFlickr-25K		IAPR-TC12		NUS-WIDE	
	8-bit	64-bit	8-bit	64-bit	8-bit	64-bit
RFDH	24.29	76.38	58.32	118.16	162.62	624.46
LCMFH	1.63	3.42	8.43	16.32	11.17	27.89
MTFH	16.53	190.24	46.37	326.51	/	/
FDDH	7.57	27.81	22.87	28.59	82.51	307.64
BATCH	0.18	0.58	0.25	0.83	1.80	3.51
EDMH	1.84	9.63	5.85	22.86	11.15	69.11
DAH	0.13	0.59	0.33	0.88	0.78	5.56
ALECH	0.62	1.15	1.06	1.92	3.31	9.15
WASH	1.65	2.92	3.58	6.65	11.54	23.39
AMSH	3.33	6.93	3.82	8.13	28.66	59.84
DSFH	0.57	1.54	0.56	1.79	4.70	17.00

Table 3: Training time (seconds) of different methods with 8-bit and 64-bit hash codes on three datasets.

4.6 Parameter Analysis

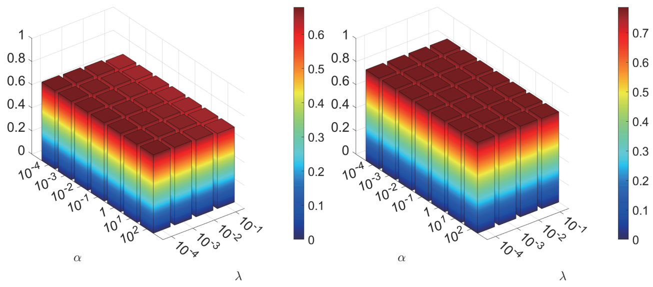
In this section, we perform experiments to explore the parameter sensitivity of our DSFH, i.e., α, λ, and k. α is the trade-off coefficient that balances cross-modal consistency and specificity. λ is the regularization parameter, and k is the number of clusters for K-Means. Experimentally, we set the parameters α and λ to vary from 10^{-4} to 10^{2} and from 10^{-4} to 10^{-1}, respectively. In addition, we set the k value as $\{100, 200, 300, 400, 500\}$. To study the effects of these parameters, we adopt the grid search method to obtain the optimal parameters. Firstly, we employ the grid search method to vary the α and λ values. The corresponding mAP results on NUS-WIDE are shown in Fig.5. Notably, DSFH shows strong retrieval performance under a large range of α and λ values. In addition, The mAP results with various k values on NUS-WIDE are shown in Fig.6. It can be observed that the mAP curves exhibit a uni-modal trend. If the value k is too small, the clustering result could not capture the structure of the label distribution. Conversely, if the value k is too large, the clustering results may be too fine-grained, which leads to meaningless clustering redundancy. Therefore, we can choose the optimum k to obtain the best performance.

4.7 Time Cost Analysis

To further demonstrate the training efficiency of our DSFH, we show the training time of all methods with 8 and 64 bits on the three datasets in Table 3. Compared to most baselines, DSFH offers some computation advantages, mainly for two reasons: 1) Only using simple matrix operations to optimize the overall hash model. 2) Avoiding the usage of the large $n \times n$ pairwise similarity matrix. However, the training

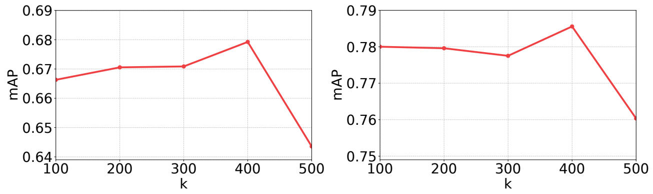
Task	Method	MIRFlickr-25K				IAPR-TC12				NUS-WIDE			
		8-bit	16-bit	32-bit	64-bit	8-bit	16-bit	32-bit	64-bit	8-bit	16-bit	32-bit	64-bit
Img2Txt	DSFH-w/o ISP	60.12	61.70	59.86	58.99	34.31	33.41	33.77	33.99	52.87	52.18	52.97	52.90
	DSFH-w/o CSP	72.87	74.00	74.50	74.79	49.60	52.58	53.85	55.70	66.46	67.54	67.88	68.20
	DSFH-full	**74.62**	**75.41**	**75.13**	**76.66**	**50.53**	**52.81**	**54.64**	**56.73**	**67.92**	**68.07**	**67.94**	**69.34**
Txt2Img	DSFH-w/o ISP	61.15	62.55	61.82	61.38	35.72	35.22	36.01	36.96	62.25	60.83	60.94	61.38
	DSFH-w/o CSP	79.73	81.41	82.27	82.90	55.40	60.71	63.89	66.91	76.99	79.71	80.73	81.29
	DSFH-full	**80.89**	**82.60**	**82.58**	**84.25**	**56.46**	**61.03**	**64.56**	**67.55**	**78.56**	**80.69**	**81.19**	**82.58**

