Pattern Recognition 160 (2025) 111225

Contents lists available at ScienceDirect

Pattern Recognition

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

Semantic decomposition and enhancement hashing for deep cross-modal retrieval

Lunke Fei [a], Zhihao He [a], Wai Keung Wong [b,*], Qi Zhu [c], Shuping Zhao [a], Jie Wen [d]

[a] The School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
[b] The Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong, China
[c] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[d] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China

ARTICLE INFO

Keywords:
Cross-modal retrieval
Semantic decomposition
Deep cross-modal hashing
Multi-label semantic learning

ABSTRACT

Deep hashing has garnered considerable interest and has shown impressive performance in the domain of retrieval. However, the majority of the current hashing techniques rely solely on binary similarity evaluation criteria to assess the semantic relationships between multi-label instances, which presents a challenge in overcoming the feature gap across various modalities. In this paper, we propose semantic decomposition and enhancement hashing (SDEH) by extensively exploring the multi-label semantic information shared by different modalities for cross-modal retrieval. Specifically, we first introduce two independent attention-based feature learning subnetworks to capture the modality-specific features with both global and local details. Subsequently, we exploit the semantic features from multi-label vectors by decomposing the shared semantic information among multi-modal features such that the associations of different modalities can be established. Finally, we jointly learn the common hash code representations of multimodal information under the guidelines of quadruple losses, making the hash codes informative while simultaneously preserving multilevel semantic relationships and feature distribution consistency. Comprehensive experiments on four commonly used multimodal datasets offer strong support for the exceptional effectiveness of our proposed SDEH.

1. Introduction

With the rapid advancement of internet technology, there has been a remarkable proliferation of different modalities of information, such as text, audio, and image. How to conduct similarity searches across different modalities has become a foundational requirement for media users, which however still faces the primary challenges of the large heterogeneous feature and semantic gaps between different modalities. Due to the binary representation of hash codes with high computational efficiency and low storage requirement, cross-modal hashing (CMH) [1] offers an efficient way to retrieve related instances from one modality to another by mapping high-dimensional semantic features into low-dimensional hash codes while preserving the similarity of the original space.

In recent decades, numerous cross-modal hashing methods have been proposed, including unsupervised and supervised methods. Unsupervised methods [2,3] aim to learn hash codes by exploring the intrinsic structure of the data in Euclidean space while simultaneously preserving the original similarity. For example, Ding et al. [2] utilized collective matrix factorization for hash learning, and Wang et al. [3] utilized individual and joint matrix factorization techniques to simultaneously preserve the specific and shared properties of data for hash learning. In general, unsupervised methods usually exhibit inferior performances as compared with supervised methods because supervised methods [6] can use label guidance to establish correlations across modalities to increase retrieval effectiveness. For example, Huang et al. [4] utilized the complementary information between modality-specific features and semantic labels for hash learning, and Qin et al. [5] introduced an asymmetric learning framework for generating hash codes with supervised label information.

Over the past few years, with the remarkable performance of deep neural networks (DNNs), deep supervised cross-modal hashing methods [9–12] have emerged and achieved promising performances in retrieval tasks. For example, Jiang et al. [7] developed a deep hashing model by integrating feature learning and hashing learning processes, and Zhu et al. [8] introduced a deep hashing model by comprehensively

* Corresponding author.
E-mail address: calvin.wong@polyu.edu.hk (W.K. Wong).

https://doi.org/10.1016/j.patcog.2024.111225
Received 3 June 2024; Received in revised form 16 September 2024; Accepted 24 November 2024
Available online 26 November 2024

capturing cross-modal semantic-consistency features. Notably, most existing deep cross-modal hashing methods only measure the similarity of two instances based on whether they share at least one category label, ignoring the fact that two cross-modal paired instances can share multiple similar labels. To better reflect the complex relationships among instances during cross-modal hashing, many self-supervised hashing methods have been proposed. For example, Shu et al. [13] introduced a novel deep hashing model to learn class centers to guide the multi-modal information hashing process. Zuo et al. [14] integrated a multi-label semantic affinity preserving module into a semantic network to explore the semantic correlations within multi-labels for hash learning. Due to the sparsity of the original multi-label vectors, most existing self-supervised deep hashing methods are usually hard to directly capture the common semantic information within multi-labels, making it still a challenge to precisely measure the cross-modal semantic similarity.

In this paper, we propose a new semantic decomposition and enhancement hashing (SDEH), which consists of three modality-specific feature learning subnetworks and a hash learning subnetwork. First, we employ ResNet-based and BOW-based backbones embedded with parallel channels and spatial attention blocks to capture modality-specific features with both global and local details. Then, we develop a joint feature learning subnetwork to capture the semantic information from multi-label vectors by decomposing the associations shared by the image and text modalities. Finally, we jointly convert multimodal and multi-label feature descriptors into common hash codes guided by quadruple loss functions. Comprehensive experiments on four prominent multi-modal datasets convincingly demonstrate the superiority of the proposed SDEH.

The primary contributions of this paper can be outlined as follows:

- We develop a new joint feature learning network, which first exploits the modality-specific feature descriptors and then jointly learns the common semantic information from shared multi-label vectors such that the gap between different modalities can be bridged.
- We develop quadruple losses, including cross-modal triple, intra-modal multilevel semantics, quantization, and balancing losses, to make the hash codes informative by comprehensively preserving the multilevel semantic relationships within each modality and reducing information loss during hashing.
- We perform comprehensive experiments on four challenging cross-modal datasets, and the experimental results demonstrate the superiority of our proposed SDEH over most existing state-of-the-art methods.

The rest of this paper is structured as follows. Section 2 offers an overview of related work. Section 3 provides a detailed explanation of our proposed SDEH method. Section 4 presents an analysis of the experimental results. Section 5 concludes this paper.

2. Related work

In this section, we briefly review three related topics, including self-supervised semantic learning, multi-label semantic learning, and the attention mechanism.

2.1. Self-supervised semantic learning

Self-supervised semantic learning [18,19] aims to project label information into a higher-dimensional semantic space to capture latent semantic information, thereby alleviating the heterogeneity gap across modalities in cross-modal hashing tasks. For example, Li et al. [15] integrated self-supervised semantic learning into a cross-modal hashing framework to maintain semantic consistency across modalities. Ma et al. [16] introduced a global and local semantic alignment mechanism to learn hash codes in a self-supervised manner. In addition, Zou et al. [17] integrated an attention module into self-supervised semantic learning to compensate for the semantic information of multi-label vectors for hash learning. In this study, we design a new joint feature learning network based on self-supervised learning by enhancing multi-label semantic information with shared multi-modal information.

2.2. Multi-label semantic learning

The multi-label of an instance usually contains more complex semantic information than a single label does; however, measuring the semantic information from multi-labels with a simple binary similarity evaluation is difficult. To address this, there have been numerous multi-label semantic learning studies presented in the literature, which aim to associate an instance with multiple label categories for a better exploration of multi-label semantic information. For example, Huo et al. [20] introduced semantic-aware proxy loss to fully exploit semantic correlations in multi-labels for hash learning. Liu et al. [21] integrated semantic ranking with feature learning to fully leverage the multilevel semantic correlations of multi-labels for hashing. Sun et al. [22] designed a bidirectional relation reasoning scheme to explore the multilevel semantic correlations within multiple labels based on the consistent and inconsistent relationships of multi-labels for hash learning. In addition, Zhan et al. [23] introduced a semantic correction similarity matrix to fully preserve the multi-label semantic relevance of hash codes, and Zou et al. [24] proposed a multi-label semantic similarity criterion to evaluate the complex semantic relationships between multi-label instances for hash learning. In this study, with the aim of preserving multilevel semantic relationships within each modality, we propose an intra-modal multilevel semantic loss to leverage the complex semantic relationships within multi-labels for cross-modal hashing.

2.3. Attention mechanism

Attention mechanisms [25] can dynamically focus on learning important information (such as feature channels) of the input data and effectively capture the critical features (such as local details of modality-specific information) for downstream tasks. In the last few decades, several cross-modal hashing studies have employed attention mechanisms to specifically learn crucial features of multimodal data. For example, Yao et al. [26] introduced an attention-aware deep cross-modal hashing method by highlighting useful information and simultaneously suppressing irrelevant information. Zhang et al. [27] developed an attention-aware deep adversarial hashing method to learn hash codes by intentionally focusing on important regions in multimodal data. In addition, Zhang et al. [28] designed a deep medical cross-modal attention hashing method to exploit both global and local information in multimodal features for hash learning. In this work, we develop an attention-based network to focus primarily on the multi-modal collaborative and task-specific feature learning, making the cross-modal retrieval relatively easy.

3. Proposed SDEH method

In this section, we first present the notations used in this paper. Then, we elaborate our proposed SDEH method, including the overall architecture and the main components of the SDEH.

3.1. Notation

Without loss of generality, in this paper, two widely used image and text modalities are considered for cross-modal retrieval. Suppose we have a multi-label cross-modal dataset $O = \{v_i, t_i, l_i\}_{i=1}^{N}$ with Ninstances, where v_i and t_i represent the image-modality and text-modality data of the $i-$th instance, respectively. $l_i \in \{0,1\}^C$ denotes the multi-label vector, where C represents the total number of categories, and the $j-$th

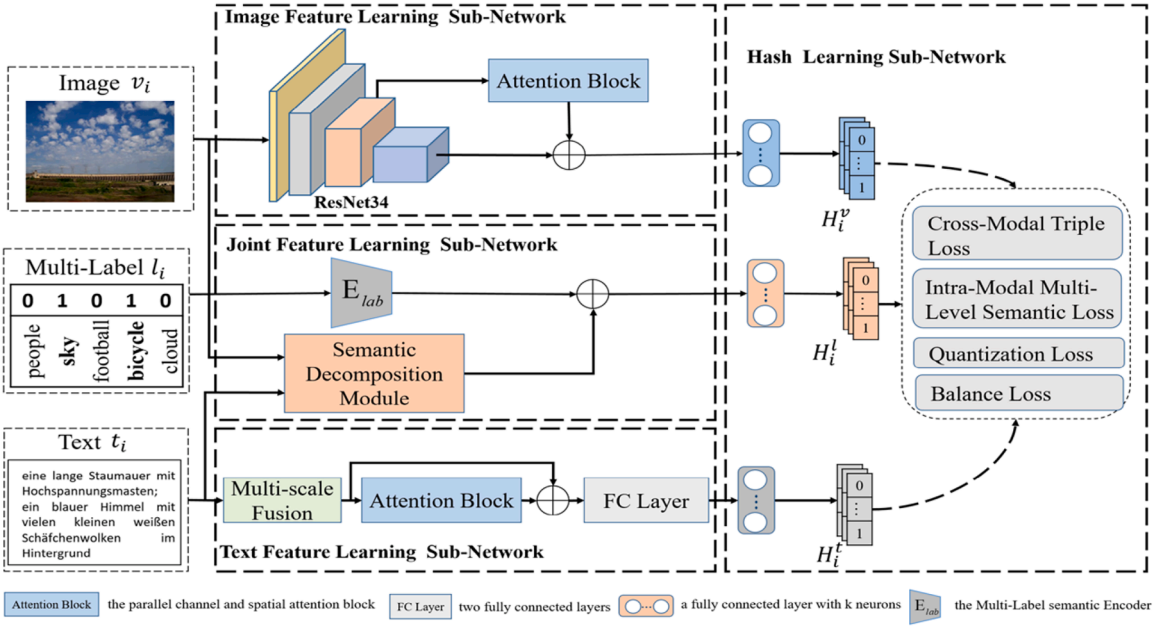


Fig. 1. The main framework of the proposed SDEH consists of three feature learning and one hash learning subnetworks.

