# Deep Neighborhood-preserving Hashing with Quadratic Spherical Mutual Information for Cross-modal Retrieval

Qibing Qin, Yadong Huo, Lei Huang, *Member, IEEE*, Jiangyan Dai, Huihui Zhang and Wenfeng Zhang

***Abstract*—Driven by the high nonlinearity of deep neural networks, deep hashing has achieved the pictured great potential in cross-modal retrieval applications, significantly bridging the modality gap. Current deep cross-modal hashing usually utilizes affinity matching or local ranking to capture the local semantic relationships in the learned common space, leading to high neighborhood ambiguity. Simultaneously, most of these frameworks utilize additional regularization terms or margin thresholds to enhance the overall performance, in which searching the model's hyper-parameters under mass training data would have a substantial overhead. In this paper, with a novel extension of information-theoretic measures, a novel deep cross-modal hashing method, named Deep Neighborhood-preserving Hashing (DNpH), is designed to learn a highly separable discrete space, effectively mitigating the semantic gap across different modalities. Specifically, to minimize neighborhood ambiguity, the Quadratic Spherical Mutual Information (QSMI) is first introduced into deep cross-modal hashing to separate neighbors and non-neighbors well, while it is free of tuning parameters during model training compared with other similarity measures. To optimize quadratic mutual information loss smoothly, a square clamping method is developed to improve the stability of model optimization, avoiding converging on bad local optimum. Besides, two transformer encoders are exploited as feature extractors for multi-modal samples to learn the informative semantic representations. Finally, we compare our proposed DNpH framework with various state-of-the-art cross-modal hashing on four public datasets, and large amounts of experiment results demonstrate our contributions and show that DNpH outperforms the compared baselines on different evaluation metrics. The corresponding code is available at https://github.com/QinLab-WFU/DNpH.**

***Index Terms*—Cross-modal Retrieval, Deep Hashing, Quadratic Mutual Information, Neighborhood Structure, Transformer Encoder.**

This work was supported by the National Natural Science Foundation of China (No.62302338, No.62006174), Natural Science Foundation of Shandong Province (No.ZR2022QF046, No.ZR2021MF026, No.ZR2023MF033), Natural Science Foundation of Chongqing (No.2023NSCQ-MSX1645) and Science and Technology Research Program of Chongqing Municipal Education Commission (No.KJQN202200551). (*Corresponding author: Huihui Zhang, Wenfeng Zhang*)

Qibing Qin is with the School of Computer Engineering, Weifang University, Weifang, China, and is also an Academic Visitor of the Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China (email: qinbing@wfu.edu.cn).

Yadong Huo is with the School of Computer Science, Qufu Normal University, Rizhao, China (e-mail: hyd199810@163.com).

Lei Huang is with the Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China (e-mail: huangl@ouc.edu.cn).

Jiangyan Dai and Huihui Zhang are with the School of Computer Engineering, Weifang University, Weifang, China (e-mail: daijy@wfu.edu.cn; huihui@wfu.edu.cn).

Wenfeng Zhang is with the College of Computer and Information Science, Chongqing Normal University, Chongqing, China, and is with the Lishui Institute of Hangzhou Dianzi University, Lishui, China (e-mail: itzhangwf@cqnu.edu.cn).

## I. INTRODUCTION

Over the last decades, with the rapid development of modern mobile devices and Internet techniques, high-dimensional multimedia data has been growing exponentially, such as images, videos, and text [1], [2]. With the rapid expansion of multimedia data coming from different media types, traditional unimodal retrieval methods could hardly apply to real-world scenarios [3]. Therefore, how to retrieve related objects from massive samples from different modalities has become one of the current research hot directions in the field of multimedia processing [4]. Along with the data quantity unceasing increase, the storage and search costs of traditional cross-modal has risen sharply, which could hardly meet the real retrieval requirements. Due to its preferable efficiency in storage and computation, cross-modal retrieval-based hash learning has attracted wide interest in both academic and industry communities [5], [6]. Cross-modal hashing aims to project high-dimensional samples into a unified discrete common space, where semantic relationships between multi-modal samples can be preserved [7]. These approaches provide effective solutions to the challenges posed by cross-modal retrieval and have the potential to significantly improve the efficiency of multimedia data storage and retrieval [8]–[10].

According to whether the annotated information is needed, cross-modal hashing could be broadly categorized as supervised approaches and unsupervised approaches [11], [12]. Generally speaking, unsupervised cross-modal hashing typically relies on the predefined similarity metric to construct the hash project and learn the binary codes [13]–[16]. In contrast, supervised cross-modal hashing usually employs human-annotated labels to perform effective compact representation learning, obtaining informative representations and compact codes [17]–[21]. With the introduction of annotated information, cross-modal hashing methods typically yield an excellent representation compared with unsupervised approaches [22]. In this paper, we focus primarily on supervised cross-modal hashing, which has shown great promise in cross-modal retrieval applications.

With the amazing achievement of deep neural networks in the fields of computer vision [23], natural language processing [24], and information retrieval [25], emerging deep cross-modal hashing integrates feature representation with discrete

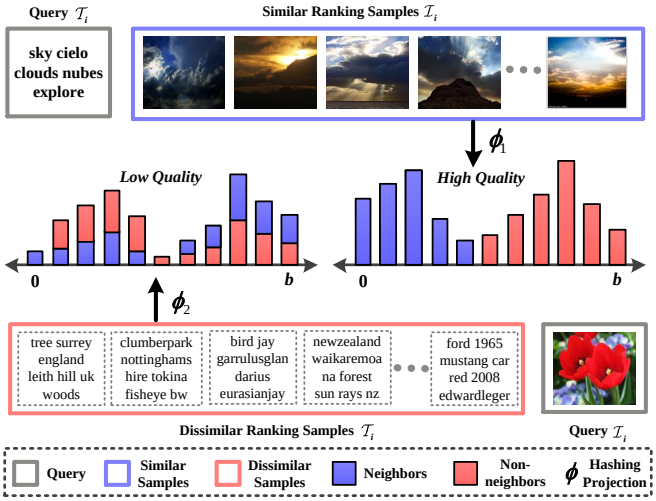



Fig. 1: For the given hash projection, the samples from different modalities $\mathcal{I}_i$, $\mathcal{T}_i$, along with the neighbors and non-neighbors, are mapped into compact discrete codes, generating two distributions of Hamming distances. As shown in the example, due to the limitation of local similarity, current hashing projection $\phi_2$ could not separate neighborhood samples and non-neighborhood samples, leading to high neighborhood ambiguity. By contrast, the quadratic spherical mutual information is introduced into cross-modal hashing retrieval to learn a higher quality Hamming distance distribution, as illustrated in $\phi_1$, maintaining the neighborhood structure of the samples.

code learning by an end-to-end framework, which has been proven to be more powerful in real-world retrieval applications [26]–[28]. By virtue of the distinctive representation capability of neural networks, deep cross-modal hashing methods could capture abundant semantic information from the samples of different modalities [29], resulting in mitigating the modality gap and achieving superior retrieval performance.

By developing the affinity matching strategy [30], [31] and local ranking strategy [32], [33], several researchers have suggested deep semantic-preservation cross-modal hashing to maintain the semantic structure of the samples from different modalities and their similarity relationships. Specifically, by calculating the distance between sample pairs, the affinity matching mechanism is introduced into cross-modal hashing to make sure that Hamming distance between similar (dissimilar) items is small (large). Meanwhile, based on triplet-wise learning, the local ranking method is proposed to minimize the distance between anchor points and positive samples and maximize the distance between anchor points and negative samples, maintaining the local semantic relationships. What's more, to mitigate the heterogeneity between different modalities, both intra-modal and inter-modal semantic relations are taken into account in cross-modal retrieval [34].

Despite the promising performance of deep semantic-preservation cross-modal hashing, there are still several under-explored limitations. The ultimate goal of deep hashing is to learn the more powerful and robust projection, capturing the original neighborhood relations and returning the separable nearest neighbors, such as $\phi_1$ in Fig 1. However, based on the simple and local semantic relationships, current methods, including affinity matching or local ranking strategy, are usually only indirectly related to retrieval performance, leading to higher neighborhood ambiguity, as indicated in $\phi_2$. And, it means the learned binary codes through the hashing projection $\phi_2$ in Fig 1 (such as affinity matching hashing or local ranking hashing) are insufficient to determine the neighbor relationships between samples from different modalities [35]. Besides, to improve the overall performance, available deep hashing usually introduces additional regularization terms or margin thresholds to optimize the model, resulting in the extra overhead of searching hyper-parameters [36], [37].

To tackle this problem, we exploit a novel extension of information-theoretic measures, Quadratic Spherical Mutual Information (QSMI) [38], to quantize neighborhood ambiguity and design a novel solution tailored for cross-modal hashing retrieval. Our key observation is that the learned discriminative and stable binary embedding could separate neighbors and non-neighbors well in a common discrete space, reducing the nonzero overlaps of the projection function $\phi_2$ in Fig 1. In this paper, by utilizing quadratic mutual information as a learning objective, a novel supervised deep cross-modal hashing method, named Deep Neighborhood-preserving Hashing framework (DNpH), is proposed to learn a highly separable Hamming space, as illustrated in Fig 2. To the best of our knowledge, our work is one of the pioneers to introduce quadratic mutual information into deep cross-modal hashing to minimize neighborhood ambiguity. Different from others that may need margins or thresholds, another advantage of quadratic mutual information loss is free of tuning parameters. Besides, to optimize quadratic mutual information smoothly, a square clamping optimization method is developed to deal with stochastic gradient descent in deep hash learning.

To sum up, our primary contributions to this paper are briefly presented as follows:

(1) Based on the information-theoretic quantity, the Quadratic Spherical Mutual Information (QSMI) is first introduced to cross-modal hashing to minimize neighborhood ambiguity, capturing the original neighborhood structures. Compared with other similarity measures, our proposed QSMI is more powerful and robust for cross-modal retrieval applications, while it is free of tuning parameters.

(2) To avoid bad local minima, a square clamping optimization method is developed to optimize quadratic spherical mutual information loss, enhancing the stability of model optimization.

(3) To obtain informative representations, two transformer encoders are adopted as the feature learning module for heterogeneous modalities, where the global semantic features of the raw images are obtained by capturing the long-range visual dependencies, and with multi-head self-attention, critical semantic characteristics are learned from raw text.

