Deep Semantic-Aware Proxy Hashing for Multi-Label Cross-Modal Retrieval

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

***Abstract*— Deep hashing has attracted broad interest in cross-modal retrieval because of its low cost and efficient retrieval benefits. To capture the semantic information of raw samples and alleviate the semantic gap, supervised cross-modal hashing methods that utilize label information which could map raw samples from different modalities into a unified common space, are proposed. Although making great progress, existing deep cross-modal hashing methods are suffering from some problems, such as: 1) considering multi-label cross-modal retrieval, proxy-based methods ignore the data-to-data relations and fail to explore the combination of the different categories profoundly, which could lead to some samples without common categories being embedded in the vicinity; 2) for feature representation, image feature extractors containing multiple convolutional layers cannot fully obtain global information of images, which results in the generation of sub-optimal binary hash codes. In this paper, by extending the proxy-based mechanism to multi-label cross-modal retrieval, we propose a novel Deep Semantic-aware Proxy Hashing (DSPH) framework, which could embed multi-modal multi-label data into a uniform discrete space and capture fine-grained semantic relations between raw samples. Specifically, by learning multi-modal multi-label proxy terms and multi-modal irrelevant terms jointly, the semantic-aware proxy loss is designed to capture multi-label correlations and preserve the correct fine-grained similarity ranking among samples, alleviating inter-modal semantic gaps. In addition, for feature representation, two transformer encoders are proposed as backbone networks for images and text, respectively, in which the image transformer encoder is introduced to obtain global information of the input image by modeling long-range visual dependencies. We have conducted extensive experiments on three baseline multi-label datasets, and the experimental results show that our DSPH framework achieves better performance than state-of-the-art cross-modal hashing methods. The code for the implementation of our DSPH framework is available at https://github.com/QinLab-WFU/DSPH.**

***Index Terms*— Deep hashing, semantic-aware proxy, cross-modal retrieval, semantic relationships, transformer encoder.**

I. Introduction

With the growing complexity of Internet technology, large numbers of users upload and share social data such as images, texts, audio, and videos to the Internet daily, which leads to an exponential growth in multi-media data with preserving the underlying heterogeneity and high-level semantic correlation [1]. How do you quickly retrieve the data you need from massive amounts of multi-media data? How to save multi-media data efficiently? These are the two critical challenges that we are facing.

Traditional retrieval is limited since it is too costly to retrieve when the data is enormous and high-dimensional data. In order to solve the problems caused by traditional retrieval, hashing methods for approximate nearest neighbor search techniques are proposed, which have the advantages of low cost and efficient retrieval and cause considerable interest in industry and academia [2]. Hashing methods could map high-dimensional data into binary hash codes, in which irrelevant data could generate dissimilar binary codes [3]. For cross-modal retrieval applications, researchers proposed the cross-modal hashing methods [4], [5], [6], [7], which map high-dimensional data from different modalities into a common space to minimize semantic gaps and utilize compact binary codes to represent high-dimensional data [8]. The Hamming distance between the hash codes is then computed by an exclusive OR (XOR) operation to achieve similarity retrieval [9]. Hence, cross-modal hashing methods could decrease the storage space and achieve a fast search speed.

Cross-modal hashing methods are usually divided into shallow cross-modal hashing and deep cross-modal hashing [10]. By leveraging hand-crafted features, shallow hashing methods perform learning to the hash, which can result in poorly captured semantic information [11], [12]. In recent years, with the rapid development of deep neural networks, deep cross-modal hashing methods have become a popular subject of extensive studies [13], [14], [15]. Compared to shallow hashing methods, deep hashing methods employ the powerful

Manuscript received 25 March 2023; revised 28 May 2023; accepted 6 June 2023. Date of publication 12 June 2023; date of current version 8 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62006174, in part by the Shandong Provincial Natural Science Foundation under Grant ZR2022QF046, and in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant KJQN202200551. This article was recommended by Associate Editor H. Zhang. *(Corresponding authors: Qibing Qin; Chengduan Wang.)*

Yadong Huo is with the School of Computer Science, Qufu Normal University, Rizhao 276826, China (e-mail: hyd199810@163.com).

Qibing Qin is with the School of Computer Engineering, Weifang University, Weifang 261061, China, and also with the Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266100, China (e-mail: qinbing@wfu.edu.cn).

Jiangyan Dai, Lei Wang, and Chengduan Wang are with the School of Computer Engineering, Weifang University, Weifang 261061, China (e-mail: daijy@wfu.edu.cn; 20111281@wfu.edu.cn; 20111182@wfu.edu.cn).

Wenfeng Zhang is with the College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China (e-mail: itzhangwf@cqnu.edu.cn).

Lei Huang is with the College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China (e-mail: huangl@ouc.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3285266.

Digital Object Identifier 10.1109/TCSVT.2023.3285266

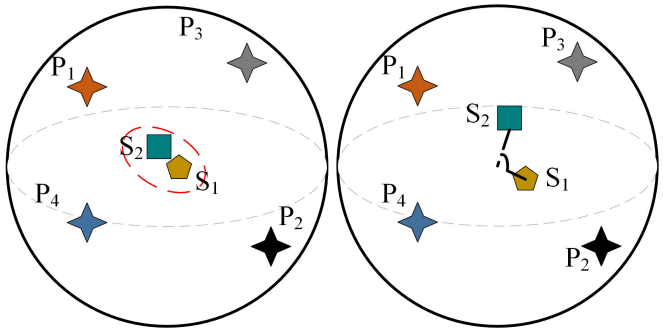


Fig. 1. The quadrangles represent the proxies for each category, and the other different shapes represent irrelevant data. P_1 denotes the sky category, P_2 represents the grass category, P_3 means the sea category, and P_4 represents the mountain category. Sample S_1 is associated with P_1 and P_2. Sample S_2 is associated with P_3 and P_4. In contrast to proxy-based approaches, the semantic-aware proxy method not only captures abundant multi-label correlations but pulls away irrelevant data pairs, including intra-modal irrelevant pairs and inter-modal irrelevant pairs, maintaining the correct fine-grained ranking of returned samples.

non-linear representation capabilities of deep neural networks to extract feature representations of high-dimensional data [16]. Generally, deep cross-modal hashing is categorized into unsupervised hashing and supervised hashing. By learning hash functions from the distributions and topological information of the raw samples, unsupervised deep cross-modal hashing is limited by a lack of strong prior knowledge, which could affect the quality of generated hash codes [17], [18], [19]. In contrast, supervised deep cross-modal hashing leverages extensive label information to learn hash mapping, resulting in better retrieval performance than unsupervised frameworks [20], [21], [22]. As a consequence, supervised deep cross-modal hashing has been the studied hotspot of multi-media retrieval. Since data attributes are not unique in the real world [23], multi-label cross-modal retrieval systems based on deep hashing are facing many challenges. By exploring the association of samples with multiple labels, some cross-modal hashing frameworks are proposed to obtain more powerful and robust binary codes for multi-label retrieval applications [24], [25].

Despite considerable progress in supervised deep cross-modal hashing, they still encounter some limitations. 1) Proxy-based approaches usually design corresponding proxies for each category and learn the hash function by exploring the simple relations between samples and proxies during model training. However, for multi-label cross-modal retrieval, proxy-based approaches force multi-label samples to be embedded in the middle of these proxies to ensure that the retrieved samples are relevant to the query samples. Specifically, as shown in Fig. 1, suppose that four proxies are embedded in the common space well, where P_1 denotes the sky category, P_2 represents the grass category, P_3 means the sea category and P_4 represents the mountain category. Sample S_1 is associated with P_1 and P_2, and sample S_2 is associated with P_3 and P_4. Proxy-based approaches embed S_1 into the middle of P_1 and P_2 and project S_2 into the middle of P_3 and P_4. However, S_1 and S_2 are embedded in nearby locations because of ignoring data-to-data relations, and proxy-based methods alienate the distance between S_1 and P_3, P_4, which could not guarantee that S_1 is far from the middle of P_3 and P_4. The problem will become increasingly serious in multi-label cross-modal datasets. In addition, with the number of category combinations growing constantly and exponentially, proxy-based methods cannot capture inclusive(i.e., relevant categories) and exclusive (i.e., irrelevant categories) relations well [26]; 2) Existing cross-modal hashing methods generally use encoder with multiple convolutional layers as the image feature extractor [27], [28]. Due to the limited fixed size of convolution kernels, such models normally fail to capture the global information of the image well [29]. For the foregoing reasons, the performance improvement of deep cross-modal hashing is far from satisfactory and adequate.

To remedy these problems, a novel Deep Semantic-aware Proxy Hashing (DSPH) framework is proposed for multi-label cross-modal hashing retrieval in this paper by extending the proxy-based mechanisms to multi-label cross-modal retrieval, which could alleviate inter-modal semantic gaps effectively. The overall framework of our proposed framework is illustrated in Fig. 2, which consists of Image Feature Learning Network (ImgNet), Text Feature Learning Network (TxtNet), and Hash Learning. To the best of our knowledge, our proposed DSPH is the first attempt to introduce semantic-aware proxy mechanisms into multi-label cross-modal retrieval to preserve the fine-grained similarity ranking of sample pairs. Specifically, for the feature extraction module, the image transformer encoder is designed to capture global information of the image by modeling long-range visual dependencies, meanwhile, the text feature extractor employs a transformer encoder to obtain the semantic features of the text. For hash learning, by learning multi-modal multi-label proxy terms and multi-modal irrelevant terms jointly, the semantic-aware proxy loss is designed to capture multi-label correlations and preserve the correct fine-grained similarity ranking among samples, alleviating inter-modal semantic gaps.

In summary, the contributions of the paper are as follows.

(1) Firstly, to the best of our knowledge, our proposed DSPH framework is the first attempt to introduce the semantic-aware proxy mechanism to achieve multi-label cross-modal hash, which could embed multi-label data into discrete spaces via learnable proxies with semantic relations preserved sufficiently and alleviate inter-modal semantic gaps effectively.

