Deep Evidential Hashing for Trustworthy Cross-Modal Retrieval

Yuan Li[1], Liangli Zhen[2], Yuan Sun[1,3], Dezhong Peng[1,3], Xi Peng[1], Peng Hu[1]*

[1]College of Computer Science, Sichuan University
[2]Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore
[3]Sichuan National Innovation New Vision UHD Video Technology Co., Ltd., Chengdu 610095, China
{blackant1997, llzhen}@outlook.com, sunyuan_work@163.com, pengdz@scu.edu.cn, {pengx.gm, penghu.ml}@gmail.com

Abstract

Cross-modal hashing provides an efficient solution for retrieval tasks across various modalities, such as images and text. However, most existing methods are deterministic models, which overlook the reliability associated with the retrieved results. This omission renders them unreliable for determining matches between data pairs based solely on Hamming distance. To bridge the gap, in this paper, we propose a novel method called Deep Evidential Cross-modal Hashing (DECH). This method equips hashing models with the ability to quantify the reliability level of the association between a query sample and each corresponding retrieved sample, bringing a new dimension of reliability to the cross-modal retrieval process. To achieve this, our method addresses two key challenges: i) To leverage evidential theory in guiding the model to learn hash codes, we design a novel evidence acquisition module to collect evidence and place the evidence captured by hash codes on a Beta distribution to derive a binomial opinion. Unlike existing evidential learning approaches that rely on classifiers, our method collects evidence directly through hash codes. ii) To tackle the task-oriented challenge, we first introduce a method to update the derived binomial opinion, allowing it to present the uncertainty caused by conflicting evidence. Following this manner, we present a strategy to precisely evaluate the reliability level of retrieved results, culminating in performance improvement. We validate the efficacy of our DECH through extensive experimentation on four benchmark datasets. The experimental results demonstrate our superior performance compared to 12 state-of-the-art methods.

Code — https://github.com/blackant-dev/DECH

Introduction

Cross-modal retrieval, a crucial task in information retrieval, involves searching for semantically related data across heterogeneous modalities, such as images and texts (Hu et al. 2023b,a; Feng et al. 2023). To this end, cross-modal hashing methods have gained popularity due to their computational and storage efficiency in handling large-scale multimodal data (Hu et al. 2021; Sun et al. 2023, 2024a,b). These methods typically map high-dimensional data into compact binary codes, enabling fast retrieval via simple logical operations (e.g., XOR) (Jiang and Li 2017; Fang, Zhang, and Ren 2019; Hu, Peng, and Peng 2024). Although these methods achieve promising performance, almost all of them are deterministic, limiting their ability to account for retrieval reliability. This limitation would lead to unreliable matching results based solely on the Hamming distance as shown in Figure 1, particularly in practical scenarios where deep neural networks tend to overestimate their predictions (Guo et al. 2017).

Figure 1: An example of our observation and our basic idea. The figure displays three image results retrieved from the MS-COCO dataset using a sentence query. The Hamming similarity between each image and the sentence, which is the ranking criterion for retrieval, is shown in the lower-left corner of the image. The reliability level, estimated by our method, is shown below the image. From the figure, one can see that the second image is the only one that truly matches the query text, even though the other images have higher Hamming similarity scores (i.e., smaller Hamming distance). This shows the limitation of using similarity as the only criterion for cross-modal retrieval. However, our method can provide a reliability level for each retrieved result, which reflects the confidence/reliability of the match, thereby embracing more reliable and accurate retrieval.

To address reliability estimation, numerous methods have been presented by introducing various uncertainty modeling techniques into deep neural networks, such as Bayesian networks (Sahlin, Helle, and Perepolkin 2021), deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017), evidential deep learning (EDL) (Sensoy, Kaplan, and Kandemir 2018; Jøsang 2016). Among these approaches, EDL has gained attention due to its outstanding efficiency, as it views the class predictions of neural networks as subjective opin-

*Corresponding author.

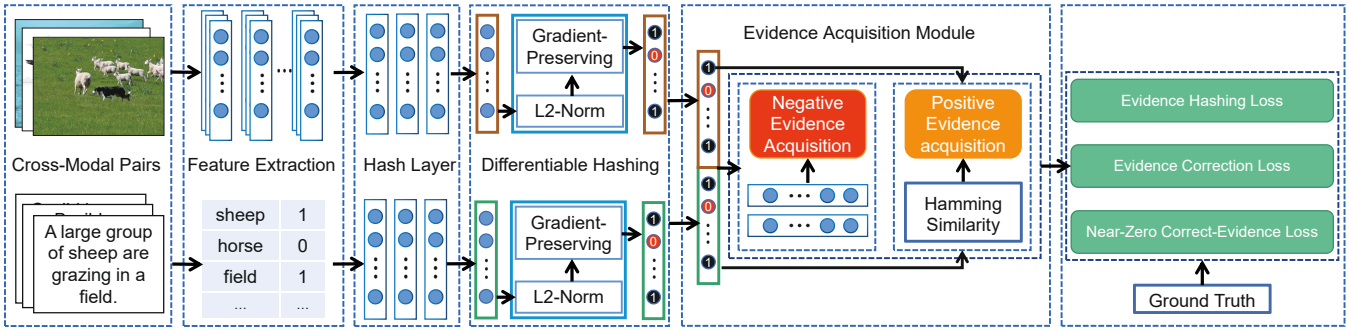


Figure 2: The overview of our Deep Evidential Cross-modal Hashing framework. It consists of four main components: feature extraction, differentiable hashing, evidence acquisition module, and our proposed loss functions. First, image and text samples are fed into modality-specific networks to obtain their continuous-valued representations, which are then binarized by our differentiable hashing module. Next, positive evidence and negative evidence for each cross-modal pair are extracted and used to update the parameters of a Beta distribution, which models the binomial opinion. Finally, DECH uses our proposed loss functions to learn the parameters of the Beta distribution.

ions and directly infers uncertainty (Sensoy, Kaplan, and Kandemir 2018; Jøsang 2016). Although EDL has shown considerable improvement in uncertainty estimation, it is still challenging and less touched in cross-modal hashing due to the difference between the tasks. Specifically, existing EDL methods (Sensoy, Kaplan, and Kandemir 2018) mainly focus on classification tasks, viewing the output as evidence — a quantifiable metric derived from data, indicating the amount of support for assigning a given sample to a particular class. In contrast, cross-modal hashing aims at learning common binary representations for retrieval, which makes it difficult for traditional EDL techniques to model the evidence without a classifier. Furthermore, directly optimizing binary representations is an NP-hard and non-differentiable problem. Existing methods (Zhang, Peng, and Yuan 2018; Zhang and Peng 2019; Kumar and Udupa 2011) often relax the hash codes to continuous values to train the model, which results in a gap between upstream training and downstream inference, leading to hashing performance degradation. As a result, these fundamental differences present significant obstacles in applying EDL to cross-modal hashing.