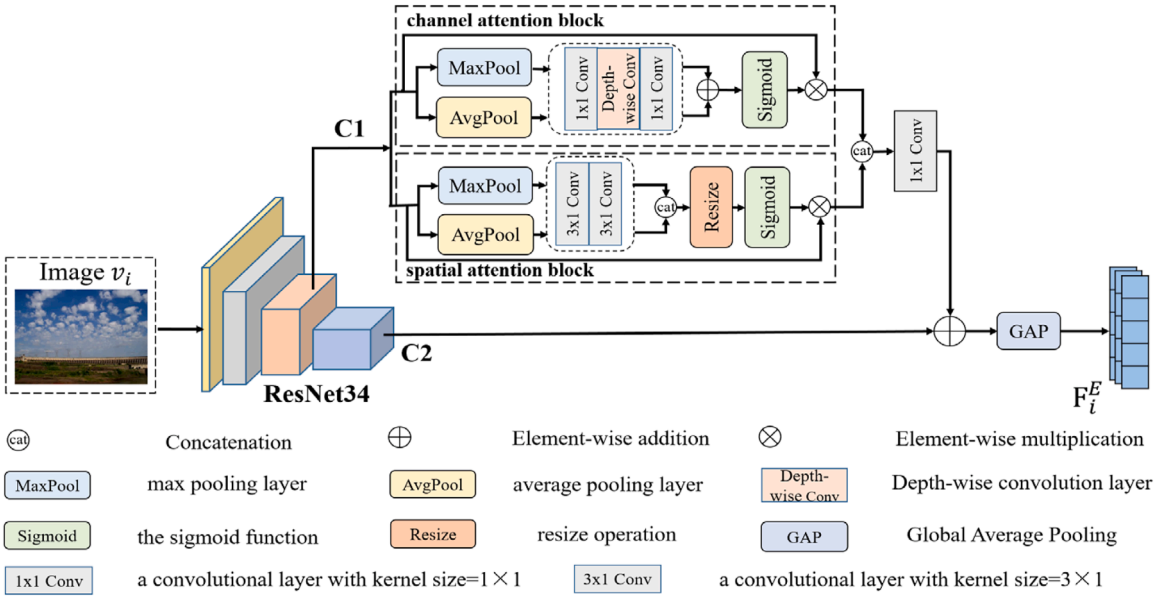


Fig. 2. The basic architecture of the image feature learning subnetwork.

element of l_i (i.e., l_{ij}) is 1 if the $i-$th instance (e.g., O_i) belongs to the $j-$th class, and 0 otherwise. In this paper, we construct a similarity matrix $S \in R^{N\times N}$ on the basis of the multi-label vectors (i.e., $\{l_i\}_{i=1}^N$). Specifically, if the image v_i and the text t_j do not share any labels, S_{ij} is set to 0, and otherwise S_{ij} is a real value between 0 and 1, representing the degree of similarity between the two instances (seen in Section 3.6).

3.2. SDEH architecture

To mitigate the inherent distribution differences across modalities in feature space, our proposed SDEH network aims to learn the common representation of the two different modalities while preserving their semantic similarity. To achieve this, we first learn the modality-specific features from both image and text data, which aim to essentially characterize the discriminative information associated with image and text modalities. To better establish the bridge crossing image to text modalities, we further learn the multi-label feature shared by both the image and text modalities. The modality-specific and multi-label feature learning can be functioned as follows:

$$\begin{cases} F^E = f^v\left(\{v_i\}_{i=1}^N, \theta^v\right), \\ G^E = f^t\left(\{t_i\}_{i=1}^N, \theta^t\right), \\ L^E = f^l\left(\{v_i, t_i, l_i\}_{i=1}^N, \theta^l\right), \end{cases} \tag{1}$$

where f^v, f^t, and f^l represent the feature learning functions of the image-modality, text-modality, and multi-label information, respectively, and where F^E, G^E, and L^E represent the corresponding feature maps. θ denotes the parameters of the related feature learning functions. Furthermore, via hash learning, we convert the multi-modal and multi-label feature maps into their hash code representations (denoted as H^v, H^t,

Algorithm 1
Image feature learning.

Input: Image data X.
Output: Enhanced text features F^E.
1: Initialize the parameters of the image feature learning subnetwork.
2:**repeat**
3: **for** each mini-batch $\{X_i\}$ **do**
4: Extract two-scale image feature maps with the scaled steps of 16 and 32 through ResNet34:
C_1, C_2=ResNet34(X_i)
5: Enhance local details of the feature map C_1 using parallel spatial and channel attention modules:
$F_i^{PSCA} = W_{1\times1}\text{Concat}(SAB(C_1), CAB(C_1))$,
6: Combine the enhanced C_1 with C_2 to generate the enhanced image feature descriptor:
$F_i^E = GAP(C_2 + F_i^{PSCA})$,
7: Update the subnetwork parameters with $(\mathscr{L}_{total} - \mathscr{L}_{decomposition})$ by using back propagation.
8: **until** convergence

and H^l) and enforce a bank of loss constraints on these hash codes, making them as similar as possible. To achieve this, we develop an SDEH network, as presented in Fig. 1, which comprises four subnetworks, including three feature extractors and a hash code generator.

3.3. Image feature learning

Fig. 2 depicts the basic architecture of the image feature learning subnetwork. Following [8,17], to extensively exploit the semantic information from images, we employ the widely used ResNet-34 [29] as the backbone for extracting the two-scale image feature maps with scaled steps of 16 and 32 (referred to as C_1 and C_2), respectively. Notably, the deeper-layer feature map C_2 usually focuses more on global features while the local details of images are ignored. To address this, we develop a parallel channel and spatial attention block to enhance the local details of the shallow feature map C_1.

To be more specific, in the channel attention block, we first use max and average pooling to simultaneously extract channel-specific information from C_1, and further map it into the same semantic space via a three-layer convolution, such that two channel-specific weight relationship maps can be obtained. After that, we integrate the two weight relationship maps via a summation operation and normalize them via the sigmoid function. Moreover, in the spatial attention block, we exploit the spatial information from the parallel max and average pooling maps via two 3×1 convolution layers and further integrate these two pooled feature maps via concatenation to generate the spatial attention feature map. During feature learning, we resize the spatial attention feature map via bilinear interpolation and a 1×1 convolution layer, making it consistent with the original spatial and channel dimensions. The feature learning processes of the parallel channel and spatial attention module can be formulated as follows:

$$F_i^{SA} = C_1 * Sigmoid(W_1 * AvgPool(C_1) + W_1 * MaxPool(C_1)), \tag{2}$$

and

$$F_i^{CA} = C_1 * Sigmoid\big(H_{up} * W_3 * Concat(W_2 * AvgPool(C_1), W_2 * MaxPool(C_1))\big), \tag{3}$$

where $AvgPool()$ and $MaxPool()$ represent the average pooling and maximum pooling operations, respectively. W_1 represents the weight of the convolution group, which contains two 1×1 convolution layers and one depth-separable convolution layer. W_2 represents the weights of the two 3×1 convolutional layers, while W_3 represents the weights of the 1×1 convolutional layer. $Concat()$ represents the concatenation operation, H_{up} represents the up-sampling operation, and $Sigmoid()$ represents the sigmoid function. Finally, we fuse the enhanced C_1 (i.e., F_i^{SA} and F_i^{CA}) with C_2 to obtain the image-modality feature descriptor (referred to as F_i^E) as follows:

$$F_i^E = GAP\big(C_2 + W_3 * Concat\big(F_i^{SA}, F_i^{CA}\big)\big). \tag{4}$$

In this manner, we can learn the image feature descriptor that simultaneously highlights both the global and local details. Algorithm 1 summarizes the procedure of the image-modality feature learning.

3.4. Text feature learning

Fig. 3 shows the basic architecture of the text feature learning subnetwork, which seeks to extensively exploit the semantic information from the text modality. Specifically, we first encode the raw text information into bag-of-words (BOW) vectors [7]. After that, with the aim of determining the relevance of text data, we extract multiscale feature maps from the text BOW via a multiscale fusion module (MS) [15], which comprises five average pooling layers with sizes of 50×50, 30×30, 15×15, 10×10, and 5×5, adding a 1×1 convolutional layer. Moreover, we utilize parallel channel and spatial attention blocks with similar structures as the image attention subnetwork to enhance the local details of the multiscale text feature maps. Finally, we convert the enhanced text feature maps into the final text descriptor via two fully connected layers with 4096 and 512 neurons, enhancing the expressive capability of the text feature descriptor. Algorithm 2 presents the detailed procedure of the text-modality feature learning.

3.5. Joint feature learning

The multi-label vector usually preserves important common semantic information shared by the intraclass image and text modalities.

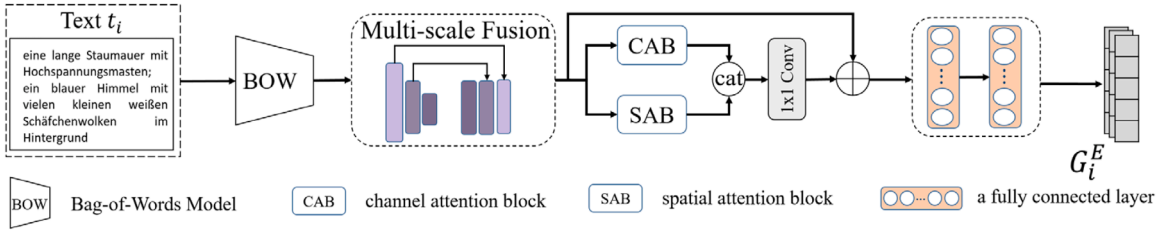


Fig. 3. The basic architecture of the text feature learning subnetwork.

Algorithm 2
Text Feature Learning.

Input: Text data Y.
Output: Enhanced text features G^E.
1: Initialize the parameters of the text feature learning subnetwork.
2:**repeat**
3: **for** each mini-batch $\{Y_i\}$ **do**
4: Convert the input text Y_i into a Bag-of-Words (BOW) vector t_i:
$BOW(Y_i) \rightarrow Vector\ representation\ of\ t_i$,
5: Use multi-scale convolution operations to extract multi-scale features as:
$G_i^{ms} = MS(BOW(t_i))$,
6: Enhancing local details of multi-scale features using parallel spatial and channel attention modules:
$G_i^{PSCA} = G_i^{ms} + W_{1\times1}\mathrm{Concat}\left(SAB(G_i^{ms}), CAB(G_i^{ms})\right)$,
7: Apply a fully connected layer to generate the final enhanced text features G_i^E.
8: Update the subnetwork parameters with $(\mathscr{L}_{total} - \mathscr{L}_{decomposition})$ by using back propagation.
9: **until** convergence

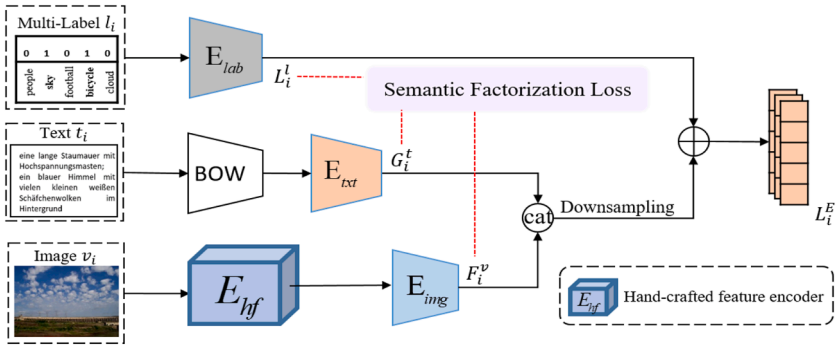


Fig. 4. Basic architecture of the semantic decomposition module.