(4) Extensive experiments on four public datasets demonstrate that our proposed DNpH framework can learn a highly separable Hamming space and outperforms the state-of-the-art cross-modal hashing.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3349075



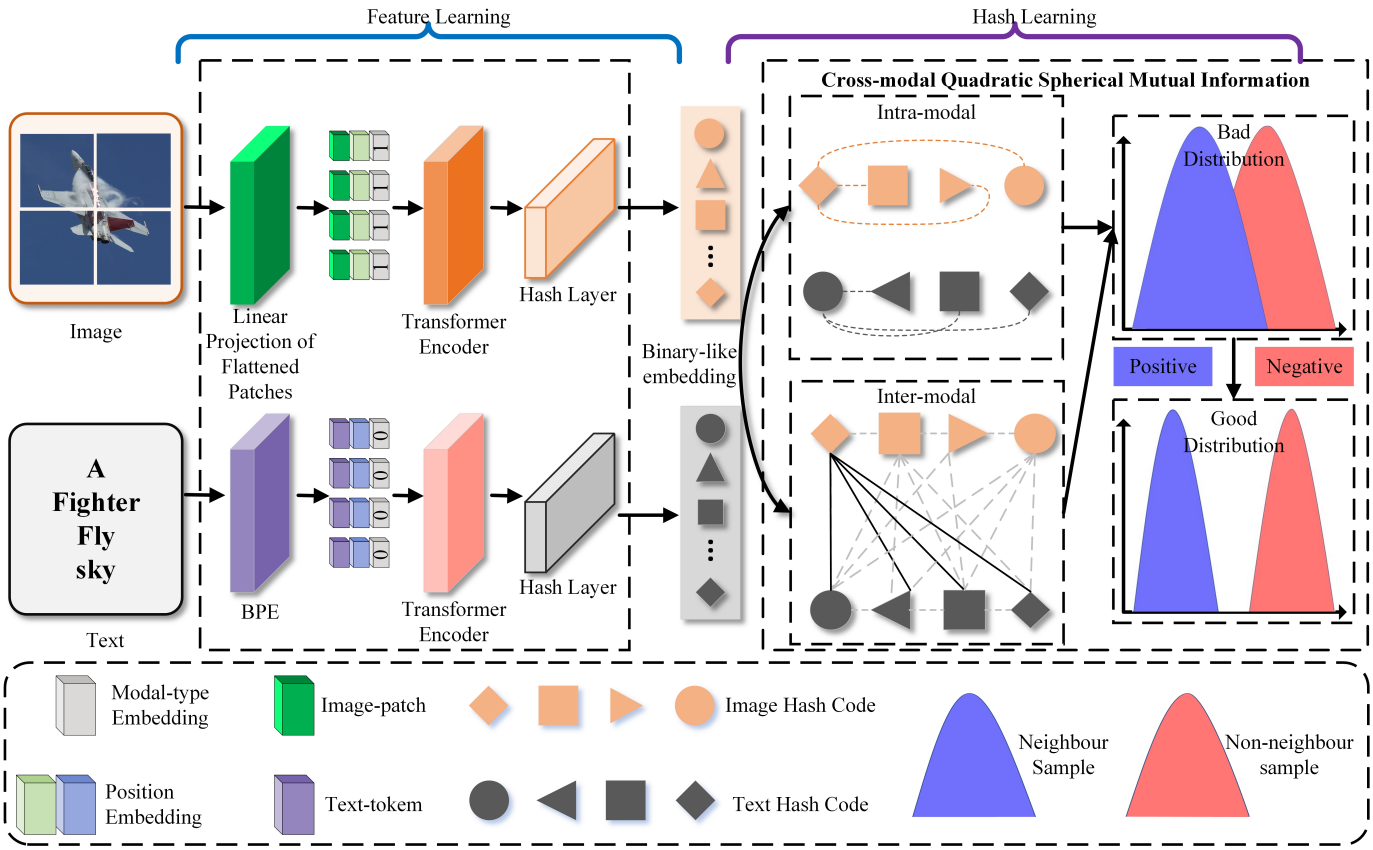



Fig. 2: An overview of DNpH, including two parts: (1) Feature Learning: The image feature extractor with transformer encoder architecture is designed to capture the global semantic information of the image. The text feature encoder incorporates the BPE to encode the text and then employs the text transformer encoder to obtain the semantic features of the text. (2) Hash Learning: Quadratic Spherical Mutual Information is first introduced to cross-modal hashing to quantize the neighborhood ambiguity, separating neighborhood and non-neighborhood samples well, and a square clamping strategy is developed to optimize mutual information loss effectively.

The rest of the paper is organized as follows. In Section II, representative deep cross-modal hashing methods are briefly reviewed. We elaborate on the implementation of our DNpH framework in Section III. Section IV shows the comparison and analysis of our proposed DNpH algorithm with other state-of-the-art hashing. The conclusions and future directions are drawn in Section V.

## II. Related Works

In this section, we provide a brief overview of representative cross-modal hashing algorithms, including supervised hashing and unsupervised hashing, and mutual information in hashing retrieval applications.

### A. Supervised Cross-modal Hashing

With preferable efficiency in storage and computation, several cross-modal hashing based on shallow architecture have been proposed, such as SMH [39], DCH [40], SMFH methods [41], and so on, but these methods are based on hand-crafted features, which do not sufficiently preserve the semantic relationships between multi-modal data and lead to poor retrieval performance. With the amazing achievement of deep neural networks, deep cross-modal hashing has been made toward substantial progression. Specifically, by incorporating feature extraction and hash learning into an end-to-end framework for the first time, Jiang *et al.* propose a novel Deep Cross-Modal Hashing to achieve satisfactory performance [30]. Based on the generative adversarial strategy, the Self-Supervised Adversarial Hashing framework is proposed to bridge the semantic gap across different modalities [31]. By introducing probabilistic asymmetric learning, the Modality-Invariant Asymmetric Network is presented to maintain the intra-modal and inter-modal semantic relationships, effectively relieving the semantic and heterogeneity gap [42]. Since the aforementioned pairwise methods easily converge to local minima, some researchers attempt to introduce the triplet-wise strategy into cross-modal retrieval to explore the neighborhood relations across different modalities. To capture the relative neighbourship between heterogeneous samples, Deng *et al.* introduce the triplet-wise ranking into the inter-modal and the intra-modal learning and propose the Triplet-based Deep Hashing (TDH) framework to preserve the original neighborhood structure between compact binary codes and achieve superior retrieval performance [32]. With efficient asymmetric online supervised strategies, the Flexible Online Multi-modal Hashing architecture is designed

to fuse heterogeneous modalities and narrow the semantic gap, generating discriminative binary codes.

### B. Unsupervised Cross-modal Hashing

Without human-annotated information, unsupervised cross-modal hashing is more flexible and applicable for real-world retrieval situations by learning the predefined similarity signals between heterogeneous data. By jointly calculating intra-similarity and intra-similarity across different modals, the Linear Cross-Modal Hashing (LCMH) framework is proposed to achieve scalable indexing for cross-modal retrieval [43]. With the high nonlinearity of Generative Adversarial Networks, Zhang *et al.* develop an Unsupervised Generative Adversarial Cross-modal Hashing to effectively mitigate the semantic gap across heterogeneous modalities and capture more semantic information [44]. By learning a joint-semantics affinity matrix and exploring joint-semantics relationships jointly, Deep Joint-Semantics Reconstructing Hashing is presented to maintain the original neighborhood structures from different modalities [45]. Recently, several researchers introduce contrastive learning mechanisms into unsupervised hashing and also have shown promising performance. By designing a novel momentum-based optimizer, Hu *et al.* integrate contrastive loss and cross-modal ranking into a unified framework to realize the Unsupervised Contrastive Cross-modal Hashing method architecture, minimizing the gap between contrastive learning and deep hashing [46].

### C. Mutual Information

In recent years, mutual information theory has been successfully applied to hashing retrieval, making substantial progress. Specifically, by utilizing mutual information to measure the quality of binary codes, the MIHash algorithm is designed to address the problem of hash table updating in online hashing [35]. With regarding mutual information as the learning objective, Cakir *et al.* expand and improve the MIHash algorithm by setting the batch learning, to maintain the original neighborhood relations [36]. By maximizing the mutual information between the hash-like features and feature representations, Hoang *et al.* propose the Cross-Modal Info-Max Hashing (CMIMH) framework to capture intramodal and intermodal similarities simultaneously and bridge the modality gaps [47]. Based on information theory, unsupervised cross-modal retrieval applications have shown promising results and provide a valuable contribution to the field. By extending the Quadratic Divergence Measures, Torkkola *et al.* introduce quadratic mutual information into feature extraction applications [38]. By setting quadratic mutual information as the kernel function, Bouzas *et al.* propose a new dimensionality reduction to optimize the graph embedding framework [48]. Based on knowledge distillation and quadratic mutual information, Passalis *et al.* design a novel representation learning method by matching the probability distribution of samples to replace actual representation [49]. Subsequently, for image retrieval application, the above authors propose an effective deep supervised hashing algorithm that utilizes an information-theoretic measure, the quadratic mutual information, to learn compact binary codes and achieve satisfactory performance [37]. Furthermore, Tzelepi *et al.* present a regularization strategy based on quadratic mutual information to improve the generalization of the model [50].

Compared with previous studies [35]–[37], by exploiting a novel extension of information-theoretic measures from the inter-modal view and the intra-modal view, the Quadratic Spherical Mutual Information is introduced into cross-modal hashing at the first attempt to quantize neighborhood ambiguity, maintaining initial neighborhood relationships among heterogeneous data from different modalities, while it is free of tuning parameters. To optimize quadratic mutual information loss effectively, a square clamping method is developed for Stochastic Gradient Descent in deep hashing. Besides, two transformer encoders are used for feature extractors to learn representative and informative characteristics from raw samples.

## III. METHODOLOGY

### A. Preliminaries

Suppose there is a dataset $\mathcal{D} = \{d_i\}_{i=1}^{N}$ with $N$ sample pairs, where $d_i = \{\mathcal{I}_i, \mathcal{T}_i\}$. $L = \{l_1 \ldots, l_i\}_{i=1}^{N}$ denotes the corresponding label, where $C$ in $l_i \in \{0,1\}^C$ denotes the number of categories. $F$ indicates the latent feature representation, which contains $F^{\mathcal{I}}$ and $F^{\mathcal{T}}$ to represent the latent semantic features of images and texts respectively. For given samples from image and textual modalities $\mathcal{I}$ and $\mathcal{T}$, and the corresponding labels $L$, our cross-modal hash aims to map high-dimensional samples to the discriminative discrete space by the hash function $\mathcal{H}^{\mathcal{I}}$ and $\mathcal{H}^{\mathcal{T}}$, in which the Hamming distance between similar samples is much smaller than the distance of semantically irrelevant samples. Since discretized optimization during model training is hard, the continuous relaxation strategy is introduced to hashing learning to achieve approximate optimization. In our proposed DNpH framework, the binary-like codes are converted to compact discrete codes by utilizing the $sign$ function ($B^{\mathcal{I}} = sign(\mathcal{H}^{\mathcal{I}})$ and $B^{\mathcal{T}} = sign(\mathcal{H}^{\mathcal{T}})$), formulated as follows.

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

In this paper, by treating quadratic spherical mutual information as the learning objective, we propose a supervised cross-modal hashing approach to minimize neighborhood ambiguity. Suppose that sample $d_i$ involves the semantically relevant subset $\mathcal{D}'_p$ and semantically irrelevant subset $\mathcal{D}'_n$. In our proposed DNpH framework, we employ an anchor point and $(\mathcal{D}'_p, \mathcal{D}'_n)$ to describe the neighborhood structure of $d_i$. If the two samples share at least one category, then the two samples are considered semantically relevant, otherwise, they are semantically irrelevant. Therefore, the key to cross-modal hashing is that the obtained compact binary codes from heterogeneous samples could preserve the original neighbor structure, reducing neighborhood ambiguity. For the obtained hash codes $B^{\mathcal{I}}$ and $B^{\mathcal{T}}$, the Hamming distance between sample $d_i$ and the data point in $\mathcal{D}'_p$ is less than the Hamming distance between the data point in $\mathcal{D}'_n$ and sample $d_i$, which could satisfy the following constraint.

$$\mathrm{Hd}(d_i, \mathcal{D}'_p) < \mathrm{Hd}(d_i, \mathcal{D}'_n) \tag{2}$$

### B. Feature Learning

For the feature learning module, most existing cross-modal hashing commonly adopts CNN-based networks to obtain the latent features from raw samples. Due to the limited fixed size of convolution kernels, these models only capture local semantic representations by convolution and loop operations, leading to the degradation of retrieval accuracy [51], [52]. In our work, by modeling the long-range dependencies, the transformer encoder is developed into cross-modal hashing as the feature extraction model to learn the global semantic information from raw heterogeneous samples.