(2) Secondly, the multi-modal multi-label proxy loss is proposed to ensure that the relevant multi-label data can be embedded in suitably nearby locations and pull away irrelevant data-proxy pairs. To obtain the fine-grained semantic relations between data, the multi-modal irrelevant loss is developed to alienate irrelevant data pairs, which include intra-modal irrelevant pairs and inter-modal irrelevant pairs.

(3) Thirdly, two transformer encoders are proposed as backbone networks for image and text, respectively, in which the image transformer encoder is introduced to obtain global information of the input image by modeling long-range visual dependencies, and the text transformer

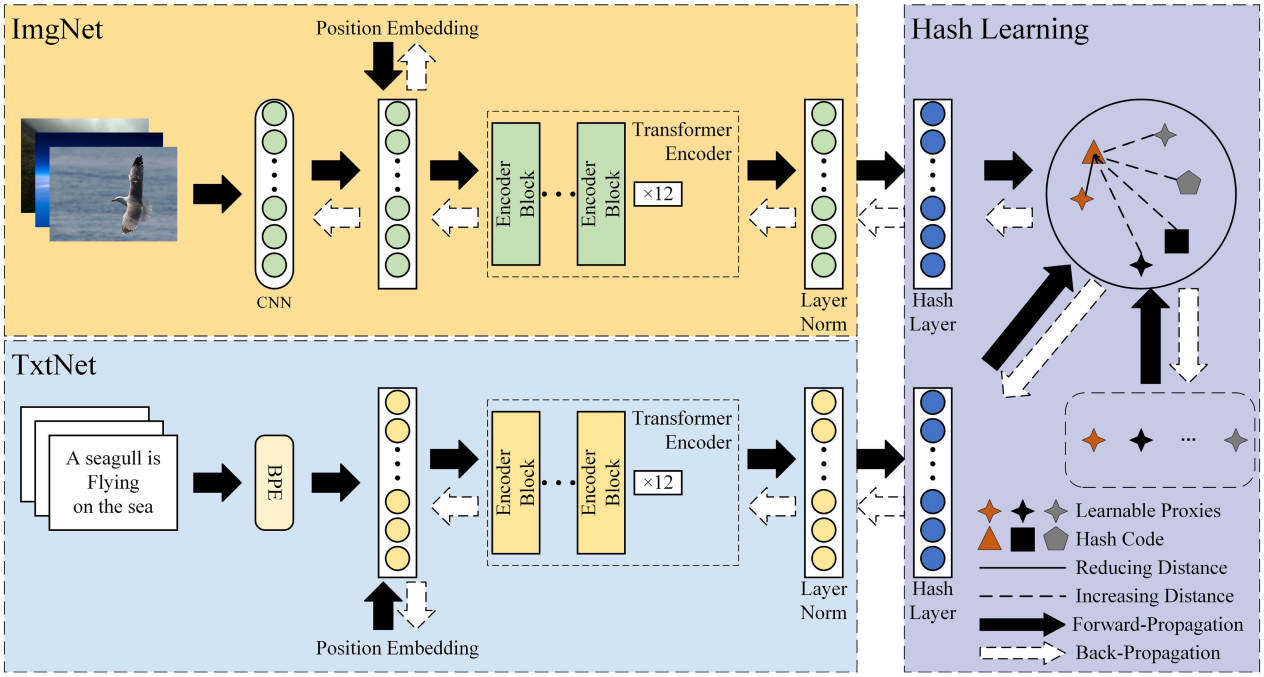


Fig. 2. An overview of DSPH, including three parts: (1) ImgNet: Image transformer encoder is designed to capture the global information from raw images by modeling long-range visual dependencies. (2) TxtNet: By employing BPE encoding to represent the text, the text transformer encoder is used as the text feature extractor. (3) Hash Learning: Multi-modal multi-label proxy loss is designed to ensure that the relevant multi-label data can be embedded in suitably nearby locations and pull away irrelevant data-proxy pairs. Besides, to get the fine-grained semantic relations among data, the multi-modal irrelevant loss is proposed to alienate semantically irrelevant data.

encoder is proposed to learn critical semantic information by multi-headed self-attention.

(4) Finally, extensive experiments on three popular datasets demonstrate that our proposed DSPH framework achieves excellent performance compared to several other state-of-the-art cross-modal hashing works.

The remainder of this paper is organized as follows. Representative supervised cross-modal hashing methods, unsupervised cross-modal hashing methods, and transformer networks on hashing retrieval are introduced in Section II. In Section III, we provide detailed overviews of DSPH methods. In Section IV, we demonstrate the effectiveness of our method through comprehensive experiments. Relevant conclusions and future works are summarized in Section V.

II. Related Works

Hash methods are receiving extensive interest in retrieval applications [30]. Cross-modal hashing methods learn the hash functions to project high-dimensional data from different modalities into a unified common space and reduce gaps among modalities so that Hamming distances of semantically relevant data are smaller than those of semantically irrelevant data. In this section, we briefly introduce supervised cross-modal hashing methods, unsupervised cross-modal hashing methods, and works about transformer on hashing retrieval.

A. Supervised Cross-Modal Hashing

Although unsupervised cross-modal hashing does not require label information, it lacks strong semantic correlations, and retrieval performance is generally inferior to supervised cross-modal hashing methods. Supervised cross-modal hashing methods mainly leverage label information to learn hash functions and can yield satisfying performance. In particular, SePH method converts similarity matrices into probability distributions and learns binary hash codes for multi-modal data by minimizing KL scatter to maintain the semantic similarity of multi-modal data better and solves the problem of out-of-sample expansion [31]. For example, DCMH approach integrates feature learning and hash code learning into an end-to-end framework, which uses a constructed similarity matrix and a negative log-likelihood loss function to maintain the similarity of the original data [21]. The SSAH method combines self-supervised and adversarial learning to bridge the heterogeneity gap better. It uses the semantic information obtained from labels to supervise the captured semantics of the data in the way of GAN [32], [33]. Through introducing graph convolutional networks to cross-modal hashing and exploring semantic structural information, the graph convolutional hashing (GCH) method is proposed, which guides feature learning networks through semantic encoders to implement the learning of hash functions [34]. To capture the multi-scale semantic information of data, DMFH approach applies multi-scale fusion models to multi-modal data to generate hash codes with discriminative power [35]. TDH proposed intra-modal and inter-modal triplet loss and graph regular loss to capture similarity relations between different modal data by using constructed cross-modal triples to obtain richer semantic information [36]. FCMH approach captures

fine-grained semantic information via exploring similarity relationships between global and local information of data and employs a discrete hash learning scheme to learn binary hash codes discretely [37]. In addition, to address the difficulty of optimizing binary hash codes during training, DCHMT proposed differentiable hashing methods, which use a selection mechanism to generate hash codes that can be optimized using gradient descent [29]. To enable cross-modal hash models to handle data from different domains in real-time, VLKD method is proposed [38]. The method leverages the hierarchical recurrent network to learn global and fine-grained information about images and text. It employs knowledge distillation to explore semantic information of data and uses inter-modal and intra-modal attention transfer to make models capable of life-long correlation learning.

In recent studies, researchers have investigated the multi-label deep cross-modal hashing approach. To reduce the modal gap, by employing semantic information from multi-label annotations, MS^2GAH framework is proposed to explore the neighborhood relations between nodes through graph attention networks (GAT) [39]. By designing semantic fusion methods to maintain multi-label semantic information, RMSH is proposed to make the distance between dissimilar hash codes greater than a specific value, and margin-adaptive triplet loss is introduced to capture the fine-grained relationships of raw samples [40]. For the MMACH framework [24], a multi-label modality enhanced attention module is proposed to alleviate the sparse feature representation of multi-labels. In addition, multi-label cross-modal triplet loss is proposed to maintain the semantic similarity of the original samples.

B. Unsupervised Cross-Modal Hashing

Unsupervised cross-modal hashing methods learn hash functions directly by computing similarity relations from data features without using label information. IMH studies inter-media and intra-media data consistency to project data into a Hamming space and introduces linear regression to learn hash functions [41]. LSSH proposed an iterative strategy to narrow the heterogeneity gap between modal data. It leverages sparse coding and matrix decomposition to obtain latent semantic information about data, which improves the quality of hash codes [42]. Meanwhile, unsupervised cross-modal hashing methods based on deep learning are attracting the great interest of researchers. For example, DGPCN improves the quality of hash codes by considering graph neighbor consistency, intra-modal and inter-modal data consistency, and designing a half-real-valued half-binary optimization strategy [19]. Several deep unsupervised hashing methods use a constructed similarity matrix to supervise the learning of hash codes. For example, DJRSH exploited multi-modal data features to build a unified semantic affinity matrix and proposed a semantic reconstruction framework that allows hash codes to maintain a specific similarity value rather than just similar order, which is more suitable for batch training methods [17]. UKD method introduces knowledge distillation, which uses the similarity matrix learned from the teacher network to supervise the learning of hash codes in the student network [18]. Subsequently, JDSH proposed a weighted cross-modal similarity matrix construction method and studied similarity relationships between cross-modal samples by using four similarities: Self-similarity, Extreme dissimilarity, Relative similarity, and Relative dissimilarity, which allowed to generate discriminative hash codes [43]. To enhance feature representation in unsupervised hashing and minimize the semantic gap between cross-modal data, UCCH introduces contrastive learning into unsupervised cross-modal hashing. In order to make the similarities of semantically relevant pairs greater than semantically irrelevant pairs, Cross-modal Ranking Learning loss is proposed, and a hashing memory bank is designed to solve the difficult problem of binary optimization [44].

C. Transformer on Hashing Retrieval

Recently, the transformer mechanism has gained extensive popularity in computer vision [45]. For the BTH framework, by capturing correlations between video frames through bidirectional transformers, three self-supervised learning tasks are proposed to learn similarity relationships between data [46]. However, the above method only focuses on unimodal data. Besides, CLIP4Hashing implements video-text hash learning in cross-modal hashing by introducing the CLIP model. A dynamic weighting strategy and a parameter-free min-max hash layer approach are proposed to improve the retrieval performance [47].