To overcome these challenges, we present a novel framework, as depicted in Figure 2, aimed at enhancing the learning of effective hash codes while quantifying the reliability of retrieved results. Specifically, 1) We introduce an evidence acquisition module that models cross-modal matching as a binomial opinion, equivalent to a Beta distribution. This module collects positive and negative evidence to train the Beta distribution parameters, effectively guiding hash code learning through evidential theory. 2) To address the task-oriented challenges, we adjust the derived binomial opinion associated with the cross-modal pair, enabling it to quantify the uncertainty caused by conflicting evidence. Based on this modified binomial opinion, we propose a novel approach that allows us to estimate the reliability of retrieved results both accurately and explicitly. 3) To handle the binarization challenge, we propose a novel module called Differentiable Hashing (DH), which enables the discrete optimization of binary codes without continuous-value relaxation during the network training process. Our contributions are summarized as follows:

- We propose a novel framework to endow cross-modal hashing with the ability to capture the reliability of retrieved results. To the best of our knowledge, our DECH might serve as the first approach to explore trustworthy retrieval by cross-modal hashing.
- We present a novel evidence acquisition module specifically designed to collect both positive and negative evidence for cross-modal pairs, and use it to train the Beta distribution, thus enabling evidence quantization in cross-modal retrieval.
- We conduct extensive experiments on four widely recognized datasets and rigorously compare our approach with 12 state-of-the-art approaches. The results demonstrate that our proposed approach outperforms these methods across various metrics.

Related Work

Cross-modal hashing methods are generally classified into supervised and unsupervised types. A key challenge in unsupervised cross-modal hashing is learning shared semantics without class labels. To address this, Zhang et al. explore the underlying manifold structure across modalities to facilitate hashing learning (Zhang, Peng, and Yuan 2018). However, the lack of semantic labels limits the potential of these methods. In contrast, supervised methods leverage labeled data to improve performance. For example, Huang et al. design an objective function to penalize samples that do not preserve the consistency of the label (Huang et al. 2017). Despite their effectiveness, they are almost all deterministic models that cannot capture the reliability of trustworthy retrieval. To measure the reliability of results, Evidential Deep Learning (EDL) has gained widespread attention for its ability to explicitly quantify uncertainty (Sensoy, Kaplan, and Kandemir 2018; Han et al. 2021, 2022). However, these methods typically rely on classifiers to achieve the objectives. In contrast, cross-modal hashing, which focuses on discrete rep-

resentation learning, often lacks a specific classifier, posing unique challenges in integrating EDL techniques.

Methodology

Problem Formulation

Consider a training dataset $\mathcal{D}$, comprising n paired samples $(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, \ldots, n$, drawn from different modalities. In this paper, we focus on image-text pairs, where $\mathbf{x}_i$ and $\mathbf{y}_i$ represent the i-th image and text samples, respectively. Additionally, each pair is also associated with a multi-label vector $\mathbf{l}_i = [l_{i1}, l_{i2}, \ldots, l_{ic}] \in \mathbb{R}^c$, where c denotes the total number of categories. Here, $l_{ik} = 1$ indicates that the i-th pair belongs to the k-th category, while $l_{ik} = 0$ otherwise.

The objective of cross-modal hashing is to learn two hashing functions that map image and text samples to binary codes. Specifically, the hash codes for the i-th image sample are denoted as $\mathbf{b}_i^x \in \{-1, +1\}^L$, and the hash codes for the j-th text sample are $\mathbf{b}_j^y \in \{-1, +1\}^L$, where L represents the length of the hash codes. The Hamming Distance is used as an efficient metric for quantifying the similarity between different modalities. The Hamming Distance between $\mathbf{b}_i^x$ and $\mathbf{b}_j^y$ is computed as follows:

$$H_d\left(\mathbf{b}_i^x, \mathbf{b}_j^y\right) = \frac{1}{2}\left(L - \langle \mathbf{b}_i^x, \mathbf{b}_j^y \rangle\right), \tag{1}$$

Subsequently, the Hamming similarity between the i-th image sample, represented by its hash code $\mathbf{b}_i^x$, and the j-th text sample, represented by its hash code $\mathbf{b}_j^y$, can be quantitatively expressed through the inner product as follows:

$$H_s(\mathbf{b}_i^x, \mathbf{b}_j^y) = \frac{1}{L}\langle \mathbf{b}_i^x, \mathbf{b}_j^y \rangle \in [-1, 1]. \tag{2}$$

To learn the hashing functions and estimate the reliability, we introduce a Generalized Deep Evidential Cross-modal Hashing framework. This framework comprises two independent networks for learning hash codes from the different modalities, along with a network dedicated to evidence acquisition. The overall objective function is:

$$\underset{\Theta^x, \Theta^y, \Theta^z}{\arg\min} \left(\mathcal{L}_e + \lambda \mathcal{L}_{kl} + \gamma \mathcal{L}_{nzce}\right), \tag{3}$$

where $\mathcal{L}_e$ is the evidence hashing loss, $\mathcal{L}_{kl}$ is the evidence correction loss, and $\mathcal{L}_{nzce}$ is the near-zero correct-evidence loss. The parameters λ and γ are hyperparameters that control the trade-off between these loss functions, and Θ^x, Θ^y, and Θ^z represent the weight parameters of the networks. The subsequent sections will provide a detailed explanation of these loss functions.

