Neurocomputing 511 (2022) 366–379

Contents lists available at ScienceDirect

Neurocomputing

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

Discrete asymmetric zero-shot hashing with application to cross-modal retrieval

Zhenqiu Shu [a,*], Kailing Yong [a], Jun Yu [b], Shengxiang Gao [a], Cunli Mao [a], Zhengtao Yu [a]

[a] School of Faculty of Information Engineering and Automation, Kunming University of Science and Technology, China
[b] College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, China

ARTICLE INFO

Article history:
Received 6 June 2022
Revised 24 August 2022
Accepted 4 September 2022
Available online 13 September 2022
Communicated by Zidong Wang

Keyword:
Zero-shot hashing
Asymmetric
Cross-modal retrieval
Class attributes
Pairwise similarity

ABSTRACT

In recent years, cross-modal retrieval technology has attracted extensive attention with the massive growth of multimedia data. However, most cross-modal hashing methods mainly focus on exploring the retrieval of seen classes, while ignoring the retrieval of unseen classes. Therefore, traditional cross-modal hashing methods cannot achieve satisfactory performances in zero-shot retrieval. To mitigate this challenge, in this paper, we propose a novel zero-shot cross-modal retrieval method called discrete asymmetric zero-shot hashing(DAZSH), which fully exploits the supervised knowledge of multimodal data. Specifically, it integrates pairwise similarity, class attributes and semantic labels to guide zero-shot hashing learning. Moreover, our proposed DAZSH method combines the data features with the class attributes to obtain a semantic category representation for each category. Therefore, the relationships between seen and unseen classes can be effectively captured by learning a category representation vector for each instance. Therefore, the supervised knowledge can be transferred from the seen classes to the unseen classes. In addition, we develop an efficient discrete optimization strategy to solve the proposed model. Massive experiments on three benchmark datasets show that our proposed approach has achieved promising results in cross-modal retrieval tasks. The source code of this paper can be obtained from https://github.com/szq0816/DAZSH.

1. Introduction

Over the past decade, cross-modal retrieval tasks have become a great challenge owing to the exponential growth of multimodal data. In general, multimodal data are usually dependent and have essential connections in most cases. Therefore, it is fundamental to learn the correlation information between multimodalities in pattern recognition and machine learning, which is referred to as the heterogeneity gap. To resolve this discrepancy, traditional approaches try to project multimedia data into a common semantic space and then perform the retrieval task. However, real-value projections require more storage space and expensive computational costs due to the increase in multimedia data. This has become a significant obstacle in cross-modal retrieval applications. Therefore, hashing is an effective way to perform the retrieval task in large-scale datasets due to its low storage and high computational efficiency. It aims to project original samples into compact binary codes, which preserves their similarity in Hamming space. Recently, researchers have made many efforts to bridge the heterogeneity gap between multiple modalities and have achieved promising performances in many real applications [1–4].

To the best of our knowledge, most existing cross-modal hashing retrieval methods are studied in the seen classes dataset [5–7]. However, with the explosive growth of multimedia data, some new concepts, such as unseen classes, have emerged in the past few years. Retraining the existing cross-modal hashing model after collecting new concept data is high cost and requires much storage space. Therefore, it is necessary to adopt a new cross-modal hashing model to deal with training data containing new concepts. Therefore, zero-shot learning aims to identify previously unseen data categories. Specifically, the trained classifier not only identifies the existing data categories in the training set, but also distinguishes the data from the unseen categories [8].

In the past several years, many zero-shot learning approaches have been applied for cross-modal retrieval [9–11]. Yang et al. [9] realized the potential semantic transfer by projecting the labels into the word embedding space. Shi et al. [10] proposed a zero-shot hashing method based on the asymmetric ratio similarity matrix, which can significantly improve the ability of knowledge transfer from seen classes to unseen classes. Transductive zero-shot hash-

* Corresponding author.
E-mail address: shuzhenqiu@163.com (Z. Shu).

https://doi.org/10.1016/j.neucom.2022.09.037
0925-2312/

ing (T-MLZSH) [11] was proposed as a multilabel image retrieval model based on zero-shot learning. In this model, the labels of unseen classes are predicted by the instance-concept coherence ranking. Nevertheless, all the abovementioned works were applied to single-modality retrieval tasks, and there are still few efforts on unseen cross-modality retrieval tasks. With the continuous emergence of new concepts, existing cross-modal retrieval methods have the following limitations. (1) They only consider data from the seen categories and ignore the unseen cases. Therefore, these models are unsuitable for cross-modal retrieval with mixed unseen data. (2) Most of them neglect the class attribute information in hash code learning and thus are inconducive to knowledge transfer from seen classes to unseen classes. (3) Existing zero-shot hashing approaches fail to consider the pairwise similarity, class labels and class attributes to train models at the same time.

To address the above challenges, in this work, a novel zero-shot hashing method, called discrete asymmetric zero-shot hashing (DAZSH), is proposed for cross-modal retrieval. It integrates pairwise similarity, semantic labels and category attributes into a framework to fully explore semantic information. Specifically, we combine the data features with category attributes to obtain the category representation vector of each instance. In addition, the relationship between seen classes and unseen classes can be better captured, and the supervision knowledge can be transferred from the seen classes to the unseen classes. Fig. 1 shows the framework of our proposed DAZSH method in zero-shot cross-modal retrieval. The experimental results on three datasets show that our proposed DAZSH method can achieve better retrieval performances in dealing with unseen data.

The contributions of this work are as follows:

(1) We propose a unified discrete asymmetric zero-shot framework for learning hash codes. It combines data features with class attributes to learn an attribute space for each modality. Therefore, we explore the relationship between the seen classes and the unseen classes, which can transfer the supervision information from the seen classes to the unseen classes. In addition, our proposed approach aims at embedding the labels into the attribute space to improve retrieval accuracy. Therefore, our proposed model can generate more discriminative hash codes than traditional hashing methods.

(2) To maintain the characteristics of each modality, we generate different hash codes for different modalities using the asymmetric similarity strategy. Furthermore, we employ the maximum likelihood estimation algorithm to explore the pairwise similarity of multimodality data. To the best of our knowledge, this study is the first to utilize the pairwise similarity of different modalities using the maximum likelihood estimation algorithm in the cross-modal zero-shot learning field. Compared with traditional cross-modal zero-shot learning methods, our proposed method can effectively alleviate the heterogeneity gap between different modalities and more closely connect the seen and unseen classes, simultaneously.

(3) We develop a discrete optimization scheme to solve our proposed model, and then give its complexity analysis. Comprehensive experimental results on three benchmark datasets have shown the superiority of our proposed DAZSH method in different retrieval tasks.

The remainder of this paper is organized as follows: Section 2 introduces the previous work on cross-modal retrieval. Section 3 details our approach and its optimization scheme. Section 4 gives the experimental results and their analysis. Section 5 draws the conclusion of this work.

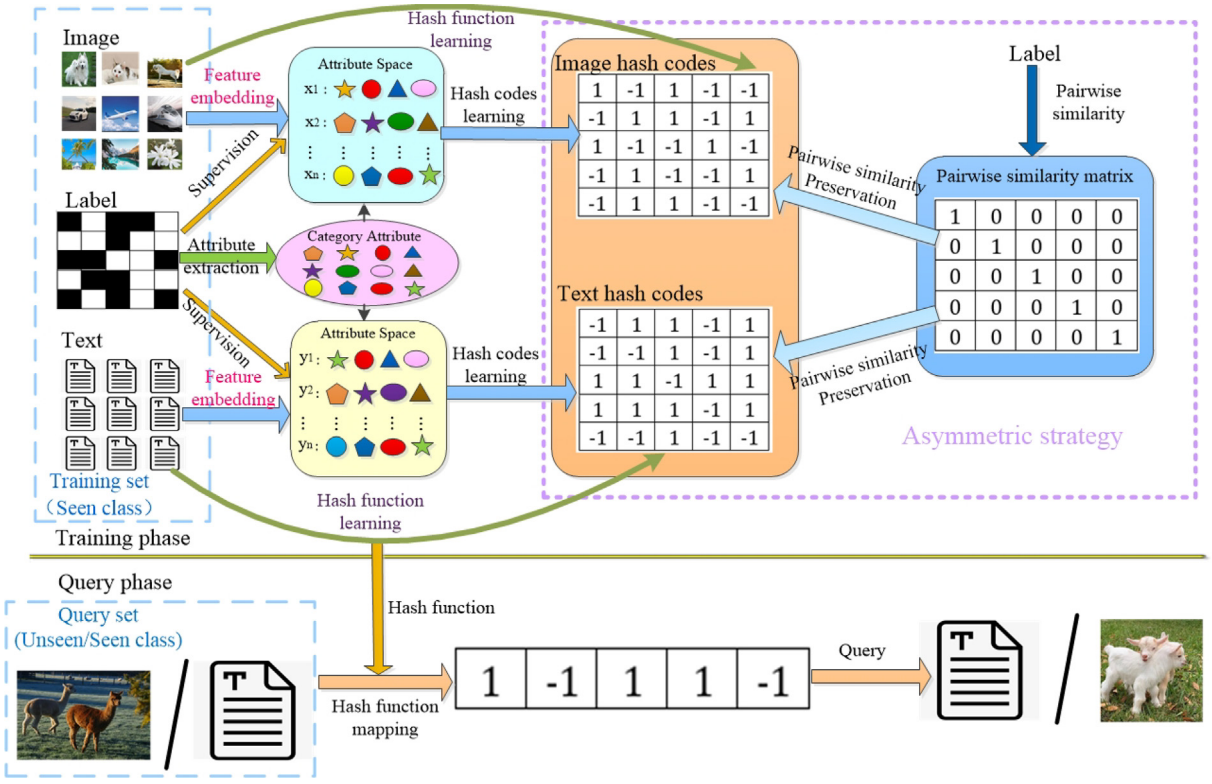

Fig. 1. The framework of our proposed DAZSH approach.

2. Related works

In this section, we give a preliminary introduction to works on traditional cross-modal retrieval and zero-shot cross-modal retrieval.

2.1. Cross-modal retrieval

Cross-modal retrieval uses one-modality data as a query to retrieve similar data in another modality. Due to the "heterogeneous gap" between different modalities, cross-modal retrieval becomes a challenging task in large-scale datasets. To solve this issue, the hashing-based cross-modal methods try to project multimodal data into Hamming space and then perform retrieval operations. Existing hashing methods are mainly divided into unsupervised methods and supervised methods.

In the real world, most multimodal data are unlabelled, so they require substantial labour and time for labelling, which is impractical. Therefore, unsupervised hashing approaches have attracted extensive attention in cross-modal retrieval [12–16]. In reference [12], the common representation space is obtained by using collaborative matrix factorization, and then the unified hash codes can be learned. Semantic topic multimodal hashing (STMH) [13] adopts robust matrix factorization to generate hash codes. Latent semantic sparse hashing (LSSH) [14] was proposed to learn hash codes by integrating matrix factorization and sparse coding. However, the aforementioned methods tend to keep the shared property, while ignoring the modal specific property. Therefore, the performances of the algorithms in cross-modal retrieval tasks are affected. To solve this issue, Yao et al. [16] proposed the discrete robust matrix factorization hashing (DRMFH) method to explore the shared and modal-specific properties of multimodal data and learn the discrete hash codes. However, the ahove methods are unsupervised hashing methods, which cannot fully take advantage of the supervised information to further improve the retrieval performance.