Unlike previous approaches, the image transformer encoder is proposed to obtain global information about the image by modeling long-range visual dependencies. The text transformer encoder is designed to represent the semantic features of the text. The semantic-aware proxy loss is developed to capture the multi-label correlations and the fine-grained semantic relations between sample pairs.

III. Methodology

In this section, we describe the DSPH in detail. The definition is given in Subsection III-A. ImgNet and TxtNet are described in more detail in Subsection III-B. The details of multi-modal multi-label proxy loss and multi-modal irrelevant loss are introduced in subsection III-C. The summary and analysis of the DSPH algorithm are shown in Subsection III-D. Subsection III-E shows how to generate binary hash codes from trained DSPH models.

A. Notation

In this work, we mainly consider the image-text cross-modal retrieval task. Suppose that given an image-text training set $D = \{d_i\}_{i=1}^{M}$ containing M samples. x_i and y_i in $d_i = \{x_i, y_i, l_i\}$ denote the i-th example of the image and text modalities, respectively. $l_i = [l_{i1}, l_{i2}, \ldots, l_{ic}]$ is a multi-label annotation d_i and C means the number of categories. The goal of deep cross-modal hashing is to learn hash functions H^x and H^y that project image data and text data into binary hash codes $B^x \in \{-1, 1\}^K$ and $B^y \in \{-1, 1\}^K$, respectively, K indicate the hash code length. The specific symbols and their corresponding meanings are shown in Table I.

TABLE I
THE MAIN SYMBOLS AND CORRESPONDING MEANINGS IN DSPH

Notation	Definition
D	Training Dataset
D'	Irrelevant Pairs subset
M	Number of Samples
M'	Number of Irrelevant Pairs
C	Number of Categories
G	Feature Extractor
F	Features of Samples
K	Hash Code Length
H	Hash Functions
B	Hash Code
x_i	i-th Image Sample
y_i	i-th Text Sample
l_i	Label Corresponding to i-th Sample
P	Learnable Proxies
e_{pos}	Positional Encoding
I	Indicator Function

B. Semantic Feature Extraction

This subsection describes the detailed structure of ImgNet and TxtNet as follows.

1) ImgNet: Existing cross-modal hashing methods generally use an encoder with multiple convolutional layers as the image feature extractor. However, due to the limited fixed size of convolution kernels, such models normally fail to capture the global information of the image nicely, which results in the generation of sub-optimal binary hash codes. Hence, in order to obtain the global information of raw images, we introduce the image transformer encoder to obtain the semantic feature descriptors of images by modeling long-range visual dependencies.

For long-range visual dependencies, the semantic relations between two pixels of an image that are far apart are expressed. By obtaining long-range visual dependencies, the model can capture rich semantic relations.

The image transformer encoder has the same structure as the VIT encoder [48], consisting of 12 layers of encoder blocks in stacks. The structure of each encoder block is the same, which contains Layernorm (LN), multi-headed self-attention (MSA), and MLP blocks. The number of the MSA is 12. The MSA is explained in detail as follows. Assume that the vector a_i, by the linear transformation of the three matrices W^q, W^k, W^v gets the vectors q_i, k_i, v_i as input to the attention blocks.

$$q_i = a_i W^q, k_i = a_i W^k, v_i = a_i W^v \tag{1}$$

$$attention_i = softmax(\frac{q_i * k_i^T}{\sqrt{e_k}}) * v_i \tag{2}$$

where e_k is the dimension of the input. Connecting the vectors of attention of the 12 heads to get the vector b_i according to Eq.(3)

$$b_i = concat(attention_1, attention_2, \ldots, attention_{12}) w^o \tag{3}$$

Specifically, suppose we have image x_i. First, we feed x_i into a convolutional layer. The output features of the convolutional layer are treated by the flatten operation to get the middle feature vector x_r. Then, trainable positional encoding e^x_{pos} is added to obtain an intermediate feature vector x_{pos}, with the following Eq.(4).

$$x_{pos} = x_r + e^x_{pos} \tag{4}$$

x_{pos} is fed into the image transformer encoder, and the output of the MLP layer in the last encoder block is the feature representation f_i^x of the image. $F^x = G^x(X, \theta_x)$, where G^x denotes the image feature extractor and θ_x represents the image network parameters.

2) TxtNet: For the text representation, to generate more discriminative hash codes, we tokenize the text by using the BPE method and then extract the text features via the text transformer encoder [49]. BPE encoding is a popular subword encoding method, and BPE encoding relies on a corpus. The first step is to set the number of tokens, and then start an iterative process of combining the most frequent pairs of characters until the pre-set number of tokens is achieved. The text transformer encoder consists of 12 encoder blocks each of which has 8 MSA.

Specifically, the text y_i is encoded y_r using BPE encoding, and the trainable positional encoding e^y_{pos} is added to obtain the vector y_{pos}.

$$y_{pos} = y_r + e^y_{pos} \tag{5}$$

where y_{pos} is fed into to the text transformer encoder, and the text feature is finally represented f_i^y. $F^y = G^y(Y, \theta_y)$, in which G^y denotes the text feature extractor and θ_y represents the parameters of text network.

C. Semantic-Aware Proxy Loss

Most cross-modal hashing methods consider only single-label retrieval and ignore multi-label scenarios, which leads to insufficient capture of the semantic relations between data. Although proxy-based methods can achieve satisfactory performance in single-label cross-modal retrieval, considering multi-label cross-modal retrieval, proxy-based methods have been shown to produce poor performance with limited hash bits. Specifically, with failing to express multi-label correlations profoundly and ignoring data-to-data relations, the proxy-based approach is unsuitable for multi-label cross-modal retrieval [50]. Therefore, by learning multi-modal multi-label proxy loss and multi-modal irrelevant loss, we introduce the semantics-aware proxy mechanism into cross-modal hashing to capture the multi-label correlations and explore the fine-grained semantic relations between data, alleviating inter-modal semantic gaps.

1) Multi-Modal Multi-Label Proxy Loss: To ensure that the relevant multi-label data are embedded in the vicinity and pull away irrelevant data-proxy pairs, we propose multi-modal multi-label proxy loss to optimize the model parameters. For $P = \{p_1, p_2, \ldots\ldots p_C\}$, P is the learnable proxies for each category, where p_i is a K-bits vector. When the samples and the proxy are relevant, we could calculate the cosine distance between the binary-like hash codes and relevant proxies by using Eq.(6).

$$\cos_+(h, p) = -\cos\langle h, p\rangle = -\frac{|h \cdot p|}{|h| \cdot |p|} \tag{6}$$

In addition, we not only consider the relations between the data and the relevant proxies but pull the data away from the irrelevant proxies. The distance between binary-like hash codes and irrelevant proxies is pushed away by calculating Eq.(7).

$$\begin{aligned}\cos_{-}(h, p) &= (\cos(h, p) - \sigma)_{+} \\ &= \max(\frac{|h \cdot p|}{|h| \cdot |p|} - \sigma, 0)\end{aligned} \tag{7}$$

where $\sigma = \sigma(C, K)$ is a margin term, which can be set by reference to [51].

Hence, the multi-label proxy loss of image $\mathcal{L}^{x}_{proxy}$ could be calculated as shown in Eq.(8).

$$\begin{aligned}\mathcal{L}^{x}_{proxy} = &\frac{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=1)\cos_{+}(h^{x}_{i}, p_j)}{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=1)} \\ &+ \frac{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=0)\cos_{-}(h^{x}_{i}, p_j)}{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=0)}\end{aligned} \tag{8}$$

where I is an indicator function. The denominator term is designed to reduce the gradient bias due to many negative pairs. The multi-label proxy loss for text $\mathcal{L}^{y}_{proxy}$ is calculated to alleviate the semantic gap and enable better embedding of multi-modal samples into discrete space.

$$\begin{aligned}\mathcal{L}^{y}_{proxy} = &\frac{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=1)\cos_{+}(h^{y}_{i}, p_j)}{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=1)} \\ &+ \frac{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=0)\cos_{-}(h^{y}_{i}, p_j)}{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=0)}\end{aligned} \tag{9}$$

The total multi-modal multi-label proxy loss $\mathcal{L}_{proxy}$ is calculated, as shown in Eq.(10).

$$\mathcal{L}_{proxy} = \mathcal{L}^{x}_{proxy} + \mathcal{L}^{y}_{proxy} \tag{10}$$

2) Multi-Modal Irrelevant Loss: To capture fine-grained semantic relations among data and pull away irrelevant pairs, multi-modal irrelevant loss is proposed to train the network. We refer to the approach for the definition of irrelevant pairs [26]. If $x_i(y_i)$ and $x_j(y_j)$ correspond to labels $|l_i \cdot l_j| = 0$, and $|l_i| > 1$, $|l_j| > 1$, then $x_i(y_i)$ and $x_j(y_j)$ are called an irrelevant pair. Suppose multi-modal irrelevant pairs subset $D' = \{d'_i\}_{i=1}^{M'} \subseteq D$. In multi-label datasets, the number of irrelevant pairs M' is greatly smaller than the total number of samples M. The irrelevant loss of the image $\mathcal{L}_{x\_neg\_pair}$ could be calculated by Eq.(11).

$$\mathcal{L}_{x\_neg\_pair} = \frac{\sum_{i=1}^{M'}\sum_{j=1}^{M'} I(|l_i \cdot l_j| = 0)\cos_{-}(h^{x}_{i}, h^{x}_{j})}{\sum_{i=1}^{M'}\sum_{j=1}^{M'} I(|l_i \cdot l_j| = 0)} \tag{11}$$

Similarly, irrelevant loss of text $\mathcal{L}_{y\_neg\_pair}$ is calculated by Eq.(12).

$$\mathcal{L}_{y\_neg\_pair} = \frac{\sum_{i=1}^{M'}\sum_{j=1}^{M'} I(|l_i \cdot l_j| = 0)\cos_{-}(h^{y}_{i}, h^{y}_{j})}{\sum_{i=1}^{M'}\sum_{j=1}^{M'} I(|l_i \cdot l_j| = 0)} \tag{12}$$

In summary, the multi-modal irrelevant loss considers both intra-modal irrelevant loss and inter-modal irrelevant loss.