Beta Evidential Learning

Inspired by the Dempster–Shafer Theory of Evidence (Dempster 1968), we employ evidential theory to form binomial opinions within cross-modal data pairs. Cross-modal hashing seeks to learn common binary representations for retrieval, a task that poses challenges for traditional EDL techniques, which typically rely on classifiers to quantify evidence. To address this, we design a novel evidence acquisition module that can estimate positive and negative evidence for any data pair without using any classifier.

Specifically, we utilize the Hamming similarity between data pairs to quantify the positive evidence, denoted as $PE(\mathbf{b}^x, \mathbf{b}^y)$, representing the amount of support for the pair $(\mathbf{x}, \mathbf{y})$ to match. Additionally, a subnetwork quantifies the negative evidence, denoted as $NE(\mathbf{b}^x, \mathbf{b}^y)$, indicating the degree of support for their non-matching. These are calculated as follows:

$$\begin{aligned} PE(\mathbf{b}^x, \mathbf{b}^y) &= e^{\frac{H_s(\mathbf{b}^x, \mathbf{b}^y)}{\tau}}, \\ NE(\mathbf{b}^x, \mathbf{b}^y) &= e^{\frac{g(\mathbf{b}^x, \mathbf{b}^y; \Theta^z)}{\tau}}, \end{aligned} \tag{4}$$

where $e(\cdot)$ is an evidence activation function that ensures the generated evidence is non-negative, and $\tau \in (0, 1]$ controls the magnitude of the generated evidence. The functions $H_s(\mathbf{b}^x, \mathbf{b}^y)$ and $g(\mathbf{b}^x, \mathbf{b}^y; \theta^z)$ are evidence generation functions, with $H_s(\mathbf{b}^x, \mathbf{b}^y)$ computing the Hamming similarity between $\mathbf{b}^x$ and $\mathbf{b}^y$ and $g(\mathbf{b}^x, \mathbf{b}^y; \theta^z)$ being a parameterized function with θ^z as its parameters.

Using the evidence, belief and disbelief masses are assigned to the binomial opinion:

$$b = \frac{PE(\mathbf{b}^x, \mathbf{b}^y)}{\phi}, d = \frac{NE(\mathbf{b}^x, \mathbf{b}^y)}{\phi}, \text{ and } u = \frac{2}{\phi}, \tag{5}$$

where b is belief mass, d is disbelief mass, u is uncertainty mass, and $\phi = PE(\mathbf{b}^x, \mathbf{b}^y) + NE(\mathbf{b}^x, \mathbf{b}^y) + 2$. According to subjective logic (Jøsang 2016) and EDL (Sensoy, Kaplan, and Kandemir 2018), the belief assignment corresponds to the Beta distribution $Beta\left(p|\alpha(\mathbf{b}^x, \mathbf{b}^y), \beta(\mathbf{b}^x, \mathbf{b}^y)\right)$, where $p \in [0, 1]$ represents the matching probability, parameterized by the respective evidence. The correspondence between the parameters of the Beta distribution and the evidence is as follows:

$$\begin{aligned} \alpha(\mathbf{b}^x, \mathbf{b}^y) &= PE(\mathbf{b}^x, \mathbf{b}^y) + 1, \\ \beta(\mathbf{b}^x, \mathbf{b}^y) &= NE(\mathbf{b}^x, \mathbf{b}^y) + 1, \end{aligned} \tag{6}$$

For clarity, we abbreviate $\alpha(\mathbf{b}^x, \mathbf{b}^y)$ as α, and $\beta(\mathbf{b}^x, \mathbf{b}^y)$ as β. By collecting the positive and negative evidence, the parameters of $Beta\left(p|\alpha, \beta\right)$ could be learned in a data-driven manner, thereby quantifying the corresponding binomial opinion in retrieval.

The matching measurement between two samples could be simplified as a binary classification problem, where they are either matched or unmatched. Therefore, this problem could be modeled by a Bernoulli distribution, with its parameter representing the matching probability p between the two points. The corresponding likelihood function is formulated as follows:

$$Bernoulli(S|p) = p^S * (1 - p)^{(1-S)}, \tag{7}$$

where S is the matching ground truth of the two points. Specifically, $S = 1$ when the points have a shared label, and $S = 0$ otherwise. The Beta distribution $Beta\left(p|\alpha, \beta\right)$ could model a probability distribution for the matching probabilities p. Notably, the likelihood p is proportional to the discrimination (similarity) between the two samples. That is to say, p should be large when $S = 1$ and small when $S = 0$. To this end, $Beta\left(p|\alpha, \beta\right)$ is treated as a prior on the likelihood

$Bernoulli(S|p)$, and the negative log-marginal-likelihood is employed to maximize the discrimination by integrating the matching probabilities as follows:

$$\begin{aligned}\mathcal{L}_e&\left(\mathbf{b}^x, \mathbf{b}^y, S\right)\\ &= -\log\left(\int Bernoulli(S|p)\ Beta\left(p|\alpha,\beta\right)dp\right)\\ &= S*\log\left(\frac{\alpha+\beta}{\alpha}\right)+(1-S)*\log\left(\frac{\alpha+\beta}{\beta}\right).\end{aligned} \tag{8}$$

This loss function drives the model to generate more correct evidence than incorrect ones, but it does not guarantee the generation of zero incorrect evidence. To address this, following (Sensoy, Kaplan, and Kandemir 2018), we introduce a Kullback-Leibler (KL) divergence term:

$$\begin{aligned}\mathcal{L}_{kl}&\left(\mathbf{b}^x, \mathbf{b}^y, S\right)\\ &= KL\left[Beta\left(p\mid\tilde{\alpha},\tilde{\beta}\right)\|Beta\left(p\mid 1,1\right)\right]\\ &= \log\left(\frac{\Gamma\left(\tilde{\alpha}+\tilde{\beta}\right)}{\Gamma(\tilde{\beta})\Gamma\left(\tilde{\alpha}\right)}\right)+(\tilde{\alpha}-1)\left[\psi\left(\tilde{\alpha}\right)-\psi\left(\tilde{\beta}+\tilde{\alpha}\right)\right]\\ &\quad+\left(\tilde{\beta}-1\right)\left[\psi\left(\tilde{\beta}\right)-\psi\left(\tilde{\beta}+\tilde{\alpha}\right)\right],\end{aligned} \tag{9}$$

where $\tilde{\alpha}=S+(1-S)*\alpha$, $\tilde{\beta}=(1-S)+S*\beta$, $\tilde{\alpha}$ and $\tilde{\beta}$ are the Beta parameters after removing the correct evidence, and $\Gamma(\cdot)$ and $\psi(\cdot)$ are the gamma and digamma functions, respectively.