Unlike unsupervised methods, supervised hashing methods try to obtain more semantic relevance from supervised information to improve retrieval accuracy [17–22]. Semantic preserving hashing algorithm (SePH) [17] learns hash codes by minimizing the Kullback–Leibler (KL) divergence of a probability distribution. Fast discrete cross-modal hashing (FDCH) [18] regresses the semantic labels of training examples to the corresponding hash codes with a drift, and thus generates stable hash codes to improve retrieval performance. Label consistent matrix factorization hashing (LCMFH) [19] aims to constrain matrix factorization using label information. Thus, its hash codes can embed more semantic information of multimodality data. Label consistent flexible matrix factorization hashing (LFMH)[20] can jointly learn modality-specific latent semantic spaces with similar semantics through flexible matrix factorization. Two-stage discrete cross-media hashing (WATCH) [21] integrates smooth matrix factorization and label relaxation into the model. Unlike the previous methods, it adopts a two-stage strategy to enhance the flexibility of the model. In recent years, the pairwise similarity of cross-modal data, as important prior information, has been widely used in cross-modal retrieval. Li et al. [22] proposed learning the hash codes through the asymmetric similarity strategy. Meanwhile, it employs semantic label correlation learning to embed more semantic information in hash codes. However, most existing cross-modal retrieval methods mainly focus on seen data retrieval. Therefore, we cannot achieve excellent performances in the retrieval tasks of unseen classes.

2.2. Zero-shot cross-modal retrieval

Zero-shot learning is a powerful learning tool in which the class covered by the training instance is disjointed from the class we seek to classify [23]. It transfers the knowledge from seen classes to unseen classes, thus realizing the classification of unseen classes. The zero-shot cross-modal retrieval method aims to solve the retrieval challenges of unseen classes in cross-modal data by zero-shot learning. Liu et al. [24] proposed integrating similarity preservation, class attribute space learning and hash function learning to solve unseen classes problem. Zhong et al. [25] proposed learning hash codes by connecting features, hash codes and class attributes at the same time. In addition, the local structure information of each modality is embedded into the Hamming space. In the abovementioned methods, the pairwise similarity of multimodal data is completely ignored in zero-shot hash learning. To solve this issue, cross-modal hashing with orthogonal projection (CHOP) [26] was proposed to map the cross-modal features and the class attributes to different Hamming spaces. In addition, orthogonal constraints are applied to learn the discriminant hash codes of each modality. Furthermore, Shi et al.[10] proposed an asymmetric ratio similarity matrix-based zero-shot hashing. It avoids excessive learning of the seen classes and thus enhances the knowledge transferring ability from the seen to the unseen class.

Over the past decade, many studies have paid attention to the powerful representation ability of deep learning. At present, deep learning is applied to zero-shot cross-modal retrieval, which improves the knowledge transfer ability. Attribute-guided network for cross-modal zero-shot hashing (AgNet) aligns different modal data into a semantically rich attribute space, which bridges the gap caused by modal heterogeneity and zero-shot setting [27]. In reference [28], correlated features synthesis and alignment (CFSA) jointly maps the synthesized multimodal features and the real multimodal features to a common semantic space by a distributed alignment scheme, which transfers the knowledge to unseen classes. Furthermore, generative adversarial networks (GANs) have shown powerful data distribution modeling capabilities through adversarial learning and are widely used in zero-shot cross-modal retrieval. Xu et al. proposed the ternary adversarial networks with self-supervision (TANSS) [29] model, which promotes the semantic relevance between different modalities by constructing three parallel subnetworks in adversarial learning. Reference [30] designed an improved GAN structure, namely cWGAN, which can simultaneously synthesize related multimodal features with the guidance of class embedding. This method is beneficial for generating meaningful synthetic features and learning an efficient common embedding space. These methods can achieve encouraging results in cross-modal retrieval applications. However, these methods are usually limited by the number of training samples and the computing resources.

3. Discrete asymmetric zero-shot hashing (DAZSH)

This section introduces the proposed DAZSH model in detail.

3.1. Notations

In this paper, we take image modality and text modality as an example. Given a set of seen multimodal data, $O_s = \{o_i\}_{i=1}^{n_s}$, where $o_i = (x_i, y_i)$ is a pair of multimodal data, and x_i and y_i denote the feature vectors of the i-th instance of two modalities, respectively. $X = \{x_1, x_2, \ldots, x_{n_s}\} \in R^{n_s \times d_x}$ and $Y = \{y_1, y_2, \ldots, y_{n_s}\} \in R^{n_s \times d_y}$ are the feature matrices of the two modalities, respectively, where $n_s$de-

notes the number of seen class samples. d_x and d_y are the dimensions of the image samples and text samples, respectively, and $d_x \neq d_y$. Besides, we employ $L_s = \{l_1, l_2, \ldots l_{n_s}\} \in R^{n_s \times c}$to represent the label matrix, where c is the total number of categories. If the i-th sample belongs to class j, then $l_{ij} = 1$; Otherwise, $l_{ij} = 0$. Let $A = \{a_1, a_2, \ldots, a_c\} \in R^{c \times d_A}$ be the class attribute corresponding to each class, where d_A represents the dimension of the attribute. There is also a set of unseen multimodal data, $O_u = \{o_j\}_{j=1}^{n_u}$, where $o_j = (x_j, y_j)$ is a pair of multimodal data of unseen classes and $n_u$denotes the number of unseen class samples. Analogously, let $L_u = \{l_1, l_2, \ldots l_{n_u}\} \in R^{n_u \times c}$denote the label matrix. In the zero-shot setting, L_s and L_u have no intersection, that is, $L_s \cap L_u = \varnothing$.

3.2. Proposed method

Our proposed DAZSH method is divided into two steps: training and retrieval. In the training step, data features are combined with class attributes to learn an attribute space for each modality. Then, the proposed DAZSH approach learns the hash codes of each modality from the pairwise similarity and the attribute space, and obtains the mapping matrix of each modality. In the second step, we project the query samples to generate the hash codes by the mapping matrix learned in the training step and then perform the retrieval task based on these hash codes. Fig. 1 depicts the framework of the proposed DAZSH approach.

1) Attribute space learning. First, we extract the word vectors for each class name as the attribute of each class. In this procedure, we adopt the GloVe method [31] to extract a 300-dimensional real-valued vector for each class, named the class attribute matrix A. Then, we aim to learn an attribute vector-guided embedding class space for cross-modal zero-shot hashing. To preserve the properties of each modality, our goal is to establish the attribute space for each modality. Specifically, the attribute space is constructed by combining multimodal features with class attributes to obtain a semantic class representation for each class in the instance. Since the class attributes contain the attribute information of the seen and unseen classes, the attribute space guided by the class attributes can realize the transfer of knowledge from the seen classes to the unseen classes through the attribute connection between the classes. Besides, the relationship between seen and unseen classes can be better captured by learning the category representation of each instance. In addition, to improve the semantic information of the attribute space, we embed the supervision information into the attribute space. Therefore, we obtain the attribute space by optimizing the following problem:

$$\begin{aligned} &\mathscr{L}_1 = \alpha_1 \|X - E_1 A\|_F^2 + \alpha_2 \|Y - E_2 A\|_F^2 \\ &s.t. \quad E_1 = Z_1 L_s, E_2 = Z_2 L_s, \end{aligned} \tag{1}$$

where E_1 and E_2 denote the attribute spaces corresponding to the image modality and text modality, respectively. α_1 and α_2 represent the nonnegative parameters. To consider the supervision information, we introduce auxiliary matrices Z_1 and Z_2 to impose the label constraints on E_1 and E_2, respectively.

After learning the discriminant attribute space, we aim to generate the hash codes of each modality in the attribute space. By introducing the projection matrix, the attribute representations of different modalities are mapped to their corresponding hash codes. Therefore, the loss function is expressed as follows:

$$\begin{aligned} &\mathscr{L}_2 = \beta_1 \|B - Z_1 L_s P_1\|_F^2 + \beta_2 \|V - Z_2 L_s P_2\|_F^2 \\ &s.t. \quad B, V \in \{-1, 1\}^{n_s \times k}, \end{aligned} \tag{2}$$

where B and V represent hash codes of the image modality and the text modality, respectively. In addition, P_1 and P_2 are the projection matrices of the two modalities, which are used to realize the projection of the category attribute space to the hash space. β_1 and β_2 denote the nonnegative parameters and k is the hash code length.

2) Discrete asymmetric pairwise similarity preserving. Inspired by the literatures [32–34], we generate more discriminating hash codes by adopting the following model:

$$\mathscr{L}(B) = \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} F\left(B_{i*} B_{j*}^T; S_{ij}\right), \tag{3}$$

where $B_{i*}B_{j*}^T$ is the symmetric binary inner product of two instances and F(.) is a certain loss function. S_{ij}denotes the similarity between the i-th instance and the j-th instance. Furthermore, the paired similarity matrix is constructed by the label matrix, whose element is defined as $S_{ij} \in \{0, 1\}^{n_s \times n_s}$. If X_{i*} and X_{j*} are similar, then $S_{ij} = 1$; otherwise, $S_{ij} = 0$.

Then we can learn the hash codes for the image modality and the text modality, and obtain the binary code pairs $\{B_{i*}, V_{j*}\}$. We define Θ_{ij}as: $\Theta_{ij} = \frac{\lambda}{k} B_{i*} V_{j*}^T$, where λ is a nonnegative hyperparameter and k denotes the hash code length. According to logistic function $C_{ij} = \frac{1}{1+e^{-\Theta_{ij}}}$, the likelihood of the cross-modal similarity can be defined as follows:

$$p(S|B, V) = \prod_{i,j=1}^{n_s} p(S_{ij}|B, V), \tag{4}$$

where $p(S_{ij}|B, V)$ is expressed as follows:

$$p(S_{ij}|B, V) = \begin{cases} C_{ij}, & if\ S_{ij} = 1, \\ 1 - C_{ij}, & otherwise. \end{cases} \tag{5}$$

Therefore, the log-likelihood of B and V can be derived as follows:

$$\mathscr{L}_3 = \log p(S|B, V) = \sum_{i,j=1}^{n_s} \left[S_{ij}\Theta_{ij} - \log\left(1 + e^{\Theta_{ij}}\right)\right] \tag{6}$$

The Hamming distances between similar points can be reduced by maximizing the likelihood function of Eq. (6). Hash codes B and V are generated for the image modality and text modality by solving Eq. (6). To the best of our knowledge, most existing cross-modal hashing methods employ continuous variables instead of discrete codes to solve the symmetry problem [35–37]. However, this optimization scheme may lead to a large quantization error. Therefore, we calculate the discrete asymmetric inner product of B and V instead of solving the discrete symmetric inner product directly.

3) Hash function learning. Our DAZSH approach includes two stages: hash function learning and hash codes learning. We adopt the above two steps to obtain the hash codes B and V of the two modalities and then learn the modality-specific hash function to deal with the out-of-sample problem. Generally, the hash functions are learned by minimizing the following least-squares regression problems:

$$\mathscr{L}_4 = \mu_1 \|B - XW_1\|_F^2 + \mu_2 \|V - YW_2\|_F^2, \tag{7}$$

where W_1 and W_2 denote the projection matrices of the image modality and the text modality, respectively. μ_1 and μ_2 denote the nonnegative parameters.