Algorithm 1 Learning Algorithm for DSPH

Input:
Training dataset D; Binary codes length K; Hyper-parameters α.

Output:
Binary code B, Parameters Θ_x and Θ_y.

1: Initialize network parameters Θ_x and Θ_y, maximum iteration number $epoch$, mini-batch size 128.
2: **while** $iter < epoch$ **do**
3: Compute feature vector F^x and F^y by forward propagation.
4: Compute multi-modal multi-label proxy loss $\mathcal{L}_{proxy}$ by Eq.(10).
5: Compute multi-modal irrelevant loss $\mathcal{L}_{neg\_pair}$ by Eq.(14).
6: Update proxies $P = \{p_1, p_2, p_3, \ldots, p_c\}$ by back-propagations.
7: Update the network parameters θ_x and θ_y by back-propagation.
8: **end while**
9: **return** The trained DSPH model.

TABLE II
TIME COMPLEXITY

Method	Time Complexity.
DCPH [50]	$O(MC)$
SSAH [32]	$O(M^2)$
GCH [34]	$O(M^2)$
DSPH(Ours)	$O(MC + \lambda M^2)$

Therefore, we could compute the inter-modal irrelevant loss $\mathcal{L}_{xy\_neg\_pair}$ according to Eq.(13).

$$\mathcal{L}_{xy\_neg\_pair} = \frac{\sum_{i=1}^{M'}\sum_{j=1}^{M'} I(|l_i \cdot l_j| = 0)\cos_{-}(h^{x}_{i}, h^{y}_{j})}{\sum_{i=1}^{M'}\sum_{j=1}^{M'} I(|l_i \cdot l_j| = 0)} \tag{13}$$

Overall multi-modal irrelevant loss $\mathcal{L}_{neg\_pair}$ could be calculated as shown in Eq.(14).

$$\mathcal{L}_{neg\_pair} = \mathcal{L}_{x\_neg\_pair} + \mathcal{L}_{y\_neg\_pair} + \mathcal{L}_{xy\_neg\_pair} \tag{14}$$

As a result, by learning multi-modal multi-label proxy loss and multi-modal irrelevant loss, the semantic-aware proxy loss of DSPH method $\mathcal{L}_{sp}$ could be calculated according to Eq.(15).

$$\mathcal{L}_{sp} = \mathcal{L}_{proxy} + \alpha\mathcal{L}_{neg\_pair} \tag{15}$$

where α is a hyper-parameter to balance multi-modal multi-label proxy Loss and multi-modal irrelevant loss.

D. Overview of Algorithm

1) Training Algorithm: The algorithm for the DSPH model is summarised in Algorithm 1. The DSPH model is optimized by standard backpropagation algorithms and mini-batch gradient descent methods.

Algorithm 2 Learning Hash Codes for DSPH

Input:
Query samples q_i, Parameters for DSPH.
Output:
Binary hash code for q_i.
1: Calculate binary-like hash codes by feeding the query data q_i into the trained DSPH model.
2: Calculating hash codes by using the $sign$ function.

2) Time Complexity Analysis: As shown in Table II. M denotes the number of training samples, and C is the number of categories. λ represents the ratio of irrelevant pairs of data over the number of training samples. Although DCPH has lower time complexity, the previous study has demonstrated that the proxy-based approach is only suitable for single-label retrieval and not applicable for multi-label retrieval [26]. The first term of the time complexity of DSPH denotes the time complexity of the data-to-proxy pairs, and the second term represents the time complexity between irrelevant pairs. The time complexity of the other two methods shows a squared relation with the number of training samples. By analyzing and comparing, the DSPH method has a lower training cost because λ is far less than 1.

E. Out-of-Sample Extension

The algorithm for generating the hash code is shown in Algorithm 2. For a query sample, the binary-like hash code is generated by the DSPH model that has been trained, and then the $sign$ function is employed to generate the binary hash code. The retrieval results are returned through ranking using the Hamming distance by calculating the Hamming distance between the hash code of query samples and database samples.

$$sign(x) = \begin{cases} +1, x > 0 \\ -1, x < 0 \end{cases} \tag{16}$$

Specifically, for given the query data $x_i(y_i)$, the compact hash code could be generated by Eq.(17).

$$\begin{aligned} b_i^x &= sgn(H^x(x_i)) \\ b_i^y &= sgn(H^y(y_i)) \end{aligned} \tag{17}$$

IV. Experiments

To demonstrate the validity of our proposed DPSH framework, we conducted comprehensive experiments on three publicly available cross-modal multi-label datasets MIRFLICKR-25K,[1] NUS-WIDE,[2] and MS COCO.[3] Besides, we introduce the datasets for the experiments and explain the details of the implementation of DSPH and evaluate metrics.

[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/

A. Datasets

MIRFLICKR-25K is a small cross-modal multi-label dataset collected on the Flickr website containing 24,581 image-text pairs. It has a total of 24 different categories. Each image-text pair has a corresponding multi-label annotation and belongs to at least one of the 24 categories.

NUS-WIDE dataset is the frequently used larger cross-modal multi-label dataset which contains 269,648 image-text pairs, each labeled by at least one of the different 81 categories. Follow the previous method to divide this dataset [21]. We removed the categories that contained very little data and selected 21 general categories. This sub-set contains 195,834 image-text pairs, each of which belongs to one of the categories at least.

MS COCO dataset is a popular dataset used in object detection and is also a multi-label dataset with 80 different categories. It is divided into training and validation sets containing 82,785 images in the training set and 40,504 images in the validation set, each with 5 captions. In our experiments, the training and validation sets are combined, each instance contains both image and text modalities of data, and each sample belongs to at least one category.

We adopt the same sampling strategy for the three datasets in our experiments and randomly selected 5000 image-text pairs from this dataset as the query set and the remainder as the database set. In training the model, 10,000 image-text pairs were randomly chosen as the training set. For the different datasets, we handle both images and text in the same way, where the images are resized to 224×224, and the text is denoted by using BPE encoding.

B. Baseline Methods

In our experiments, we selected 8 state-of-the-art deep cross-modal hashing methods for comparison, which contain Deep Cross-Modal Hashing (DCMH) [21], Self-Supervised Adversarial Hashing (SSAH) [32], Cross-Modal Hamming Hash (CMHH) [52], Adversary Guided Asymmetric Hashing (AGAH) [53], Deep Adversarial Discrete Hashing (DADH) [54], Self-Constraining Attention Hashing Network (SCAHN) [55], Multi-label Enhancement Self-supervised Deep Cross-modal Hashing (MESDCH) [25], Differentiable Cross-modal Hashing via Multimodal Transformers (DCHMT) [29]. We implemented these methods with code the author published online, and the parameter settings of the above-mentioned cross-modal hashing refer to the original papers.

C. Experimental Details

We implemented the proposed DSPH in the framework of PyTorch [56], with GPU leveraging NVIDIA RTX 3090. Our model employs an adaptive moment estimation (Adam) optimizer to update the network parameters until it converges [57]. In experiments, the hyper-parameter is 0.8, the initial learning rate is 0.001, and the batch size is 128.

TABLE III
COMPARISON WITH BASELINES IN TERMS OF mAP W.R.T. 16BITS, 32BITS, 64BITS ON MIRFLICKR-25K, NUS-WIDE, AND MS COCO

Task	Method	MIRFLICKR-25K			NUS-WIDE			MS COCO		
		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
	SSAH	0.8079	0.8129	0.8220	0.6032	0.6058	0.6095	0.5411	0.4855	0.5395
	CMHH	0.6932	0.6979	0.6984	0.5439	0.5546	0.5520	0.5145	0.4509	0.5209
	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
	SCAHN	0.7955	0.8248	0.8297	0.6463	0.6616	0.6645	0.6376	0.6475	0.6519
	MESDCH	0.7258	0.7438	0.7421	0.5638	0.5801	0.5875	0.5759	0.5670	0.5636
	DCHMT	**0.8201**	0.8253	0.8222	0.6596	0.6706	0.6863	0.6309	0.6216	0.6553
	DSPH	0.8129	**0.8482**	**0.8541**	**0.6830**	**0.6979**	**0.7162**	**0.7044**	**0.7510**	**0.7694**
T2I	DCMH	0.7857	0.7998	0.8029	0.5747	0.5810	0.5853	0.5271	0.5424	0.5450
	SSAH	**0.8089**	0.8127	0.8017	0.6011	0.6058	0.6167	0.4901	0.4798	0.5053
	CMHH	0.7181	0.7104	0.7294	0.4956	0.4831	0.4820	0.4910	0.4930	0.4889
	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
	SCAHN	0.7826	0.8066	0.8064	0.6587	0.6626	0.6648	0.6377	0.6512	0.6493
	MESDCH	0.7385	0.7387	0.7437	0.5562	0.5720	0.5803	0.5044	0.5083	0.5085
	DCHMT	0.7983	0.8048	0.8031	0.6761	0.6837	0.6943	0.6241	0.6212	0.6486
	DSPH	0.8000	**0.8238**	**0.8294**	**0.6997**	**0.7153**	**0.7304**	**0.7040**	**0.7556**	**0.7713**

D. Evaluation Metrics

Hamming ranking and hash lookup are the two most popular retrieval protocols for cross-modal retrieval. In our experiments, 5 evaluation metrics to evaluate the performance of retrieval are used, including mean Average Precision (mAP), mean Average Precision within Hamming radius 2(mAP@H≤2), Precision-Recall (PR) curves,TopN-precision curves, and NDCG@rank. mAP is the most commonly used evaluation indicator in cross-modal hashing. The mAP is calculated following Eq.(18), where N_q is the number of query samples and $AP(i)$ is the average precision of the i-th sample. mAP@H≤2 calculates the mAP within the Hamming radius 2. PR curves represent the relationship between precision and recall, a measure of overall retrieval performance. TopN-precision curves show precision for the top N samples returned. The NDCG@rank metric is a general metric to evaluate multi-label retrieval. It gets a higher score if samples with high similarity are ranked highly. In our experiments, the rank is 1000.