Due to the sparsity, it is usually difficult to directly capture the common semantic information within multi-labels. To address this, we jointly learn the multi-label semantic information by decomposing the shared semantic information from the image and text modalities. Fig. 4 illustrates the basic architecture of the semantic decomposition module.

First, we convert text modality into word vectors via BOW, and simultaneously convert the image modality into hand-crafted feature vector via a encoder E_{hf}. Then, we design three encoders for image, text, and multi-label data (referred to as E_{img}, E_{txt}, and E_{lab}), each of which consists of two fully connected layers with 4096 and 512 neurons, and a ReLU activation function, to collaboratively exploit the consistent semantic information from diverse modalities. Specifically, via the encoders, the image, text and multi-label data can be encoded into 512-dimensional feature descriptors, referred to as F_i^v, G_i^t, and L_i^l, respectively. After that, we decompose the common semantic information shared by the image and text feature descriptors (i.e., F_i^v and G_i^t) to capture the multi-label semantic feature, which can be achieved via the following loss function:

$$\mathscr{L}_{decomposition} = \sum_{i=1}^{N} (\alpha \parallel F_i^v - G_i^t\|_2^2 + \beta(\ \|F_i^v - L_i^l\|_2^2 + \|G_i^t - L_i^l\|_2^2)), \tag{5}$$

where α and β are balancing parameters. $\|\cdot\|_2$ is the L_2-norm. L_i^l denotes the semantic information learned from multi-labels. In the loss function, the first term minimizes the mean square error between F_i^v and G_i^t, aiming to exploit the shared information of the image and text modalities, and the second item aims to explore the common multi-label semantic information (i.e., L_i^l) via F_i^v and G_i^t. Finally, we generate the shared multi-label semantic feature descriptor as follows:

$$L_i^E = L_i^l + \mathrm{downSample}\left(\mathrm{concat}\left(F_i^v, G_i^t\right)\right), \tag{6}$$

where concat() is the concatenation operation, and where downSample() represents the dimensionality reduction. L_i^E represents the enhanced multi-label semantic feature of the $i-$th instance, which

Algorithm 3
Joint feature learning.

Input: Image data X, Text data Y, Multi-Label data L.
Output: Enhanced multi-label semantic features L^E.
1: Initialize the parameters of three encoders E_{img}, E_{txt}, and E_{lab}, respectively.
2:**repeat**
3: **for** each mini-batch $\{X_i, Y_i, L_i\}$ **do**
4: Compute the image, text, and multi-label feature descriptors via the three encoders as:

$$\begin{cases} F_i^v = E_{img}(E_{hf}(X_i)), \\ G_i^t = E_{txt}(BOW(Y_i)), \\ L_i^l = E_{lab}(L_i), \end{cases}$$

5: Decompose the shared semantic information of image and text feature descriptors by semantic decomposition loss function $\mathscr{L}_{decomposition}$.
6: Fuse the shared semantic information with the multi-label feature descriptors via:
$L_i^E = L_i^l + downSample(concat(F_i^v, G_i^t))$,
7: Update the three encoders parameters with $(\mathscr{L}_{total} - \mathscr{L}_{inter})$ by using back propagation.
8: **until** convergence

Algorithm 4
SDEH.

Input: Training set $\{v_i, t_i, l_i\}_{i=1}^{N}$, batch size bs, and hash code length K.
Output: Binary hash code matrix B, and optimized network parameters θ^l, θ^v, θ^t.
1: Calculate the multi-level semantic similarity matrix S by Eqs. (8), (9), and (10).
3: Initialize the parameters of networks θ^v, θ^t, and θ^l, and hyper-parameters α, β, λ, φ, γ, η, and m.
4: **repeat**
5: **for** $j = 1{:}\frac{N}{bs}$ **do**
6: Compute the hash codes for the image-modality, text-modality, and label information separately by performing forward propagation using a mini-batch of samples;
7: Update θ^l with $(\mathscr{L}_{total} - \mathscr{L}_{\text{inter}})$ by using back propagation;
8: Update $\theta^*, * \in \{v, t\}$ with $(\mathscr{L}_{total} - \mathscr{L}_{decomposition})$by using back propagation;
9: **end for**
10: Update the binary hash code matrix B by $B = \text{sign}(H^v + H^t + H^l)$.
11: **until** convergence

conveys both the multi-label semantic and the common semantic information shared by both the image- and text-modality data. In addition, in this way, the image and text feature learning subnetworks can be trained in a self-supervised manner. A detailed description of the joint feature learning is summarized in Algorithm 3.

3.6. Hash learning

Having obtained multimodal data and label feature maps, we further map them into K-bit hash code vectors via a fully connected layer with K neurons, referred to as H^v, H^t, and H^l, respectively. To make the hash code representations informative, we enforce four constraints during the learning of hash codes, including the cross-modal triple, intra-modal multilevel semantics, quantization, and balancing losses.

Cross-modal triple loss aims to minimize the cross-modal distance of instances with similar semantics while simultaneously maximizing the distances of instances with dissimilar semantics. Suppose that the $j-$th instance is semantically related to the $i-$th instance with more similar label information, and that the $k-$th instance is semantically unrelated to the $i-$th instance with no similar label information. We define the following cross-modal triple loss function:

$$\begin{aligned}\mathscr{L}_{\text{inter}} &= \mathscr{L}_{VL} + \mathscr{L}_{TL} \\ &= \sum_{i,j,k}^{N} \max\left(0, \cos\left(H_i^l, H_k^v\right) - \cos\left(H_i^l, H_j^v\right) + m\right) + \sum_{i,j,k}^{N} \max\left(0, \cos\left(H_i^v, H_k^l\right) - \cos\left(H_i^v, H_j^l\right) + m\right) \\ &+ \sum_{i,j,k}^{N} \max\left(0, \cos\left(H_i^l, H_k^t\right) - \cos\left(H_i^l, H_j^t\right) + m\right) + \sum_{i,j,k}^{N} \max\left(0, \cos\left(H_i^t, H_k^l\right) - \cos\left(H_i^t, H_j^l\right) + m\right),\end{aligned} \tag{7}$$

where H^v, H^t, and H^l represent the corresponding hash code representations of the image-modality, text-modality, and label information, respectively. m represents the margin parameter. $\mathscr{L}_{VL}$ and $\mathscr{L}_{TL}$ denote the cross-modal triple losses of the image and text modalities, respectively. By minimizing the overall cross-modal triple loss $\mathscr{L}_{\text{inter}}$, the instance pairs with greater semantic similarity usually have a smaller feature distance.

Intra-Modal multilevel semantic loss aims to comprehensively preserve the multilevel semantic relationships within each modality. To achieve this, we first employ the relevant information to calculate the similarity of two instances (referred to as S_{ij}^{re}) as follows:

$$S_{ij}^{re} = \frac{\| l_i \cap l_j \|}{\| l_i \| + \| l_j \| - \| l_i \cap l_j \|}, \tag{8}$$

where $\| l_i \|$ (or $\| l_j \|$) denotes the number of categories to which o_i (or o_j) belongs, and where $\| l_i \cap l_j \|$ represents the number of categories to which both instance o_i and instanceo_j belong. Furthermore, we use the irrelevant information of two instances to compute their similarity (referred to as S_{ij}^{unre}) as follows:

$$S_{ij}^{unre} = \begin{cases} \dfrac{\| \widehat{l}_i \cap \widehat{l}_j \|}{\| \widehat{l}_i \| + \| \widehat{l}_j \| - \| \widehat{l}_i \cap \widehat{l}_j \|}, & \text{if instances } i \text{ and } j \text{ share one or more categories}, \\ 0, & \text{else}, \end{cases} \tag{9}$$

where $\widehat{l}_i$ and $\widehat{l}_j$ are the inverses of l_i and l_j, respectively, i.e., $\widehat{l}_* = \sim l_* (* = i, j)$. After that, we combine S_{ij}^{re} and S_{ij}^{unre} to form the multilevel similarities of two instances as follows:

$$S_{ij} = \lambda S_{ij}^{re} + (1 - \lambda) S_{ij}^{unre}, \tag{10}$$

where λ is the balance parameter.

With the multilevel semantic similarity matrix (i.e., S_{ij}), we formulate the following intra-modal multilevel semantic loss function:

$$\mathscr{L}_{\text{intra}} = \mathscr{L}_L + \mathscr{L}_V + \mathscr{L}_T = \sum_{i=1,j=1}^{N} \left(S_{ij} - \cos\left(H_i^l, H_j^l\right)\right)^2 + \sum_{i=1,j=1}^{N} \left(S_{ij} - \cos\left(H_i^v, H_j^v\right)\right)^2 + \sum_{i=1,j=1}^{N} \left(S_{ij} - \cos\left(H_i^t, H_j^t\right)\right)^2, \tag{11}$$

By minimizing $\mathscr{L}_{\text{intra}}$, we can fully explore the rich multilevel semantic relationships within multi-label vectors for each modality.

Balancing loss aims to maximize the information in each bit of the generated real-value hash codes, which takes the following form:

$$\mathscr{L}_{balance} = \sum_{i=1}^{N} \| H_i^l 1 \|_2^2 + \| H_i^v 1 \|_2^2 + \| H_i^t 1 \|_2^2, \tag{12}$$

where I is a vector of all ones. The balancing loss function (9) ensures that the hash codes of $+1$ and -1 are consistent with training samples, so that more information can be conveyed in each bit.

Quantization loss aims to minimize the information loss between the binary hash codes and the corresponding feature descriptors during hash learning, which has the following form:

$$\mathscr{L}_{quantization} = \sum_{i=1}^{N} \left(\| B_i - H_i^l \|_2^2 + \| B_i - H_i^v \|_2^2 + \| B_i - H_i^t \|_2^2 \right), \tag{13}$$

where $B_i = \text{sign}(H_i^v + H_i^t + H_i^l)$ denotes the unified binary hash code of the i−th instance, obtained via the sign function.