4) Overall objective function. The kernel trick can effectively deal with the linear non-separable problem[38,39]. In particular, $\phi(X) = [\phi(x_1),\ \phi(x_2), \ldots, \phi(x_n)]$ and $\phi(Y) = [\phi(y_1),\ \phi(y_2), \ldots, \phi(y_n)]$ are the kernel feature matrices of the two modalities. Here, $\phi(\cdot)$ denotes the RBF kernel function. Therefore, kernel features $\phi(x_i)$ and $\phi(y_i)$ are given as follows:

$$\begin{aligned}\phi(x_i) &= \left[exp\left(-\frac{\left\|x_i-\delta_1^{(1)}\right\|^2}{2\vartheta_{(1)}^2}\right),\ldots,exp\left(-\frac{\left\|x_i-\delta_q^{(1)}\right\|^2}{2\vartheta_{(1)}^2}\right)\right],\\ \phi(y_i) &= \left[exp\left(-\frac{\left\|y_i-\delta_1^{(2)}\right\|^2}{2\vartheta_{(2)}^2}\right),\ldots,exp\left(-\frac{\left\|y_i-\delta_q^{(2)}\right\|^2}{2\vartheta_{(2)}^2}\right)\right],\end{aligned} \tag{8}$$

where $\left\{\delta_j^{(t)}\right\}_{j=1}^{q}(t=1,2)$ denotes m anchor points. $\vartheta_{(1)}=\frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q}\left\|x_i-\delta_j^1\right\|$and $\vartheta_{(2)}=\frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q}\left\|y_i-\delta_j^2\right\|$denote the kernel widths of the two modalities. In this paper, we empirically set $q=d_A$.

By integrating Eqs. (1), (2), (6) and (7) into a framework, the overall objective function of the proposed DAZSH approach is given as follows:

$$\begin{aligned}&\min_{Z_1,Z_2,P_1,P_2,W_1,W_2,B,V}\mathscr{L}(Z_1,Z_2,P_1,P_2,W_1,W_2,B,V)\\ &=\sum_{i,j=1}^{n_s}\left[\log\left(1+e^{\Theta_{ij}}\right)-S_{ij}\Theta_{ij}\right]\\ &+\alpha_1\|\phi(X)-Z_1L_sA\|_F^2+\alpha_2\|\phi(Y)-Z_2L_sA\|_F^2\\ &+\beta_1\|B-Z_1L_sP_1\|_F^2+\beta_2\|V-Z_2L_sP_2\|_F^2\\ &+\mu_1\|B-\phi(X)W_1\|_F^2+\mu_2\|V-\phi(Y)W_2\|_F^2\\ &+\gamma R(Z_1,Z_2,P_1,P_2,W_1,W_2)\\ &s.t. B,V\in\{-1,1\}^{n_s\times k}\end{aligned} \tag{9}$$

where $R(\cdot)$ aims to avoid overfitting and γ is the regularization parameter.

3.3. Algorithm optimization

Evidently, the overall objective function is a nonconvex optimization problem. Therefore, we develop an effective alternating iterative algorithm to solve this proposed problem. Specifically, Eq. (9) can be optimized by the following steps:

B-step: Update B by fixing Z_1,Z_2,P_1,P_2,W_1,W_2, and V. Therefore, Eq. (9) with respect to B can be simplified as follows:

$$\begin{aligned}&\min_B\Gamma=\sum_{i,j=1}^{n_s}\left[\log\left(1+e^{\Theta_{ij}}\right)-S_{ij}\Theta_{ij}\right]+\beta_1\|B-Z_1L_sP_1\|_F^2\\ &+\mu_1\|B-\phi(X)W_1\|_F^2\\ &s.t.\ B,V\in\{-1,1\}^{n_s\times k}\end{aligned} \tag{10}$$

Optimize the whole B at one time is a difficult problem. Inspired by [33], we adopt the column-wise iterative strategy to obtain matrix B. Specifically, hash codes B are updated bit by bit. We need to construct the lower bound for Eq. (10) to optimize the i-th column of B and then obtain the closed solution for B.

The gradient and Hessian of the problem (10) with respect to B can be computed as follows:

$$\begin{cases}\frac{\partial\Gamma}{\partial B_{*i}}=\frac{\lambda}{k}\sum_{j=1}^{n_s}\left(S_{j*}^T-C_{j*}^T\right)V_{ji}-2(\beta_1(B_{*i}-Z_1L_sP_{1*i})\\ +\mu_1(B_{*i}-\phi(X)W_{*i})),\\ \frac{\partial^2\Gamma}{\partial B_{*i}\partial B_{*i}^T}=-\frac{\lambda^2}{k^2}diag(c_1,c_2,\ldots,c_n)-2(\beta_1+\mu_1)I,\end{cases} \tag{11}$$

where $c_i=\sum_{j=1}^{n_s}C_{ij}(1-C_{ij})$ and $diag(\cdot)$ denotes a diagonal matrix. Here, $B_{*i}(t)$ denotes the value of B_{*i} at the t-th iteration and $\frac{\partial\Gamma}{\partial B_{*i}}(t)$ is expressed as the gradient of $B_{*i}(t)$. Therefore, we define the lower bound of $\Gamma(B_{*i})$ as follows:

$$\begin{aligned}\tilde{\Gamma}(B_{*i}) &= \Gamma(B_{*i}(t))+(B_{*i}-B_{*i}(t))^T\frac{\partial\Gamma}{\partial B_{*i}}(t)\\ &+\tfrac{1}{2}(B_{*i}-B_{*i}(t))^TH(B_{*i}-B_{*i}(t))\\ &=B_{*i}^T(\tfrac{\partial\Gamma}{\partial B_{*i}}(t)-HB_{*i}(t))-(B_{*i}(t))^T\tfrac{\partial\Gamma}{\partial B_{*i}}(t)\\ &+\Gamma(B_{*i}(t))+\tfrac{(B_{*i}(t))^THB_{*i}(t)}{2}-\tfrac{\lambda^2n_s^3}{8k^2}\\ &=B_{*i}^T(\tfrac{\partial\Gamma}{\partial B_{*i}}(t)-HB_{*i}(t))+const,\end{aligned} \tag{12}$$

where $H=-(\frac{n_s\lambda^2}{4k^2}+2(\beta_1+\mu_1))I$ and *const* represent a constant term that is independent of B_{*i}. According to Theorem 1 in [33], the optimization problem of B can be transformed into the following problem:

$$\begin{aligned}&\max_{B_{*i}}\tilde{\Gamma}(B_{*i})=B_{*thethethei}^T(\tfrac{\partial\Gamma}{\partial B_{*i}}(t)-HB_{*i}(t))\\ &s.t.\quad B_{*i}\in\{-1,1\}^{n_s}.\end{aligned} \tag{13}$$

Obviously, the solution to problem (13) is $B_{*i}=sgn(\frac{\partial\Gamma}{\partial B_{*i}}(t)-HB_{*i}(t))$. Therefore, we use this scheme to obtain $B_{*i}(t+1)$ as follows:

$$\begin{aligned}&B_{*i}(t+1)=sgn(\tfrac{\partial\Gamma}{\partial B_{*i}}(t)-HB_{*i}(t))\\ &s.t.\quad B_{*i}(t+1)\in\{-1,1\}^{n_s}.\end{aligned} \tag{14}$$

However, the gradient calculation requires a high computational cost. Therefore, we adopt the random learning strategy [40] to optimize B_{*i}. In each iteration, we randomly select m columns from S to update the gradient of B_{*i}. Therefore, we obtain the updating rule of B_{*i} as follows:

$$\begin{aligned}B_{*i}(t+1)=sgn(&\frac{\lambda}{k}\sum_{u=1}^{m}(S_{j_u*}^T-C_{j_u*}^T)V_{j_ui}+(\frac{m\lambda^2}{4k^2}+2(\beta_1\\ &+\mu_1))B_{*i}(t)).\end{aligned} \tag{15}$$

V-step: Update V by fixing Z_1,Z_2,P_1,P_2,W_1,W_2, and B. Eq. (9) with respect to V can be simplified as follows:

$$\begin{aligned}&\min_V\Gamma=\sum_{i,j=1}^{n_s}\left[\log\left(1+e^{\Theta_{ij}}\right)-S_{ij}\Theta_{ij}\right]+\beta_2\|V-Z_2L_sP_2\|_F^2\\ &+\mu_2\|V-\phi(Y)W_2\|_F^2\\ &s.t.B,V\in\{-1,1\}^{n_s\times k}.\end{aligned} \tag{16}$$

Similarly, we can obtain the closed solution to V_{*i} as follows:

$$\begin{aligned}V_{*i}(t+1)=sgn(&\frac{\lambda}{k}\sum_{u=1}^{m}(S_{*j_u}^T-C_{*j_u}^T)B_{j_ui}+(\frac{m\lambda^2}{4k^2}+2(\beta_2\\ &+\mu_2))V_{*i}(t)).\end{aligned} \tag{17}$$

Z_1-step: Update Z_1 by fixing Z_2,P_1,P_2,W_1,W_2,B, and V. Eq. (9) becomes the following minimization problem:

$$\min_{Z_1}\alpha_1\|\phi(X)-Z_1L_sA\|_F^2+\beta_1\|B-Z_1L_sP_1\|_F^2+\gamma\|Z_1\|_F^2 \tag{18}$$

By setting the partial derivative w.r.t. Z_1 to zero, we can derive the closed solution to Z_1 as follows:

$$\begin{aligned}Z_1=&(\alpha_1\phi(X)A^TL_s^T\\ &+\beta_1BP_1^TL_s^T)(\alpha_1L_sAA^TL_s^T+\beta_1L_sP_1P_1^TL_s^T+\gamma I)^{-1}.\end{aligned} \tag{19}$$

Z_2-step: Update Z_2 by fixing Z_1,P_1,P_2,W_1,W_2,B, and V. Eq. (9) becomes the following minimization optimization problem:

$$\min_{Z_2}\alpha_2\|\phi(Y)-Z_2L_sA\|_F^2+\beta_2\|V-Z_2L_sP_2\|_F^2+\gamma\|Z_2\|_F^2. \tag{20}$$

Similarly, we can obtain the closed solution to Z_2 as follows:

$$Z_2 = (\alpha_2\phi(Y)A^TL_s^T + \beta_2 VP_2^TL_s^T)(\alpha_2 L_s AA^TL_s^T + \beta_2 L_s P_2 P_2^T L_s^T + \gamma I)^{-1}. \quad (21)$$

P_1-step: Update P_1 by fixing $Z_1, Z_2, P_2, W_1, W_2, B$, and V. Eq. (9) adopts the following form:

$$\min_{P_1} \beta_1\|B - Z_1L_sP_1\|_F^2 + \gamma\|P_1\|_F^2. \quad (22)$$

Similarly, by setting the partial derivative w.r.t. P_1 to zero, we can derive the closed solution to P_1 as follows:

$$P_1 = (\beta_1 L_s^T Z_1^T B)(\beta_1 L_s^T Z_1^T Z_1 L_s + \gamma I)^{-1}. \quad (23)$$

P_2-step: Update P_2 by fixing $Z_1, Z_2, P_1, W_1, W_2, B$, and V. Eq. (9) can be written as the following problem:

$$\min_{P_2} \beta_2\|V - Z_2L_sP_2\|_F^2 + \gamma\|P_2\|_F^2. \quad (24)$$

By adopting a similar solution scheme, the closed solution to P_2 is given as follows:

$$P_2 = (\beta_2 L_s^T Z_2^T V)(\beta_2 L_s^T Z_2^T Z_2 L_s + \gamma I)^{-1}. \quad (25)$$

W_1-step: Update W_1 by fixing $Z_1, Z_2, P_1, P_2, W_2, B$, and V. We can simplify Eq. (9) as follows:

$$\min_{W_1} \mu_1\|B - \phi(X)W_1\|_F^2 + \gamma\|W_1\|_F^2. \quad (26)$$

By taking the partial derivative w.r.t. W_1 to zero, the closed solution to W_1 can be expressed as follows:

$$W_1 = (\mu_1\phi(X)^T\phi(X) + \gamma I)^{-1}(\mu_1\phi(X)^T B). \quad (27)$$

W_2-step: Update W_2 by fixing $Z_1, Z_2, P_1, P_2, W_1, B$, and V. Eq. (9) can be rewritten as follows:

$$\min_{W_2} \mu_2\|V - \phi(Y)W_2\|_F^2 + \gamma\|W_2\|_F^2. \quad (28)$$

Similarly, we can obtain the closed solution of W_2 as follows:

$$W_2 = (\mu_2\phi(Y)^T\phi(Y) + \gamma I)^{-1}(\mu_2\phi(Y)^T V). \quad (29)$$

In summary, Algorithm 1 describes the solution steps of our proposed DAZSH approach in detail.