Table 4: Performance comparison (mAP) of DSFH and its variants on three datasets.



(a) Img2Txt (b) Txt2Img

Figure 5: Sensitivity analysis of DSFH with 8 bits for parameters α and λ on NUS-WIDE.



(a) Img2Txt (b) Txt2Img

Figure 6: Sensitivity analysis of DSFH with 8 bits for parameter k on NUS-WIDE.

time of DSFH is slower than that of DAH and BATCH. This could be because DAH and BATCH directly learn the common latent representations of all modalities and ignore fine-grained semantic information. In summary, DSFH remains highly competitive in terms of training time cost.

4.8 Convergence Analysis

We conduct the experiments to analyze the convergence properties of the proposed DSFH. Specifically, we draw the curves of the objective value and mAP scores with respect to iteration steps on NUS-WIDE in Fig.7. We can conclude the following observations: 1) The objective value decreases rapidly within the five iterations, and then converges to approximately zero and remains stable after the 10-th iteration. 2) The mAP curves of Txt2Img and Img2Txt tasks both steadily ascend with the iteration steps. Then, DSFH reaches the highest mAP score within the 10-th iteration. In general, these results demonstrate that DSFH enjoys both rapid initial convergence and subsequent stable characteristics.

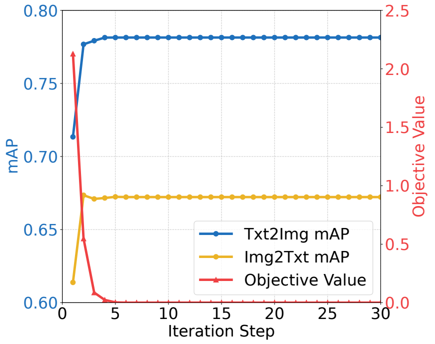


Figure 7: Convergence analysis with 8 bits on NUS-WIDE.

4.9 Ablation Experiments

Our proposed DSFH has two primary modules, i.e., instance-level semantic preservation (ISP) and cluster-level semantic preservation (CSP). Therefore, we meticulously design two variants, namely DSFH-w/o ISP and DSFH-w/o CSP. Compared to the original objective function in Eq.4, the variant DSFH-w/o ISP removes the $\boldsymbol{L}^\top\boldsymbol{L}$ term, and the DSFH-w/o CSP removes $\boldsymbol{Y}^\top\boldsymbol{Y}$ term, respectively. As shown in Table 4, we can observe that DSFH-w/o ISP exhibits a notable performance degradation of over 10%, because instance-level semantics offer the primary supervision in hash codes learning. In addition, DSFH-w/o CSP exhibits about 1% of performance degradation, which indicates the cluster-level semantic relationship can capture finer-grained semantics in multi-labels, thereby enhancing the retrieval performance.

5 Conclusion

In this paper, we propose a novel dual semantic fusion hashing (DSFH) for multi-label cross-modal retrieval. We first propose modality-specific learning and consensus hashing learning to preserve the intrinsic consistency and specificity across different modalities. Then, we utilize label distribution to construct cluster-level semantic relationships, thereby exploring finer-grained semantic similarities between all instances. Finally, we propose a dual semantic fusion learning strategy that embeds both the coarse-grained and fine-grained semantics into consensus hash codes, thereby simultaneously preserving instance-level and cluster-level semantic relationships. Extensive experiments on three benchmarks demonstrate the outstanding performance of DSFH compared with state-of-the-art CMH baselines.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 62372315), the Sichuan Science and Technology Program (Grant no. 24ZDZX0007, 2023ZYD0143, 24QYCX0375, 24QYCX0354, and 24ZHSF0545), the Mianyang Science and Technology Program (Grant no. 2022ZYDF089), and the Base Strengthening Program of National Defense Science and Technology (Grant no. 2022-JCJQ-JJ-0292).