By combining the $\mathscr{L}_{decomposition}$, $\mathscr{L}_{\text{inter}}$, $\mathscr{L}_{\text{intra}}$, $\mathscr{L}_{balance}$, and $\mathscr{L}_{quantization}$, the overall loss function for hashing learning can be written as follows:

$$\min_{\theta^l, \theta^v, \theta^t, B} \mathscr{L}_{total} = \mathscr{L}_{decomposition} + \mathscr{L}_{\text{inter}} + \varphi \mathscr{L}_{\text{intra}} + \eta \mathscr{L}_{balance} + \gamma \mathscr{L}_{quantization}, \quad s.t. B \in \{-1, 1\}^{N \times k}, \tag{14}$$

where φ, η, and γ represent balancing parameters. θ^l, θ^v, and θ^t are the learning parameters of the multi-label, image, and text subnetworks, respectively. B is the final learned unified binary hash code for all the instances.

3.7. Optimization

During model training, we employ an alternating learning strategy [7] and stochastic gradient descent (SGD) [30] to optimize the learning parameters θ^l, θ^v, θ^t, and B. In particular, we alternately update the parameters of one subnetwork through back-propagation (BP) while keeping the parameters of the other subnetworks fixed, and repeat this step until the network converges. Algorithm 4 summarizes the primary optimization procedure of the proposed method.

Optimizing θ^l: By fixing other related parameters, θ^l can be updated by using the stochastic gradient descent (SGD) with the back-propagation (BP) algorithm based on the derivations of H^l, F^v, G^t, and L^l, as follows:

$$\begin{cases} \dfrac{\partial \mathscr{L}_{total}}{\partial H^l} = \displaystyle\sum_{i,j,k}^{N} \left(\sin\left(H_i^l, H_j^v\right) + \sin\left(H_i^v, H_j^l\right) - \left(\sin\left(H_i^l, H_k^v\right) + \sin\left(H_i^v, H_k^l\right)\right)\right) \\ \quad + \displaystyle\sum_{i,j,k}^{N} \left(\sin\left(H_i^l, H_j^t\right) + \sin\left(H_i^t, H_j^l\right) - \left(\sin\left(H_i^l, H_k^t\right) + \sin\left(H_i^t, H_k^l\right)\right)\right) \\ +2\varphi \displaystyle\sum_{i,j}^{N} \left(S_{ij} - \cos\left(H_i^l, H_j^l\right)\right) \sin\left(H_i^l, H_j^l\right) + 2\eta \sum_{i}^{N} H_i^l 1 - 2\gamma \sum_{i}^{N} (B_i - H_i^l), \\ \dfrac{\partial \mathscr{L}_{total}}{\partial F^v} = 2 \displaystyle\sum_{i}^{N} \left(\alpha\left(F_i^v - G_i^t\right) + \beta\left(F_i^v - L_i^l\right)\right), \\ \dfrac{\partial \mathscr{L}_{total}}{\partial G^t} = 2 \displaystyle\sum_{i}^{N} \left(\beta\left(G_i^t - L_i^l\right) - \alpha\left(F_i^v - G_i^t\right)\right), \\ \dfrac{\partial \mathscr{L}_{total}}{\partial L^l} = -2\beta \displaystyle\sum_{i}^{N} \left(F_i^v + G_i^t - 2L_i^l\right). \end{cases} \tag{15}$$

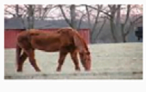
Databases	Type	Example 1	Example 2	Example 3
MIRFlickr-25K	Image Example			
	Text Example	trees, sunset, naturesfinest	vintage, piano honkytonk	branch, leaf, sky
NUS-WIDE	Image Example			
	Text Example	flowers, trees, grass, garden	new forest, foals	home, nest
IAPRTC-12	Image Example			
	Text Example	two men and two women with a lot of luggage are standing in the counter hall of an airport;	on the way through the desert on the Panamericana	a man is standing on the edge of a steep, grey canyon
MS COCO2014	Image Example			
	Text Example	A brown horse is grazing grass near a red house.	A man and a woman stand on the sidewalk lined with street lights.	A decorated table covered with plates of food and vegetables.

Fig. 5. Some typical image and text samples randomly selected from the MIRFlickr-25k, NUS-WIDE, IAPRTC-12, and MS COCO 2014 datasets.

Table 1
MAP values of different methods with diverse code lengths on the MIRFlickr-25 K, NUS-WIDE, IAPRTC-12, and MS COCO2014 datasets.

Task	Method	MIRFlickr-25K			NUS-WIDE			IAPRTC-12			MS COCO2014		
		16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I->*T*	SCM	67.85	69.62	69.96	52.12	53.24	54.42	40.24	44.13	45.98	55.02	46.61	54.49
	SePH	69.32	68.93	69.80	57.75	58.35	62.08	42.63	44.72	45.63	45.56	47.19	48.09
	DCMH	73.81	75.47	76.42	60.07	61.62	61.70	47.43	47.63	48.74	52.17	52.73	53.41
	PRDH	75.63	77.74	78.58	61.97	63.66	65.84	49.35	52.81	54.47	60.31	65.43	69.39
	CMHH	77.72	78.10	78.67	63.74	62.50	65.15	53.33	54.08	54.73	63.44	66.06	65.27
	AGAH	75.66	78.15	79.34	60.18	63.31	65.38	51.25	53.45	55.23	60.92	61.92	60.62
	MSCH	76.06	78.31	79.70	62.17	65.09	65.91	51.36	55.34	57.55	60.26	67.29	70.44
	MESDCH	80.74	82.42	83.89	64.82	66.85	68.21	53.96	57.64	60.60	67.36	72.04	75.03
	SCCGDH	77.68	79.06	79.77	64.14	65.66	66.06	49.14	48.14	54.80	56.48	57.40	61.81
	DAPH	78.53	79.29	80.44	**66.09**	67.41	68.72	50.77	54.81	58.63	67.21	72.02	74.68
	SDEH	**81.55**	**83.76**	**84.78**	65.06	**67.71**	**68.85**	**56.74**	**60.38**	**63.03**	**67.42**	**72.12**	**75.04**
T->*I*	SCM	68.61	69.85	71.42	55.42	56.48	58.82	41.18	42.40	44.83	50.19	42.22	51.20
	SePH	69.51	69.36	69.14	58.41	59.52	61.17	42.54	44.15	45.10	49.84	50.87	52.18
	DCMH	77.09	78.35	78.83	62.11	62.79	62.37	51.03	53.32	53.80	53.03	53.71	54.18
	PRDH	78.61	80.63	81.66	65.78	67.45	68.59	50.31	54.98	58.39	60.58	64.72	68.85
	CMHH	76.20	76.31	75.16	62.32	60.99	62.83	52.46	52.89	53.72	62.36	64.36	63.87
	AGAH	76.30	79.10	79.69	61.61	65.55	65.35	52.36	55.12	56.74	61.29	61.07	61.28
	MSCH	77.78	79.07	79.87	63.38	66.00	66.92	52.09	55.31	57.20	59.10	65.69	68.69
	MESDCH	78.94	80.42	81.52	64.71	66.82	67.57	52.38	55.65	58.43	**67.21**	71.96	74.49
	SCCGDH	79.25	78.65	79.33	64.36	65.52	66.74	52.00	52.48	52.93	57.00	59.48	65.66
	DAPH	78.73	80.75	82.20	**66.74**	67.50	68.41	48.72	55.76	60.24	66.75	72.02	74.33
	SDEH	**80.15**	**81.87**	**82.79**	65.82	**67.85**	**68.99**	**55.78**	**59.47**	**62.07**	66.77	**72.08**	**74.52**

Optimizing θ^v: When fixing other irrelevant parameters, θ^v can be similarly updated through the SGD with BP based on the gradient of H^v, as follows:

$$\frac{\partial \mathscr{L}_{total}}{\partial H^v} = \sum_{i,j,k}^{N} \left(\sin\left(H_i^l, H_j^v\right) + \sin\left(H_i^v, H_j^l\right) - \left(\sin\left(H_i^l, H_k^v\right) + \sin\left(H_i^v, H_k^l\right)\right)\right)$$
$$+2\varphi \sum_{i,j}^{N} \left(S_{ij} - \cos\left(H_i^v, H_j^v\right)\right)\sin\left(H_i^v, H_j^v\right) + 2\eta \sum_{i}^{N} H_i^v 1 - 2\gamma \sum_{i}^{N} \left(B_i - H_i^v\right). \tag{16}$$

Optimizing θ^t: When the other related parameters are fixed, θ^t can be similarly updated via the SGD with BP based on the gradient of H^t, computed as follows:

$$\frac{\partial \mathscr{L}_{total}}{\partial H^t} = \sum_{i,j,k}^{N} \left(\sin\left(H_i^l, H_j^t\right) + \sin\left(H_i^t, H_j^l\right) - \left(\sin\left(H_i^l, H_k^t\right) + \sin\left(H_i^t, H_k^l\right)\right)\right)$$
$$+2\varphi \sum_{i,j}^{N} \left(S_{ij} - \cos\left(H_i^t, H_j^t\right)\right)\sin\left(H_i^t, H_j^t\right) + 2\eta \sum_{i}^{N} H_i^t 1 - 2\gamma \sum_{i}^{N} \left(B_i - H_i^t\right). \tag{17}$$

Optimizing B. When θ^l, θ^v, and θ^t are fixed, the hash code B can be obtained via the following formula:

$$B = sign\left(H^l + H^v + H^t\right). \tag{18}$$

3.8. Computational complexity analysis

The computational complexity of our method mainly involves the semantic decomposition loss, cross-modal triplet loss, intra-modal multilevel semantics loss, quantization loss, and balance constraints. Of them, 1) the semantic decomposition loss primarily involves the joint feature learning subnetwork, and its computational complexity depends on the feature dimension d and the number of training samples N. As a result, the complexity of the semantic decomposition loss is $O(3dN)$, where $3d$ represents the total dimension of features processed from different modalities. 2) The computational complexity of cross-modal triplet loss mainly comes from the construction of triplets. It requires $O(N^2)$ for constructing all triplets, and $O(4dT)$ for computing the loss for all triplets, where T is the number of triplets. Thus, the overall complexity of the cross-modal triplet loss is $O(N^2 + 4dT)$. In addition, 3) the complexity of the intra-modal multilevel semantics loss involves constructing the similarity matrix and computing the loss, which require $O(N^2)$ and $O(3dN)$, respectively, such that the total complexity of intra-modal multi-level semantic loss is $O(N^2 + 3dN)$. Lastly, 4) the computational complexity of the quantization loss and balance constraints is $O(3dN)$, which is mainly determined by the feature dimension d. Therefore, by combining the complexities of these components, the total computational complexity of the proposed method is $O(2N^2 + 6dN + 4dT)$.

4. Experiments

In this section, we evaluate our proposed SDEH through comparative experiments on four commonly used multimodal datasets, namely, MIRFlickr-25 K [31], NUS-WIDE [32], IAPRTC-12 [33], and MS COCO2014 [34]. For the proposed SDEH, we utilize the PyTorch framework and employ SGD as the optimizer, with a maximum of 100 iterations. In addition, we uniformly set the initial learning rate of the three subnetworks to $10^{-1.5}$. For the image feature learning subnetwork, the learning rate gradually decreases with a step of 10^{-6} for every 100 epochs, and for the text and joint feature learning subnetworks the learning rate decreases by 40 % for every 15 epochs until it reaches 10^{-6}. The batch size is specified as 128, and the hyper-parameters are set as: $m = 0.3$, $\alpha = 0.2$, $\lambda = \beta = 0.5$, $\varphi = 1.4$, $\gamma = 0.1$, and $\eta = 0.0001$.