Near-Zero Correct-Evidence Learning

Inspired by (Pandey and Yu 2023), we observe that the aforementioned loss functions allow the model to learn effectively from most training pairs, except for those with near-zero correct evidence, which would impair performance.

To overcome this issue, we propose three specialized loss functions for near-zero correct-evidence learning to endow our model with the ability to learn from near-zero correct-evidence sample pairs:

$$\begin{aligned}&\mathcal{L}_{nzce-RA}\left(\mathbf{b}^x, \mathbf{b}^y, S\right)=\frac{1}{E_{gt}},\\ &\mathcal{L}_{nzce-CE}\left(\mathbf{b}^x, \mathbf{b}^y, S\right)=-log(tanh(E_{gt})),\\ &\mathcal{L}_{nzce-RM}\left(\mathbf{b}^x, \mathbf{b}^y, S\right)=log(1+\frac{1}{E_{gt}}),\end{aligned} \tag{10}$$

where $E_{gt}=e^{O_{gt}}$, and O_{gt} is defined as $H_s(\mathbf{b}^x,\mathbf{b}^y)$ when two cross-modal samples match, and as $g(\mathbf{b}^x,\mathbf{b}^y;\theta)$ otherwise.

Differentiable Hashing

By minimizing the loss function, we could optimize the parameters of the hashing functions (i.e., $f^x\left(\cdot,\Theta^x\right)$ and $f^y\left(\cdot,\Theta^y\right)$) using a gradient descent optimization algorithm, where Θ^x and Θ^y represent the parameters of the image and text networks, respectively. However, due to the discreteness of the hash codes, directly optimizing the model with $\mathbf{b}^x$ and $\mathbf{b}^y$ is an NP-hard problem (Kong and Li 2012). To address this problem, most existing methods relax the hash codes as continuous representations to optimize the model (Zhang, Peng, and Yuan 2018; Zhang and Peng 2019; Kumar and Udupa 2011). However, this relaxation will unavoidably result in a gap between the continuous values used in training and the binary codes required for inference, thus leading to potential performance degradation.

To tackle this issue, we present a plug-and-play module, termed Differentiable Hashing, to train the model without relaxation. To be specific, during forward propagation, this module enforces the network to generate binary codes by:

$$\begin{aligned}\mathbf{b}^x&=\frac{1}{\sqrt{L}}sgn\left(\frac{f^x(\mathbf{x})}{||f^x(\mathbf{x})||}\right),\\ \mathbf{b}^y&=\frac{1}{\sqrt{L}}sgn\left(\frac{f^y(\mathbf{y})}{||f^y(\mathbf{y})||}\right),\end{aligned} \tag{11}$$

where $||\cdot||$ denotes the ℓ_2-norm function, and sgn is the sign function. The binary codes generated by Equation (11) are then used to compute the loss to train the model. However, the sign function is non-differentiable due to its discreteness, which prevents the network from training through gradient descent. To address this, we employ the Straight-Through Estimator (STE) (Bengio, Léonard, and Courville 2013) to produce straight gradients during backward propagation, thus enabling the network to update its parameters. In other words, this module is treated as an identity function with a derivative of 1, enabling the network to update its parameters by gradient descent. Therefore, our module could guarantee that the upstream training is consistent with the downstream inference without continue-value relaxation, embracing superior performance.

Optimization

For a given mini-batch containing M image-text pairs, the matching ground-truth between the i-th text $\mathbf{x}_i$ and the j-th image $\mathbf{y}_j$ is denoted as S_{ij}. Specifically, $S_{ij}=1$ if $\mathbf{x}_i$ and $\mathbf{y}_j$ share one or more labels in common; otherwise, $S_{ij}=0$. The corresponding hash codes generated for these samples are represented by $\mathbf{b}_i^x$ and $\mathbf{b}_j^y$. To simplify the overall loss function, we introduce the notation ζ_{ij} to represent the parameters of the loss function. The overall loss for this mini-batch is then formulated as follows:

$$\mathcal{L}_{overall}=\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\mathcal{L}_e(\zeta_{ij})+\lambda\mathcal{L}_{kl}(\zeta_{ij})+\gamma\mathcal{L}_{nzce}(\zeta_{ij}), \tag{12}$$

where ζ_{ij} represents a trio of parameters consisting of $\mathbf{b}_i^x, \mathbf{b}_j^x$ and S_{ij}. The optimization process is detailed in Algorithm 1.

Reliability Estimation in Cross-Modal Hashing

Given that our model is based on binomial opinions, two significant challenges arise: 1) Even with a well-trained model, conflicting evidence may emerge, where both positive and negative indicators are abundant and roughly equivalent in quantity. In such scenarios, although the uncertainty mass

Algorithm 1: The Optimization Procedure of DECH

Input: The data of n paired samples $(\mathbf{x}_i, \mathbf{y}_i)$ from different modalities, each corresponding to a multi-label vector $\mathbf{l}_i$, the bit length L of the generated hash codes, the number N_b of samples per mini-batch, the hyperparameters τ, λ, γ, iteration number N_t and learning rate α.