References

[Bai *et al.*, 2020] Cong Bai, Chao Zeng, Qing Ma, Jinglin Zhang, and Shengyong Chen. Deep adversarial discrete hashing for cross-modal retrieval. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 525–531, 2020.

[Chatfield *et al.*, 2014] K Chatfield, K Simonyan, A Vedaldi, and A Zisserman. Return of the devil in the details: delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference 2014*. British Machine Vision Association, 2014.

[Chen *et al.*, 2020] Zhen-Duo Chen, Chuan-Xiang Li, Xin Luo, Liqiang Nie, Wei Zhang, and Xin-Shun Xu. 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, 2020.

[Chen *et al.*, 2022] Yong Chen, Hui Zhang, Zhibao Tian, Jun Wang, Dell Zhang, and Xuelong Li. Enhanced discrete multi-modal hashing: More constraints yet less time to learn. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):1177–1190, 2022.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009.

[Escalante *et al.*, 2010] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villasenor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer vision and image understanding*, 114(4):419–428, 2010.

[Huiskes and Lew, 2008] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, 2008.

[Li *et al.*, 2023] Huaxiong Li, Chao Zhang, Xiuyi Jia, Yang Gao, and Chunlin Chen. Adaptive label correlation based asymmetric discrete hashing for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1185–1199, 2023.

[Lin *et al.*, 2015] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Semantics-preserving hashing for cross-view retrieval. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3864–3872, 2015.

[Lin *et al.*, 2021] Qiubin Lin, Wenming Cao, Zhiquan He, and Zhihai He. Mask cross-modal hashing networks. *IEEE Transactions on Multimedia*, 23:550–558, 2021.

[Liu *et al.*, 2012] W. Liu, Jun Wang, R. Ji, Yu-Gang Jiang, and Shih-Fu Chang. Supervised hashing with kernels. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2074–2081, 2012.

[Liu *et al.*, 2019] Xin Liu, Zhikai Hu, Haibin Ling, and Yiuming Cheung. Mtfh: A matrix tri-factorization hashing framework for efficient cross-modal retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):964–981, 2019.

[Liu *et al.*, 2022] Xin Liu, Xingzhi Wang, and Yiu-Ming Cheung. Fddh: Fast discriminative discrete hashing for large-scale cross-modal retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11):6306–6320, 2022.

[Liu *et al.*, 2024] Haoran Liu, Ying Ma, Ming Yan, Yingke Chen, Dezhong Peng, and Xu Wang. Dida: Disambiguated domain alignment for cross-domain retrieval with partial labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3612–3620, 2024.

[Luo *et al.*, 2018] Xin Luo, Peng-Fei Zhang, Ye Wu, Zhen-Duo Chen, Hua-Junjie Huang, and Xin-Shun Xu. Asymmetric discrete cross-modal hashing. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 204–212, 2018.

[Luo *et al.*, 2023] Kaiyi Luo, Chao Zhang, Huaxiong Li, Xiuyi Jia, and Chunlin Chen. Adaptive marginalized semantic hashing for unpaired cross-modal retrieval. *IEEE Transactions on Multimedia*, 2023.

[Nie *et al.*, 2021] Xiushan Nie, Bowei Wang, Jiajia Li, Fanchang Hao, Muwei Jian, and Yilong Yin. Deep multiscale fusion hashing for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1):401–410, 2021.

[Ou *et al.*, 2023] Weihua Ou, Jiaxin Deng, Lei Zhang, Jianping Gou, and Quan Zhou. Cross-modal generation and pair correlation alignment hashing. *IEEE Transactions on Intelligent Transportation Systems*, 24(3):3018–3026, 2023.

[Qin *et al.*, 2021] Jianyang Qin, Lunke Fei, Shaohua Teng, Wei Zhang, Dongning Liu, Genping Zhao, and Haoliang Yuan. Discrete semantic matrix factorization hashing for cross-modal retrieval. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1550–1557, 2021.

[Ranjan *et al.*, 2015] Viresh Ranjan, Nikhil Rasiwasia, and C. V. Jawahar. Multi-label cross-modal retrieval. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4094–4102, 2015.