Algorithm 1 DAZSH

Trainingstage

Input : Hash code length k, label matrix L_s, feature matrices X and Y of two modalities, attribute matrix A and parameters $\alpha_1, \alpha_2, \beta_1, \beta_2, \mu_1, \mu_2$, and γ.

Output : Hash codes B and V, mapping matrices W_1 and W_2 .

Procedure :

1. Calculate $\phi(X), \phi(Y)$ and similar matrix S.
2. Initialize Z_1, Z_2, P_1, P_2, W_1, and W_2.

Repeat

(1) Update B by Eq. (15).
(2) Update V by Eq. (17).
(3) Update Z_1 by Eq. (19).
(4) Update Z_2 by Eq. (21).
(5) Update P_1 by Eq. (23).
(6) Update P_2 by Eq. (25).
(7) Update W_1 by Eq. (27).
(8) Update W_2 by Eq. (29).

Until reaching the maximum iteration or convergence.

RetrievalStage

Input :Feature matrices X_{query} and Y_{query} of the retrieved data (seen class or unseen class), mapping matrices W_1 and W_2.

a (*continued*)

Algorithm 1 DAZSH

Output : B_x and V_y.

Procedure :

1. Calculate $\phi(X_{query})$ and $\phi(Y_{query})$.
2. For X_{query}: Calculate the hash code by $B_x = sgn(\phi(X_{query})W_1)$.
3. For Y_{query}: calculate the hash code by $V_y = sgn(\phi(Y_{query})W_2)$.

3.4. Complexity analysis

In this subsection, we give the complexity analysis of our DAZSH approach. In Section 3.3, the overall complexity of the proposed optimization scheme consists of updating $Z_1, Z_2, P_1, P_2, W_1, W_2, V, B$. Specifically, the computational complexity of updating Z_1 and Z_2 is $O((2n_s^2d_Ac + 2n_s^2c^2d_A + 2n_s^3)t)$. We need $O((n_s^2ck + c^2k + n_s^3c^2 + c^3)t)$ to update P_1 and P_2. The cost of updating W_1 and W_2 is $O((d_An_sk + d_A^2k + d_A^3 + d_A^2n_s)t)$. For variables V and B, the computational complexity is $O(mn)$. Here, m is the number of randomly selected columns from S and c is the number of categories. k is the hash code length and t is the updating iteration times. Since $n_s \gg d_A > c, k, m$, the overall complexity of our DAZSH method is linear with n_s (the size of the training dataset).

4. Experiments

In this section, we carry out experiments to verify the effectiveness of DAZSH in cross-modal retrieval. Specifically, we set up two common query tasks: image query text and text query image. In addition, the experiments are conducted on query sets in two different scenarios (seen and unseen).

4.1. Datasets

Wiki [41]: It contains 2866 image-text pairs collected from Wikipedia. AlexNet and the latent Dirichlet allocation (LDA) model are used to extract the features of the image modality and the text modality, respectively. Therefore, each image and each text can be represented by a 128-dimensional vector and a 10-dimensional topic vector, respectively. The training and testing datasets contain 2173 image-text pairs and 693 image-text pairs, respectively. In addition, we randomly select two classes from this dataset as unseen classes and the remaining classes as seen classes.

LabelMe [42]: This dataset consists of 2686 outside scenes from eight categories. We use a 366-dimensional phrase frequency and a 512-dimensional GIST feature to describe each sample of the text and image modalities, respectively. In our experiments, 2014 image-text pairs are randomly selected as the training dataset, and the remaining 672 image-text pairs are used as the testing dataset. Similarly, we randomly select two classes as unseen classes on this dataset.

Pascal VOC [43]: The dataset includes 5619 image-text pairs from 20 categories. Each image modality and text modality sample is described by a 512-dimensional vector and a 399-dimensional vector, respectively. In the experiments, we only choose single-label data and then randomly sample 2799 image-text pairs for training and the remaining 2820 image-text pairs are used as the testing dataset. Here, we randomly set four classes as unseen classes on this dataset.

4.2. Baselines and implementation details

To evaluate the effectiveness of DAZSH, we select several state-of-the-art cross-modal hashing methods, such as 1) CMFH [12]; 2) JIMFH [15]; 3) DRMFH [16]; 4) LCMFH [19]; 5) ASCSH [34]; 6) TSK [9]; 7) AH [44]; 8) CMAH [25]; 9) CHOP [26], as the comparison algorithms. Among them, DRMFH and JIMFH are unsupervised learning methods, and LCMFH, ASCSH, TSK, AH, CMAH, CHOP and our proposed DAZSH approach fully consider the supervised information. In addition, among these supervised hashing methods, both CMAH and CHOP are zero-shot cross-modal hashing approaches. In addition, both TSK and AH are zero-shot unimodal hashing approaches, and we train hash codes separately according to the image modality and the text modality. Regarding the baselines, we set the parameter values according to the suggestion of the original paper.

First, we randomly select several classes from the original datasets as unseen classes. To avoid the instability caused by random selection, we conduct the experiments 20 times and report the average values. For the zero-shot cross-modal retrieval task, our dataset is empirically set as follows. The training data are selected from the seen classes, and then, the query data are selected from the unseen classes. Then, the remaining data of the unseen classes and the samples of all seen classes are used as the retrieval set.

Due to the particularity of datasets and application scenes, we set different parameters for each dataset to ensure the best performances. For zero-shot retrieval, the parameters of the three datasets are set as follows: for the Wiki dataset, $[\alpha_1, \alpha_2, \beta_1, \beta_2, \mu_1, \mu_2, \gamma]$ = [1e-3,1e-4,1e3,1e2,1e-4,1e2,1e-4]; for the LabelMe dataset, $[\alpha_1, \alpha_2, \beta_1, \beta_2, \mu_1, \mu_2, \gamma]$ = [1e3,1e5,1e5,1e6,1e5,1e6,1e6]; and for the Pascal VOC dataset, $[\alpha_1, \alpha_2, \beta_1, \beta_2, \mu_1, \mu_2, \gamma]$ = [1e-1,1e-1,1e2,1e3,1e2,1e3,1e3]. For retrieval of the seen class, the parameters of the three datasets are set as follows: for the Wiki dataset, $[\alpha_1, \alpha_2, \beta_1, \beta_2, \mu_1, \mu_2, \gamma]$ = [1e-3,1e-4,1e-4,1e-6,1e-3,1e-4,1e-4]; for the LabelMe dataset, $[\alpha_1, \alpha_2, \beta_1, \beta_2, \mu_1, \mu_2, \gamma]$ = [1e-3,1e-5,1e-5,1e-6,1e-6,1e-6,1e-8]; and for the Pascal VOC dataset, $[\alpha_1, \alpha_2, \beta_1, \beta_2, \mu_1, \mu_2, \gamma]$ = [1e-1,1e-1,1e-5,1e-6,1e-4,1e-6,1e-6].

4.3. Evaluation metrics

In the retrieval task, the most common evaluation metric is the mean of average precision (mAP). Given the query samples and their retrieval results, mAP is calculated as:

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{N}\sum_{r=1}^{N}P_q(r)\xi_q(r), \tag{30}$$

where N denotes the number of relevant instances in the query set, Q is the query instances, and $P_q(r)$ is the precision of the top-r retrieval instances of the q-th query. $\xi_q(r) = 1$ if the r-th retrieval instance is a neighbour of the q-th query; otherwise, $\xi_q(r) = 0$.

To evaluate the performances of the retrieval methods more comprehensively, we also employ the top-N precision curve and mean precision within Hamming radius 2 (PH2)[45] as supplementary evaluation metrics.

4.4. Experimental results

In the experiments, we adopt four different hash code lengths to verify the retrieval performance of different algorithms on three benchmark multimodal datasets. Specifically, the lengths of the hash codes are set to 8 bits, 12 bits, 16 bits and 32 bits. Table 1.

Table 1
Statistics of the three datasets.

Statistics	Wiki	LabelMe	Pascal VOC
Total size	2866	2686	5619
Training dataset size	2173	2014	2799
Query dataset size	693	672	2829
Total classes	10	8	20
Unseen classes	2	2	4
Seen classes	8	6	16
Image feature	128	512	512
Text feature	10	366	399

4.4.1. Cross-modal retrieval of unseen classes

Our DAZSH method mainly focuses on zero-shot cross-modal retrieval. Note that the query set in our experiments comes from unseen classes. Then we evaluate the retrieval performance of DAZSH and the baselines on unseen classes from two aspects, mAP and PH2. The mAP values of both DAZSH and the other baselines on the three datasets are reported in Table 2, and their PH2 values are shown in Fig. 2. From the experimental results, we can make the following observations:

(1) As seen from Table 2, our proposed DAZSH approach is superior to other competitors in text-to-Image and image-to-text tasks on three datasets. In addition, from Fig. 2, the DAZSH method also achieves the best results in most cases. The main reason is that the DAZSH method makes full use of pairwise similarity, class attribute characteristics and label information at the same time. Therefore, our proposed method can learn more knowledge from seen classes and then effectively transfer them to unseen classes.

(2) It is obvious that the mAP values of JIMFH are lower than those of both LCMFH and ASCSH on the Wiki dataset. This is because the supervision information, such as label information and pairwise similarity, are used to constrain hash code learning in the LCMFH and ASCSH methods. This indicates that the supervision information is helpful in improving the discriminative ability of hash codes. In addition, the performances of DRMFH are superior to those of JIMFH on the three datasets. One possible reason is that the discrete solution scheme in the DRMFH model plays an important role in the optimization process.

(3) The performances of single-modal zero-shot hashing methods, such as TSK, and AH, are lower than those of cross-modal zero-shot hashing methods (e.g. CMAH, and CHOP) on the LabelMe dataset. The main reason is that the single modality methods cannot effectively narrow the heterogeneity gap between different modalities when they are applied in multimodal retrieval. Therefore, they ignore the correlation between different modal data. In addition, the mAP values of our proposed DAZSH method outperform those of CHOP on three datasets. The main reason is that our DAZSH method integrates the label constraint in attribute space learning, which can obtain more semantic knowledge of attributes from seen classes.