The transformer encoder is introduced into computer vision applications for the first time in [53], achieving satisfactory performance. Benefiting from Multi-head Self-Attention (MSA) mechanism, the encoders with Transformer structures could capture excellent global representations from the raw samples by modeling long-range dependencies. Specifically, for the given vector $a_i$, the vectors $q_i$, $k_i$, and $v_i$ are obtained by linear transformations of the three matrices $W^q$, $W^k$, and $W^v$. After that, $q_i$, $k_i$, and $v_i$ are fed into attention blocks.

$$q_i = a_i W^q, k_i = a_i W^k, v_i = a_i W^v \tag{3}$$

$$attention_i = softmax(\frac{q_i * k_i^T}{\sqrt{d_k}}) * v_i \tag{4}$$

where $d_k$ is the dimension. The vector $b_i$ is obtained by concatenating the vectors of multi-head self-attention according to Eq 5.

$$b_i = concat(attention_1, attention_2, \ldots, attention_{head}) w^o \tag{5}$$

In this paper, we utilize the transformer encoder architecture as an image feature extractor to obtain global information about the image, including 12 layers of encoder blocks. And, each encoder block shares the same structure, which consists of Layernorm (LN), twelve Multi-head Self-Attention (MSA), and Multi-Layer Perceptron (MLP) blocks. Specifically, for the given image sample $\mathcal{I}_i$, the image is patched by convolution operation, and each patch is added to the corresponding position encoding $e^{\mathcal{I}}_{pos}$.

$$\mathcal{I}_e = \mathcal{I}_i + e^{\mathcal{I}}_{pos} \tag{6}$$

Finally, patches are fed into the image transformer encoder to get the image feature vector $F^{\mathcal{I}} = E^{\mathcal{I}}(\mathcal{I}_i, \theta_{\mathcal{I}})$, which $E^{\mathcal{I}}$ represents the image feature encoder and $\theta_{\mathcal{I}}$ represents the network parameters.

For text data sources, a multi-layer fully-connected layer is typically used as the feature extractor, often resulting in insufficient semantic information [34]. To optimize text feature learning, several researchers develop the transformer encoder structure in the field of natural language processing, which shows remarkable properties [54]. Therefore, we exploit the transformer encoder as the text feature encoder to learn critical semantic features from the text. Specifically, for the given text sample $\mathcal{T}_i$, each sample is encoded into tokens by introducing BPE encoding [55], and the corresponding position encoding $e^{\mathcal{T}}_{pos}$ is added to each token.

$$\mathcal{T}_e = \mathcal{T}_i + e^{\mathcal{T}}_{pos} \tag{7}$$

Subsequently, $\mathcal{T}_e$ is fed into the text feature encoder, which contains the same 12 layers of encoder blocks and involves 8 multi-head attentions in each layer. Finally, the text semantic features $F^{\mathcal{T}} = E^{\mathcal{T}}(\mathcal{T}_i, \theta_{\mathcal{T}})$ are generated by the text feature encoder, where $E^{\mathcal{T}}$ represents the text feature encoder and $\theta_{\mathcal{T}}$ indicates the parameters of the text network.

### C. Hash Learning

*1) Mutual Information:* Existing cross-modal hashing generally takes affinity matching or local ranking to eliminate the semantic gaps across heterogeneous modalities, while overlooking the original neighborhood relationships between samples [36], [37]. Therefore, these methods could produce the "Low Quality" Hamming space, as shown in Fig 1 and limit the retrieval performance of the network. In contrast to previous methods [30]–[33], we introduce mutual information into the cross-modal hashing framework to measure the quality of obtained binary codes and learn a highly separable Hamming space, as exhibited in 'High Quality' of Fig 1. Besides, it is free of tuning parameters compared with other similarity measures.

Cross-modal retrieval aims to use a sample from one modality as a query to retrieve similar data points from another modality, such as text-image retrieval, text-video retrieval, and so on. In this paper, we mainly direct our attention to cross-modal hashing retrieval between visual and textual modalities, which could learn a highly separable Hamming space. Suppose $Q^{\mathcal{I}} = \{q^{\mathcal{I}}_1 \ldots, q^{\mathcal{I}}_i\}^M_{i=1}$ and $Q^{\mathcal{T}} = \{q^{\mathcal{T}}_1 \ldots, q^{\mathcal{T}}_i\}^M_{i=1}$ denote the individual query requirements for the image and text modalities. Each $q_i$ in $Q^{\mathcal{I}}$ and $Q^{\mathcal{T}}$ has the corresponding binary code subsets from visual and textual modalities $B^{\mathcal{I}}_i = \{b^{\mathcal{I}}_{1(i)} \ldots, b^{\mathcal{I}}_{N'_i(i)}\}$ and $B^{\mathcal{T}}_i = \{b^{\mathcal{T}}_{1(i)} \ldots, b^{\mathcal{T}}_{N'_i(i)}\}$, in which $b^{\mathcal{I}(\mathcal{T})}_{j(i)}$ indicates the $j$-th image and text sample that satisfies query requirement $q_i$. Thus, the distributions of samples satisfying the $i$-th query can be represented by the conditional probability density function $p(b|q_i) = p(b^{\mathcal{I}}|q_i) + p(b^{\mathcal{T}}|q_i)$.

Based on Shannon's entropy, the uncertainty of randomly selected samples in returned ranking list is defined as $H(Q) = -\sum_q P(q)\log(P(q))$, where $P(q)$ represents the prior probability of the samples satisfying the query requirement $q$, which is called as the prior probability of $q$. If the query $q$ is known, by counting conditional entropy, the uncertainty of information needs that it meets could be computed as follows.

$$H(Q|B) = -\int_b p(b)(\sum_q p(q|b)\log(p(q|b)))db \tag{8}$$

Thus, mutual information between query requirement $Q$ and binary codes $B$ can be defined as follows.

$$I(Q,B) = H(Q) - H(Q|B) = \sum_q \int_b p(q,b) \log(\frac{p(q,b)}{P(q)p(b)}) db \tag{9}$$

It is observed from Eq.( 9) that the mutual information (MI) between the query requirement $Q$ and discrete codes $B$ in the returned ranking list can be maximized by maximizing the KL divergence between the joint probability density $p(q,b)$ and $P(q)p(b)$, where $p(b)$ is the density of all samples. However, during the optimization modeling, it is very difficult to calculate $p(b|q_i)$ and integration of Eq.( 9) directly [37]. Therefore, in this paper, we propose an algorithm to efficiently estimate $p(b|q_i)$, and design an effective mechanism for cross-modal retrieval based on deep hashing.

*2) Cross-modal Quadratic Spherical Mutual Information:* By calculating the conditional probability density function between queries and returned result lists, the mutual information could be exploited to model the inherent uncertainty of query vectors. However, the computation of the probability density in mutual information is still a difficult problem, leading to application limitations in the real retrieval scenario. Different from previous works [35], [36], [47], quadratic mutual information is introduced into cross-modal hashing to replace mutual information, in which the Parzen window estimation mechanism is suggested to estimate the probability density. Therefore, the quadratic mutual information in cross-modal retrieval is formulated as follows.

$$\begin{aligned} I_q(Q,B) = &\sum_q \int_{b^{\text{intra}}} \left(p(q,b^{\text{intra}}) - P(q)p(b^{\text{intra}})\right)^2 db \\ &+ \sum_q \int_{b^{\text{inter}}} \left(p(q,b^{\text{inter}}) - P(q)p(b^{\text{inter}})\right)^2 db \end{aligned} \tag{10}$$

where $b^{\text{intra}}$ represents samples from same modality in query $q$ and $b^{\text{inter}}$ indicates samples from different modality in query requirement. Based on quadratic divergence [38], the quadratic mutual information could be re-formulated as follows.

$$I_q(Q,B) = V_{IN}(Q,B) + V_{ALL}(Q,B) - 2V_{BTW}(Q,B) \tag{11}$$

$$\begin{aligned} V_{IN}(Q,B) = &\sum\nolimits_q \int_{b^{\text{intra}}} p(q,b^{\text{intra}})^2 db \\ &+ \sum\nolimits_q \int_{b^{\text{inter}}} p(q,b^{\text{inter}})^2 db \end{aligned} \tag{12}$$

$$\begin{aligned} V_{ALL}(Q,B) = &\sum\nolimits_q \int_{b^{\text{intra}}} P(q)^2 p(b^{\text{intra}})^2 db \\ &+ \sum\nolimits_q \int_{b^{\text{inter}}} P(q)^2 p(b^{\text{inter}})^2 db \end{aligned} \tag{13}$$

$$\begin{aligned} V_{BTW}(Q,B) = &\sum\nolimits_q \int_{b^{\text{intra}}} p(q,b^{\text{intra}}) P(q) p(b^{\text{intra}}) db \\ &+ \sum\nolimits_q \int_{b^{\text{inter}}} p(q,b^{\text{inter}}) P(q) p(b^{\text{inter}}) db \end{aligned} \tag{14}$$

$V_{IN}(Q,B)$ expresses the interaction of Intra-modal and inter-modal samples in query requirement, $V_{ALL}(Q,B)$ represents the interaction between all samples from different modalities, and $V_{BTW}(Q,B)$ shows the interaction of all intra-modal and inter-modal samples, which meets a specific query. By analyzing the above equation, it can be observed that $P(q)$, $p(b)$, and $p(q,b)$ need to be calculated. $P(q_i) = \frac{N'_i}{N}$, where $N'_i$ is the number of samples in the $i$-th query requirement. The Parzen window estimation method is proposed to calculate $p(q,b)$ and $p(b)$, as shown below.

$$\begin{aligned} p(b^{\text{intra}}|q_i) &= \frac{1}{{N'}_i} \sum_{j=1}^{N'_i} K(b^{\text{intra}} - b^{\text{intra}}_{j(i)}; \sigma^2) \\ p(b^{\text{inter}}|q_i) &= \frac{1}{{N'}_i} \sum_{j=1}^{N'_i} K(b^{\text{inter}} - b^{\text{inter}}_{j(i)}; \sigma^2) \end{aligned} \tag{15}$$

where $K(b;\sigma^2)$ is a Gaussian kernel with width $\sigma$.

Hence, $p(q_i,b) = p(b^{\text{intra}}|q_i)p(q_i) + p(b^{\text{inter}}|q_i)p(q_i) = \frac{1}{N}\sum_{j=1}^{N'_i} K(b^{\text{intra}} - b^{\text{intra}}_{j(i)}; \sigma^2) + \frac{1}{N}\sum_{j=1}^{N'_i} K(b^{\text{inter}} - b^{\text{inter}}_{j(i)}; \sigma^2)$.

In addition, $p(b)$ is calculated according to the following Eq.( 16).

$$\begin{aligned} p(b) = &\frac{1}{N} \sum_{j=1}^{N} K(b^{\text{intra}} - b^{\text{intra}}_{j(i)}; \sigma^2) \\ &+ \frac{1}{N} \sum_{j=1}^{N} K(b^{\text{inter}} - b^{\text{inter}}_{j(i)}; \sigma^2) \end{aligned} \tag{16}$$

By putting the above calculations to $V_{IN}(Q,B)$, $V_{ALL}(Q,B)$ and $V_{BTW}(Q,B)$, the quadratic mutual information could be reformulated as indicated below.

$$\begin{aligned} V_{IN}(Q,B) = &\frac{1}{N^2} \sum_{k=1}^{M} \sum_{i=1}^{N'_k} \sum_{j=1}^{N'_k} K(b^{\text{intra}}_{i(k)} - b^{\text{intra}}_{j(k)}, 2\sigma^2) \\ &+ \frac{1}{N^2} \sum_{k=1}^{M} \sum_{i=1}^{N'_k} \sum_{j=1}^{N'_k} K(b^{\text{inter}}_{i(k)} - b^{\text{inter}}_{j(k)}, 2\sigma^2) \end{aligned} \tag{17}$$

$$\begin{aligned} V_{ALL}(Q,B) = &\frac{1}{N^2} \left(\sum_{k=1}^{M} \left(\frac{N'_k}{N}\right)^2\right) \sum_{i=1}^{N} \sum_{j=1}^{N} K(b^{\text{intra}}_i - b^{\text{intra}}_j, 2\sigma^2) \\ &+ \frac{1}{N^2} \left(\sum_{k=1}^{M} \left(\frac{N'_k}{N}\right)^2\right) \sum_{i=1}^{N} \sum_{j=1}^{N} K(b^{\text{inter}}_i - b^{\text{inter}}_j, 2\sigma^2) \end{aligned} \tag{18}$$

$$\begin{aligned} V_{BTW}(Q,B) = &\frac{1}{N^2} \sum_{k=1}^{M} \left(\left(\frac{N'_k}{N}\right) \sum_{i=1}^{N'_k} \sum_{j=1}^{N} K(b^{\text{intra}}_{i(k)} - b^{\text{intra}}_j, 2\sigma^2)\right) \\ &+ \frac{1}{N^2} \sum_{k=1}^{M} \left(\left(\frac{N'_k}{N}\right) \sum_{i=1}^{N'_k} \sum_{j=1}^{N} K(b^{\text{inter}}_{i(k)} - b^{\text{inter}}_j, 2\sigma^2)\right) \end{aligned} \tag{19}$$

In our work, the Parzen window estimation with Gaussian kernel is developed to estimate the probability density, we

exploit the Euclidean distance to calculate the similarity of intra-modal and inter-modal samples. For the training procedures, although the distribution of mutual information could be maximized by introducing quadratic mutual information, there are some restrictions and limitations [37]. (1) It is a delicate matter to choose the applicable width values in the Gaussian functions, and non-optimal choices allow the model to fall into bad local minima, leading to performance degradation. (2) While utilizing Euclidean distances to measure the Hamming similarity between samples, it could cause the variations of distance consistent between binary codes and binary-like embedding, which generates suboptimal hash codes.