[Shen *et al.*, 2021] Heng Tao Shen, Luchen Liu, Yang Yang, Xing Xu, Zi Huang, Fumin Shen, and Richang Hong. Exploiting subspace relation in semantic labels for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 33(10):3351–3365, 2021.

[Sun *et al.*, 2023a] Yuan Sun, Dezhong Peng, Jian Dai, and Zhenwen Ren. Stepwise refinement short hashing for image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6501–6509, 2023.

[Sun *et al.*, 2023b] Yuan Sun, Xu Wang, Dezhong Peng, Zhenwen Ren, and Xiaobo Shen. Hierarchical hashing learning for image set classification. *IEEE Transactions on Image Processing*, 32:1732–1744, 2023.

[Sun *et al.*, 2024a] Yuan Sun, Jian Dai, Zhenwen Ren, Yingke Chen, Dezhong Peng, and Peng Hu. Dual self-paced cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15184–15192, 2024.

[Sun *et al.*, 2024b] Yuan Sun, Zhenwen Ren, Peng Hu, Dezhong Peng, and Xu Wang. Hierarchical consensus hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26:824–836, 2024.

[Teng *et al.*, 2023] Shaohua Teng, Chengzhen Ning, Wei Zhang, NaiQi Wu, and Ying Zeng. Fast asymmetric and discrete cross-modal hashing with semantic consistency. *IEEE Transactions on Computational Social Systems*, 10(2):577–589, 2023.

[Tu *et al.*, 2022] Junfeng Tu, Xueliang Liu, Zongxiang Lin, Richang Hong, and Meng Wang. Differentiable cross-modal hashing via multimodal transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 453–461, 2022.

[Wang *et al.*, 2016] Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):2010–2023, 2016.

[Wang *et al.*, 2017] Di Wang, Quan Wang, and Xinbo Gao. Robust and flexible discrete hashing for cross-modal similarity search. *IEEE transactions on circuits and systems for video technology*, 28(10):2703–2715, 2017.

[Wang *et al.*, 2018] Di Wang, Xinbo Gao, Xiumei Wang, and Lihuo He. Label consistent matrix factorization hashing for large-scale cross-modal similarity search. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2466–2479, 2018.

[Wang *et al.*, 2019] Di Wang, Xinbo Gao, Xiumei Wang, and Lihuo He. Label consistent matrix factorization hashing for large-scale cross-modal similarity search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2466–2479, 2019.

[Wang *et al.*, 2021] Yongxin Wang, Xin Luo, Liqiang Nie, Jingkuan Song, Wei Zhang, and Xin-Shun Xu. Batch: A scalable asymmetric discrete cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 33(11):3507–3519, 2021.

[Wang *et al.*, 2022] Yongxin Wang, Zhen-Duo Chen, Xin Luo, Rui Li, and Xin-Shun Xu. Fast cross-modal hashing with global and local similarity embedding. *IEEE Transactions on Cybernetics*, 52(10):10064–10077, 2022.

[Wang *et al.*, 2023] Xu Wang, Dezhong Peng, Ming Yan, and Peng Hu. Correspondence-free domain alignment for unsupervised cross-domain image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10200–10208, 2023.

[Xu *et al.*, 2017] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing*, 26(5):2494–2507, 2017.

[Xu *et al.*, 2024] Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16129–16137, 2024.

[Zhang and Li, 2014] Dongqing Zhang and Wu-Jun Li. Large-scale supervised multimodal hashing with semantic correlation maximization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), Jun. 2014.

[Zhang and Wu, 2022] Donglin Zhang and Xiao-Jun Wu. Robust and discrete matrix factorization hashing for cross-modal retrieval. *Pattern Recognition*, 122:108343, 2022.

[Zhang *et al.*, 2023a] Chao Zhang, Huaxiong Li, Yang Gao, and Chunlin Chen. Weakly-supervised enhanced semantic-aware hashing for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):6475–6488, 2023.

[Zhang *et al.*, 2023b] Donglin Zhang, Xiao-Jun Wu, Tianyang Xu, and He-Feng Yin. Dah: Discrete asymmetric hashing for efficient cross-media retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1365–1378, 2023.

[Zhu *et al.*, 2024] Lei Zhu, Chaoqun Zheng, Weili Guan, Jingjing Li, Yang Yang, and Heng Tao Shen. Multi-modal hashing for efficient multimedia retrieval: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):239–260, 2024.