$$mAP = \frac{1}{N_q}\sum_{i=1}^{N_q} AP(i) \tag{18}$$

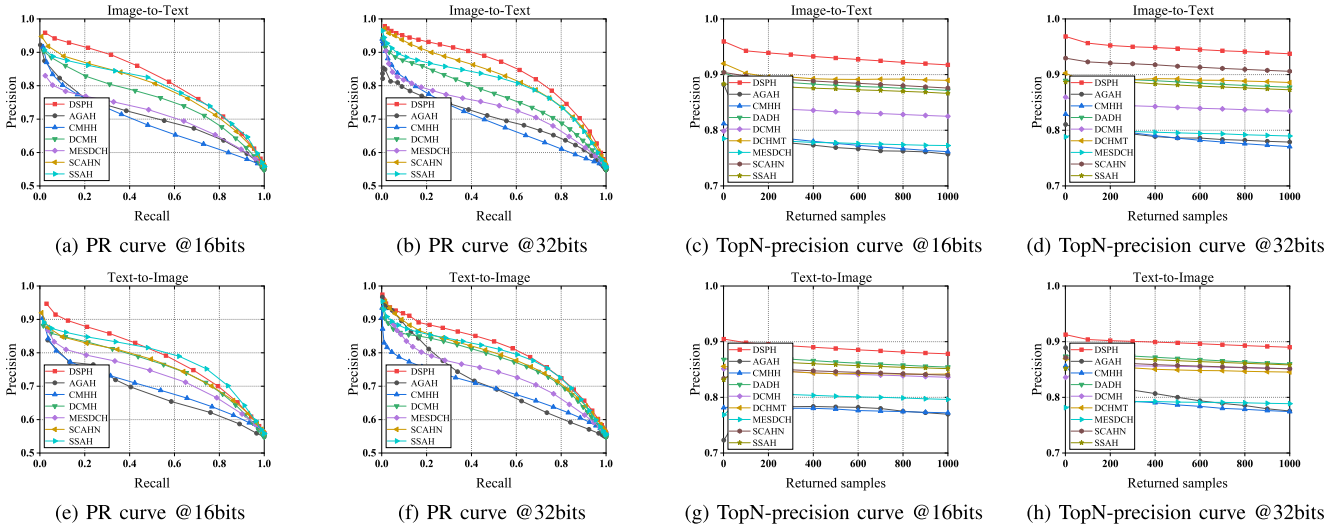
E. Comparison With the Baselines

For two retrieval tasks on three public datasets, we validate the performance of the DSPH by comparing it with the state-of-the-art deep cross-modal hashing methods.

The results of mAP are shown in Table III, where "I2T" denotes image retrieval text and "T2I" denotes text retrieval image. DSPH outperforms other baseline methods in most cases, achieving satisfactory performance. Because the multi-modal multi-label proxy loss is designed to ensure that the relevant multi-label data can be embedded in suitably nearby locations and pull away irrelevant data-proxy pairs. The fine-grained semantic relations between data are captured via multi-modal irrelevant loss. Compared to baseline methods, DSPH performs satisfactorily on the MIRFLICKR-25K dataset. Compared with the DCHMT method, our method improved on average by 2.69% on the image retrieval text task and 3.04% on the text retrieval image task for the NUS-WIDE dataset. Compared with SCAHN, the best-performing retrieval on the MS COCO dataset, DSPH improves by 6.63% to 12.2%. Another interesting observation was that many methods failed to perform as well as the NUS-WIDE for retrieval on the MS COCO. Since the number of categories in MS COCO is 80, much larger than the 21 categories in NUS-WIDE, the previous method was heavily influenced by the number of classes. However, DSPH performs better on the MS COCO than the NUS-WIDE, showing that the number of categories has a lower influence on our approach.

The results of mAP@H≤2 for different lengths of hash codes on the MIRFLICKR-25K, NUS-WIDE, and MS COCO datasets are shown in Fig. 6, where (a) (b) (c) show the results for mAP@H≤2 on the image retrieval text task. (d) (e) (f) shows the results for mAP@H≤2 on the text retrieval image task. In the task of image retrieval text, the comparison shows that our method outperforms all methods on 16-bit and 32-bit hash codes. DSPH performs excellently on the NUS-WIDE and MS COCO datasets regarding text retrieval images. We also have an exciting discovery that the results for mAP@H≤2 for most methods tend to decrease when the hash code length is 64, compared to the hash code length of 16. Our analysis is that as the hash code length increase, the discrete space becomes increasingly sparse, resulting in fewer data points falling within Hamming radius 2.

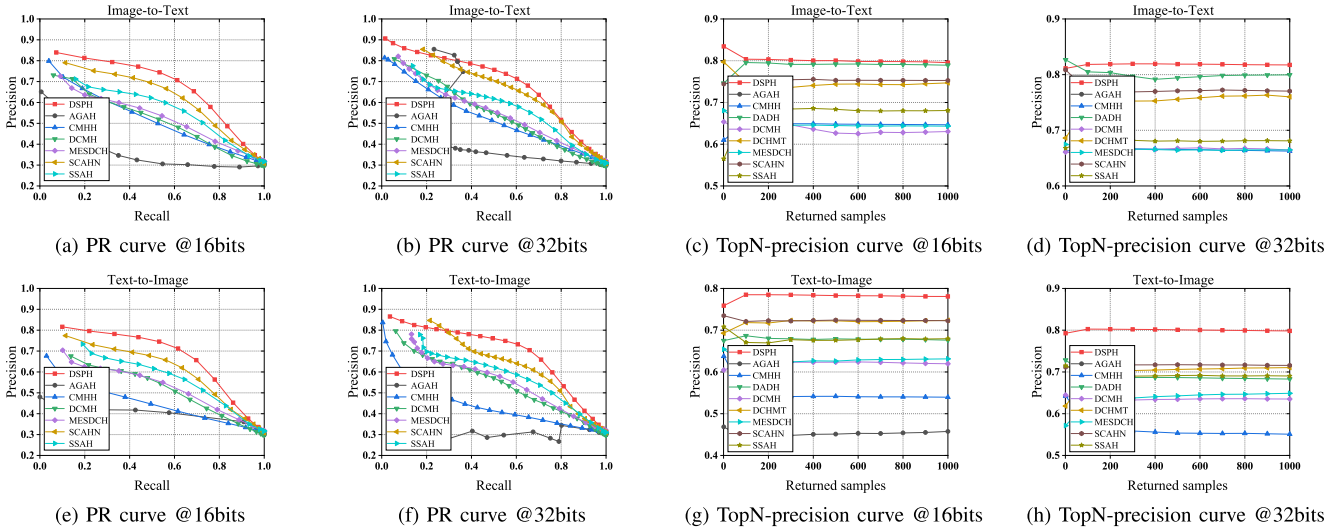
To further evaluate the performance of our method, Fig. 3, Fig. 4, and Fig. 5 show PR curves and TopN-precision curves on MIRFLICKR-25K, NUS-WIDE, and MS COCO datasets



(a) PR curve @16bits (b) PR curve @32bits (c) TopN-precision curve @16bits (d) TopN-precision curve @32bits

(e) PR curve @16bits (f) PR curve @32bits (g) TopN-precision curve @16bits (h) TopN-precision curve @32bits

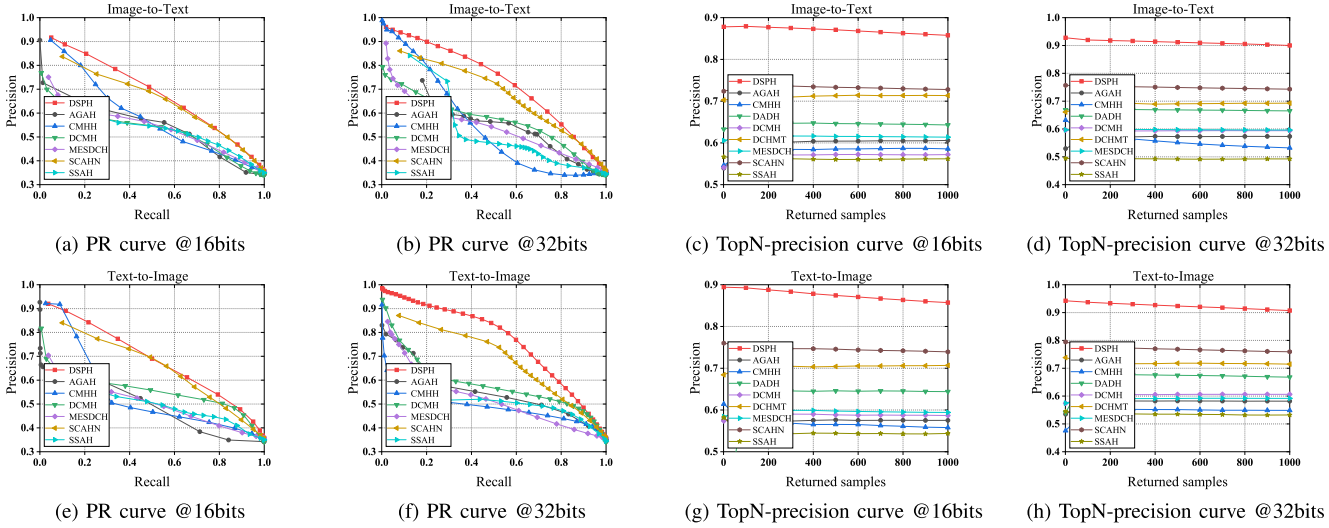
Fig. 3. Results of Precision-Recall curves, TopN-precision curves on MIRFLICKR-25K w.r.t.16bits and 32bits.