Output: Optimized DECH model

1: Randomly initialize the network parameters Θ^x, Θ^y, and Θ^z of the DECH model.
2: **repeat**
3: Randomly sample N_b data pairs to form a mini-batch.
4: Use the corresponding hash function from Equation (11) to calculate the discrete hash code for each sample in the mini-batch.
5: Use Equation (12) to calculate the overall loss.
6: Update the the network parameters by minimizing $\mathcal{L}_{overall}$ in Equation (12) with descending their stochastic gradient:

$$\Theta^* = \Theta^* - \alpha \left(\frac{\partial \mathcal{L}_{overall}}{\partial \Theta^*} \right)$$

7: **until** Convergence

might be low, the model still struggles to accurately determine whether the data pairs are a match. 2) The primary objective in cross-modal hashing retrieval is to identify samples that match the query, rather than to retrieve non-matching samples. Scenarios where positive evidence substantially outweighs negative evidence (indicating a match), or vice versa (indicating a non-match), typically result in low uncertainty mass. Consequently, we cannot directly employ uncertainty mass as a reliable metric for uncertainty estimation in cross-modal hashing retrieval.

To address the first issue, we adopt a method based on the Relative Difference, which is extensively used to compare the magnitude of differences between two quantities in a normalized way. Based on this, we propose the following definition to quantify the dissonance between PE and NE:

Definition 1 (Dissonance Uncertainty)**.**

$$Diss(PE, NE) = 1 - \frac{|PE - NE|}{\max(PE, NE)} \in [0, 1]. \quad (13)$$

When the values of PE and NE are close, it suggests that the evidence generated is conflicting, leading Dissonance Uncertainty to approach 1. Conversely, when PE and NE values are not close, Dissonance Uncertainty tends to approach 0.

In the inference stage, we first use the evidence obtained from the model to parameterize the Beta distribution and derive the corresponding binomial opinion. Following (Cho et al. 2017; Josang, Cho, and Chen 2018), we use $Diss(PE, NE)$ to adjust the binomial opinion as follows:

Definition 2 (Correct Binomial Opinion)**.**

$$\begin{aligned} b_{cor} &= b * (1 - Diss(PE, NE)), \\ d_{cor} &= d * (1 - Diss(PE, NE)), \\ u_{cor} &= 1 - b_{cor} - d_{cor}. \end{aligned} \quad (14)$$

As discussed earlier, the uncertainty mass in binomial opinion merely quantifies whether the model has sufficient evidence to determine if data pairs match. To estimate retrieval reliability, based on the projected probability (Jøsang 2016), we propose the following method to measure reliability:

Definition 3 (Retrieval Reliability)**.**

$$r_{cmh} = 1 - (d_{cor} + 0.5 * u_{cor}). \quad (15)$$

From Definition 3, one could observe that the model generates high d_{cor} or u_{cor} when it deems data pairs mismatched or uncertain, resulting in low reliability. Conversely, when the model determines sample pairs to be matched, it produces high b_{cor}, which reduces the sum of d_{cor} and u_{cor}, thereby increasing the reliability level. Thus, our proposed r_{cmh} can effectively estimate the reliability of data pair matching in cross-modal hashing.

Experiments

Datasets

We conduct our experiments on four benchmark datasets: MIRFLICKR25K(Huiskes and Lew 2008), IAPR TC-12(Escalante et al. 2010), NUS-WIDE(Rasiwasia et al. 2010), and MS-COCO(Lin et al. 2014). MIRFLICKR25K contains 20,500 image-text pairs from 24 classes, with 2,000 pairs reserved for querying, 10,000 for training, and the rest for retrieval. IAPR TC-12 comprises 20,000 pairs across 255 categories, with 2,000 pairs used for querying, 10,000 for training, and the remainder for retrieval. NUS-WIDE includes 195,834 pairs in 21 categories, with 2,000 pairs for querying, 10,500 for training, and the rest for retrieval. MS-COCO consists of 122,218 pairs in 80 classes, with 5,000 pairs for querying, 10,000 for training, and the remaining pairs forming the retrieval database. In each dataset, images and texts are represented by high-dimensional feature vectors derived from pre-trained models or bag-of-words approaches.

Baselines and Implementation

To evaluate the performance of our DECH, we compared it against 12 state-of-the-art methods: DJSRH (Su, Zhong, and Zhang 2019), JDSH (Liu et al. 2020), DGCPN (Yu et al. 2021), DCMH (Jiang and Li 2017), DADH (Bai et al. 2020), MLSPH (Zou et al. 2021), HMAH (Tan et al. 2022), SCCGDH (Shu et al. 2022), MESDCH (Zou et al. 2022), MIAN (Zhang et al. 2022), DNPH (Qin et al. 2024), and DHAPH (Huo et al. 2024).

In our experiments, we conduct two cross-modal retrieval tasks: image-to-text (I $\rightarrow$ T) and text-to-image (T $\rightarrow$ I). To ensure fair comparisons, all models employ an identical pre-trained backbone for feature extraction, with the layers of this backbone kept fixed during training. Each method is tested using the same training, retrieval, and query sets. All baselines are configured with the default settings provided by their respective authors. For our DECH, we set τ to 0.2 and γ to 1. The parameter λ is empirically determined as