To overcome the aforementioned limitations, we propose the Quadratic Spherical Mutual Information (QSMI) loss as the objective function to optimize deep hashing networks. Different from using the Gaussian kernel in previous studies, the similarity information is defined by calculating the cosine distance, which is calculated as follows.

$$S_{\cos}(b_1, b_2) = \frac{1}{2}(\frac{b_1^T b_2}{\parallel b_1\parallel_2 \parallel b_2\parallel_2} + 1) \tag{20}$$

where $\parallel \cdot \parallel_2$ represents the $l_2$ norm. Without choosing the intractable width values in the Gaussian kernel, the cosine distance is employed to measure the similarity between samples. Meanwhile, the cosine distance could be more approximate to the discrete space mathematically. Therefore, based on the cosine similarity, our proposed QSMI could be calculated as follows.

$$I_q^{\cos}(Q, B) = V_{IN}^{\cos}(Q, B) + V_{ALL}^{\cos}(Q, B) - 2V_{BTW}^{\cos}(Q, B) \tag{21}$$

$$\begin{aligned} V_{IN}^{\cos}(Q, B) = &\frac{1}{N^2}\sum_{k=1}^{M}\sum_{i=1}^{N'_k}\sum_{j=1}^{N'_k} S_{\cos}(b_{i(k)}^{\text{intra}}, b_{j(k)}^{\text{intra}}) \\ &+ \frac{1}{N^2}\sum_{k=1}^{M}\sum_{i=1}^{N'_k}\sum_{j=1}^{N'_k} S_{\cos}(b_{i(k)}^{\text{inter}}, b_{j(k)}^{\text{inter}}) \end{aligned} \tag{22}$$

$$\begin{aligned} V_{ALL}^{\cos}(Q, B) = &\frac{1}{N^2}(\sum_{k=1}^{M}(\frac{N'_k}{N})^2)\sum_{i=1}^{N}\sum_{j=1}^{N} S_{\cos}(b_i^{\text{intra}}, b_j^{\text{intra}}) \\ &+ \frac{1}{N^2}(\sum_{k=1}^{M}(\frac{N'_k}{N})^2)\sum_{i=1}^{N}\sum_{j=1}^{N} S_{\cos}(b_i^{\text{inter}}, b_j^{\text{inter}}) \end{aligned} \tag{23}$$

$$\begin{aligned} V_{BTW}^{\cos}(Q, B) = &\frac{1}{N^2}\sum_{k=1}^{M}((\frac{N'_k}{N})\sum_{i=1}^{N'_k}\sum_{j=1}^{N} S_{\cos}(b_{i(k)}^{\text{intra}}, b_j^{\text{intra}})) \\ &+ \frac{1}{N^2}\sum_{k=1}^{M}((\frac{N'_k}{N})\sum_{i=1}^{N'_k}\sum_{j=1}^{N} S_{\cos}(b_{i(k)}^{\text{inter}}, b_j^{\text{inter}})) \end{aligned} \tag{24}$$

When the query requirements are equiprobable, i.e. $P(q) = \frac{1}{M}$, QSMI can be written as $I_q^{\cos}(Q, B) = V_{IN}^{\cos}(Q, B) - V_{BTW}^{\cos}(Q, B)$. With a feature similarity matrix $S \in R^{N\times N}$, the QSMI could be reformulated as follows.

$$\begin{aligned} I_q^{\cos} = &\frac{1}{N^2}\mathbf{1}_N^T(\Delta \odot S^{\mathcal{I}} - \frac{1}{M}S^{\mathcal{I}})\mathbf{1}_N \\ &+ \frac{1}{N^2}\mathbf{1}_N^T(\Delta \odot S^{\mathcal{T}} - \frac{1}{M}S^{\mathcal{T}})\mathbf{1}_N \\ &+ \frac{1}{N^2}\mathbf{1}_N^T(\Delta \odot S^{\mathcal{IT}} - \frac{1}{M}S^{\mathcal{IT}})\mathbf{1}_N \end{aligned} \tag{25}$$

where $M$ represents the number of query requirements, and $\Delta$ is defined as shown below.

$$[\Delta] = \begin{cases} 1, & common\_categories \\ 0, & otherwise \end{cases} \tag{26}$$

*3) Square Clamp Optimization:* During the training stage, the Quadratic Spherical Mutual Information loss (QSMI) is optimized directly, which could fall into a local optimal [37]. To address this shortcoming, A square clamping strategy is proposed to optimize smoothly the proposed objective function. Besides, the feature similarity matrix $S$ from intra-modal and inter-modal is explored jointly to alleviate inter-modal heterogeneity. Based on the given similarity matrix $S$ in the unit interval, the QSMI could be reformulated, as illustrated in Eq.( 27), Eq.( 28) and Eq.( 29).

$$L_{QSMI}^{\mathcal{I}} = \frac{1}{N^2}\mathbf{1}_N^T(\Delta\odot(S^{\mathcal{I}}-1)\odot(S^{\mathcal{I}}-1)+\frac{1}{M}(S^{\mathcal{I}}\odot S^{\mathcal{I}}))\mathbf{1}_N \tag{27}$$

$$L_{QSMI}^{\mathcal{T}} = \frac{1}{N^2}\mathbf{1}_N^T(\Delta\odot(S^{\mathcal{T}}-1)\odot(S^{\mathcal{T}}-1)+\frac{1}{M}(S^{\mathcal{T}}\odot S^{\mathcal{T}}))\mathbf{1}_N \tag{28}$$

$$L_{QSMI}^{\mathcal{IT}} = \frac{1}{N^2}\mathbf{1}_N^T(\Delta\odot(S^{\mathcal{IT}}-1)\odot(S^{\mathcal{IT}}-1)+\frac{1}{M}(S^{\mathcal{IT}}\odot S^{\mathcal{IT}}))\mathbf{1}_N \tag{29}$$

Thus, by jointly calculating the QSMI from the inter-modal view and the intra-modal view, the ultimate overall loss function could be reformulated, as shown in Eq.( 30).

$$L_{QSMI} = L_{QSMI}^{\mathcal{I}} + L_{QSMI}^{\mathcal{T}} + L_{QSMI}^{\mathcal{IT}} \tag{30}$$

Remarkably, our proposed DNpH framework does not directly generate discriminative hash codes with semantic information preserved after optimizing the deep network and only obtains continuous hash-like features with different bits which have values within $[-1, 1]$. Besides, testing samples from heterogeneous modalities are fed into the trained network, and we utilize the element-wise transformation $sign(\cdot)$ to project continuous real-valued features into corresponding discrete codes. The optimization procedure of our proposed DNpH framework is listed in Algorithm 1. During the query phase, for the given query samples, the returned ranking lists are obtained by computing Hamming distance between the query and candidate database.

**Algorithm 1** Learning algorithm for DNpH

**Input:**

Training dataset $\mathcal{D} = \{d_i\}_{i=1}^{N}$; Binary codes length $K$.

**Output:**

Parameters $\Theta_{\mathcal{I}}$ and $\Theta_{\mathcal{T}}$ for DNpH.

```
Initialize network parameters Θ_I and Θ_T.
while epoch < max_epoch do
    Extract binary-like embedding by forward-propagation through image and text feature encoders, respectively.
    Calculate QSMI to to optimize the deep hashing network according to Eq.( 30).
    Update parameters Θ_I and Θ_T via back-propagation.
end while
return The trained DNpH model.
```

## IV. EXPERIMENTS

To verify and evaluate the comprehensive performance of our proposed DNpH, extensive contrastive experiments are carried out on four common benchmark datasets, including MIRFLICKR-25K[1], NUS-WIDE[2], MS COCO[3] and IAPR TC-12[4].

### A. Datasets

**MIRFLICKR-25K** is a small cross-modal retrieval dataset frequently collected on the Flickr website. It includes 24581 image-text pairs corresponding to 24 classes, in which each sample pair belongs to at least one category.

**NUS-WIDE** dataset contains 269,648 image-text pairs, where each of them belongs to at least one of the 81 categories. However, some categories include a very small number of sample pairs, we removed several categories according to [30]. The processed dataset involves 195,834 sample pairs corresponding to 21 common categories.

**MS COCO** dataset is widely used in the computer vision field, including object detection, semantic segmentation, and multimedia retrieval. Similar to the previous protocol [34], we merged 82,785 training sets and 40,504 validation sets, which contain a total of 80 different categories.

**IAPR TC-12** dataset includes different static nature images and their corresponding text descriptions, totaling 20,000 image-text pairs, and each sample pair belongs to at least one of 291 categories.

In our experiments, the four benchmark datasets are divided in the same way according to [34]. Specifically, 10,000 image-text pairs are randomly selected as the training set, 5,000 sample pairs are selected as the query set, and the remainder is used to formulate the database set.

### B. Experimental Settings

*1) Baseline Methods:* To prove the effectiveness of our designed algorithm, we compare the DNpH framework with ten typical deep cross-modal hashing, including DCMH [30], SSAH [31], CMHH [56], AGAH [33], DADH [57], SCAHN [58], MESDCH [59], DCHMT [60], Bi−NCMH [61], and MIAN [42]. The implementation of comparison methods is based on the officially published source code, and the network parameters are set according to the description of open-source codes or original papers.

*2) Implementation Details:* Our proposed DNpH method is implemented with the open-source Pytorch 1.12.1 framework [62], and all comparative experiments are carried out on a Linux server with an NVIDIA RTX 3090. The Adam optimizer is utilized to implement the optimization process, where the learning rate and the batch size are respectively set to 0.001 and 128.

*3) Evaluation Metrics:* To prove the retrieval performance of cross-modal hashing systematically, five evaluation metrics are employed in our experiments, which include mean Average Precision (mAP), Precision within Hamming radius 2 (P@H$\leq$2), Precision-Recall (PR) curves, TopN-precision curves, and Normalized Discounted Cumulative Gain (NDCG)@1000. The mAP result refers to the mean value of the average precision and is a commonly available performance indicator in the information retrieval area. P@H$\leq$2 presents the mean value of the average accuracy within the Hamming radius 2. The PR curve shows the relationship between recall rate and precise rate, which describes the overall retrieval performance, and the TopN-precision curve calculates the precision with the returned TopN samples. NDCG@1000 is a commonly used metric to evaluate retrieval applications. If the returned rank samples have high similarity, the value is increased accordingly. In our experiments, the number of rank is set to 1000.