(4) Considering the traditional hashing methods and zero-shot hashing methods, the latter is more suitable for cross-modal retrieval of unseen classes. A possible reason is that the zero-shot hashing methods explore the class attribute information in hash code learning, and can transfer the semantic knowledge of class attributes from seen classes to unseen classes.

(5) Fig. 2 shows the PH2 values of DAZSH and the comparison methods on three datasets. Obviously, the performance of the DAZSH method is superior to that of the other methods

Table 2
The mAP of cross-modal retrieval for unseen queries of all methods on three datasets with different hash code lengths..

Task	Methods	Wiki				LabelMe				Pascal VOC			
		8	12	16	32	8	12	16	32	8	12	16	32
Text to Image	CMFH	0.1579	0.1447	0.1509	0.1386	0.3196	0.2917	0.2929	0.2687	0.0913	0.1058	0.0851	0.0679
	JIMFH	0.1945	0.1583	0.1614	0.1436	0.2423	0.2494	0.2093	0.2293	0.1171	0.1074	0.0970	0.0815
	DRMFH	0.2360	0.2063	0.1975	0.1778	0.4550	0.4332	0.4456	0.4176	0.2101	0.1921	0.1820	0.1702
	LCMFH	0.2047	0.2263	0.2349	0.2151	0.3894	0.3535	0.3451	0.4058	0.1635	0.1380	0.1571	0.1400
	ASCSH	0.2459	0.2427	0.2451	0.2256	0.3254	0.3438	0.3210	0.3643	0.1665	0.1728	0.1562	0.1458
	TSK	0.2191	0.2182	0.1749	0.1737	0.2840	0.3289	0.2062	0.2294	0.1744	0.1456	0.1536	0.1427
	AH	0.2202	0.2035	0.2061	0.1740	0.2584	0.2244	0.2161	0.1982	0.1207	0.1276	0.1107	0.1125
	CMAH	0.2496	0.2174	0.2227	0.1857	0.4549	0.4051	0.4192	0.4630	0.1689	0.1180	0.1263	0.0782
	CHOP	0.2398	0.2422	0.2735	0.2781	0.4098	0.4104	0.4164	**0.5242**	0.2425	0.2423	0.3154	0.2771
	DAZSH	**0.4520**	**0.4662**	**0.4756**	**0.4676**	**0.4839**	**0.4802**	**0.4807**	0.4735	**0.3759**	**0.3625**	**0.3849**	**0.3953**
Image to Text	CMFH	0.1777	0.1616	0.1661	0.1483	0.3112	0.2857	0.2840	0.2663	0.0946	0.1160	0.0934	0.0726
	JIMFH	0.2208	0.1813	0.1953	0.1827	0.2370	0.2551	0.2249	0.2427	0.1290	0.1164	0.1175	0.0945
	DRMFH	0.2494	0.2183	0.2068	0.1959	0.4243	0.4114	0.3766	0.3985	0.1873	0.1722	0.1558	0.1441
	LCMFH	0.2291	0.2565	0.2453	0.2400	0.3907	0.3839	0.3433	0.3705	0.1750	0.1722	0.1765	0.1655
	ASCSH	0.3143	0.2942	0.3201	0.2947	0.3172	0.3232	0.2911	0.3446	0.1988	0.1829	0.1781	0.1687
	TSK	0.2863	0.2514	0.2520	0.2433	0.3017	0.3086	0.2111	0.2391	0.1799	0.1517	0.1471	0.1349
	AH	0.2000	0.2151	0.1941	0.1386	0.2433	0.2036	0.1981	0.1928	0.1560	0.1495	0.1635	0.1435
	CMAH	0.2558	0.2264	0.2185	0.2063	0.4012	0.3994	0.3566	0.4060	0.1532	0.1112	0.1259	0.0771
	CHOP	0.2465	0.2408	0.2563	0.2828	0.3311	0.3516	0.3107	0.4332	0.2026	0.2210	0.2438	0.1809
	DAZSH	**0.4178**	**0.4149**	**0.4033**	**0.4200**	**0.4751**	**0.4901**	**0.4872**	**0.4928**	**0.3080**	**0.2869**	**0.3142**	**0.3280**

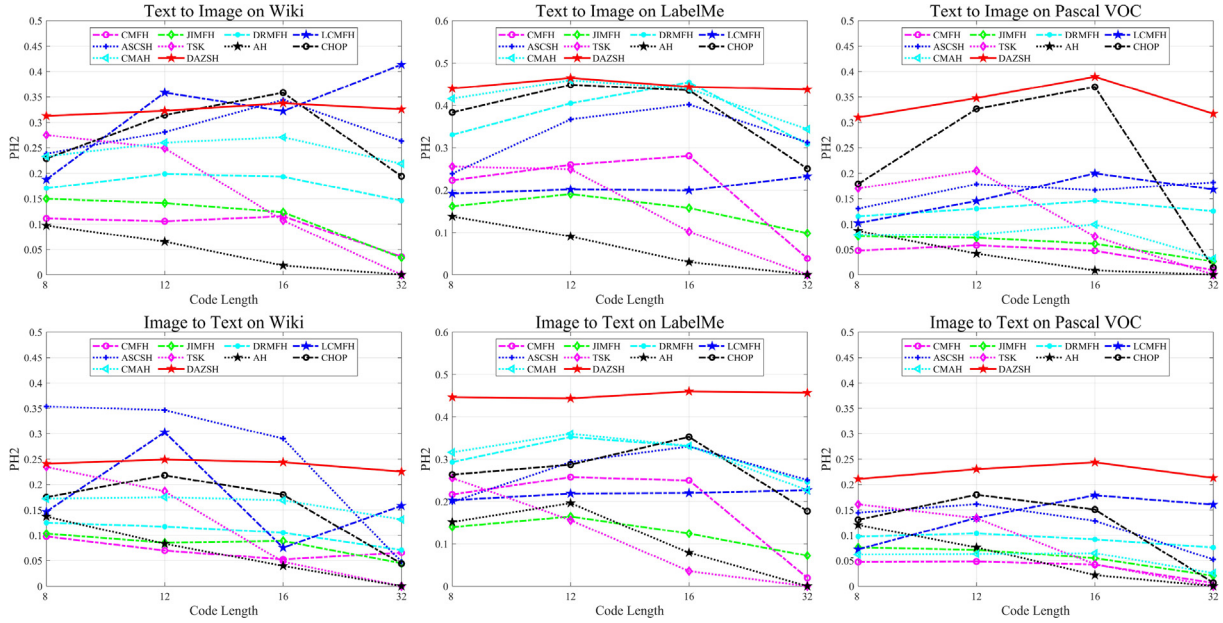

Fig. 2. The PH2 values of cross-modal retrieval for unseen classes on three datasets..

in most cases. The results show that DASH can obtain more discriminative hash codes in Hamming space. In addition, the change curves of the DAZSH method are relatively flat, while the change curves of the other comparison methods are relatively large. This indicates that our DAZSH method is more insensitive to hash code length than other competitors.

4.4.2. Cross-modal retrieval of seen classes

To evaluate the effectiveness of our proposed DAZSH method more comprehensively, we also conduct experiments on seen multimodal datasets. The mAP values of both DAZSH and other competitors on the three datasets are shown in Table 3, and their top-*N* curves are shown in Figs. 3,4,. From these results, we can make the following observations:

(1) Most retrieval methods achieve higher mAP values in text-to-image tasks than in image-to-text tasks. The main reason is that the high-dimensional original feature matrix of the image modality retains less semantic information in hash codes than the text modality. From the semantic point of view, textual information can capture semantic information more effectively than visual features.

(2) Table 3 shows- that most methods achieve better results in the seen class retrieval task than in the unseen class retrieval task. This is because the training set is the seen class. When using the seen class to query, we can directly employ the supervised information and the semantic information in the features to query without considering the transfer of attribute semantic knowledge. However, the retrieval results of both the TSK and AH methods in seen tasks cannot be significantly improved. This is because the single-modal

Table 3
The mAP of cross-modal retrieval for seen queries of all methods on three datasets with different hash code lengths.

Task	Methods	Wiki				LabelMe				Pascal VOC			
		8	12	16	32	8	12	16	32	8	12	16	32
Text to Image	CMFH	0.3756	0.3986	0.4171	0.4560	0.4995	0.5416	0.5614	0.6018	0.1659	0.1709	0.1921	0.2240
	JIMFH	0.4757	0.5179	0.5243	0.5755	0.5863	0.6526	0.6200	0.7349	0.2688	0.3358	0.3410	0.2428
	DRMFH	0.4509	0.4722	0.4924	0.5280	0.8247	0.8539	0.8803	0.9004	0.2859	0.3057	0.3254	0.3484
	LCMFH	0.6749	0.6698	0.6769	0.7060	0.5666	0.5141	0.5491	0.5324	0.4794	0.5689	0.5948	0.6659
	ASCSH	0.3436	0.5391	0.6460	0.7505	0.5816	0.8213	0.9090	0.9463	0.2328	0.4002	0.5261	0.7020
	TSK	0.2234	0.2287	0.2061	0.1983	0.2659	0.2687	0.2971	0.2488	0.1708	0.1795	0.1601	0.1811
	AH	0.1404	0.1395	0.1410	0.1381	0.2256	0.2055	0.2191	0.2085	0.1143	0.1142	0.1143	0.1127
	CMAH	0.7016	0.6789	0.6522	0.5622	0.9219	0.9190	0.9279	0.9226	0.3217	0.2609	0.2781	0.2193
	CHOP	0.5656	0.6578	0.6691	0.7358	0.7230	0.8053	0.8595	0.9054	0.4189	0.4864	0.5850	0.6777
	DAZSH	**0.7030**	**0.7139**	**0.7299**	**0.7565**	**0.9334**	**0.9426**	**0.9432**	**0.9486**	**0.6266**	**0.7392**	**0.7838**	**0.8841**
Image to Text	CMFH	0.1854	0.1819	0.1824	0.1880	0.4480	0.4699	0.4823	0.5212	0.1420	0.1422	0.1456	0.1516
	JIMFH	0.2475	0.2559	0.2571	0.2767	0.4111	0.4547	0.4147	0.5012	0.2271	0.2578	0.2679	0.2755
	DRMFH	0.2490	0.2519	0.2583	0.2684	0.7543	0.7856	0.8159	0.8331	0.2319	0.2415	0.2546	0.2683
	LCMFH	0.3306	0.3417	0.3471	0.3624	0.5057	0.4656	0.4914	0.4769	0.3318	0.3552	0.3694	0.3890
	ASCSH	0.2056	0.2579	0.2881	0.3344	0.5103	0.7102	0.7996	0.8634	0.1845	0.2425	0.2991	0.3927
	TSK	0.2217	0.2215	0.2202	0.2102	0.2782	0.2760	0.2852	0.2547	0.1697	0.1745	0.1711	0.1663
	AH	0.1695	0.1714	0.1770	0.1720	0.2292	0.2036	0.2127	0.2039	0.1289	0.1330	0.1258	0.1340
	CMAH	0.2956	0.2824	0.2716	0.2449	0.8157	0.8287	0.8431	0.8291	0.2313	0.2098	0.2154	0.1998
	CHOP	0.2673	0.2952	0.3222	0.3584	0.5948	0.6675	0.7223	0.7839	0.2708	0.2787	0.3230	0.3755
	DAZSH	**0.3388**	**0.3629**	**0.3759**	**0.4025**	**0.8210**	**0.8494**	**0.8576**	**0.8676**	**0.3443**	**0.3980**	**0.4214**	**0.4697**