4.1. Datasets

The MIRFlickr-25 K dataset consists of 25,000 image-text pairs collected from Flickr, covering 24 categories. Each instance is associated with one or more of these categories. Following the widely used protocol of DCMH [7], we exclude unannotated samples from the dataset and select 20,015 samples with complete annotations for the experiments. Specifically, we randomly select 2000 samples as the query set, and the remaining samples as the retrieval set. From the retrieval set, we randomly select 10,000 samples as training set. In addition, each text-modality sample is encoded into a 1386-dimensional BOW feature vector.

The NUS-WIDE dataset is a large-scale cross-modal dataset with 269,648 image-text pairs, and each image-text pairs is associated with 1

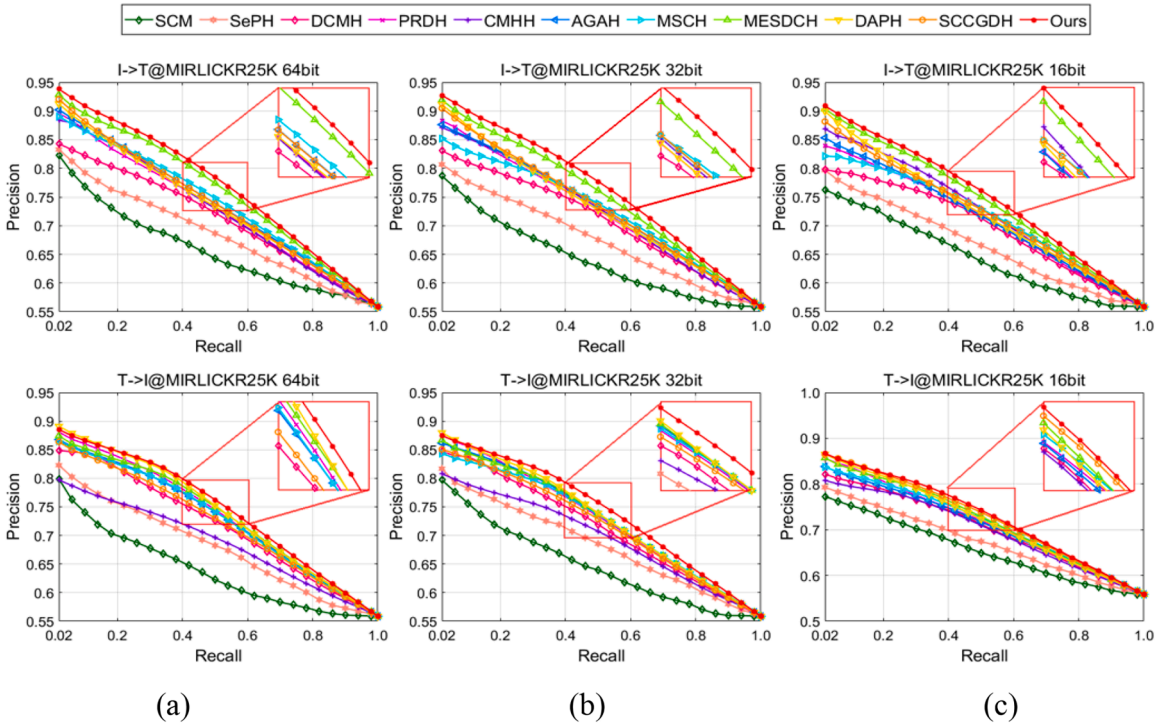


Fig. 6. ROC curves of different methods with hash codes of (a) 64, (b) 32, and (c) 16 bits on the MIRLICKR25 K dataset.

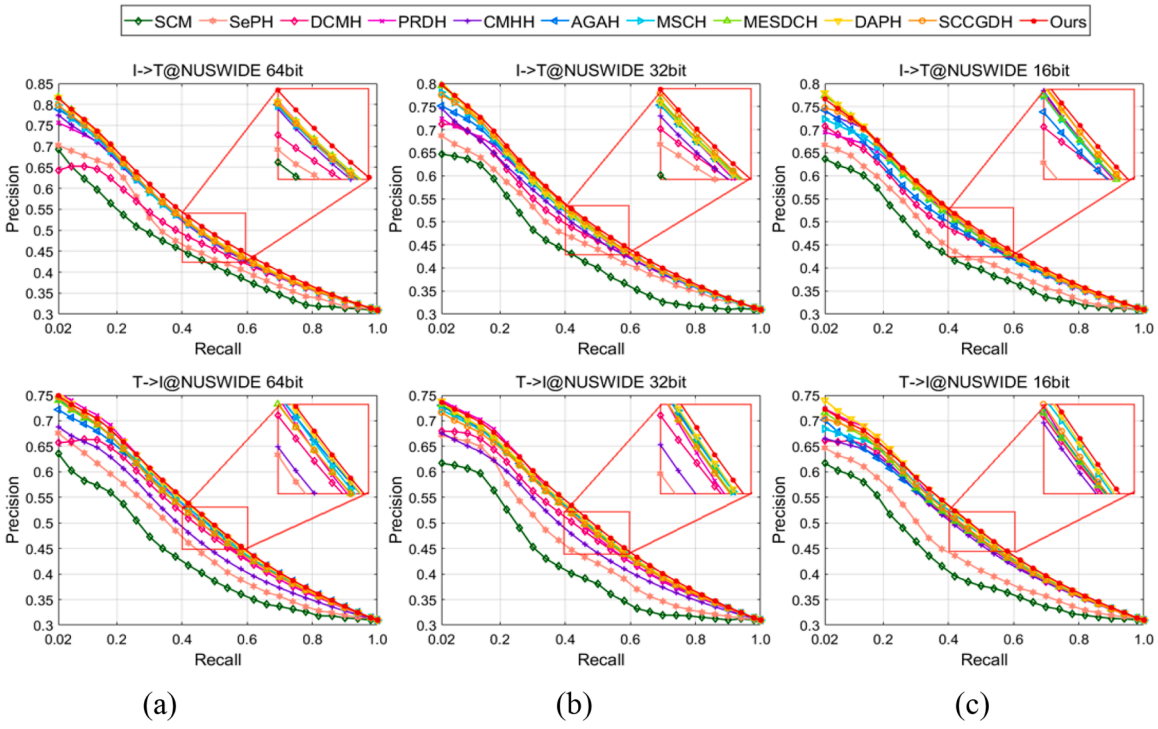


Fig. 7. ROC curves of different methods with hash codes of (a) 64, (b) 32, and (c) 16 bits on the NUSWIDE dataset.

to 81 categories. In our study, we specially select 190,421 image-text pairs from the 21 commonly used categories to conduct our experiments [7]. For the NUS-WIDE dataset, we randomly select 2100 samples as the query set with the remaining as the retrieval set, and further randomly select 10,000 retrieval samples from the training set. In the experiment, the text-modality samples are uniformly encoded into 1000-dimensional BOW feature vectors.

The IAPRTC-12 dataset contains 20,000 image-text pairs with 255 different categories. After removing the samples with no category labels, we employ 19,998 annotated image-text pairs for our experiments. Similar to the MIRFlickr-25 K, we use the randomly selected 2000 samples and the rest 17,998 samples as the query and retrieval sets, respectively, and further randomly select 10,000 samples from the retrieval set for training. In addition, each text sample is converted into a

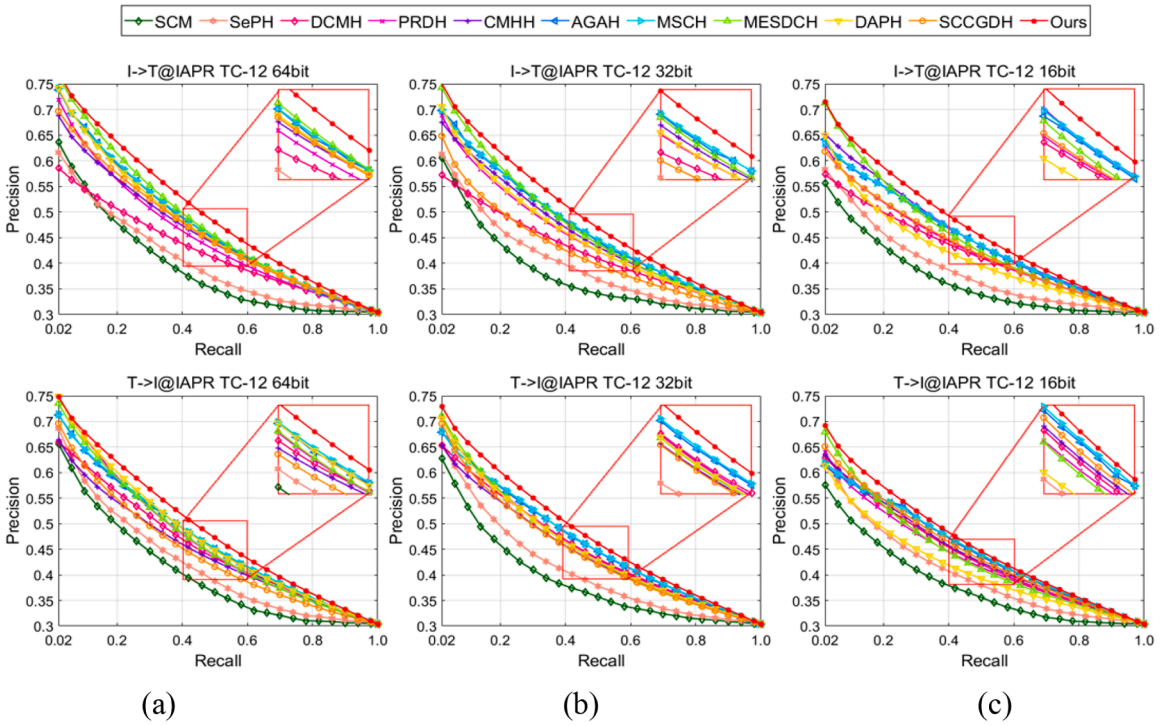


Fig. 8. ROC curves of different methods with hash codes of (a) 64, (b) 32, and (c) 16 bits on the IAPR TC-12 dataset.