(a) PR curve @16bits (b) PR curve @32bits (c) TopN-precision curve @16bits (d) TopN-precision curve @32bits

(e) PR curve @16bits (f) PR curve @32bits (g) TopN-precision curve @16bits (h) TopN-precision curve @32bits

Fig. 4. Results of Precision-Recall curves, TopN-precision curves on NUS-WIDE w.r.t.16bits and 32bits.



(a) PR curve @16bits (b) PR curve @32bits (c) TopN-precision curve @16bits (d) TopN-precision curve @32bits

(e) PR curve @16bits (f) PR curve @32bits (g) TopN-precision curve @16bits (h) TopN-precision curve @32bits

Fig. 5. Results of Precision-Recall curves, TopN-precision curves on MS COCO w.r.t.16bits and 32bits.

with 16-bits and 32-bits, and the results demonstrate that the overall performance of our approach outperforms the other methods. The NDCG@1000 evaluation metric is employed to evaluate the performance of our method in a multi-label retrieval scenario. Results for NDCG@1000 are shown in Fig. 7. By comparing the other methods, DSPH outperforms

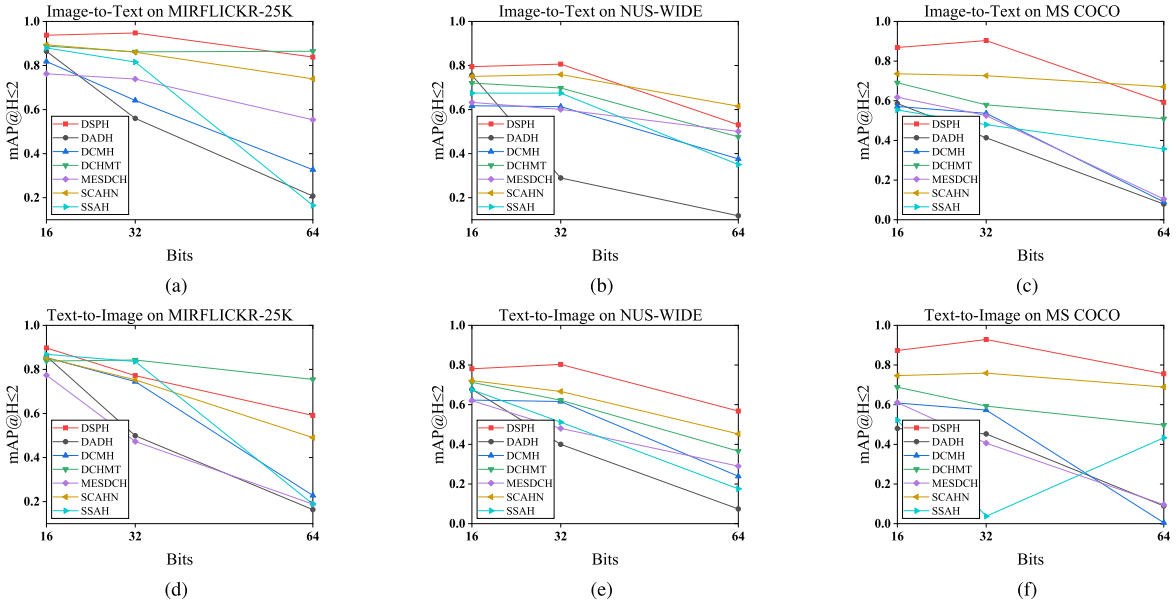


Fig. 6. The mAP@H≤2 w.r.t. different code lengths on MIRFLICKR-25K, NUS-WIDE and MS COCO datasets.

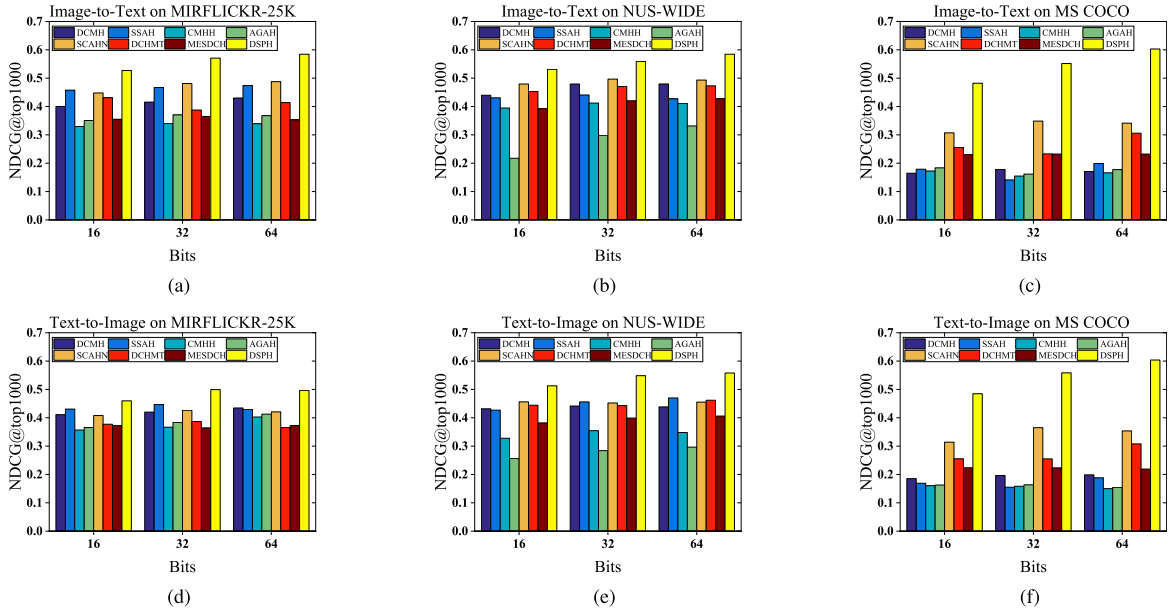


Fig. 7. The NDCG@1000 w.r.t. different code lengths on MIRFLICKR-25K, NUS-WIDE and MS COCO datasets.

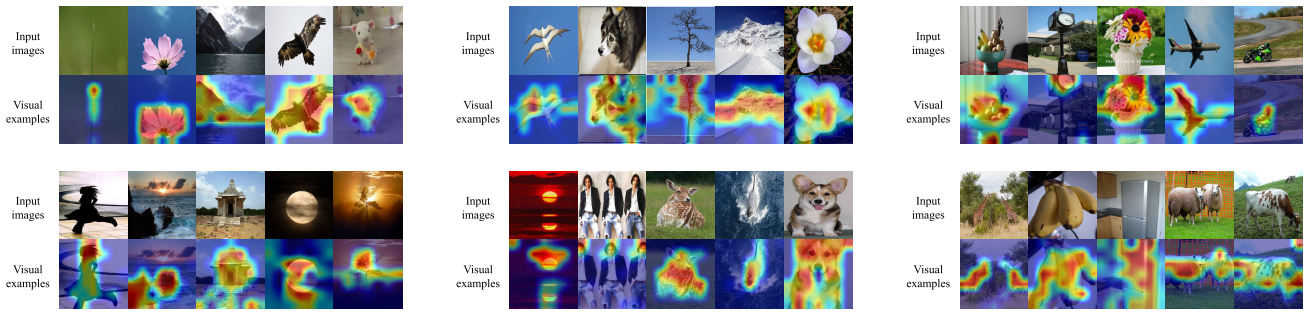
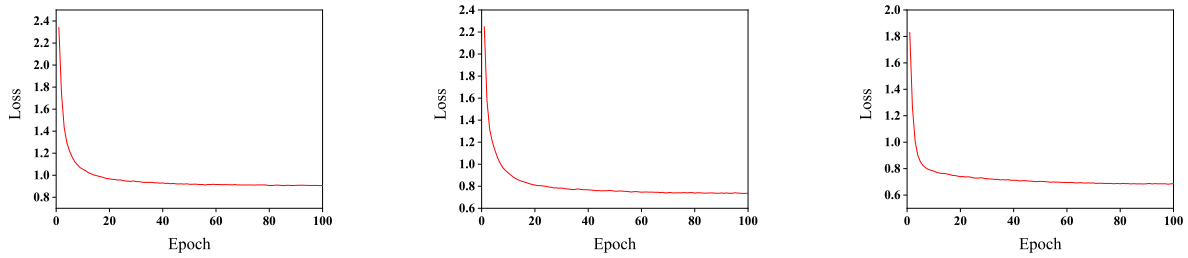


Fig. 8. Some visual examples of the image transformer encoder on MIRFLICKR-25K, NUS-WIDE, MS COCO datasets via Grad-CAM.

them all. A more significant boost is obtained on the MS COCO dataset. The semantic-aware proxy loss is proposed to embed multi-label data into discrete spaces well and obtain satisfactory retrieval results.

F. Parameter Sensitivity

By comparing mAP results of different hash coding lengths on the MIRFLICKR-25K dataset, we study the influence of the hyper-parameter α on performance. As shown in Fig. 10,

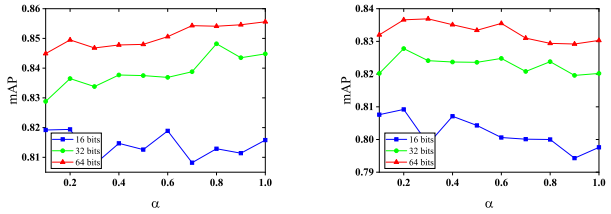


(a) MIRFLICKR-25K (b) NUS-WIDE (c) MS COCO

Fig. 9. Convergence analysis of DSPH with code length of 16.