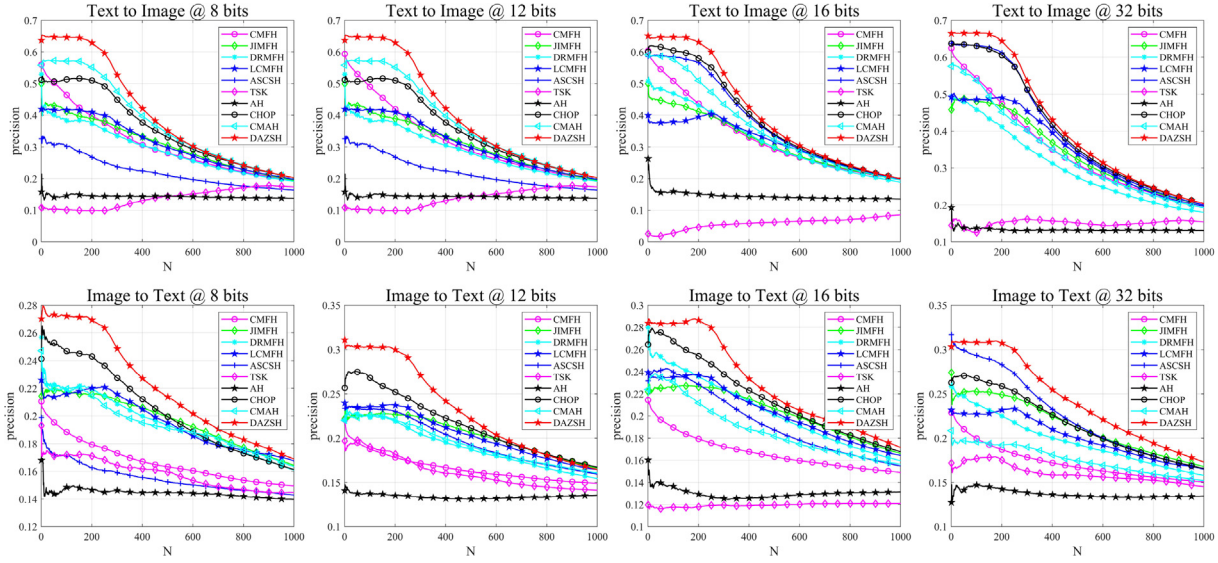

Fig. 3. The Precision@K curves of DAZSH and baselines on the Wiki dataset..

method cannot narrow the semantic gap between multi-modalities by learning the hash codes of each mode separately.

(3) From the top-*N* curves of Figs. 3–5, the top-N curves of our DAZSH method is always at the top. Therefore, DAZSH achieves better performances than other methods in the top-N curve metric. The results are consistent with the evaluation results in Table 3.

(4) Overall, the experimental results show that our DAZSH method outperforms other baselines in terms of both mAP values and top-*N* curves. Therefore, the proposed DAZSH method is more competitive than other state-of-the-art methods in seen class retrieval tasks.

4.5. Ablation Study

In this subsection, we conduct ablation experiments to verify the validity of several components of our proposed model. Therefore, four variants of DAZSH, i.e., DAZSH-S, DAZSH-A, DAZSH-L, and DAZSH-B, are constructed for comparison. Specifically, DAZSH-S removes the pairwise similarity-preserving term in hash code learning, unlike DAZSH. DAZSH-A is constructed by removing the attribute space learning terms. DAZSH-L discards the label information of cross-modal data in the original model. DAZSH-B learns the common hash codes for each modality, ignoring the characteristics of each modality. On the wiki dataset, we compare the performances of DAZSH with its four variants in two scenarios: unseen class retrieval and seen class retrieval. Table 4 reports the mAP results with different hash code lengths. Therefore, we can make the following observations:

(1) As seen from Table 4, DAZSH outperforms its variants in both retrieval scenarios. This is because our proposed DAZSH model makes full use of pairwise similarity, attribute space and labels to guide hash code learning. This demonstrates the effectiveness of our proposed DAZSH method.

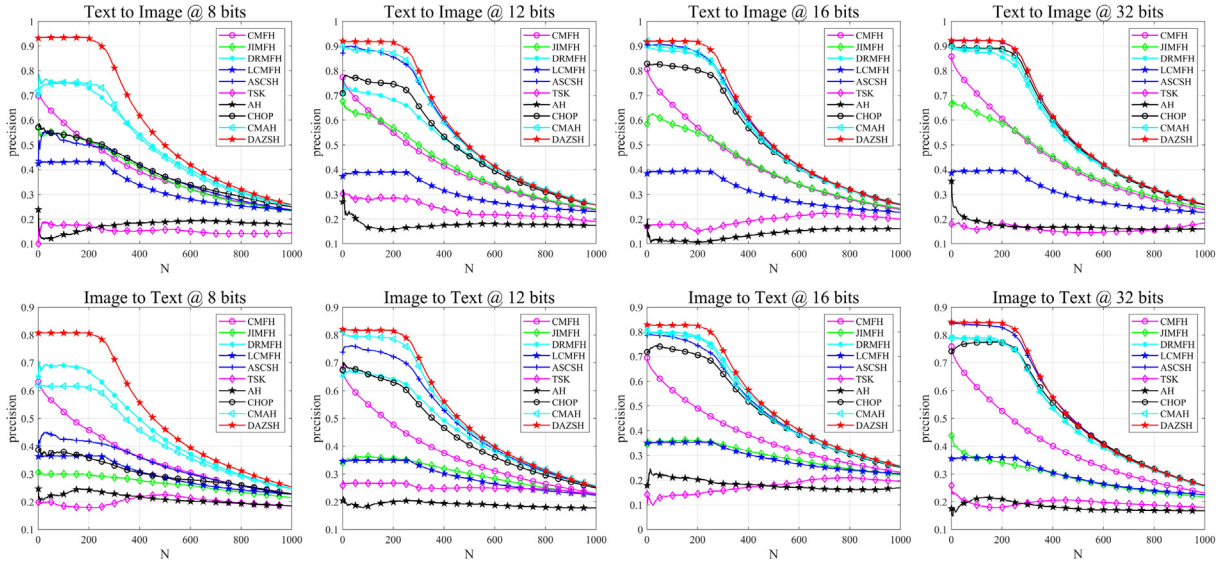

Fig. 4. The Precision@K curves of DAZSH and baselines on the LabelMe dataset..

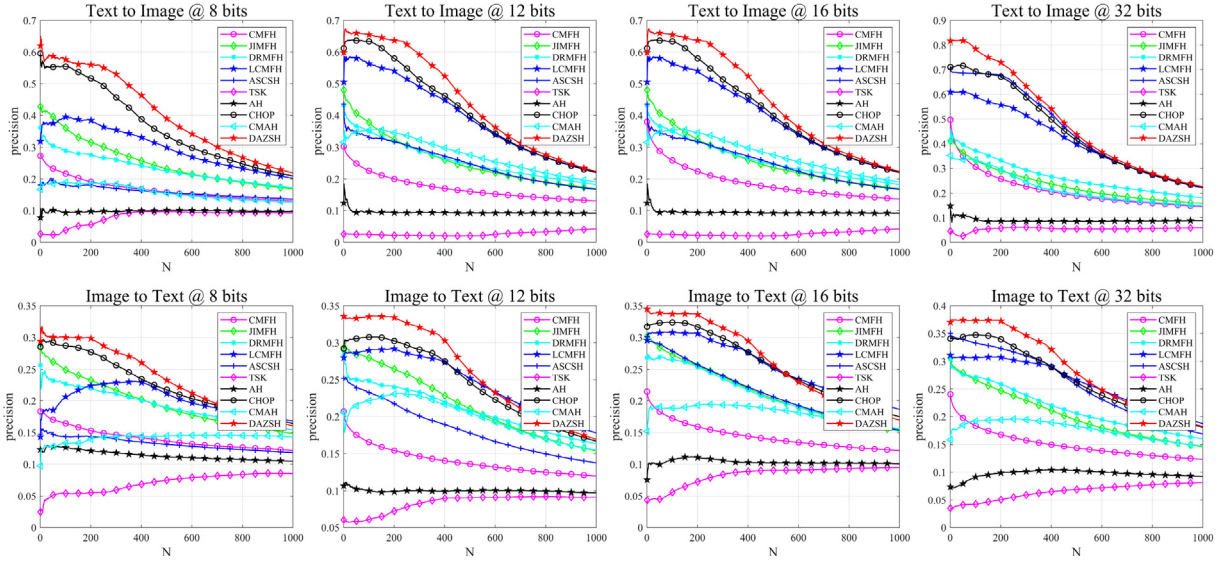

Fig. 5. The Precision@K curves of DAZSH and baselines on the Pascal VOC dataset..

Table 4
Ablation study on the Wiki dataset

Task	Methods	Unseen classes retrieval				Seen classes retrieval			
		8	12	16	32	8	12	16	32
Text to Image	DAZSH-S	0.4495	0.4499	0.4715	0.4672	0.1407	0.1432	0.1389	0.1416
	DAZSH-A	0.4467	0.4650	0.4466	0.4444	0.6856	0.7043	0.7245	0.7503
	DAZSH-L	0.2350	0.2523	0.2202	0.2070	0.6873	0.7055	0.7223	0.7489
	DAZSH-B	0.3082	0.3243	0.3217	0.3284	0.6862	0.7134	0.7289	0.7515
	DAZSH	**0.4520**	**0.4664**	**0.4756**	**0.4676**	**0.7030**	**0.7139**	**0.7299**	**0.7565**
Image to Text	DAZSH-S	0.3939	0.4024	0.4011	0.3864	0.1470	0.1489	0.1490	0.1502
	DAZSH-A	0.3859	0.4036	0.3822	0.4043	0.3203	0.3576	0.3721	0.3963
	DAZSH-L	0.2535	0.2864	0.2728	0.2882	0.3243	0.3619	0.3712	0.3940
	DAZSH-B	0.3725	0.3635	0.3695	0.3789	0.3334	0.3533	0.3719	0.4023
	DAZSH	**0.4178**	**0.4149**	**0.4033**	**0.4200**	**0.3388**	**0.3629**	**0.3759**	**0.4025**

(2) Among the unseen class retrieval tasks, DAZSH-L achieves the worst performance among all methods, which indicates that the label information in attribute space learning can improve the performances in cross-retrieval tasks.

(3) In the seen class retrieval task, the performances of DAZSH-S are inferior to those of other methods, which demonstrates that pairwise similarity preservation has an important influence on generating discriminative hash codes.

(4) The DAZSH-A approach achieves worse results on the unseen class retrieval task than on the seen class retrieval task. This is because DAZSH-A discards the learned knowledge of the attribute space. Moreover, for unseen class retrieval, it is important to transfer the attribute semantic knowledge of the seen class to the unseen class.
(5) Table 4 shows that the DAZSH method outperforms the DAZSH-B method on both unseen and seen classes retrieval tasks. The main reason is that the DAZSH-B method learns the common hash codes for different modalities. Therefore, it ignores the characteristics of each modality, which seriously affects the performance of cross-modal retrieval.

4.6. Convergence analysis

Since the proposed model is optimized by the iterative updating strategy, its convergence rate is critical in real applications. Fig. 6 shows the convergence curves of our DAZSH model on three cross-modal datasets, in which we set the hash code lengths to 8 and 12 bits. In Fig. 6, the x-axis is the number of iterations, and the y-axis is the objective function value. Fig. 6 shows that the proposed DAZSH method converges within 10 iterations for different hash code lengths. This demonstrates the efficiency of the optimization scheme in practice.