### C. Comparison with the Baselines

For cross-modal retrieval applications, Hamming Ranking and Hash Lookup are the most popular retrieval protocols. To validate the retrieval performance of our method DNpH method, five commonly-used evaluation metrics are calculated on four common benchmark datasets. Specifically, we report the mAP results of the baselines and our DNpH w.r.t. 16bits, 32bits, 64bits for two retrieval tasks, i.e. Image-to-Text retrieval (I2T), Text-to-Image retrieval (T2I), and Table I exhibits detailed results. It is worth noting that $*$ denotes the results of Bi-NCMH are directly cited from the original paper since the codes are unavailable. Since the Quadratic Spherical Mutual Information is introduced to cross-modal hashing to minimize neighborhood ambiguity, our proposed DNpH method achieves the whole optimum performance than state-of-the-art cross-modal hashing. Specifically, for the MIRFLICKR-25K dataset, our DNpH framework achieves superior performance, which achieves 1.19% and 0.88% improvement on I2T andT2I over the best baseline MIAN with three different hash lengths. The minimal improvements of the I2T and T2I tasks on the NUSWIDE dataset are 2.08% and 1.96%, respectively. For the MS COCO and IAPR TC-12 datasets, our proposed framework achieves an increase of 23.94% and 23.37% at most. Besides, there are some interesting observations that both the number

[1]https://press.liacs.nl/mirflickr/

[2]https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html

[3]https://cocodataset.org/

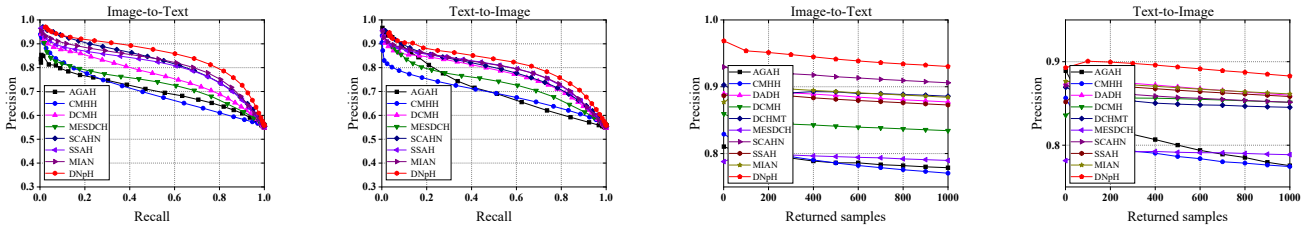
[4]https://www.imageclef.org/photodata

TABLE I: Comparison with baselines in terms of mAP results w.r.t. 16bits, 32bits, 64bits on MIRFLICKR-25K, NUS-WIDE, MS COCO and IAPR TC-12. $*$ denotes the results are directly cited from the original paper.

| Task | Method | MIRFLICKR-25K | | | NUS-WIDE | | | MS COCO | | | IAPR TC-12 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 16bits | 32bits | 64bits | 16bits | 32bits | 64bits | 16bits | 32bits | 64bits | 16bits | 32bits | 64bits |
| I2T | DCMH | 0.7687 | 0.7736 | 0.7797 | 0.5379 | 0.5513 | 0.5617 | 0.5399 | 0.5444 | 0.5627 | 0.4595 | 0.4875 | 0.5063 |
| | SSAH | 0.8079 | 0.8129 | 0.8220 | 0.6032 | 0.6058 | 0.6095 | 0.5411 | 0.4855 | 0.5395 | 0.5018 | 0.5240 | 0.5415 |
| | CMHH | 0.6932 | 0.6979 | 0.6984 | 0.5439 | 0.5546 | 0.5520 | 0.5145 | 0.4509 | 0.5209 | 0.4780 | 0.4532 | 0.4496 |
| | AGAH | 0.7248 | 0.7217 | 0.7195 | 0.3945 | 0.4107 | 0.4258 | 0.5501 | 0.5515 | 0.5518 | - | - | - |
| | DADH | 0.8098 | 0.8162 | 0.8193 | 0.6350 | 0.6568 | 0.6546 | 0.5952 | 0.6118 | 0.6237 | 0.5355 | 0.5622 | 0.5712 |
| | SCAHN | 0.7955 | 0.8248 | 0.8297 | 0.6463 | 0.6616 | 0.6645 | 0.6376 | 0.6475 | 0.6519 | 0.5083 | 0.5543 | 0.5660 |
| | MESDCH | 0.7258 | 0.7438 | 0.7421 | 0.5638 | 0.5801 | 0.5875 | 0.5759 | 0.5670 | 0.5636 | 0.5436 | 0.5698 | 0.5868 |
| | DCHMT | 0.8201 | 0.8253 | 0.8222 | 0.6596 | 0.6706 | 0.6863 | 0.6309 | 0.6216 | 0.6553 | 0.6021 | 0.6196 | 0.6254 |
| | Bi-NCMH$*$ | 0.7900 | 0.8000 | 0.8080 | 0.5150 | 0.5260 | 0.5550 | - | - | - | - | - | - |
| | MIAN | 0.8262 | 0.8444 | 0.8501 | 0.6431 | 0.6548 | 0.6625 | 0.5952 | 0.5987 | 0.6121 | 0.5410 | 0.5657 | 0.5774 |
| | **DNpH** | **0.8423** | **0.8552** | **0.8588** | **0.6921** | **0.7022** | **0.7071** | **0.6727** | **0.6903** | **0.6860** | **0.6359** | **0.6633** | **0.6797** |
| T2I | DCMH | 0.7857 | 0.7998 | 0.8029 | 0.5747 | 0.5810 | 0.5853 | 0.5271 | 0.5424 | 0.5450 | 0.5183 | 0.5376 | 0.5472 |
| | SSAH | 0.8089 | 0.8127 | 0.8017 | 0.6011 | 0.6058 | 0.6167 | 0.4901 | 0.4798 | 0.5053 | 0.5225 | 0.5383 | 0.5311 |
| | CMHH | 0.7181 | 0.7104 | 0.7294 | 0.4956 | 0.4831 | 0.4820 | 0.4910 | 0.4930 | 0.4889 | 0.4627 | 0.4449 | 0.4719 |
| | AGAH | 0.7082 | 0.7182 | 0.7344 | 0.4344 | 0.3980 | 0.4382 | 0.5012 | 0.5146 | 0.5191 | - | - | - |
| | DADH | 0.8019 | 0.8101 | 0.8137 | 0.6111 | 0.6182 | 0.6218 | 0.5649 | 0.5790 | 0.5870 | 0.5530 | 0.5801 | 0.5953 |
| | SCAHN | 0.7826 | 0.8066 | 0.8064 | 0.6587 | 0.6626 | 0.6648 | 0.6377 | 0.6512 | 0.6493 | 0.4973 | 0.5410 | 0.5498 |
| | MESDCH | 0.7385 | 0.7387 | 0.7437 | 0.5562 | 0.5720 | 0.5803 | 0.5044 | 0.5083 | 0.5085 | 0.5251 | 0.5423 | 0.5500 |
| | DCHMT | 0.7983 | 0.8048 | 0.8031 | 0.6761 | 0.6837 | 0.6943 | 0.6241 | 0.6212 | 0.6486 | 0.6068 | 0.6286 | 0.6387 |
| | Bi-NCMH$*$ | 0.7500 | 0.7600 | 0.7650 | 0.5500 | 0.5730 | 0.5740 | - | - | - | - | - | - |
| | MIAN | 0.8139 | 0.8180 | 0.8216 | 0.6733 | 0.6922 | 0.6936 | 0.5947 | 0.6000 | 0.6067 | 0.5383 | 0.5712 | 0.5865 |
| | **DNpH** | **0.8147** | **0.8292** | **0.8361** | **0.6992** | **0.7137** | **0.7139** | **0.6562** | **0.6860** | **0.6928** | **0.6480** | **0.6786** | **0.6978** |

TABLE II: Comparison with baselines in terms of NDCG@1000 results w.r.t. 16bits, 32bits, 64bits on MIRFLICKR-25K, NUS-WIDE, MS COCO and IAPR TC-12.

| Task | Method | MIRFLICKR-25K | | | NUS-WIDE | | | MS COCO | | | IAPR TC-12 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 16bits | 32bits | 64bits | 16bits | 32bits | 64bits | 16bits | 32bits | 64bits | 16bits | 32bits | 64bits |
| I2T | DCMH | 0.4002 | 0.4156 | 0.4300 | 0.4396 | 0.4794 | 0.4794 | 0.1646 | 0.1783 | 0.1705 | 0.2934 | 0.3125 | 0.3246 |
| | SSAH | 0.4578 | 0.4671 | 0.4738 | 0.4307 | 0.4406 | 0.4272 | 0.1788 | 0.1411 | 0.1991 | 0.3205 | 0.3477 | 0.3667 |
| | CMHH | 0.3299 | 0.3408 | 0.3393 | 0.3947 | 0.4121 | 0.4113 | 0.1725 | 0.1547 | 0.1661 | 0.2977 | 0.2771 | 0.2727 |
| | AGAH | 0.3506 | 0.3706 | 0.3677 | 0.2174 | 0.2984 | 0.3314 | 0.1836 | 0.1614 | 0.1775 | - | - | - |
| | SCAHN | 0.4480 | 0.4811 | 0.4876 | 0.4793 | 0.4966 | 0.4933 | 0.3070 | 0.3485 | 0.3413 | 0.3644 | 0.4134 | 0.4222 |
| | MESDCH | 0.3550 | 0.3650 | 0.3529 | 0.3934 | 0.4202 | 0.4279 | 0.2311 | 0.2324 | 0.2322 | 0.3735 | 0.4026 | 0.4190 |
| | DCHMT | 0.4310 | 0.3876 | 0.4137 | 0.4534 | 0.4708 | 0.4729 | 0.2557 | 0.2328 | 0.3059 | 0.3231 | 0.3403 | 0.3390 |
| | MIAN | 0.4881 | 0.5126 | 0.5202 | 0.4962 | 0.4716 | 0.5069 | 0.2218 | 0.2424 | 0.2698 | 0.3770 | 0.3997 | 0.4064 |
| | **DNpH** | **0.5022** | **0.5266** | **0.5430** | **0.5086** | **0.5337** | **0.5526** | **0.3388** | **0.4006** | **0.4107** | **0.4032** | **0.4371** | **0.4566** |
| T2I | DCMH | 0.4111 | 0.4200 | 0.4345 | 0.4316 | 0.4414 | 0.4383 | 0.1853 | 0.1962 | 0.1986 | 0.3334 | 0.3573 | 0.3656 |
| | SSAH | 0.4307 | 0.4465 | 0.4298 | 0.4269 | 0.4560 | 0.4700 | 0.1689 | 0.1551 | 0.1883 | 0.3273 | 0.3429 | 0.3297 |
| | CMHH | 0.3567 | 0.3668 | 0.4028 | 0.3277 | 0.3544 | 0.3480 | 0.1606 | 0.1583 | 0.1499 | 0.2639 | 0.2563 | 0.2990 |
| | AGAH | 0.3657 | 0.3837 | 0.4121 | 0.2567 | 0.2841 | 0.2965 | 0.1626 | 0.1635 | 0.1547 | - | - | - |
| | SCAHN | 0.4079 | 0.4264 | 0.4208 | 0.4561 | 0.4523 | 0.4552 | 0.3139 | 0.3650 | 0.3536 | 0.3522 | 0.3949 | 0.4030 |
| | MESDCH | 0.3731 | 0.3645 | 0.3728 | 0.3820 | 0.3990 | 0.4060 | 0.2236 | 0.2234 | 0.2191 | 0.3651 | 0.3926 | 0.4038 |
| | DCHMT | 0.3772 | 0.3869 | 0.3663 | 0.4442 | 0.4431 | 0.4618 | 0.2554 | 0.2550 | 0.3079 | 0.3539 | 0.3715 | 0.3836 |
| | MIAN | **0.4427** | 0.4454 | 0.4541 | 0.4749 | 0.4889 | 0.5114 | 0.2402 | 0.2589 | 0.2760 | 0.3728 | 0.3915 | 0.3908 |
| | **DNpH** | 0.4230 | **0.4517** | **0.4615** | **0.4974** | **0.5067** | **0.5287** | **0.3332** | **0.4004** | **0.4126** | **0.4384** | **0.4731** | **0.4982** |