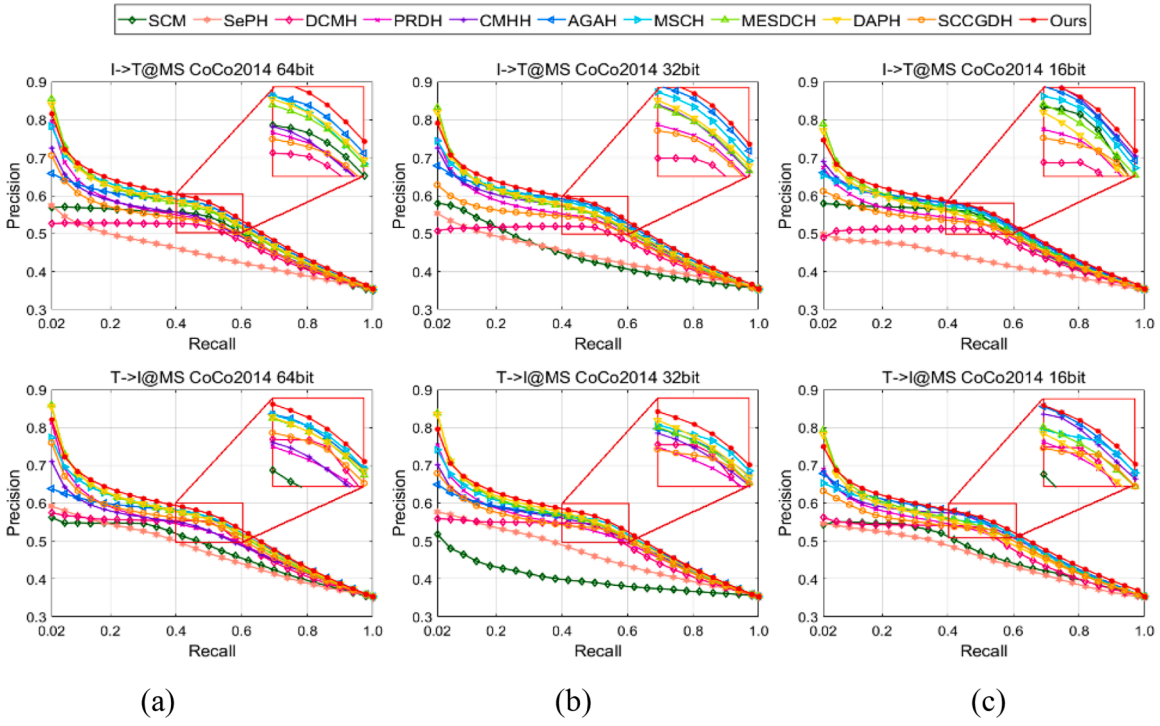


Fig. 9. ROC curves of different methods with hash codes of (a) 64, (b) 32, and (c) 16 bits on the MS COCO2014 dataset.

2885-dimensional BOW vector.

The MS COCO2014 dataset contains 122,218 image samples, each of which is annotated with at least one of 80 categories. In this dataset, each image is accompanied by an average of 5 descriptive captions, which are considered as the text modality to conduct our cross-modal retrieval experiments. For the MS COCO2014 dataset, we randomly select 5000 samples as the query samples, and use the remaining as retrieval samples, and we further randomly select 10,000 samples from the retrieval set for model training. Similarly, all text-modality data are encoded into 1251-dimensional BOW vectors.

Fig. 5 depicts representative examples randomly selected from the four datasets.

Table 2
MAP values acquired by different SDEH variants on the IAPR TC-12 dataset.

Methods	I->*T*			T->*I*		
	16bits	32bits	64bits	16bits	32bits	64bits
SDEH	**56.74**	**60.38**	**63.03**	**55.78**	**59.47**	**62.07**
SDEH—CSA	56.69	60.22	62.65	55.74	59.15	61.50
SDEH-SD	54.21	58.38	61.82	53.35	57.41	60.35
SDEH- RIS	56.05	60.01	62.26	55.14	59.10	61.63
SDEH- BS	55.93	60.16	62.40	55.59	58.98	61.34

Table 3
MAP values of the parallel channel and spatial attention blocks at different positions in ResNet34 on the IAPR TC-12.

Task	SDEH_Res4	SDEH_Res3	SDEH_Res2	SDEH_All
I->*T*	62.85	**63.03**	62.44	62.52
T->*I*	61.74	**62.07**	61.55	61.68

4.2. MAP results of cross-modal retrieval

To evaluate the proposed SDEH, we conduct two cross-modal retrieval tasks: using images for text retrieval (denoted as I→T) and using texts for image retrieval (denoted as T→I), and compare the proposed SDEH with the state-of-the-art methods, including two shallow hashing methods, i.e., SCM [35] and SePH [36], and eight deep hashing methods, including DCMH [7], PRDH [12], CMHH [9], AGAH [10], MSCH [23], MESDCH [14], SCCGDH [13], and DAPH [19]. All methods are implemented under the same environment with an Intel Core i5 13600KF processor, 32 GB of RAM, a 1 TB SSD, and ResNet-34 [29] pretrained on ImageNet [37] as the backbone for image feature learning to ensure the fairness of the experiment. Additionally, we use the widely adopted mean average precision (MAP) [8] as the cross-modal retrieval metric to evaluate the performance of all the methods.

Table 1 presents the MAP results of different methods with various bit lengths of 16, 32, and 64 bits on the four datasets, which show that our proposed SDEH outperforms most baseline methods by achieving a significantly higher MAP. For example, our proposed SDEH achieves substantial improvements over similarity-supervision-based learning methods such as DCMH, PRDH, CMHH, AGAH, and MSCH. This is because most existing similarity-supervision-based methods (such as DCMH, PRDH, CMHH, and AGAH) often utilize a binary similarity matrix as supervision information, which makes it challenging to comprehensively explore the complex relationships in multi-labels. In contrast, our proposed SDEH deeply explores the abundant multilevel semantic relationships within multi-labels, and further introduces a self-supervision mechanism for maximizing the semantic relevance between modalities, such that higher cross-modal retrieval performances can be obtained.

Additionally, our proposed SDEH is superior to other self-supervised methods (such as MESDCH, DAPH, and SCCGDH), with a minimal MAP improvement of 0.64. The possible reason is that the MESDCH and SCCGDH learn semantic information with only label information, which usually cannot effectively capture the correlation between images and texts with limited label information. In addition, while DAPH can utilize semantic information from images and texts, it neglects the possible negative impact of modality-specific private information. In contrast, our method further eliminates private information from images and texts via the semantic decomposition module.

4.3. Precision–recall curves of cross-modal retrieval

Furthermore, we show the ROC curves, i.e., precision versus recall rates, of the proposed SDEH compared with the existing state-of-the-art methods on four datasets in Figs. 6–9. Our proposed SDEH consistently outperforms all the baseline methods by achieving higher precisions against the same recall rates. This superiority is attributed to our proposed method embedding parallel channel and spatial attention blocks into the image and text feature learning subnetworks, which enhances the local details of the image and text feature descriptors. Furthermore, the semantic decomposition module of our proposed SDEH can enrich the semantic information of the multi-label vectors by decomposing the shared semantic information of the images and texts such that more compensation for the heterogeneous differences across different modalities can be obtained, and higher precision can be achieved.

4.4. Ablation studies

Impact of the individual subnetworks. To evaluate the effectiveness of the various components involved in the four subnetworks, we form four different variants of the proposed SDEH by modifying each of the subnetworks, including 1) the SDEH variant with no parallel channel and spatial attention block (referred to as SDEH—CSA); 2) the SDEH variant with no semantic decomposition module as the SSAH [15] (referred to as SDEH-SD); 3) the SDEH variant with only multi-label related information for supervised similarity matrix construction (referred to as SDEH-RIS); and 4) the SDEH variant with only the binary similarity matrix as intra-modal supervision information (referred to as SDEH-BS). Table 2 shows the MAP values acquired by the proposed SDEH and its four variants on the IAPR TC-12 dataset. The comparative results suggest the following three observations. 1) After removing the parallel channel and spatial attention block, the SDEH—CSA results in a slightly lower MAP than does the SDEH because of the loss of local details during feature extraction via ResNet-34 only. 2) The performance of the SDEH-SD significantly decreases because the semantic information cannot be effectively captured from the multi-label via traditional feature learning subnetworks such as SSAH, whereas our proposed SDEH uses a semantic decomposition module to decompose shared semantic information from images and texts, enabling joint learning of the multilabel semantic information. 3) The SDEH consistently outperforms the SDEH-RIS and SDEH-BS variants. This is because the intra-modal multilevel semantic loss can comprehensively capture the latent semantic associations within multi-labels with both relevant and irrelevant information.

Impact of attention mechanism. To investigate the optimal placement of the parallel channel and spatial attention block in enhancing the local details of image feature descriptors within ResNet34, we embedded it behind the second, third, fourth, and all the residual blocks within ResNet34 (referred to as SDEH-Res2, SDEH-Res3, SDEH-Res4, and SDEH-All, respectively). Table 3 shows the MAP values of the parallel channel and spatial attention block at different positions within ResNet34, with a bit length of 64. Placing the parallel channel

Table 4
MAP results obtained by the proposed method with removing the parallel channel and spatial attention blocks.

Image-attention	Text-attention	I->*T*			T->*I*		
		16	32	64	16	32	64
×	×	56.69	60.22	62.65	55.74	59.15	61.50
×	√	56.71	60.35	62.74	55.66	59.38	61.96
√	×	56.16	60.36	62.90	55.48	59.42	61.66
√	√	**56.74**	**60.38**	**63.03**	**55.78**	**59.47**	**62.07**

Table 5
MAP values obtained by the proposed method with removing different loss constraints on the IAPR TC-12 dataset.

$\mathcal{L}_{inter}$	$\mathcal{L}_{intra}$	$\mathcal{L}_{quantization}$	$\mathcal{L}_{balance}$	I->*T*			T->*I*		
				16	32	64	16	32	64
×	×	×	×	41.82	39.86	38.47	45.64	45.27	44.68
√	×	×	×	47.04	51.41	52.62	47.54	51.34	53.27
√	√	×	×	53.17	56.76	59.61	53.61	55.88	59.60
√	√	√	×	56.56	60.56	62.67	55.64	59.46	61.98
√	√	√	√	**56.74**	**60.58**	**63.03**	**55.78**	**59.52**	**62.07**

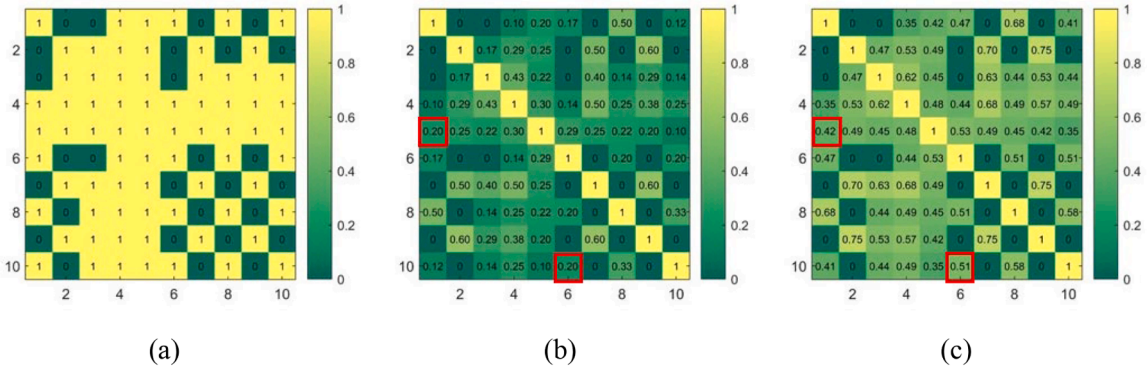


Fig. 10. The visualized distributions of different similarity matrices on the MIRFLICK25 K dataset, including (a) the original binary similarity, (b) the proportion-based similarity, and (d) multi-level semantic similarity calculated by our proposed method.