4.7. Parameter sensitivity analysis

In this subsection, we carry out experiments on the LabelMe dataset and analyse the sensitivity of the parameters in the proposed DAZSH model. We fix the length of the hash codes to 8 bits and apply them to the unseen class retrieval scenario. Specifically, we vary the value of one parameter while fixing the value of the other parameters. Fig. 7 shows the performance influence of the DAZSH model with different parameter values. From Fig. 7, we can make the following observations:

(1) α_1 and α_2 aim to control the attribute space learning of two modalities, whose values are set from $1e-3$ to $1e5$. It can be observed that the results are insensitive to α_1 and α_2.
(2) β_1 and β_2 represent the weight parameters, which are used to control hash codes learning. The mAP values of our proposed method are increasing when β_1 and β_2 are increased from $1e-1$ to $1e5$, and the mAP values remain relatively stable when β_1 and β_2 reach $1e5$. The main reason is that the hash code learning terms from the attribute space are controlled by β_1 and β_2. When β_1 and β_2 are too small, the semantic knowledge in the attribute space may be ignored.
(3) μ_1 and μ_2 aim to control the hash function learning items. Our proposed DAZSH method can achieve good performance when μ_1 and μ_2 are in the range $[1e-1, 1e7]$. The performances of our proposed methods are insensitive to the parameters μ_1 and μ_2.
(4) γ denotes the weight parameter of the regularization term and its range is set to $[1e-1, 1e7]$. Our proposed method can maintain relatively stable performance over a large range.

4.8. Visualization analysis

To better verify the effectiveness of the DAZSH method, we employ the t-SNE tool to visualize the distribution of the original features and the learned representations. Specifically, we randomly select 600 image-text pairs from the LabelMe dataset for visualization experiments. The results are shown in Fig. 8, where different colours represent different categories and different shapes represent different modes. Figs. 8 (a)-(c) show the visual distribution of the original image features, the original text features and the mixed features of the two modes. The results show that the original features of images and texts are scattered, and it is difficult to separate the categories. In addition, Fig. 8(c) shows that the scatterplots from the same category cannot correspond, indicating that the distributions of the two modalities are also very different.

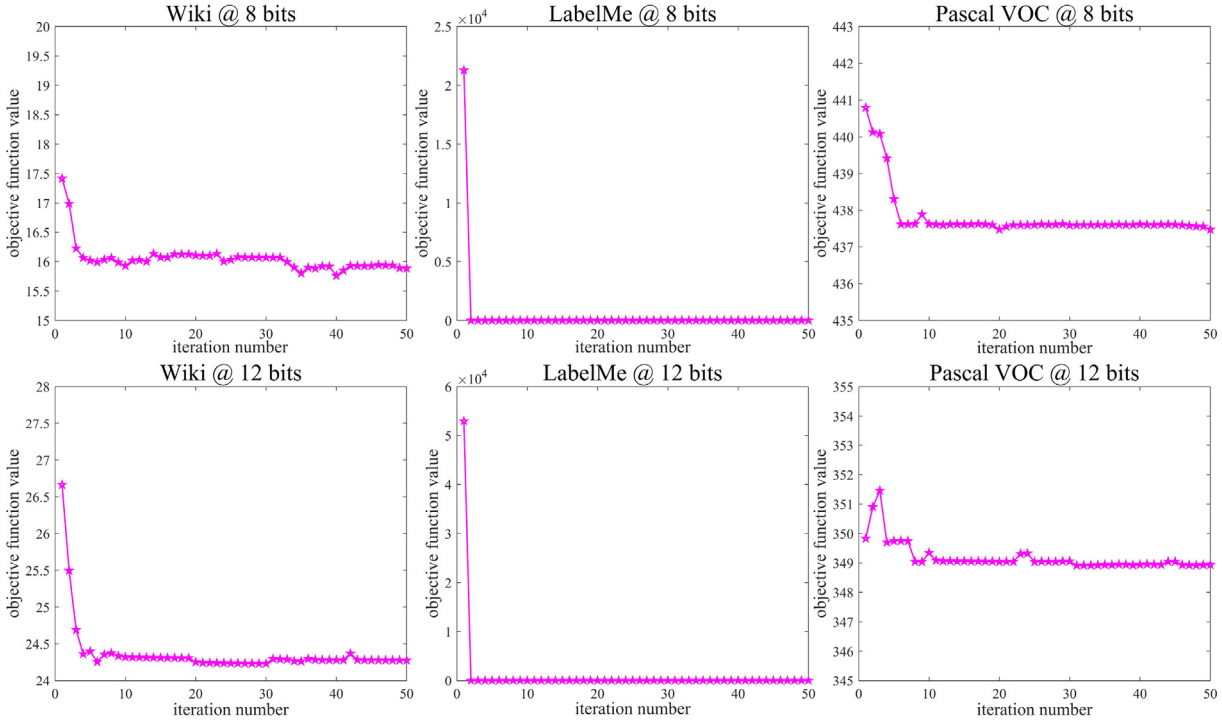

Fig. 6. Convergence curves of our proposed method on three datasets..

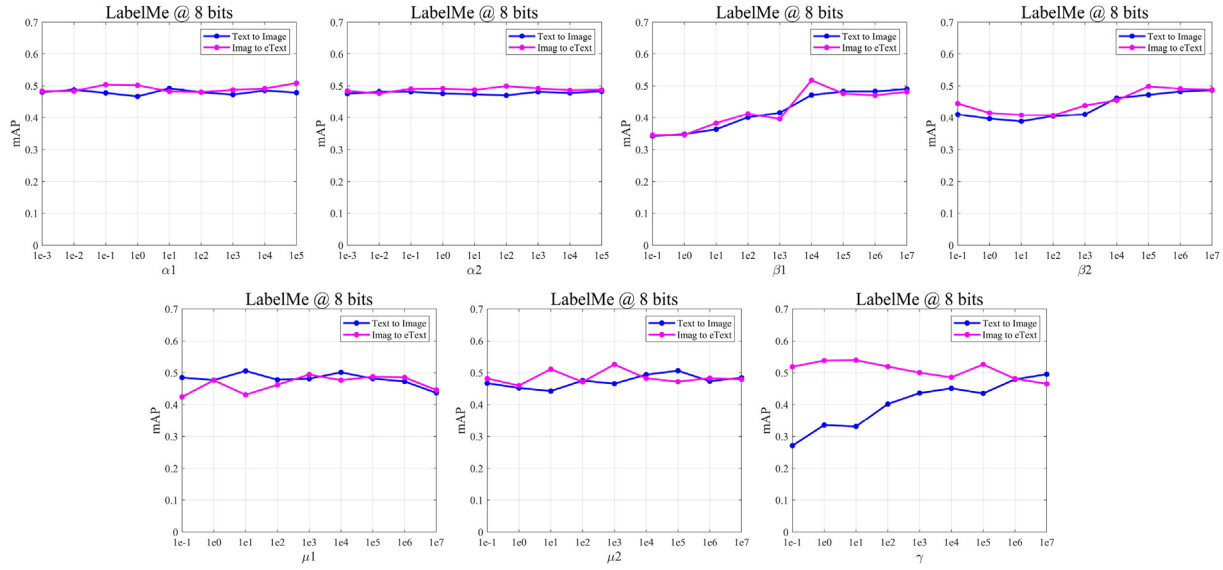

Fig. 7. The retrieval performances of our proposed methods varied with different values of the parameters on three datasets..

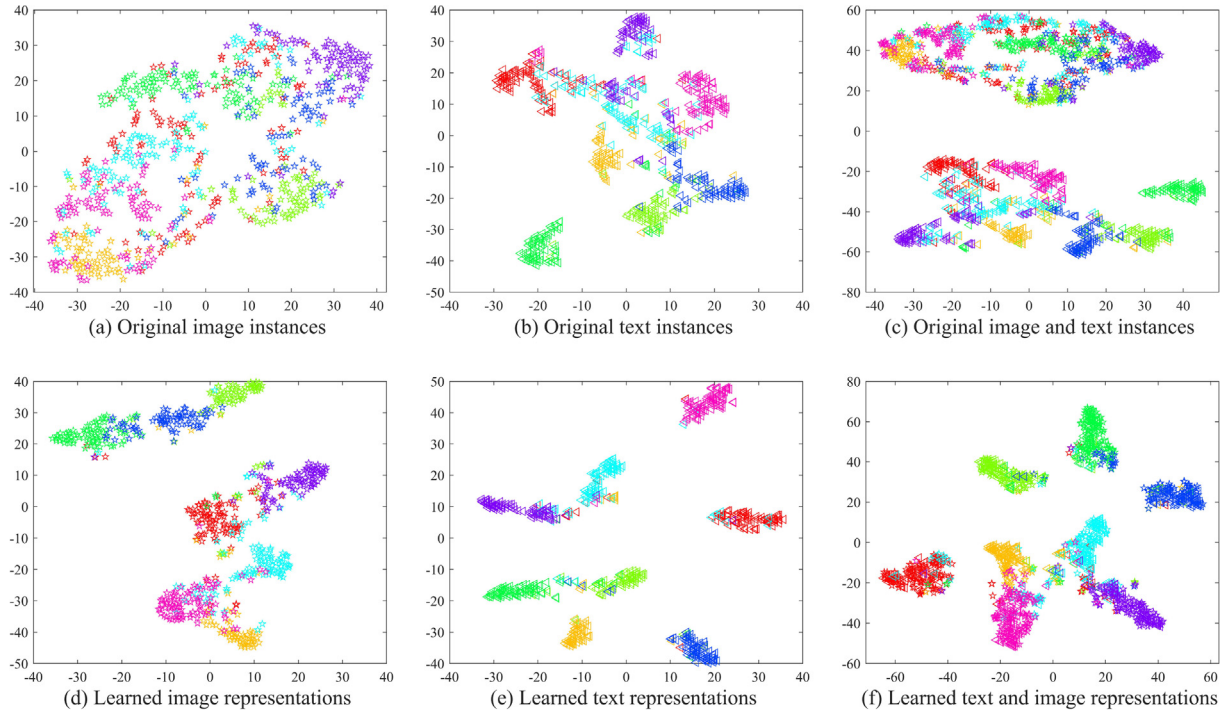

Fig. 8. t-SNE visualization of the raw features and the learned semantic features..

Figs. 8 (d)-(e) show the representations of the image and text modalities, respectively. This shows that the DAZSH method can learn the discriminative semantic representation. In Fig. 8 (f), the learned image and text representations are mixed together. From Figs. 8 (c) and (f), know that the representations of the multi-modality data obtained by our DAZSH method have stronger discriminative ability than those of the original multimodality data. In addition, the image and text samples from the same category are close, which indicates that our proposed model can effectively narrow the gap between different modalities.

5. Conclusions

In this paper, we propose a novel zero-shot hashing method, called DAZSH, for cross-modal retrieval. The method adopts an asymmetric discrete coding structure and pairwise similarity to guide hash code learning, which can significantly improve the discriminative ability of hash codes. Meanwhile, the DAZSH method constructs an attribute space for each modality by combining the feature matrix and the class attribute matrix, and thus achieves knowledge transfer from seen classes to unseen classes. In addi-

tion, we impose the label constraint on attribute space learning to improve the discriminability of the attribute space. As a result, our proposed DAZSH method can obtain more discriminative hash codes than traditional zero-shot hashing methods. The extensive experimental results show that the DAZSH method achieves better results than traditional methods in zero-shot cross-modal retrieval. In the future, we will extend the proposed method to semi-supervised zero-shot cross-modal retrieval, which is more practical than the supervised learning zero-shot case.