(a) PR curve @32bits (b) PR curve @32bits (c) TopN Precision @32bits (d) TopN Precision @32bits

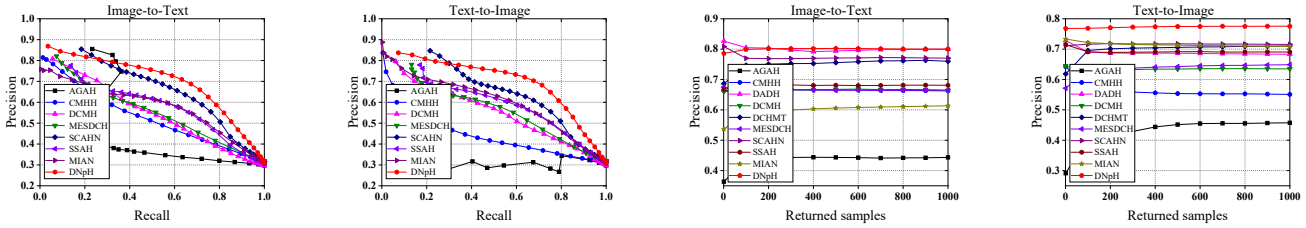
Fig. 3: Results of Precision-Recall curve and TopN Precision on MIRFLICKR-25K dataset w.r.t.32bits.

of categories and the total number of samples could impact the performance of cross-modal hashing. For example, the number of categories involved in the MIRFLICKR-25K dataset and the NUS-WIDE dataset are 24 and 21, there is a considerable difference (large differences in mAP results are up to 15.3%) in retrieval performance. We consider that the number of training samples could affect retrieval performance, in which the MIRFLICKR-25K dataset only includes 25,000 sample pairs, and the NUS-WIDE contains 195,834 sample pairs.

In addition, we also calculate NDCG@1000 values w.r.t.

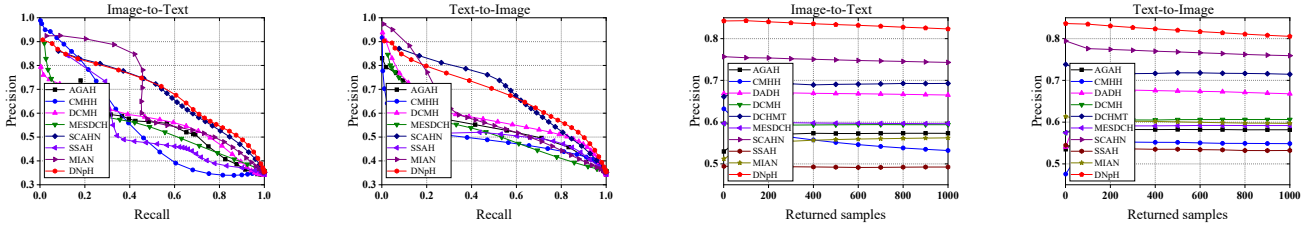
This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3349075





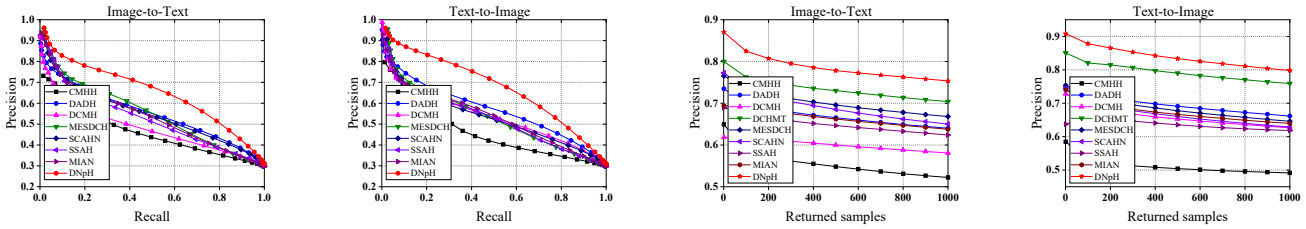

(a) PR curve @32bits (b) PR curve @32bits (c) TopN Precision @32bits (d) TopN Precision @32bits

Fig. 4: Results of Precision-Recall curve and TopN Precision on NUS-WIDE dataset w.r.t.32bits.



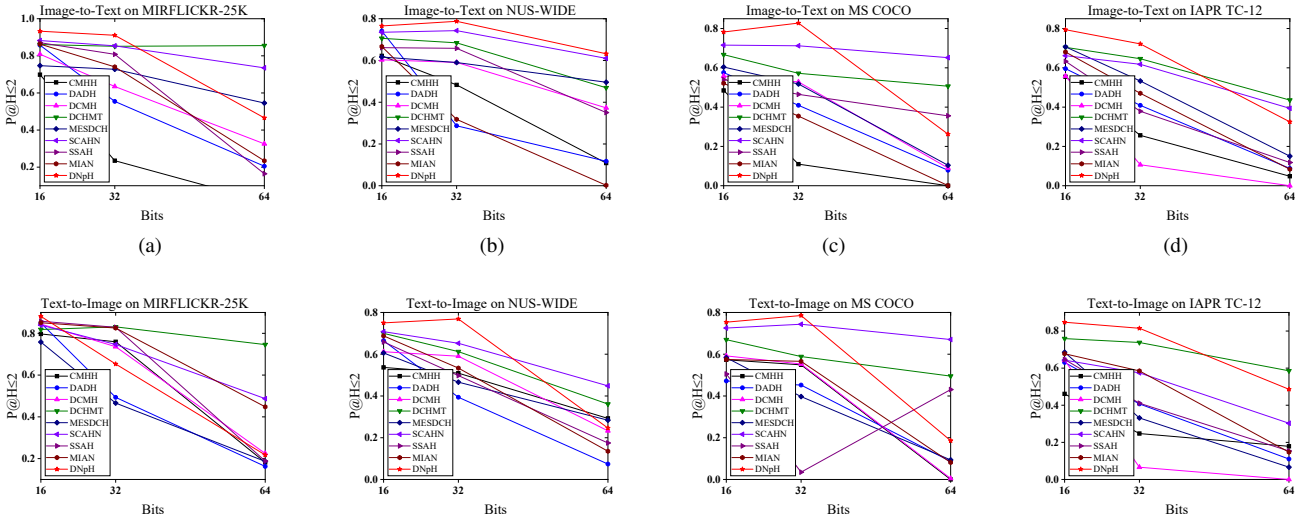

(a) PR curve @32bits (b) PR curve @32bits (c) TopN Precision @32bits (d) TopN Precision @32bits

Fig. 5: Results of Precision-Recall curve and TopN Precision on MS COCO dataset w.r.t.32bits.




(a) PR curve @32bits (b) PR curve @32bits (c) TopN Precision @32bits (d) TopN Precision @32bits

Fig. 6: Results of Precision-Recall curve and TopN Precision on IAPR TC-12 dataset w.r.t.32bits.

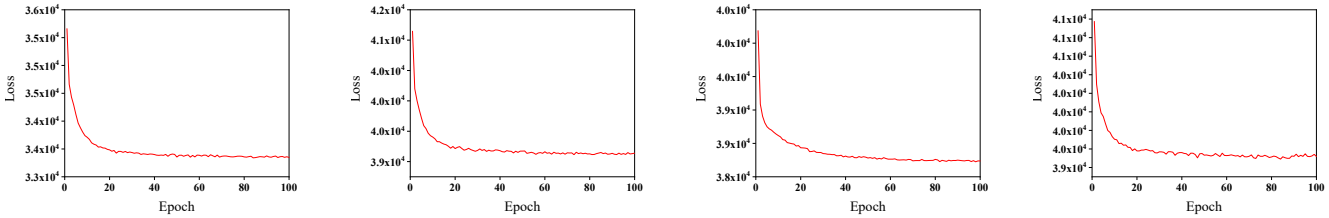



(a) (b) (c) (d)

(e) (f) (g) (h)

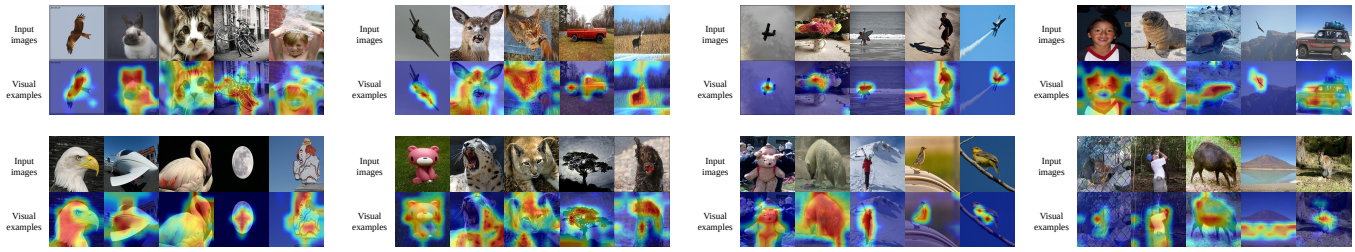
Fig. 7: P@H$\leq$2 results of DNpH and comparison methods on MIRFLICKR25K, NUS-WIDE, MS COCO and IAPR TC-12.

16bits, 32bits, 64bits to further demonstrate the effectiveness of our method, and the results are shown in Table II. From the experimental results on MIRFLICKR-25K, NUS-WIDE, MS COCO, and IAPR TC-12, we can see that our DNpH method achieves the best performance compared to other baselines. That is because two feature extractors based on the transformer encoder are introduced to learn informative semantic representations and the QSMI loss is employed to learn a highly separable Hamming embedding. Meanwhile, for a fair comparison, we replace the backbone of several comparison methods with the transformer encoder designed in our method to prove the effectiveness of our proposed QSMI loss. As displayed in Table III and Table IV, '-T' indicates that comparison methods employ the transformer encoders from our proposed DNpH framework as feature extractors. According to the experimental results, based on the same




(a) MIRFLICKR-25K (b) NUS-WIDE (c) MS COCO (d) IAPR TC-12

Fig. 8: Convergence analysis of DNpH w.r.t. 64bits on MIRFLICKR-25K, NUS-WIDE, MS COCO and IAPR TC-12.




(a) MIRFLICKR-25K (b) NUS-WIDE (c) MS COCO (d) IAPR TC-12

Fig. 9: Some visual examples of the learned global semantic maps from the image feature encoder in DNpH on MIRFLICKR-25K, NUS-WIDE, MS COCO, and IAPR TC-12.

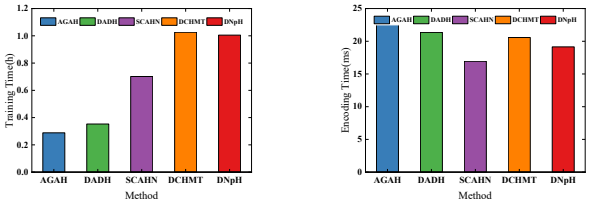
TABLE III: Comparison with the baseline based on transformer encoder in terms of mAP results w.r.t. 16bits, 32bits, 64bits on MIRFLICKR-25K and IAPR TC-12.

| Task | Method | MIRFLICKR-25K | | | IAPR TC-12 | | |
|---|---|---|---|---|---|---|---|
| | | 16bits | 32bits | 64bits | 16bits | 32bits | 64bits |
| I2T | DCMH-T | 0.8278 | 0.8423 | 0.8401 | 0.6223 | 0.6487 | 0.6543 |
| | CMHH-T | 0.8111 | 0.8204 | 0.8252 | 0.5710 | 0.6015 | 0.6108 |
| | DADH-T | 0.8230 | 0.8323 | 0.8458 | 0.6100 | 0.6396 | 0.6638 |
| | DCHMT-T | 0.8201 | 0.8253 | 0.8222 | 0.6021 | 0.6196 | 0.6254 |
| | **DNpH** | **0.8423** | **0.8552** | **0.8588** | **0.6359** | **0.6633** | **0.6797** |
| I2T | DCMH-T | 0.8088 | 0.8201 | 0.8311 | 0.6281 | 0.6582 | 0.6780 |
| | CMHH-T | 0.7833 | 0.7951 | 0.8016 | 0.5856 | 0.6165 | 0.6254 |
| | DADH-T | 0.8142 | 0.8232 | 0.8320 | 0.5931 | 0.6243 | 0.6350 |
| | DCHMT-T | 0.7983 | 0.8048 | 0.8031 | 0.6068 | 0.6286 | 0.6387 |
| | **DNpH** | **0.8147** | **0.8292** | **0.8361** | **0.6480** | **0.6786** | **0.6978** |

TABLE IV: Comparison with the baseline based on transformer encoder in terms of NDCG@1000 results w.r.t. 16bits, 32bits, 64bits on MIRFLICKR-25K and IAPR TC-12.