Task	Method	MIRFLICKR25K				NUS-WIDE				IAPR TC-12				MS-COCO			
		16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
I→T	DJSRH	0.620	0.630	0.645	0.660	0.467	0.461	0.497	0.517	0.368	0.396	0.419	0.439	0.489	0.519	0.543	0.565
	JDSH	0.711	0.711	0.708	0.717	0.538	0.536	0.559	0.574	0.435	0.459	0.473	0.484	0.611	0.635	0.648	0.651
	DGCPN	0.714	0.722	0.726	0.735	0.569	0.574	0.594	0.602	0.463	0.473	0.481	0.481	0.608	0.637	0.637	0.634
	DCMH	0.739	0.755	0.764	0.771	0.629	0.649	0.668	0.677	0.423	0.439	0.456	0.463	0.548	0.575	0.606	0.625
	DADH	0.812	0.826	0.833	0.840	0.640	0.653	0.657	0.669	0.517	0.530	0.549	0.573	0.517	0.621	0.664	0.674
	MLSPH	0.804	0.821	0.833	0.838	0.473	0.488	0.490	0.493	0.463	0.482	0.508	0.536	0.583	0.627	0.657	0.667
	HMAH	0.783	0.813	0.821	0.825	0.522	0.561	0.571	0.598	0.472	0.493	0.511	0.523	0.508	0.572	0.598	0.607
	SCCGDH	0.783	0.814	0.814	0.800	0.644	0.656	0.581	0.523	0.489	0.498	0.475	0.422	0.633	0.663	0.667	0.640
	MESDCH	0.811	0.829	0.836	0.842	0.455	0.465	0.475	0.476	0.504	0.526	0.544	0.539	0.599	0.637	0.657	0.669
	MIAN	0.815	0.824	0.834	0.835	0.637	0.647	0.643	0.651	0.485	0.510	0.534	0.543	0.587	0.603	0.599	0.627
	DHAPH	0.782	0.791	0.793	0.796	0.666	0.671	0.675	0.679	0.491	0.508	0.520	0.525	0.652	0.673	0.683	0.688
	DNPH	0.763	0.779	0.791	0.799	0.643	0.666	0.673	0.681	0.485	0.505	0.517	0.525	0.624	0.646	0.660	0.665
	DECH	0.833	0.846	0.853	0.860	0.686	0.706	0.716	0.727	0.527	0.572	0.590	0.606	0.660	0.710	0.732	0.736
	DECH$_{r=0.5}$	**0.869**	**0.880**	**0.885**	**0.891**	**0.802**	**0.835**	**0.849**	**0.867**	**0.623**	**0.664**	**0.685**	**0.694**	**0.763**	**0.813**	**0.835**	**0.845**
T→I	DJSRH	0.620	0.626	0.645	0.649	0.449	0.473	0.480	0.487	0.371	0.400	0.424	0.437	0.472	0.524	0.549	0.566
	JDSH	0.685	0.687	0.677	0.698	0.515	0.561	0.526	0.561	0.441	0.462	0.478	0.489	0.616	0.641	0.651	0.656
	DGCPN	0.698	0.700	0.709	0.718	0.585	0.606	0.594	0.613	0.466	0.478	0.489	0.485	0.610	0.632	0.632	0.629
	DCMH	0.752	0.760	0.763	0.770	0.653	0.656	0.685	0.703	0.449	0.464	0.476	0.481	0.571	0.594	0.642	0.664
	DADH	0.791	0.796	0.805	0.816	0.632	0.628	0.663	0.668	0.506	0.539	0.564	0.591	0.522	0.617	0.662	0.675
	MLSPH	0.772	0.788	0.798	0.800	0.484	0.504	0.511	0.521	0.465	0.485	0.506	0.527	0.580	0.624	0.648	0.658
	HMAH	0.773	0.796	0.805	0.814	0.556	0.594	0.598	0.623	0.487	0.508	0.541	0.556	0.514	0.574	0.600	0.616
	SCCGDH	0.749	0.783	0.761	0.772	0.608	0.629	0.554	0.520	0.480	0.506	0.487	0.435	0.631	0.668	0.672	0.631
	MESDCH	0.767	0.782	0.791	0.795	0.477	0.482	0.487	0.494	0.503	0.511	0.515	0.507	0.594	0.624	0.643	0.657
	MIAN	0.777	0.787	0.804	0.801	0.708	0.720	0.736	0.727	0.483	0.509	0.525	0.538	0.612	0.660	0.660	0.682
	DHAPH	0.771	0.781	0.784	0.786	0.720	0.727	0.732	0.736	0.506	0.524	0.535	0.540	0.622	0.647	0.655	0.662
	DNPH	0.764	0.782	0.792	0.801	0.700	0.710	0.724	0.726	0.505	0.527	0.542	0.551	0.630	0.655	0.667	0.672
	DECH	0.812	0.825	0.831	0.833	0.731	0.755	0.763	0.766	0.531	0.575	0.600	0.615	0.666	0.704	0.724	0.735
	DECH$_{r=0.5}$	**0.829**	**0.843**	**0.849**	**0.851**	**0.787**	**0.823**	**0.835**	**0.839**	**0.628**	**0.676**	**0.699**	**0.708**	**0.777**	**0.810**	**0.829**	**0.851**

Table 1: mAP@ALL comparison of DECH and baselines on the four Datasets. **Bold** highlights the highest results; underlined marks the second highest.

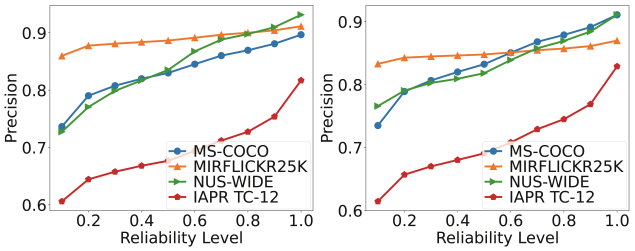
per (Sensoy, Kaplan, and Kandemir 2018). Additionally, we also employ $\mathcal{L}_{nzce-RM}$ as the near-zero correct-evidence loss during training. Our method is implemented with PyTorch(Paszke et al. 2019) on a single NVIDIA GEFORCE RTX 3090 Ti GPU.

Comparison with State-of-the-Art Methods

Table 1 reports the mean Average Precision (mAP) results for our DECH and other baselines on the four datasets. DECH$_{r=0.5}$ represents our model incorporating the confidence assessment approach with a reliability level of 0.5. The results lead to the following key observations: 1) Our DECH consistently outperforms all other baselines in retrieval performance. This can be attributed to the effective use of evidential deep learning in cross-modal hashing, which enables the model to generate discriminative hash codes. 2) The proposed differentiable hashing, which learns discrete codes directly without resorting to continuous-value relaxation, remarkably enhances retrieval performance. 3) The consistently superior performance across four different bit settings and four diverse datasets further validates the effectiveness of our approach. 4) When applying our proposed reliability to filter out less trustworthy sample pairs, we observe a marked improvement in retrieval performance. This improvement underscores the efficacy of our proposed reliability estimation method in enhancing the model's retrieval capabilities, ultimately achieving more reliable and trustworthy retrieval results.

Quantitative Analysis of Reliability Estimation

We quantitatively evaluate the impact of our DECH on retrieval performance using mAP scores. By setting various reliability levels as thresholds, we filter out samples whose reliability falls below these thresholds. As illustrated in Figure 3, higher reliability consistently corresponds to improved retrieval performance across all four datasets. These results demonstrate that our reliability estimation approach effectively eliminates mismatched sample pairs, thereby leading to more reliable retrieval outcomes. Furthermore, this finding indicates that our method maintains high performance across various datasets.