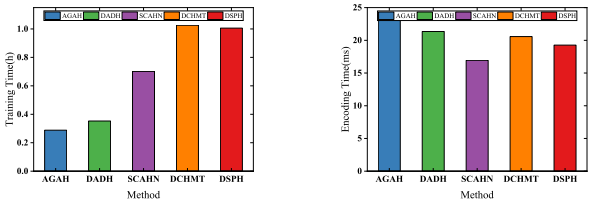
TABLE IV
mAP OF DSPH AND ITS VARIANTS ON MIRFLICKR-25K, NUS-WIDE AND MS COCO

Task	Method	MIRFLICKR-25K			NUS-WIDE			MS COCO		
		16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I2T	DSPH-P	0.8124	0.8317	0.8414	0.6724	0.6965	0.7167	0.6756	0.7116	0.7447
	DSPH-I	0.7851	0.8093	0.8125	0.6501	0.6620	0.6777	0.6290	0.5997	0.6414
	DSPH-R	0.7951	0.8128	0.8179	0.6496	0.6636	0.6755	0.6462	0.6839	0.7010
	DSPH-F	0.6953	0.7165	0.7295	0.5858	0.5966	0.6265	0.6408	0.6988	0.7267
	DSPH	**0.8129**	**0.8482**	**0.8541**	**0.6830**	**0.6979**	0.7162	**0.7044**	**0.7510**	**0.7694**



(a) I2T (b) T2I

Fig. 10. Parameter Analysis of DSPH on MIRFLICKR-25K.



(a) Training time (b) Encoding time

Fig. 11. Training time and Encoding time of DSPH on MIRFLICKR-25K.

by increasing the value of α from 0.1 to 1.0, with setting $\alpha = 0.6$ or $\alpha = 0.8$, the image retrieval text task has the best retrieval performance. By setting $\alpha = 0.2$, our proposed framework achieves the best results in the text retrieval image task. By comparing and analyzing, the hyper-parameter $\alpha = 0.8$ is set in our experiments.

G. Training and Encoding Time

We compare the training time and encoding time of several methods including AGAH, DADH, SCAHN, DCHMT and our DSPH at hash length 32 for the MIRFLICKR-25K dataset. The results of the experiment are shown in Fig. 11. The experimental results show that our proposed DSPH method requires more time to train the model. The reason for the long training time in our analysis is that we use the transformer encoder as a feature extractor for images and text respectively. Since the training process of the model is offline, the training time does not have an impact on the retrieval performance. For hash-based cross-modal retrieval methods, we are more focused on query response times. The encoding time and the time to calculate the Hamming distance are important components of the query response time. The encoding times are shown in Fig. 11, and satisfactory encoding times are obtained by our DSPH method. The encoding times for all methods are in the range of 17-23 milliseconds, with minimal impact on the efficiency of cross-modal retrieval.

H. Ablation Study

To verify the effectiveness of the components in DSPH, we implement four variations to calculate the mAP values on the image retrieval text task. Specifically: (1) DSPH-P: using only multi-modal multi-label proxy loss to train the model. (2) DSPH-I: parametric optimization by using multi-modal irrelevant loss. (3) DSPH-R: replace the image feature extractor with the ResNet18 network. (4) DSPH-F: BoW vector is used to denote text, and the text feature extractor is replaced with a three-layer fully connected layer.

The ablation experiment results are shown in Table IV. Comparison of the DSPH with the DSPH-P method shows degraded performance on the MIRFLICKR-25K and MS COCO datasets. However, in the NUS-WIDE dataset, mAP performance changes minimally with increasing numbers of hash code bits. Our analysis of the NUS-WIDE dataset has 195,834 pairs of samples, but the number of categories is only 21. As a result, it contains far fewer irrelevant pairs than the other two datasets, so there are no significant performance changes when removing multi-modal irrelevant loss. Comparison of the DSPH-I method with the DSPH method shows a decrease in mAP results, which shows that the multi-modal multi-label proxy loss is introduced to ensure that the relevant multi-label data can be embedded in suitably nearby locations and pull away irrelevant data-proxy pairs. On the three public datasets, the mAP results of DSPH-F showed a significant decrease compared to the DSPH method. This demonstrates the effectiveness of the DSPH method in employing the text

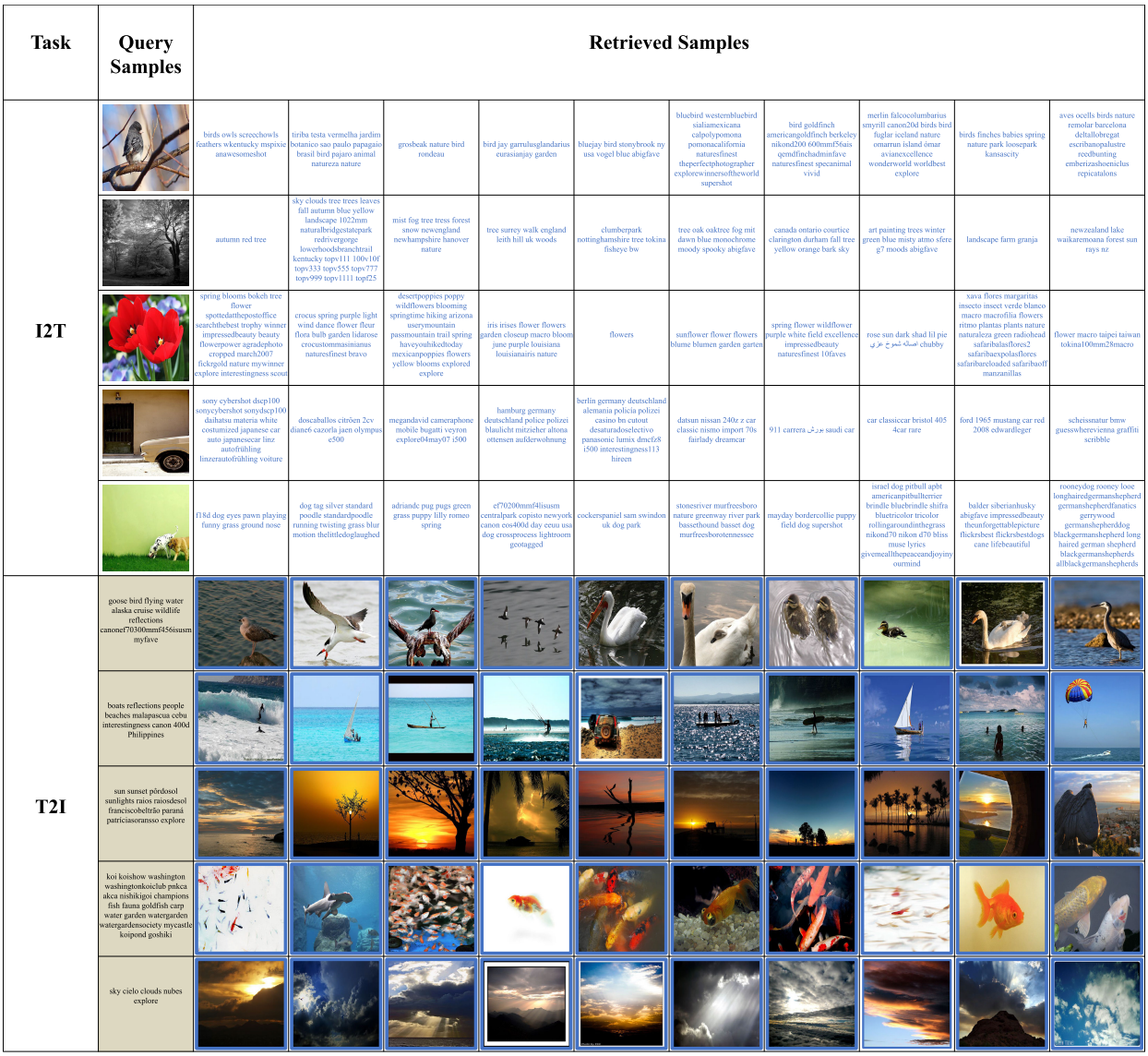


Fig. 12. Top-10 retrieval results on the MIRFLICKR-25K dataset with 64-bit hashing codes by exploiting our DSPH framework to encode raw samples. The blue markers mean that the samples returned are relevant to the query samples, and the red markers denote that the samples returned are irrelevant to the query samples.

transformer encoder proposed in the DSPH method to obtain the feature descriptors of the text. By comparing the DSPH and DSPH-*R* methods, it can be shown that the image transformer encoder proposed in the paper gets rich semantic features by modeling long visual dependencies. The effectiveness of each component of the DSPH method is demonstrated by comparing the four variations. The ImgNet and the TxtNet could learn the semantic information of the data well. By learning multi-modal multi-label proxy terms and multi-modal irrelevant terms jointly, the semantic-aware proxy loss is designed to capture multi-label correlations and preserve the correct fine-grained similarity ranking among samples.

I. Visualisation

To further study the ability of image feature extractors to capture global information. Fig. 8 shows some examples of feature visualization. Specifically, in the three datasets, we selected 10 images each and visualized feature maps via the Grad-CAM method. We visualize output features of the Layer-Norm layer of the last encoder block of the image feature extractor. We find that the image feature extractor could locate key descriptive objects in images well. For example, on the MS COCO dataset, some images contain two same targets, and in the proposed DSPH framework, the image feature extractor captures both targets accurately to represent the semantic information.

J. Convergence Analysis

Fig. 9 shows the dynamics between the loss and the number of training epochs in training when the hash code length is 16. By learning multi-modal multi-label proxy terms and multi-modal irrelevant terms jointly, the semantic-aware proxy loss is developed to decrease the training complexity, and a satisfactory convergence speed is obtained.

K. Top-10 Retrieval Results

To illustrate the actual retrieval performance of our method, we display the Top-10 samples returned by the two

retrieval tasks according to the Hamming ranking on the MIRFLICKR-25K dataset with 64-bit binary codes, as shown in Fig. 12, The retrieved samples returned from our DPSH framework are all relevant. This indicates that the DSPH method can represent the semantic information of multi-label data well and generate discriminative hash codes via the overall deep cross-modal hashing framework.

V. Conclusion and Future Work

In this paper, we propose a new Deep Semantic-aware Proxy Hashing framework (DSPH), where semantic-aware proxy loss is designed to extend the proxy-based mechanism to multi-label cross-modal retrieval. The image transformer encoder is introduced as the image feature extractor to represent the global information of input images by modeling long-range visual dependencies, meanwhile, the text transformer encoder learns rich semantic features of the text. In hash learning, by learning multi-modal multi-label proxy terms and multi-modal irrelevant terms jointly, the semantic-aware proxy loss is designed to capture multi-label correlations and preserve the correct fine-grained similarity ranking among samples and alleviate inter-modal semantic gaps. DSPH could embed multi-label data into discrete space well and generates discriminative binary hash codes. Extensive experiments on the MIRFLICKR-25K, NUS-WIDE, and MS COCO datasets prove that the DSPH method shows excellent retrieval performance.