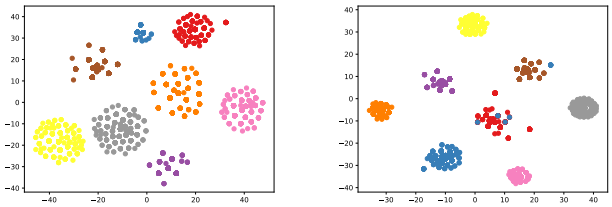
| Task | Method | MIRFLICKR-25K | | | IAPR TC-12 | | |
|---|---|---|---|---|---|---|---|
| | | 16bits | 32bits | 64bits | 16bits | 32bits | 64bits |
| I2T | DCMH-T | 0.4511 | 0.4782 | 0.5054 | 0.3520 | 0.3884 | 0.4310 |
| | CMHH-T | 0.4280 | 0.4324 | 0.4585 | 0.3151 | 0.3471 | 0.3589 |
| | DADH-T | 0.4853 | 0.5123 | 0.5255 | 0.3715 | 0.4140 | 0.4320 |
| | DCHMT-T | 0.4310 | 0.3876 | 0.4137 | 0.3231 | 0.3403 | 0.3390 |
| | **DNpH** | **0.5022** | **0.5266** | **0.5430** | **0.4032** | **0.4371** | **0.4566** |
| I2T | DCMH-T | 0.4156 | 0.4323 | 0.4574 | 0.3859 | 0.4216 | 0.4728 |
| | CMHH-T | 0.4067 | 0.4168 | 0.4428 | 0.3239 | 0.3527 | 0.3981 |
| | DADH-T | 0.4209 | 0.4492 | 0.4508 | 0.3572 | 0.3943 | 0.4053 |
| | DCHMT-T | 0.3772 | 0.3869 | 0.3663 | 0.3539 | 0.3715 | 0.3836 |
| | **DNpH** | **0.4230** | **0.4517** | **0.4615** | **0.4384** | **0.4731** | **0.4982** |

backbone, our proposed deep hashing framework achieves satisfactory results by introducing the quadratic spherical mutual information loss to minimize the neighbor ambiguity of heterogeneous modalities. To demonstrate our method more comprehensively, PR curves and TopN-precision curves are drawn on the four public datasets with 32-bit binary codes,




(a) Training time (b) Encoding time

Fig. 10: Comparisons of training and encoding time for different cross-modal hashing on MIRFLICKR-25K dataset.



(a) NUS-WIDE (b) MS COCO

Fig. 11: t-SNE visualization of learned 32-bit binary codes by using DNpH framework on NUS-WIDE and MS COCO.

as shown in Fig 3, Fig 4, Fig 5 and Fig 6. What's more, Precision results within Hamming radius 2 are represented in Fig 7. Compared to other comparison methods, our proposed DNpH framework obtains promising retrieval performance on PR curves, precision@topk curves, and P@H$\leq$2.

### *D. Training and Encoding Time*

To evaluate the time complexity of the model, we compare the training time and encoding time of our proposed DNpH framework with other typical methods with 64-bit hashing codes on the MIRFLICKR-25K dataset, as shown in Fig 10. Compared with the CNN-based models, we design the image

TABLE V: mAP results of DNpH and its variants w.r.t. 16bits, 32bits, 64bits on MIRFLICKR-25K, NUS-WIDE, MS COCO and IAPR TC-12.

| Task | Method | MIRFLICKR-25K | | | NUS-WIDE | | | MS COCO | | | IAPR TC-12 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 16bits | 32bits | 64bits | 16bits | 32bits | 64bits | 16bits | 32bits | 64bits | 16bits | 32bits | 64bits |
| I2T | DNpH-R | 0.7954 | 0.8103 | 0.8222 | 0.6582 | 0.6703 | 0.6775 | 0.6135 | 0.6439 | 0.6640 | 0.5671 | 0.5962 | 0.6121 |
| | DNpH-S | 0.7698 | 0.7797 | 0.7822 | 0.6017 | 0.6161 | 0.6127 | 0.5392 | 0.5647 | 0.5717 | 0.4808 | 0.4843 | 0.4780 |
| | DNpH-C | 0.8310 | 0.8478 | 0.8537 | 0.6596 | 0.6851 | 0.6974 | 0.5958 | 0.6608 | **0.7094** | 0.6234 | 0.6333 | 0.6797 |
| | DNpH-P | 0.8111 | 0.8204 | 0.8252 | 0.6741 | 0.6888 | 0.6928 | 0.5490 | 0.5946 | 0.6430 | 0.5710 | 0.6015 | 0.6108 |
| | **DNpH** | **0.8423** | **0.8552** | **0.8588** | **0.6921** | **0.7022** | **0.7071** | **0.6727** | **0.6903** | 0.6860 | **0.6359** | **0.6633** | **0.6797** |
| T2I | DNpH-R | 0.7687 | 0.7832 | 0.7941 | 0.6791 | 0.6938 | 0.6995 | 0.5994 | 0.6173 | 0.6317 | 0.5289 | 0.5608 | 0.5776 |
| | DNpH-S | 0.7787 | 0.7851 | 0.7875 | 0.6134 | 0.6156 | 0.6245 | 0.5308 | 0.5458 | 0.5581 | 0.4537 | 0.4510 | 0.4477 |
| | DNpH-C | 0.7995 | 0.8109 | 0.8222 | 0.6732 | 0.6929 | 0.7068 | 0.6048 | 0.6550 | **0.6973** | 0.6258 | 0.6518 | 0.6610 |
| | DNpH-P | 0.7833 | 0.7951 | 0.8016 | 0.6853 | 0.6967 | 0.7004 | 0.5533 | 0.5817 | 0.6481 | 0.5856 | 0.6165 | 0.6254 |
| | **DNpH** | **0.8147** | **0.8292** | **0.8361** | **0.6992** | **0.7137** | **0.7139** | **0.6562** | **0.6860** | 0.6928 | **0.6480** | **0.6786** | **0.6978** |

and text extractors based on the transformer encoder to extract semantic feature descriptors from raw samples, capturing the informative semantic representations. Hence, as illustrated in Fig 10-(a), it can be shown that our deep hashing would take a longer training time. However, the training and optimization process of the model is offline, which has no negative impact on retrieval performance. Hashing-based retrieval systems include two main components, i.e. the encoding time and the time to compute the Hamming distance. And, the encoding times of the different methods are calculated as illustrated in Fig 10-(b). It is known from the experiment that the encoding times of the different hashing methods are all at the millisecond level, so the retrieval efficiency is also acceptable.

### *E. Ablation Study*

To verify the validity of each component in our DNpH framework, we implement four variant models. (1) DNpH-R: Adopting the Resnet18 network as an image feature encoder. (2) DNpH-S: Removing the square clamp strategy. (3 DNpH-C: Replacing the cosine distance with Gaussian kernel function. (4) DNpH-P: The pair-wise loss function that was proposed in [56] is employed to replace the QSMI loss.

Table V describes the mAP results with different code lengths of DNpH and its variants on MIRFLICKR-25K, NUS-WIDE, MS COCO, and IAPR TC-12 datasets. From the experimental results, we can see that the image feature encoder based on transformer architecture could capture the global information by capturing the long-range visual dependencies, which can help improve the retrieval performance. The effectiveness of the square clamp strategy is further demonstrated by comparing the DNpH-S method with our approach. The mAP results of the DNpH-C method show that the QSMI loss based on cosine distance is more practical in the training process, which could avoid the choice of Gaussian kernel widths. Compared with to variant framework DNpH-P, based on quadratic mutual information, our proposed QSMI loss is introduced into deep cross-modal hashing to minimize neighbor ambiguity well and learn a highly separable discrete space.

### *F. Convergence Analysis*

By setting different training epochs, we evaluate the training convergence efficiency of our DNpH method. The links between the loss change and training epochs are drawn on MIRFLICKR-25K, NUS-WIDE, MS COCO, and IAPR TC-12 datasets with 64-bit hashing codes, as shown in Fig 8. To optimize the objective function effectively, a square clamping strategy is introduced into the process of DNpH training, avoiding converging on the local optimum. It is known from the convergence experiments that our DNpH method converges to a stable level at around 20 epochs, which obtains a favorable convergence speed.

### *G. Visualization*

*1) Grad-CAM Visualization:* To illustrate the superiority of the image feature encoder intuitively, we utilize the Grad-CAM technique in [63] to visualize the features of raw images. Specifically, we selected 10 images from each dataset and conducted several visual experiments, which are exhibited in Fig 9. The experiments indicate that our image feature encoder with transformer encoder could capture the global information by modeling long-range dependencies, which extract the key descriptive content of the raw samples. This is mainly due to the ability of multi-head self-attention for gaining semantic information.

*2) t-SNE Visualization:* To validate the discriminative capability of learned hash codes, we project the obtained 32-dimensional hash codes into the two-dimensional space via the t-SNE technique [64]. Specifically, since t-SNE can only process single-label samples, we select several single-label samples from eight different categories in the NUS-WIDE and the MS COCO dataset, and the visualization results are shown in Fig 11. It can be noticed that the hash codes learned by the DNpH method on the NUS-WIDE dataset have a distinctive representation capability. In addition, for the MS COCO dataset, satisfactory performance is also achieved. In summary, our proposed DNpH framework could separate neighbors and non-neighbors well and maintain the original neighborhood structure, so that similar samples are pulled close to each other, meanwhile, dissimilar data points are kept apart.

## V. CONCLUSION AND FUTURE WORK

In this paper, by introducing the quadratic mutual information theory, we propose a novel deep cross-modal hashing framework, called Deep Neighborhood-preserving Hashing

(DNpH), to learn a highly separable Hamming embedding, bridging the neighborhood ambiguity. For feature learning, to learn informative semantic representation from heterogeneous modalities, two feature extractors based on the transformer encoder are employed as image and text extractors respectively. For hashing learning, the Quadratic Spherical Mutual Information (QSMI) is first introduced into deep cross-modal retrieval to produce similarity-preserving binary codes. Meanwhile, a square clamping method is developed to enhance the stability of model optimization. Extensive experiments on four benchmark datasets show that the DNpH method achieves encouraging retrieval performance.

Our proposed DNpH framework only considers the scenario of supervised cross-modal retrieval. However, due to the expensive cost of human annotation, it is unrealistic to make huge amounts of multimedia data well-labeled. In future work, we would aim to extend quadratic mutual information theory to unsupervised cross-modal retrieval, making the learned embedding discriminative and stable.