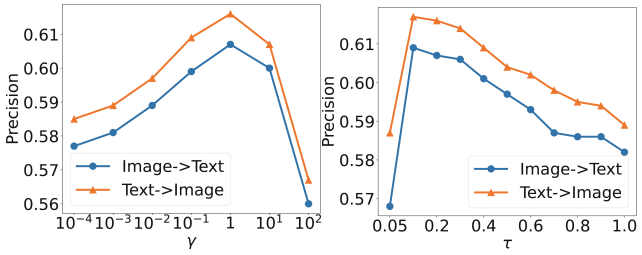
(a) Image → Text (b) Text → Image

Figure 3: Quantitative analysis of four datasets at various reliability levels with a code length of 128.

Parameter Analysis

Our model primarily encompasses two key parameters: τ in Equation (4), and γ in Equation (12). To investigate the influence of these parameters on the performance of the model, we conduct experiments by varying the values of τ and γ on the IAPR TC-12 dataset, with a hash code length of 128 bits. For τ, its value is varied from 0 to 1 while γ is kept constant at 1. Conversely, when adjusting γ, τ is maintained at 0.2. The corresponding mAP scores are illustrated in Figure 4. In the experiments varying γ, we observe that the optimal retrieval performance occurs when $\gamma = 1$. If γ is too small, the model struggles to learn from the near-zero correct-evidence sample pairs, leading to a performance decline. Conversely, if γ is too large, the influence of other components is diminished, resulting in a performance drop. This indicates that each component of our loss function contributes to the overall performance of the model.

Regarding τ, we observe that when it is too small, the evidence activation function grows exponentially, negatively impacting performance. On the other hand, if τ is too large, the activation of positive evidence becomes restricted to a very narrow range, which also adversely affects retrieval performance. The optimal retrieval performance is achieved when τ is around 0.1.



(a) Image → Text (b) Text → Image

Figure 4: Parameter sensitivity analysis for γ and τ on IAPR TC-12.

Ablation Analysis

We perform ablation experiments on the MIRFLICKR25K and IAPR TC-12 datasets to investigate the impact of individual components on retrieval performance. To this end, we compare our method with its three variants: DECH without $\mathcal{L}_{kl}$, DECH without $\mathcal{L}_{nzce}$, and DECH without DH. The experimental results, depicted in table Table 2, lead to the following conclusions: 1) The model's performance remarkably degrades when either $\mathcal{L}_{kl}$ or $\mathcal{L}_{nzce}$ is omitted, indicating that both terms are instrumental for enhancing the model's ability to learn and acquire evidence, thereby improving retrieval performance. 2) The DH module plays a crucial role in improving performance by enabling discrete optimization without sacrificing differentiability.

Additionally, we also evaluate the performance of three different forms of near-zero correct-evidence loss on these two datasets. The results, shown in Table 3, reveal that all three loss functions achieve similar and excellent performance. This is due to their ability to extract valuable information from near-zero correct-evidence sample pairs, allowing the model to learn effectively from all sample pairs and thus ensure superior performance. These observations also validate the effectiveness of all three loss functions.

Method	Image → Text			Text → Image		
	16	32	64	16	32	64
MIRFLICKR25K						
w/o $\mathcal{L}_{kl}$	0.800	0.817	0.823	0.777	0.793	0.798
w/o $\mathcal{L}_{nzce}$	0.826	0.840	0.847	0.803	0.816	0.824
w/o DH	0.813	0.808	0.811	0.795	0.791	0.792
DECH	**0.832**	**0.846**	**0.853**	**0.812**	**0.825**	**0.831**
IAPR TC-12						
w/o $\mathcal{L}_{kl}$	0.489	0.507	0.513	0.497	0.511	0.522
w/o $\mathcal{L}_{nzce}$	0.473	0.504	0.527	0.478	0.502	0.529
w/o DH	0.503	0.518	0.528	0.511	0.522	0.536
DECH	**0.527**	**0.572**	**0.590**	**0.531**	**0.575**	**0.600**

Table 2: Comparison of the mAP scores of DECH and its variants.

Method	Image → Text			Text → Image		
	16	32	64	16	32	64
MIRFLICKR25K						
$\mathcal{L}_{nzce-CE}$	**0.834**	**0.849**	0.856	0.809	0.824	**0.833**
$\mathcal{L}_{nzce-RA}$	0.833	0.847	0.853	0.809	0.824	0.830
$\mathcal{L}_{nzce-RM}$	0.832	0.846	**0.860**	**0.812**	**0.825**	0.831
IAPR TC-12						
$\mathcal{L}_{nzce-CE}$	**0.531**	0.562	0.591	**0.532**	0.566	0.597
$\mathcal{L}_{nzce-RA}$	0.527	0.564	**0.595**	0.531	0.568	0.598
$\mathcal{L}_{nzce-RM}$	0.527	**0.572**	0.590	0.531	**0.575**	**0.600**

Table 3: mAP of various forms of $\mathcal{L}_{nzce}$ across different bit lengths on the two datasets.

Conclusion

In this paper, we present a novel method termed Deep Evidential Cross-modal Hashing (DECH), designed to quantify the reliability between query samples and each retrieved result in cross-modal retrieval scenarios. Our approach consists of three novel modules: 1) Deep Evidential Cross-modal Hashing module that collects evidence and derives a binomial opinion for each cross-modal pair, 2) Refined Binomial Opinion module that allows the model to quantify dissonance uncertainty while specifically estimating reliability for the cross-modal hashing task, and 3) Differentiable Hashing module that enables the discrete optimization of binary codes without continuous-value relaxation. Extensive experiments are conducted to verify the effectiveness of our method.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2024YFB4710604; in part by NSFC under Grant 62472295, 62176171 and U21B2040; in part by Sichuan Science and Technology Planning Project under Grant 2024NSFTD0047 and 2024NSFTD0038; in part by System of Systems and Artificial Intelligence Laboratory pioneer fund grant; in part by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: [AISG2-GC-2023-007]); and in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303 and CJ202403.