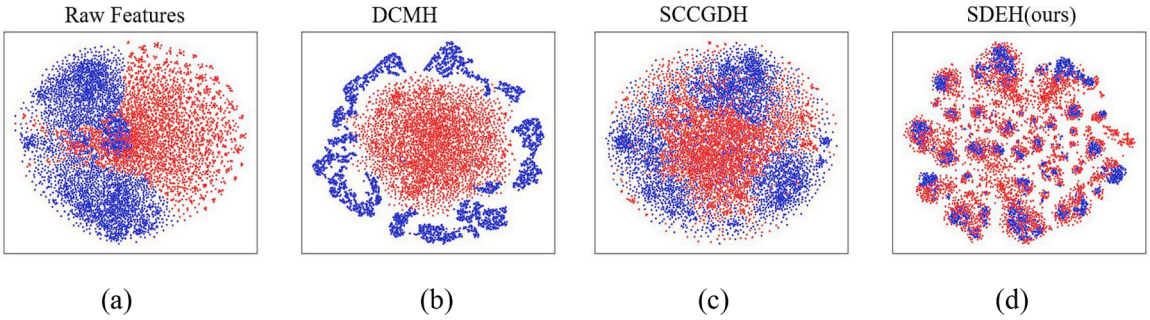


Fig. 11. Feature distributions obtained based on the (a) raw features, and the learned features by the (b) DCMH, (c) SCCGDH and (d) our proposed SDEH method.

and spatial attention block behind the third residual block in ResNet34 yields the best results; this is because the third residual block is located in the middle of ResNet34 and can capture mid-level features without excessive abstraction, retaining considerable local details.

In addition, to evaluate the impact of parallel channels and spatial attention blocks on image and text feature learning, we respectively remove the parallel channel and spatial attention blocks from the image and text feature learning subnetworks, and test the retrieval performance of the proposed method on the IAPR TC-12 dataset, as reported in Table 4. It is seen that the performance of the proposed method consistently outperforms its two variants with removing the attention blocks, showing the effectiveness of the parallel channel and spatial attention blocks for specific feature learning of both image and text modalities.

Impact of loss functions. To further evaluate the effectiveness of four loss functions of the proposed method, we remove each of these loss functions individually and test the MAP of our method as presented in Table 5. It can be seen that our proposed method (including the four loss functions) obviously outperforms the four variants with removing different loss functions from the proposed method, demonstrating the effectiveness of the loss functions for cross-modal feature learning.

4.5. Multilevel semantic relationships evaluation

To evaluate the effectiveness of our proposed method in preserving multi-level semantic relationships, we visualized the multi-level semantic similarity matrix that we proposed, as shown in Fig. 10. Fig. 10 (a) shows the binary similarity matrix with rigid similarity degree of samples, and Fig. 10(b) shows the proportion-based similarity matrix [24], which usually ignores the semantic differences between multi-labels. By contrast, Fig. 10(c) shows the multi-level semantic similarity matrix generated by our proposed method, in which the relevant information and semantic differences in the multi-labels can be comprehensively described with more accurate similarity values, such as the subtle semantic similarity values reflected in the red boxes.

4.6. Feature distribution consistency evaluation

We use t-SNE [38] to visualize the feature distributions generated by various hashing methods, including the ResNet34, DCMH [7], SCCGDH [13], and our proposed SDEH, on the MIRFLICKR25 K dataset, as shown in Fig. 11 (in which image and text features are marked in red and blue colors, respectively). It is seen that our proposed SDEH method obtains more compact clustering results when compared with other methods by

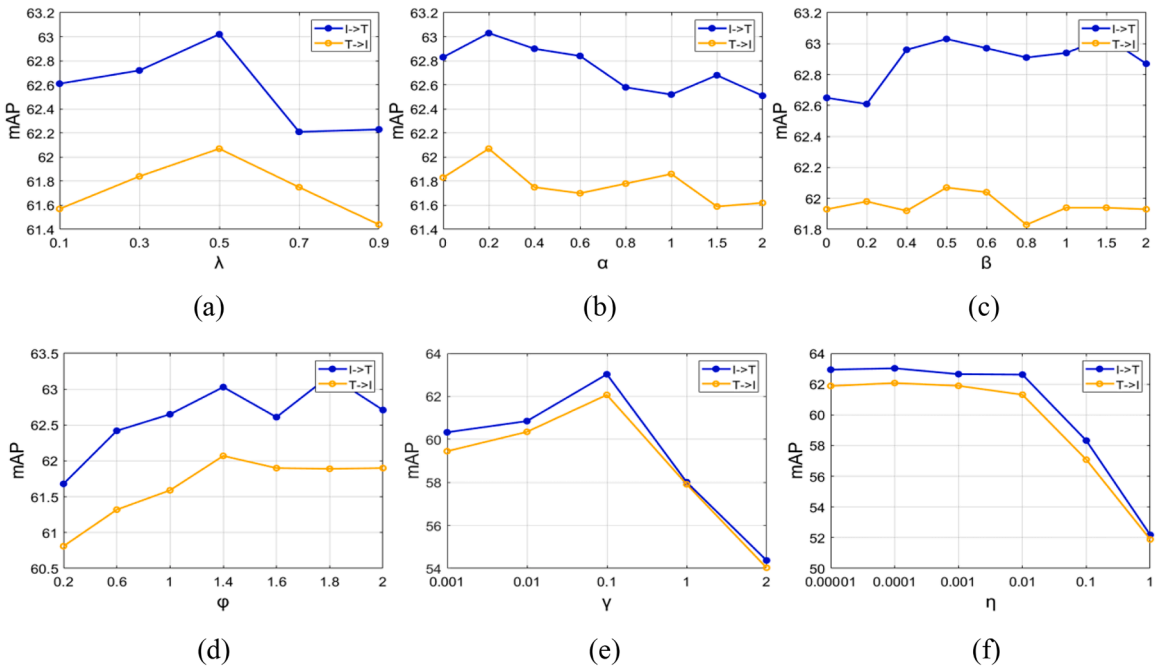


Fig. 12. MAP results of our proposed SDEH versus different values of (a) α, (b) β, (c) λ, (d) φ, (e) γ, and (f) η on the IAPR TC-12 dataset.

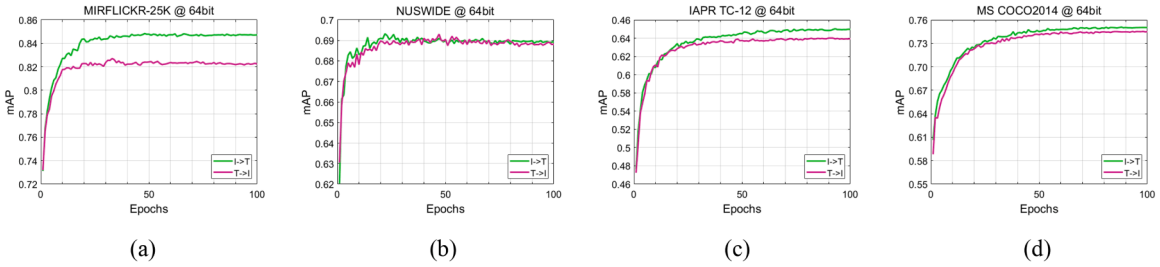


Fig. 13. MAP results of our proposed SDEH versus varying epochs on the (a) MIRFlickr-25 K, (b) NUS-WIDE, (c) IAPRTC-12, and (d) MS COCO2014 datasets.

Table 6
Comparison of SDEH with different methods in terms of the number of the floating-point operations and model parameters.

DCMH		MSCH		MESDCH		SDEH	
Flops (G)	Params (M)	Flops (G)	Params (M)	Flops (G)	Params (M)	Flops (G)	Params (M)
3.70	33.04	3.69	32.64	3.90	93.74	3.88	76.31

effectively narrowing the distribution gap between image-modality and text-modality features, ensuring the feature alignment across different modalities.

4.7. Parameter analysis

Our proposed SDEH consists of six hyper-parameters, i.e., α, β, λ, φ, γ, and η. To the best of our knowledge, there are no feasible methods for directly determining their optimal values. To achieve local optimal solutions, we first empirically fix the initial parameter ranges as: α: [0, 2], β: [0, 2], λ: [0.1, 0.9], φ: [0.2, 2], γ: [0.001, 2], and η: [0.0001, 1], and then conduct comparative experiments by calculating MAP of the proposed method with different parameter settings within the corresponding ranges. Fig. 12 depicts the MAP results of our proposed SDEH with a bit length of 64 versus different values of the parameters on the IAPR TC-12 dataset. The results indicate that the performance of the

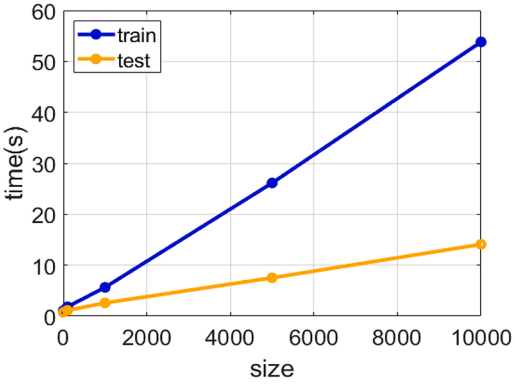


Fig. 14. The training and testing times of the proposed method at different scales of the samples.

SDEH slightly fluctuates with different parameter values of α, β, and φ, showing less sensitiveness of the SDEH to these parameters. In contrast, our proposed SDEH is more sensitive to λ, γ, and η. It is possible that γ controls the information loss of quantization, and a large value leads to more information loss. In addition, a large value of η possibly produces more redundant information during hash learning, such that the performance of the SDEH decreases. In this study, we empirically set $\alpha = 0.2$, $\beta = 0.5$, $\lambda = 0.5$, $\varphi = 1.4$, $\lambda = 0.5$, $\gamma = 0.1$, and $\eta = 0.0001$.

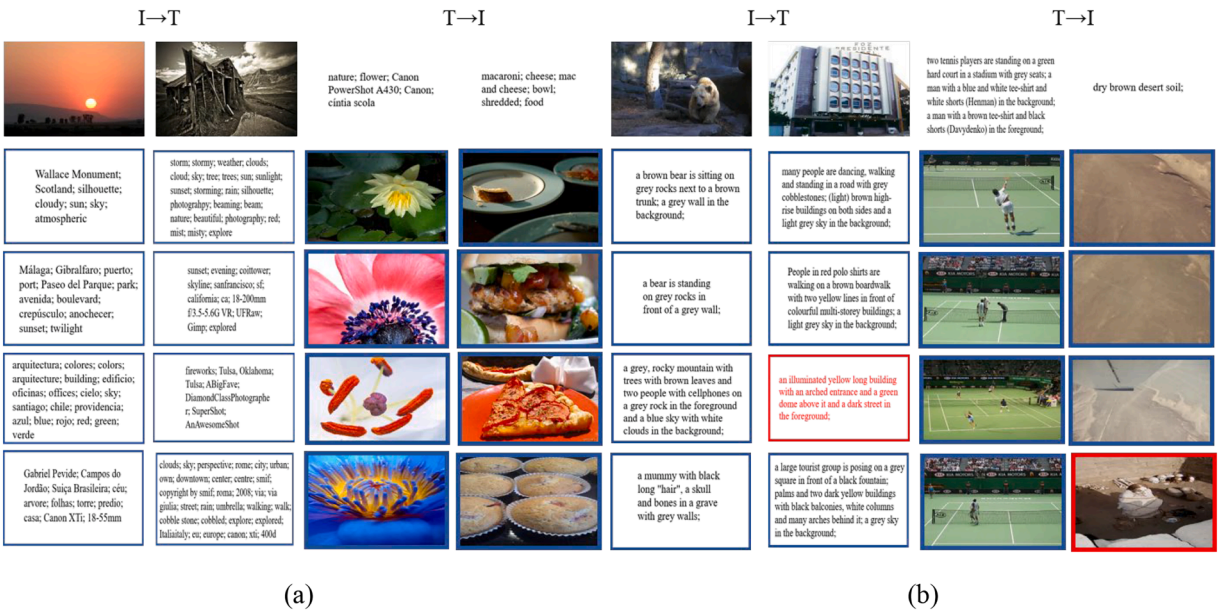


Fig. 15. The top-4 cross-modal retrieval results of the proposed SDEH on the (a) MIRFLICK25 K and (b) IAPR TC-12 datasets.