## REFERENCES


[1] W. Chen, Y. Liu, W. Wang, E. M. Bakker, T. Georgiou, P. W. Fieguth, L. Liu, and M. S. Lew, "Deep learning for instance retrieval: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7270–7292, 2023. 1

[2] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 769–790, 2018. 1

[3] Y. Liu, J. Wu, L. Qu, T. Gan, J. Yin, and L. Nie, "Self-Supervised Correlation Learning for Cross-Modal Retrieval," *IEEE Transactions on Multimedia*, vol. 25, pp. 2851–2863, 2023. 1

[4] L. Qu, M. Liu, J. Wu, Z. Gao, and L. Nie, "Dynamic Modality Interaction Modeling for Image-Text Retrieval," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1104–1113. 1

[5] X. Luo, H. Wang, D. Wu, C. Chen, M. Deng, J. Huang, and X.-S. Hua, "A survey on deep hashing methods," *ACM Transactions on Knowledge Discovery from Data*, vol. 17, no. 1, pp. 1–50, 2023. 1

[6] L. Wang, Y. Pan, C. Liu, H. Lai, J. Yin, and Y. Liu, "Deep hashing with minimal-distance-separated hash centers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 455–23 464. 1

[7] M. Cao, S. Li, J. Li, L. Nie, and M. Zhang, "Image-text retrieval: A survey on recent research and development," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2022, pp. 5410–5417. 1

[8] Q. Qin, K. Xie, W. Zhang, C. Wang, and L. Huang, "Deep neighborhood structure-preserving hashing for large-scale image retrieval," *IEEE Transactions on Multimedia*, early access, 2023, doi: 10.1109/TMM.2023.3289765. 1

[9] L. Zhu, X. Lu, Z. Cheng, J. Li, and H. Zhang, "Deep collaborative multi-view hashing for large-scale image Search," *IEEE Transactions on Image Processing*, vol. 29, pp. 4643–4655, 2020. 1

[10] H. T. Shen, L. Liu, Y. Yang, X. Xu, Z. Huang, F. Shen, and R. Hong, "Exploiting subspace relation in semantic labels for cross-modal hashing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 10, pp. 3351–3365, 2021. 1

[11] L. Zhu, C. Zheng, W. Guan, J. Li, Y. Yang, and H. T. Shen, "Multi-modal hashing for efficient multimedia retrieval: A survey," *IEEE Transactions on Knowledge and Data Engineering*, early access, 2023, doi: 10.1109/TKDE.2023.3282921. 1

[12] W. Tan, L. Zhu, J. Li, H. Zhang, and J. Han, "Teacher-student learning: Efficient hierarchical message aggregation hashing for cross-modal retrieval," *IEEE Transactions on Multimedia*, early access, 2022, doi: 10.1109/TMM.2022.3177901. 1

[13] P.-F. Zhang, Y. Li, Z. Huang, and X.-S. Xu, "Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 24, pp. 466–479, 2022. 1

[14] H. Hu, L. Xie, R. Hong, and Q. Tian, "Creating Something From Nothing: Unsupervised Knowledge Distillation for Cross-Modal Hashing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3120–3129. 1

[15] J. Yu, H. Zhou, Y. Zhan, and D. Tao, "Deep Graph-neighbor Coherence Preserving Network for Unsupervised Cross-modal Hashing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 4626–4634. 1

[16] L. Li, B. Zheng, and W. Sun, "Adaptive Structural Similarity Preserving for Unsupervised Cross Modal Hashing," in *Proceedings of the ACM International Conference on Multimedia*, 2022, pp. 3712–3721. 1

[17] D. Wang, C. Zhang, Q. Wang, Y. Tian, L. He, and L. Zhao, "Hierarchical Semantic Structure Preserving Hashing for Cross-Modal Retrieval," *IEEE Transactions on Multimedia*, vol. 25, pp. 1217–1229, 2023. 1

[18] X. Liu, Z. Hu, H. Ling, and Y.-M. Cheung, "MTFH: A Matrix Tri-Factorization Hashing Framework for Efficient Cross-Modal Retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 964–981, 2021. 1

[19] L. Wang, M. Zareapoor, J. Yang, and Z. Zheng, "Asymmetric Correlation Quantization Hashing for Cross-Modal Retrieval," *IEEE Transactions on Multimedia*, vol. 24, pp. 3665–3678, 2022. 1

[20] J. Qin, L. Fei, Z. Zhang, J. Wen, Y. Xu, and D. Zhang, "Joint Specifics and Consistency Hash Learning for Large-Scale Cross-Modal Retrieval," *IEEE Transactions on Image Processing*, vol. 31, pp. 5343–5358, 2022. 1

[21] R.-C. Tu, X.-L. Mao, R.-X. Tu, B. Bian, C. Cai, H. Wang, W. Wei, and H. Huang, "Deep Cross-Modal Proxy Hashing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 7, pp. 6798–6810, 2023. 1

[22] M. Su, G. Gu, X. Ren, H. Fu, and Y. Zhao, "Semi-Supervised Knowledge Distillation for Cross-Modal Hashing," *IEEE Transactions on Multimedia*, vol. 25, pp. 662–675, 2023. 1

[23] H. Cui, L. Zhu, J. Li, Y. Yang, and L. Nie, "Scalable deep hashing for large-scale social image retrieval," *IEEE Transactions on image processing*, vol. 29, pp. 1271–1284, 2019. 1

[24] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022. 1

[25] Q. Qin, L. Huang, Z. Wei, K. Xie, and W. Zhang, "Unsupervised Deep Multi-Similarity Hashing With Semantic Structure for Image Retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 7, pp. 2852–2865, 2021. 1

[26] Y. Shi, Y. Zhao, X. Liu, F. Zheng, W. Ou, X. You, and Q. Peng, "Deep Adaptively-Enhanced Hashing With Discriminative Similarity Guidance for Unsupervised Cross-Modal Retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 7255–7268, 2022. 2

[27] Y. Chen, S. Wang, J. Lu, Z. Chen, Z. Zhang, and Z. Huang, "Local Graph Convolutional Networks for Cross-Modal Hashing," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 1921–1928. 2

[28] Y. Hu, M. Liu, X. Su, Z. Gao, and L. Nie, "Video Moment Localization via Deep Cross-Modal Hashing," *IEEE Transactions on Image Processing*, vol. 30, pp. 4667–4677, 2021. 2

[29] Q. Qin, L. Huang, K. Xie, Z. Wei, C. Wang, and W. Zhang, "Deep Adaptive Quadruplet Hashing with Probability Sampling for Large-Scale Image Retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, early access, 2023, doi: 10.1109/TCSVT.2023.3281868. 2

[30] Q.-Y. Jiang and W.-J. Li, "Deep Cross-Modal Hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3270–3278. 2, 3, 5, 8

[31] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4242–4251. 2, 3, 5, 8

[32] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-Based Deep Hashing Network for Cross-Modal Retrieval," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3893–3903, 2018. 2, 3, 5

[33] W. Gu, X. Gu, J. Gu, B. Li, Z. Xiong, and W. Wang, "Adversary Guided Asymmetric Hashing for Cross-Modal Retrieval," in *Proceedings of the ACM SIGMM International Conference on Multimedia Information Retrieval*, 2019, pp. 159–167. 2, 5, 8

[34] Y. Huo, Q. Qin, J. Dai, L. Wang, W. Zhang, L. Huang, and C. Wang, "Deep Semantic-aware Proxy Hashing for Multi-label Cross-modal Retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, early access, 2023, doi: 10.1109/TCSVT.2023.3285266. 2, 5, 8

[35] F. Çakir, K. He, S. A. Bargal, and S. Sclaroff, "MIHash: Online Hashing with Mutual Information," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 437–445. 2, 4, 6

[36] F. Cakir, K. He, S. A. Bargal, and S. Sclaroff, "Hashing with mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 10, pp. 2424–2437, 2019. 2, 4, 5, 6

[37] N. Passalis and A. Tefas, "Deep supervised hashing using quadratic spherical mutual information for efficient image retrieval," *Signal Processing: Image Communication*, vol. 93, p. 116146, 2021. 2, 4, 5, 6, 7

[38] K. Torkkola, "Feature Extraction by Non-Parametric Mutual Information Maximization," *Journal of Machine Learning Research*, vol. 3, pp. 1415–1438, 2003. 2, 4, 6

[39] D. Zhang and W.-J. Li, "Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014, pp. 2177–2183. 3

[40] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning Discriminative Binary Codes for Large-scale Cross-modal Retrieval," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2494–2507, 2017. 3

[41] J. Tang, K. Wang, and L. Shao, "Supervised Matrix Factorization Hashing for Cross-Modal Retrieval," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3157–3166, 2016. 3

[42] Z. Zhang, H. Luo, L. Zhu, G. Lu, and H. T. Shen, "Modality-Invariant Asymmetric Networks for Cross-Modal Hashing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 5, pp. 5091–5104, 2023. 3, 8

[43] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, "Linear cross-modal hashing for efficient multimedia search," in *Proceedings of the ACM International Conference on Multimedia*, 2013, pp. 143–152. 4

[44] J. Zhang, Y. Peng, and M. Yuan, "Unsupervised Generative Adversarial Cross-Modal Hashing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 539–546. 4

[45] S. Su, Z. Zhong, and C. Zhang, "Deep Joint-Semantics Reconstructing Hashing for Large-Scale Unsupervised Cross-Modal Retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3027–3035. 4

[46] P. Hu, H. Zhu, J. Lin, D. Peng, Y.-P. Zhao, and X. Peng, "Unsupervised Contrastive Cross-Modal Hashing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3877–3889, 2023. 4

[47] T. Hoang, T.-T. Do, T. V. Nguyen, and N.-M. Cheung, "Multimodal Mutual Information Maximization: A Novel Approach for Unsupervised Deep Cross-Modal Hashing," *IEEE Transactions on Neural Networks and Learning Systems*, early access, 2022, doi: 10.1109/TNNLS.2021.3135420. 4, 6

[48] D. Bouzas, N. Arvanitopoulos, and A. Tefas, "Graph Embedded Nonparametric Mutual Information for Supervised Dimensionality Reduction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 951–963, 2015. 4

[49] N. Passalis and A. Tefas, "Learning Deep Representations with Probabilistic Knowledge Transfer," in *Proceedings of the European Conference on Computer Vision*, vol. 11215, 2018, pp. 268–284. 4

[50] M. Tzelepi and A. Tefas, "Improving the performance of lightweight CNNs for binary classification using quadratic mutual information regularization," *Pattern Recognit*, vol. 106, p. 107407, 2020. 4

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations*, 2015. 5

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 5

[53] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proceedings of the International Conference on Learning Representations*, 2021, pp. 1–22. 5

[54] J. E. Zini and M. Awad, "On the explainability of natural language processing deep models," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–31, 2022. 5

[55] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. The Association for Computer Linguistics, 2016. 5

[56] Y. Cao, B. Liu, M. Long, and J. Wang, "Cross-Modal Hamming Hashing," in *Proceedings of the European Conference on Computer Vision*, vol. 11205, 2018, pp. 207–223. 8, 12

[57] C. Bai, C. Zeng, Q. Ma, J. Zhang, and S. Chen, "Deep Adversarial Discrete Hashing for Cross-Modal Retrieval," in *Proceedings of the ACM SIGMM International Conference on Multimedia Information Retrieval*, 2020, pp. 525–531. 8

[58] X. Wang, X. Zou, E. M. Bakker, and S. Wu, "Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval," *Neurocomputing*, vol. 400, pp. 255–271, 2020. 8

[59] X. Zou, S. Wu, E. M. Bakker, and X. Wang, "Multi-label enhancement based self-supervised deep cross-modal hashing," *Neurocomputing*, vol. 467, pp. 138–162, 2022. 8

[60] J. Tu, X. Liu, Z. Lin, R. Hong, and M. Wang, "Differentiable Cross-modal Hashing via Multimodal Transformers," in *Proceedings of the ACM International Conference on Multimedia*, 2022, pp. 453–461. 8

[61] C. Sun, H. Latapie, G. Liu, and Y. Yan, "Deep Normalized Cross-Modal Hashing with Bi-Direction Relation Reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 4937–4945. 8

[62] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2019, pp. 8024–8035. 8

[63] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020. 12

[64] V. D. M. Laurens and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 2605, pp. 2579–2605, 2008. 12