References

Bai, C.; Zeng, C.; Ma, Q.; Zhang, J.; and Chen, S. 2020. Deep adversarial discrete hashing for cross-modal retrieval. In *Proceedings of the 2020 international conference on multimedia retrieval*, 525–531.

Bengio, Y.; Léonard, N.; and Courville, A. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Cho, J.-H.; Cook, T.; Rager, S.; O'Donovan, J.; and Adali, S. 2017. Modeling and analysis of uncertainty-based false information propagation in social networks. In *GLOBECOM 2017-2017 IEEE Global Communications Conference*, 1–7. IEEE.

Dempster, A. P. 1968. A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2): 205–232.

Escalante, H. J.; Hernández, C. A.; Gonzalez, J. A.; López-López, A.; Montes, M.; Morales, E. F.; Sucar, L. E.; Villasenor, L.; and Grubinger, M. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer vision and image understanding*, 114(4): 419–428.

Fang, Y.; Zhang, H.; and Ren, Y. 2019. Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing. *Knowledge-Based Systems*, 171: 69–80.

Feng, Y.; Zhu, H.; Peng, D.; Peng, X.; and Hu, P. 2023. RONO: robust discriminative learning with noisy labels for 2D-3D cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11610–11619.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2021. Trusted multi-view classification. In *International Conference on Learning Representations*.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 2551–2566.

Hu, P.; Huang, Z.; Peng, D.; Wang, X.; and Peng, X. 2023a. Cross-Modal Retrieval with Partially Mismatched Pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9595–9610.

Hu, P.; Peng, X.; and Peng, D. Z. 2024. Anchor-based Unsupervised Cross-modal Hashing. *Journal of Software*, 35(8): 3739–3751. In Chinese.

Hu, P.; Peng, X.; Zhu, H.; Lin, J.; Zhen, L.; and Peng, D. 2021. Joint versus independent multiview hashing for cross-view retrieval. *IEEE Transactions on Cybernetics*, 51(10): 4982–4993.

Hu, P.; Zhu, H.; Lin, J.; Peng, D.; Zhao, Y.-P.; and Peng, X. 2023b. Unsupervised Contrastive Cross-Modal Hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3877–3889.

Huang, H.-J.; Yang, R.; Li, C.-X.; Shi, Y.; Guo, S.; and Xu, X.-S. 2017. Supervised cross-modal hashing without relaxation. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 1159–1164. IEEE.

Huiskes, M.; and Lew, M. 2008. Lew. The mir flickr retrieval evaluation. In *Proceed-ing of the 1st ACM international conference on Multimedia information retrieval-MIR*.

Huo, Y.; Qin, Q.; Zhang, W.; Huang, L.; and Nie, J. 2024. Deep Hierarchy-aware Proxy Hashing with Self-paced Learning for Cross-modal Retrieval. *IEEE Transactions on Knowledge and Data Engineering*.

Jiang, Q.-Y.; and Li, W.-J. 2017. Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3232–3240.

Jøsang, A. 2016. *Subjective logic*, volume 4. Springer.

Josang, A.; Cho, J.-H.; and Chen, F. 2018. Uncertainty characteristics of subjective opinions. In *2018 21st International Conference on Information Fusion (FUSION)*, 1998–2005. IEEE.

Kong, W.; and Li, W.-J. 2012. Double-bit quantization for hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, 634–640.

Kumar, S.; and Udupa, R. 2011. Learning hash functions for cross-view similarity search. In *Twenty-second international joint conference on artificial intelligence*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, S.; Qian, S.; Guan, Y.; Zhan, J.; and Ying, L. 2020. Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 1379–1388.

Pandey, D. S.; and Yu, Q. 2023. Learn to Accumulate Evidence from All Training Samples: Theory and Practice. In *International Conference on Machine Learning*, 26963–26989. PMLR.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Qin, Q.; Huo, Y.; Huang, L.; Dai, J.; Zhang, H.; and Zhang, W. 2024. Deep Neighborhood-Preserving Hashing With Quadratic Spherical Mutual Information for Cross-Modal Retrieval. *IEEE Transactions on Multimedia*, 26: 6361–6374.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, 251–260.

Sahlin, U.; Helle, I.; and Perepolkin, D. 2021. "This Is What We Don't Know": Treating epistemic uncertainty in bayesian networks for risk assessment. *Integrated Environmental Assessment and Management*, 17(1): 221–232.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Shu, Z.; Bai, Y.; Zhang, D.; Yu, J.; Yu, Z.; and Wu, X.-J. 2022. Specific class center guided deep hashing for cross-modal retrieval. *Information sciences*, 609: 304–318.

Su, S.; Zhong, Z.; and Zhang, C. 2019. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3027–3035.

Sun, Y.; Dai, J.; Ren, Z.; Chen, Y.; Peng, D.; and Hu, P. 2024a. Dual Self-Paced Cross-Modal Hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15184–15192.

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

Sun, Y.; Ren, Z.; Hu, P.; Peng, D.; and Wang, X. 2023. Hierarchical consensus hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26: 824–836.

Tan, W.; Zhu, L.; Li, J.; Zhang, H.; and Han, J. 2022. Teacher-Student Learning: Efficient Hierarchical Message Aggregation Hashing for Cross-Modal Retrieval. *IEEE Transactions on Multimedia*, 1–1.

Yu, J.; Zhou, H.; Zhan, Y.; and Tao, D. 2021. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4626–4634.

Zhang, J.; and Peng, Y. 2019. Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Multimedia*, 22(1): 174–187.

Zhang, J.; Peng, Y.; and Yuan, M. 2018. Unsupervised generative adversarial cross-modal hashing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhang, Z.; Luo, H.; Zhu, L.; Lu, G.; and Shen, H. T. 2022. Modality-invariant asymmetric networks for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*.

Zou, X.; Wang, X.; Bakker, E. M.; and Wu, S. 2021. Multi-label semantics preserving based deep cross-modal hashing. *Signal Processing: Image Communication*, 93: 116131.

Zou, X.; Wu, S.; Bakker, E. M.; and Wang, X. 2022. Multi-label enhancement based self-supervised deep cross-modal hashing. *Neurocomputing*, 467: 138–162.