4.8. Convergence analysis

Fig. 13 depicts the retrieval results of our proposed SDEH versus varying epochs on four datasets with the hash code length of 64. We can observe that the MAP of our proposed SDEH increases rapidly within the first 20 epochs, and converges in about 30 to 40 epochs, demonstrating the rapid convergence speed of the proposed SDEH.

4.9. Complexity and scalability analysis

To further evaluate the complexity of our proposed method, we compare the number of floating-point operations (FLOPs) and model parameters (Params) of our proposed method in comparison with the state-of-the-art methods, including the DCMH [7], MSCH [23], and MESDCH [14], as reported in Table 6. It is seen that the FLOPs number of SDEH and MESDCH is slightly higher than that of the DCMH and MSCH. This is because both SDEH and MESDCH introduce a multi-label network, resulting in a higher model complexity. Considering the significant improvement of the retrieval performance and slightly more time taken, our proposed SDEH can make a good trade-off when the effectiveness and efficiency are both concerned.

Moreover, to verify the scalability of our method, we measure the training and testing time per iteration of the proposed method at different scales of datasets, as shown in Fig. 14. It is observed that the training and testing times are approximately linear to the scales of the datasets, showing the scalability of the proposed method in handling large-scale datasets. This is because our proposed model benefits from its streamlined and efficient batch processing strategies, ensuring the high efficiency in forward and backward propagation during model training.

4.10. Visualization of retrieval results

We show some Top-4 retrieval results of the proposed method on the MIRFLICKR25 K and IAPR TC-12 datasets in Fig. 15. It is seen that our proposed method usually obtain the correct results on both image-to-text and text-to-image retrieval tasks. Notably, there are still some failure retrieval cases, as shown in the red boxes of Fig. 14. This is possibly because our proposed model emphasizes more on certain common visual features learning, such as the shape and structure features, while the informative color and background features are ignored during feature learning, resulting in such a failure case.

5. Conclusion

In this paper, we propose a new deep cross-modal hashing method called semantic decomposition and enhancement hashing (SDEH). We first extract the modality-specific features from different modalities via two attention-based learning subnetworks and capture the latent multi-label semantic features with the help of shared information from the two different modalities, effectively reducing the heterogeneous gap between different modalities. Moreover, we design a group of multi-label semantic losses to guide our proposed method to comprehensively preserve the multilevel semantic relationships within and between modalities, making the final cross-modal hash code informative. The experimental results on four commonly used datasets demonstrate the promising effectiveness of our proposed SDEH. It is noted that our proposed method shows some limitations in retrieval scenarios with complex scenes, which suggest a future work to exploit the modality-invariant features with complex background changes of wide scenes for robust cross-modal retrieval.

CRediT authorship contribution statement

Lunke Fei: Writing – review & editing, Formal analysis. **Zhihao He:** Writing – original draft. **Wai Keung Wong:** Writing – review & editing, Supervision, Investigation. **Qi Zhu:** Visualization, Validation. **Shuping Zhao:** Software, Resources, Methodology. **Jie Wen:** Writing – review & editing, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFF0905602, in part by the National Natural Science Foundation of China under Grant 62176066, and in part by the Natural Science Foundation of Guangdong Province under Grant 2023A1515012717.

Data availability

The authors do not have permission to share data.

References

[1] P. Kaur, H. Singh Pannu, A.K. Malhi, Comparative analysis on cross-modal information retrieval: a review, Comput. Sci. Rev. 39 (2021) 100336.
[2] G. Ding, Y. Guo, J. Zhou, Y. Gao, Large-scale cross-modality search via collective matrix factorization hashing, IEEE Trans. Image Process. 25 (11) (2016) 5427–5440.
[3] Di Wang, Q. Wang, L. He, X. Gao, Y. Tian, Joint and individual matrix factorization hashing for large-scale cross-modal retrieval, Pattern Recognit. 107 (2020) 107479.
[4] J. Huang, P. Kang, X. Fang, Na Han, S. Xie, H. Gao, Efficient discriminative hashing for cross-modal retrieval, IEEE Trans. Syst. Man Cybern. Syst. (2024).
[5] J. Qin, L. Fei, Z. Zhang, J. Wen, Y. Xu, D. Zhang, Joint specifics and consistency hash learning for large-scale cross-modal retrieval, IEEE Trans. Image Process. 31 (2022) 5343–5358.
[6] F. Yang, Y. Liu, X. Ding, F. Ma, J. Cao, Asymmetric cross–modal hashing with high–level semantic similarity, Pattern Recognit. 130 (2022) 108823.
[7] Q.-Y. Jiang, Wu-J Li, Deep cross-modal hashing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3232–3240.
[8] L. Zhu, L. Cai, J. Song, X. Zhu, C. Zhang, S. Zhang, MSSPQ: multiple semantic structure-preserving quantization for cross-modal retrieval, in: Proceedings of the International Conference on Multimedia Retrieval, 2022, pp. 631–638.
[9] Y. Cao, B. Liu, M. Long, J. Wang, Cross-modal hamming hashing, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 202–218.
[10] W. Gu, X. Gu, J. Gu, Bo Li, Z. Xiong, W. Wang, Adversary guided asymmetric hashing for cross-modal retrieval, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, 2019, pp. 159–167.
[11] Ge Song, H. Su, K. Huang, F. Song, M. Yang, Deep self-enhancement hashing for robust multi-label cross-modal retrieval, Pattern Recognit. 147 (2024) 110079.
[12] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, X. Gao, Pairwise relationship guided deep hashing for cross-modal retrieval, in: Proceedings of the AAAI Conference on Artificial Intelligence 31, 2017.
[13] Z. Shu, Y. Bai, D. Zhang, J. Yu, Z. Yu, X.-J. Wu, Specific class center guided deep hashing for cross-modal retrieval, Inf. Sci. 609 (2022) 304–318.
[14] X. Zou, S. Wu, E.M. Bakker, X. Wang, Multi-label enhancement based self-supervised deep cross-modal hashing, Neurocomputing. 467 (2022) 138–162.
[15] C. Li, C. Deng, N. Li, W. Liu, X. Gao, 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.
[16] X. Ma, T. Zhang, C. Xu, Multi-level correlation adversarial hashing for cross-modal retrieval, IEEE Trans. Multimed. 22 (12) (2020) 3101–3114.
[17] X. Zou, S. Wu, N. Zhang, E.M. Bakker, Multi-label modality enhanced attention based self-supervised deep cross-modal hashing, Knowl. Based Syst. 239 (2022) 107927.
[18] Y. Duan, N. Chen, P. Zhang, N. Kumar, L. Chang, Wu Wen, MS2GAH: multi-label semantic supervised graph attention hashing for robust cross-modal retrieval, Pattern Recognit. 128 (2022) 108676.
[19] R.-C. Tu, X.-L. Mao, W. Ji, W. Wei, H. Huang, Data-aware proxy hashing for cross-modal retrieval, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 686–696.
[20] Y. Huo, Q. Qin, J. Dai, L. Wang, W. Zhang, L. Huang, C. Wang, Deep semantic-aware proxy hashing for multi-label cross-modal retrieval, IEEE Trans. Circu. Syst. Video Technol. (2023).
[21] X. Liu, G. Yu, C. Domeniconi, J. Wang, Y. Ren, M. Guo, Ranking-based deep cross-modal hashing, in: Proceedings of the AAAI Conference on Artificial Intelligence 33, 2019, pp. 4400–4407.
[22] C. Sun, H. Latapie, G. Liu, 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, 2022, pp. 4941–4949.
[23] J. Zhan, S. Liu, Z. Mo, Y. Zhu, Multi-similarity semantic correctional hashing for cross modal retrieval, in: 2020 IEEE International Conference on Multimedia and Expo (ICME), 2020, pp. 1–6.
[24] X. Zou, X. Wang, E.M. Bakker, S. Wu, Multi-label semantics preserving based deep cross-modal hashing, Signal Process. Image Commun. 93 (2021) 116131.
[25] S. Woo, J. Park, J.-Y. Lee, So Kweon, Cbam: convolutional block attention module, in: Proceedings of the European conference on Computer Vision (ECCV), 2018, pp. 3–19.
[26] H.-L. Yao, Yu-W Zhan, Z.-D. Chen, X. Luo, X.S. Xu, Teach: attention-aware deep cross-modal hashing, in: Proceedings of the 2021 International Conference on Multimedia Retrieval, 2021, pp. 376–384.
[27] Xi Zhang, H. Lai, J. Feng, Attention-aware deep adversarial hashing for cross-modal retrieval, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 591–606.
[28] Y. Zhang, W. Ou, Y. Shi, J. Deng, X. You, A. Wang, Deep medical cross-modal attention hashing, World Wide Web. 25 (4) (2022) 1519–1536.
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[30] H. Robbins, S. Monro, A stochastic approximation method, Ann. Mathemat. Stat. (1951) 400–407.
[31] M.J. Huiskes, M.S. Lew, The MIR FLICKR retrieval evaluation, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, 2008, pp. 39–43.
[32] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: Proceedings of the ACM International Conference on Image and Video Retrieval, 2009, pp. 1–9.
[33] H.J. Escalante, C.A. Hernández, J.A. Gonzalez, A. López-López, M. Montes, E. F. Morales, L. Enrique Sucar, L. Villasenor, M. Grubinger, The segmented and annotated IAPR TC-12 benchmark, Comput. Vis. Image Understand. 114 (4) (2010) 419–428.
[34] T-Yi Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. Lawrence Zitnick, Microsoft coco: common objects in context, in: European Conference on Computer Vision, 2014, pp. 740–755.
[35] D. Zhang, Wu-J Li, Large-scale supervised multimodal hashing with semantic correlation maximization, in: Proceedings of the AAAI Conference on Artificial Intelligence 28, 2014.
[36] Z. Lin, G. Ding, M. Hu, J. Wang, Semantics-preserving hashing for cross-view retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3864–3872.
[37] J. Deng, W. Dong, R. Socher, Li-J Li, K. Li, Li Fei-Fei, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[38] L.V. Maaten, G. Hinton, Visualizing data using T-SNE, J. Mach. Learn. Res. 9 (11) (2008).