CRediT authorship contribution statement

Zhenqiu Shu: Conceptualization, Writing – review & editing, Supervision. **Kailing Yong:** Data curation, Software, Validation, Visualization, Writing – original draft. **Jun Yu:** Writing – review & editing. **Shengxiang Gao:** Writing – review & editing. **Cunli Mao:** Writing – review & editing. **Zhengtao Yu:** Funding acquisition.

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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [Grant No. 61603159, 62162033, 62020106012, U21B2027], Yunnan Provincial Major Science and Technology Special Plan Projects [Grant No. 202002AD080001, 202103AA080015], Yunnan Foundation Research Projects [Grant No. 202201AT070154, 202101BE070001-056].

References

[1] Xin Liu, Zhikai Hu, Haibin Ling, et al., MTFH: A matrix tri-factorization hashing framework for efficient cross-modal retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (3) (2019) 964–981.
[2] Jun Yu, Donglin Zhang, Zhenqiu Shu, et al., Adaptive multi-modal fusion hashing via Hadamard matrix, Applied Intelligence (2022) 1–15.
[3] Donglin Zhang, Wu. Xiao-Jun, He-Feng Yin, et al., Moon: Multi-hash codes joint learning for cross-media retrieval, Pattern Recognition Letters 151 (2021) 19–25.
[4] Zhenqiu Shu, Yibing Bai, Donglin Zhang, et al., Specific class center guided deep hashing for cross-modal retrieval, Information Sciences 609 (2022) 304–318.
[5] Lu Wang, Masoumeh Zareapoor, Jie Yang, et al., Asymmetric correlation quantization hashing for cross-modal retrieval, IEEE Transactions on Multimedia 24 (2021) 3665–3678.
[6] Yongxin Wang, Xin Luo, Liqiang Nie, et al., BATCH: A scalable asymmetric discrete cross-modal hashing, IEEE Transactions on Knowledge and Data Engineering 33 (11) (2020) 3507–3519.
[7] Feng Xue, Wenbo Wang, Wenjie Zhou, et al., Cross-modal retrieval via label category supervised matrix factorization hashing, Pattern Recognition Letters 138 (2020) 469–475.
[8] Christoph H Lampert, Hannes Nickisch, Stefan Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: IEEE Conference on Computer Vvision and Pattern Recognition, IEEE, 2009, pp. 951–958.
[9] Yang Yang, Yadan Luo, Weilun Chen, et al., Zero-shot hashing via transferring supervised knowledge, in: ACM International Conference on Multimedia, ACM, 2016, pp. 1286–1295.
[10] Yang Shi, Xiushan Nie, Xingbo Liu, et al., Zero-shot hashing via asymmetric ratio similarity matrix, IEEE Transactions on Knowledge and Data Engineering (2022), https://doi.org/10.1109/TKDE.2022.3150790.
[11] Qin Zou, Ling Cao, Zheng Zhang, et al., Transductive zero-shot hashing for multilabel image retrieval, IEEE Transactions on Neural Networks and Learning Systems 33 (4) (2020) 1673–1687.
[12] Guiguang Ding, Yuchen Guo, Jile Zhou, Collective matrix factorization hashing for multimodal data, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2014, pp. 2075–2082.
[13] Di Wang, Xinbo Gao, Xiumei Wang, et al., Semantic topic multimodal hashing for cross-media retrieval, in: International Conference on Artificial Intelligence, IEEE, 2015, pp. 3890–3896.
[14] Jile Zhou, Guiguang Ding, Yuchen Guo, Latent semantic sparse hashing for cross-modal similarity search, in: ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2014, pp. 415–424.
[15] Di Wang, Quan Wang, Lihuo He, et al., Joint and individual matrix factorization hashing for large-scale cross-modal retrieval, Pattern Recognition 107 (2020) 1–12.
[16] Tao Yao, Yiru Li, Weili Guan, et al., Discrete robust matrix factorization hashing for large-scale cross-media retrieval, IEEE Transactions on Knowledge and Data Engineering (2021), https://doi.org/10.1109/TKDE.2021.3107489.
[17] Zijia Lin, Guiguang Ding, Mingqing Hu, et al., Semantics-preserving hashing for cross-view retrieval, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 3864–3872.
[18] Xingbo Liu, Xiushan Nie, Wenjun Zeng, et al., Fast discrete cross-modal hashing with regressing from semantic labels, in: ACM international conference on Multimedia, ACM, 2018, pp. 1662–1669.
[19] Di Wang, Xinbo Gao, Xiumei Wang, Lihuo He, Label consistent matrix factorization hashing for large-scale cross-modal similarity search, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (10) (2018) 2466–2479.
[20] Donglin Zhang, Wu Xiao-Jun, Yu Jun, Label consistent flexible matrix factorization hashing for efficient cross-modal retrieval, ACM Transactions on Multimedia Computing, Communications, and Applications 17 (3) (2021) 1–18.
[21] Donglin Zhang, Xiao-Jun Wu, Tianyang Xu, et al., WATCH: Two-stage discrete cross-media hashing, IEEE Transactions on Knowledge and Data Engineering (2022), https://doi.org/10.1109/TKDE.2022.3159131.
[22] Huaxiong Li, Chao Zhang, Xiuyi Jia, et al., Adaptive label correlation based asymmetric discrete hashing for cross-modal retrieval, IEEE Transactions on Knowledge and Data Engineering (2021), https://doi.org/10.1109/TKDE.2021.3102119.
[23] Wei Wang, Vincent W Zheng, Han Yu, et al., A survey of zero-shot learning: Settings, methods, and applications. ACM Transactions on Intelligent Systems and Technology, 10(2) (2019) 1–37.
[24] Xuanwu Liu, Zhao Li, Jun Wang, Guoxian Yu, Carlotta Domenicon, Xiangliang Zhang. Crossmodal zero-shot hashing. In 2019 IEEE International Conference on Data Mining, pages 449–458. IEEE, 2019.
[25] Fangming Zhong, Zhikui Chen, Geyong Min, in: An exploration of cross-modal retrieval for unseen concepts. In International Conference on Database Systems for Advanced Applications, Springer, 2019, pp. 20–35.
[26] Xu Yuan, Guangze Wang, Zhikui Chen, et al., CHOP: An orthogonal hashing method for zero-shot cross-modal retrieval, Pattern Recognition Letters 145 (2021) 247–253.
[27] Zhong Ji, Yuxin Sun, Yunlong Yu, et al., Attribute-guided network for cross-modal zero-shot hashing, IEEE Transactions on Neural Networks and Learning Systems 31 (1) (2019) 321–330.
[28] Xing Xu, Kaiyi Lin, Huimin Lu, et al., Correlated features synthesis and alignment for zero-shot cross-modal retrieval, in: ACM SIGIR Conference on Research and Development in Information Retrieval, Information Retrieval, ACM, 2020, pp. 1419–1428.
[29] Xing Xu, Huimin Lu, Jingkuan Song, et al., Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval, IEEE Transactions on Cybernetics 50 (6) (2019) 2400–2413.
[30] Xing Xu, Kaiyi Lin, Yang Yang, et al., Joint feature synthesis and embedding: Adversarial cross-modal retrieval revisited, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (6) (2020) 3030–3047.
[31] Jeffrey Pennington, Richard Socher, Christopher D. Manning, Glove: Global vectors for word representation, in: IEEE Conference on Empirical Methods in Natural Language Processing, IEEE, 2014, pp. 1532–1543.
[32] Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, Chang Shih-Fu, Supervised hashing with kernels, IEEE, 2012, pp. 2074–2081.
[33] Qing-Yuan Jiang, Wu-Jun Li, Discrete latent factor model for cross-modal hashing, IEEE Transactions on Image Processing 28 (7) (2019) 3490–3501.
[34] Min Meng, Haitao Wang, Jun Yu, et al., Asymmetric supervised consistent and specific hashing for cross-modal retrieval, IEEE Transactions on Image Processing 30 (2020) 986–1000.
[35] Cheng Da, Xu Shibiao, Kun Ding, et al., Shiming Xiang, and Chunhong Pan. Amvh: Asymmetric multi-valued hashing, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 736–744.
[36] Cheng Da, Gaofeng Meng, Shiming Xiang, et al., Nonlinear asymmetric multi-valued hashing, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (11) (2018) 2660–2676.
[37] Zheng Zhang, Zhihui Lai, Zi Huang, et al., Scalable supervised asymmetric hashing with semantic and latent factor embedding, IEEE Transactions on Image Processing 28 (10) (2019) 4803–4818.
[38] Chuan-Xiang Li, Zhen-Duo Chen, Peng-Fei Zhang, et al., SCRATCH A scalable discrete matrix factorization hashing for cross-modal retrieval, in: ACM International Conference on Multimedia, ACM, 2018, pp. 1–9.
[39] Fumin Shen, Chunhua Shen, Wei Liu, et al., Supervised discrete hashing, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 37–45.
[40] Peichao Zhang, Wei Zhang, Wu-Jun Li, et al., Supervised hashing with latent factor models, in: ACM SIGIR Conference on Research and development in Information Retrieval, ACM, 2014, pp. 173–182.
[41] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, et al., A new approach to cross-modal multimedia retrieval, in: ACM International Conference on Multimedia, ACM, 2010, pp. 251–260.

[42] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, et al., Labelme: a database and web-based tool for image annotation, International Journal of Computer Vision 77 (1) (2008) 157–173.
[43] Mark Everingham, Luc Van Gool, Christopher K.I. Williams, et al., The pascal visual object classes (voc) challenge, International Journal of Computer Vision 88 (2) (2010) 303–338.
[44] Yahui Xu, Yang Yang, Fumin Shen, et al., Attribute hashing for zero-shot image retrieval, IEEE, 2017, pp. 133–138.
[45] Xianglong Liu, Junfeng He, Cheng Deng, et al., Collaborative hashing, in: IEEE Conference on Computer Vsion and Pattern Recognition, IEEE, 2014, pp. 2139–2146.

Zhenqiu Shu received the Ph.D. degree in computer applications at Nanjing University of Science and Technology. In February 2021, he joined the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, where he is currently an associate professor. Before joining in Kunming University of Science and Technology University, he had been a postdoctoral in Jiangnan University for four years. His research interests include image processing, computer vision and machine learning.

Kailing Yong is currently pursuing toward Master degree at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. Her current research interests include multimedia information retrieval and machine learning.

Jun Yu received his Ph.D. degree at the school of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China. He joined the College of Computer and Communication Engineering, Zhengzhou University of Light Industry in 2021. His research interests include multimedia information retrieval, computer vision and deep learning.

Shengxiang Gao received the Ph.D. degree in computer application technology from the Kunming University of Science and Technology, Kunming, China, in 2016. She is currently an associate professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. Her current research interests include natural language processing and information retrieval.

Cunli Mao received the Ph.D. degree in computer application technology from the Kunming University of Science and Technology, Kunming, China, in 2014. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. His current research interests include natural language processing and machine learning.

Zhengtao Yu received the Ph.D. degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2005. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. His current research interests include natural language processing, information retrieval, and machine learning.