Our method introduces the semantic-aware proxy to image-text cross-modal retrieval tasks. However, multi-modal data sources involve more than just images and text. In the future, we would extend the DSPH framework to achieve image-text-video multi-modal retrieval.

References

[1] M. Cao, S. Li, J. Li, L. Nie, and M. Zhang, "Image-text retrieval: A survey on recent research and development," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 5410–5417.

[2] A. Singh and S. Gupta, "Learning to hash: A comprehensive survey of deep learning-based hashing methods," *Knowl. Inf. Syst.*, vol. 64, no. 10, pp. 2565–2597, Oct. 2022.

[3] Q. Qin, L. Huang, Z. Wei, K. Xie, and W. Zhang, "Unsupervised deep multi-similarity hashing with semantic structure for image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2852–2865, Jul. 2021.

[4] Y. Wang, Z. Chen, X. Luo, and X. Xu, "A high-dimensional sparse hashing framework for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8822–8836, Dec. 2022.

[5] X. Lu, L. Zhu, Z. Cheng, L. Nie, and H. Zhang, "Online multi-modal hashing with dynamic query-adaption," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 715–724.

[6] C. Sun, H. Latapie, G. Liu, and Y. Yan, "Deep normalized cross-modal hashing with bi-direction relation reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 4937–4945.

[7] X. Liu, Z. Hu, H. Ling, and Y. Cheung, "MTFH: A matrix tri-factorization hashing framework for efficient cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 964–981, Mar. 2021.

[8] H. Cui, L. Zhu, J. Li, Y. Yang, and L. Nie, "Scalable deep hashing for large-scale social image retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 1271–1284, 2020.

[9] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 769–790, Apr. 2018.

[10] X. Luo et al., "A survey on deep hashing methods," *ACM Trans. Knowl. Discovery From Data*, vol. 17, no. 1, pp. 1–50, Feb. 2023.

[11] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3594–3601.

[12] C. Deng, X. Tang, J. Yan, W. Liu, and X. Gao, "Discriminative dictionary learning with common label alignment for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 208–218, Feb. 2016.

[13] L. Xu, X. Zeng, B. Zheng, and W. Li, "Multi-manifold deep discriminative cross-modal hashing for medical image retrieval," *IEEE Trans. Image Process.*, vol. 31, pp. 3371–3385, 2022.

[14] X. Lu, L. Zhu, L. Liu, L. Nie, and H. Zhang, "Graph convolutional multi-modal hashing for flexible multimedia retrieval," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1414–1422.

[15] Y. Shi et al., "Deep adaptively-enhanced hashing with discriminative similarity guidance for unsupervised cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 7255–7268, Oct. 2022.

[16] L. Zhu, X. Lu, Z. Cheng, J. Li, and H. Zhang, "Deep collaborative multi-view hashing for large-scale image search," *IEEE Trans. Image Process.*, vol. 29, pp. 4643–4655, 2020.

[17] S. Su, Z. Zhong, and C. Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3027–3035.

[18] H. Hu, L. Xie, R. Hong, and Q. Tian, "Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3120–3129.

[19] J. Yu, H. Zhou, Y. Zhan, and D. Tao, "Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 4626–4634.

[20] L. Wang, M. Zareapoor, J. Yang, and Z. Zheng, "Asymmetric correlation quantization hashing for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 24, pp. 3665–3678, 2022.

[21] Q. Jiang and W. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3270–3278.

[22] S. Wang, H. Zhao, and K. Li, "Discrete joint semantic alignment hashing for cross-modal image-text search," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 8022–8036, Nov. 2022.

[23] J. Rodrigues, M. Cristo, and J. G. Colonna, "Deep hashing for multi-label image retrieval: A survey," *Artif. Intell. Rev.*, vol. 53, no. 7, pp. 5261–5307, Oct. 2020.

[24] X. Zou, S. Wu, N. Zhang, and E. M. Bakker, "Multi-label modality enhanced attention based self-supervised deep cross-modal hashing," *Knowl.-Based Syst.*, vol. 239, Mar. 2022, Art. no. 107927.

[25] 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, Jan. 2022.

[26] C. Xu, Z. Chai, Z. Xu, C. Yuan, Y. Fan, and J. Wang, "Hyp^2 loss: Beyond hypersphere metric space for multi-label image retrieval," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3173–3184.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] J. Tu, X. Liu, Z. Lin, R. Hong, and M. Wang, "Differentiable cross-modal hashing via multimodal transformers," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 453–461.

[30] Q. Qin, L. Huang, Z. Wei, J. Nie, K. Xie, and J. Hou, "Unsupervised deep quadruplet hashing with isometric quantization for image retrieval," *Inf. Sci.*, vol. 567, pp. 116–130, Aug. 2021.

[31] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3864–3872.

[32] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4242–4251.

[33] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[34] R. Xu, C. Li, J. Yan, C. Deng, and X. Liu, "Graph convolutional network hashing for cross-modal retrieval," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 982–988.

[35] X. Nie, B. Wang, J. Li, F. Hao, M. Jian, and Y. Yin, "Deep multiscale fusion hashing for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 401–410, Jan. 2021.

[36] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3893–3903, Aug. 2018.
[37] Y. Wang, Z. Chen, X. Luo, R. Li, and X. Xu, "Fast cross-modal hashing with global and local similarity embedding," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10064–10077, Oct. 2022.
[38] Y. Peng, J. Qi, Z. Ye, and Y. Zhuo, "Hierarchical visual-textual knowledge distillation for life-long correlation learning," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 921–941, Apr. 2021.
[39] Y. Duan, N. Chen, P. Zhang, N. Kumar, L. Chang, and W. Wen, "MS2GAH: Multi-label semantic supervised graph attention hashing for robust cross-modal retrieval," *Pattern Recognit.*, vol. 128, Aug. 2022, Art. no. 108676.
[40] G. Song, X. Tan, J. Zhao, and M. Yang, "Deep robust multilevel semantic hashing for multi-label cross-modal retrieval," *Pattern Recognit.*, vol. 120, Dec. 2021, Art. no. 108084.
[41] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2013, pp. 785–796.
[42] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proc. Int. Conf. Res. Develop. Inf. Retr.*, 2014, pp. 415–424.
[43] S. Liu, S. Qian, Y. Guan, J. Zhan, and L. Ying, "Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1379–1388.
[44] P. Hu et al., "Unsupervised contrastive cross-modal hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3877–3889, Mar. 2023.
[45] A. Shin, M. Ishii, and T. Narihira, "Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision," *Int. J. Comput. Vis.*, vol. 130, no. 2, pp. 435–454, Feb. 2022.
[46] S. Li, X. Li, J. Lu, and J. Zhou, "Self-supervised video hashing via bidirectional transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13544–13553.
[47] Y. Zhuo, Y. Li, J. Hsiao, C. Ho, and B. Li, "CLIP4Hashing: Unsupervised deep hashing for cross-modal video-text retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2022, pp. 158–166.
[48] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.
[49] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2016, pp. 1–11.
[50] R. Tu et al., "Deep cross-modal proxy hashing," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 6798–6810, Jul. 2023, doi: 10.1109/TKDE.2022.3187023.
[51] C. Xu et al., "HHF: Hashing-guided Hinge function for deep hashing retrieval," *IEEE Trans. Multimedia*, early access, 2022, doi: 10.1109/TMM.2022.3222598.
[52] Y. Cao, B. Liu, M. Long, and J. Wang, "Cross-modal Hamming hashing," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, vol. 11205. Cham, Switzerland: Springer, 2018, pp. 207–223.
[53] W. Gu, X. Gu, J. Gu, B. Li, Z. Xiong, and W. Wang, "Adversary guided asymmetric hashing for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2019, pp. 159–167.
[54] C. Bai, C. Zeng, Q. Ma, J. Zhang, and S. Chen, "Deep adversarial discrete hashing for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 525–531.
[55] 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, Aug. 2020.
[56] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 15–29.

Yadong Huo is currently pursuing the master's degree with Qufu Normal University, China. His research interests include multimedia content analysis and retrieval.

Qibing Qin received the Ph.D. degree in computer science from the Ocean University of China, Qingdao, China, in 2021. He is currently a Lecturer with the School of Computer Engineering, Weifang University, Weifang, China, and an Academic Visitor with the Faculty of Information Science and Engineering, Ocean University of China. His current research interests include multimedia content analysis and retrieval, image processing, machine learning, and data mining.

Jiangyan Dai received the Ph.D. degree from the School of Mathematics and Statistics, Northeast Normal University, in 2014. She is currently an Associate Professor with the School of Computer Engineering, Weifang University, China. Her research interests include image processing, computer vision, and machine learning.

Lei Wang received the Ph.D. degree from the School of Computer Science and Technology, Shandong University, in 2010. He is currently an Associate Professor with the School of Computer Engineering, Weifang University, China. His research interests include intelligent manufacturing systems and machine learning.

Wenfeng Zhang (Member, IEEE) received the Ph.D. degree in computer science from the Ocean University of China, Qingdao, China, in 2021. He is currently a Lecturer with the College of Computer and Information Science, Chongqing Normal University, Chongqing, China, and the Lishui Institute, Hangzhou Dianzi University, Lishui, China. His research interests include computer vision and machine learning.

Lei Huang (Member, IEEE) received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2013. He is currently a Professor with the Ocean University of China, Qingdao, China. His current research interests include multimedia content analysis and retrieval, computer vision, pattern recognition, and machine learning.

Chengduan Wang is currently a Professor with the School of Computer Engineering, Weifang University, Weifang, China. His current research interests include information processing, machine learning, and data mining.