feature extraction and the hash codes learning are divided into two independent parts, which cannot learn more hash codes from the multi-modal data [3–7]. Several deep hashing approaches are developed to alleviate the problems for cross-modal retrieval in the past few years [8–12]. Most deep hashing methods use pairwise or triplet loss to learn hash codes. However, large intraclass variations between the multi-media data may be caused in most cases (as illustrated in Fig. 1). Consequently, the semantic information hidden in the hash codes may be inconsistent with the labels, leading to serious performance degradation in cross-modal retrieval tasks.

In this work, we design a novel hashing learning framework, called Specific Class Center Guided Deep Hashing (SCCGDH), which significantly improves the performance in cross-modal retrieval. Our SCCGDH method uses three deep networks to learn the hash codes by utilizing the labels, image modality and text modality. Specifically, we construct a label network to learn the hash codes of each class center. In addition, we design an image network and a text network to generate the hash codes of image data and text data, respectively. Then we urge the hash codes of the image modality and the text modality to approach the hash codes of their corresponding centers learned from the labels. Our proposed SCCGDH method effectively reduces the semantic gap between different modality data. At the same time, the modality invariance loss is also used to eliminate the discrepancy of multi-modalities. Extensive experimental results indicate that our SCCGDH model is effective in cross-modal retrieval tasks.

The contributions of this work can be summarized as follows:

(1) We propose a unified neural network learning framework including the label network, image network and text network to learn three different hash codes. Moreover, these three networks can be optimized at the same time, and they can benefit from each other in the learning process.
(2) In our proposed network, the hash codes of the labels generated from the label network are used for the class-specific centers and effectively guide the hashing learning of the image and text modalities. In other words, the hash codes learned from the image network and the text network are forced to be close to the corresponding class-specific centers. Therefore, it can effectively reduce the intraclass variation of the hash codes of image modality and text modality in the same category.
(3) The specific class-centric based hashing method can better solve the multi-label dependency problem, resulting in better performance on multi-label datasets. Extensive experiments on three benchmark datasets demonstrate the effectiveness of our SCCGDH approach under different hash code lengths in real applications.

2. Related work

In this section, we briefly introduce the related works from two aspects, i.e., traditional hashing approaches and deep hashing approaches.

2.1. Traditional hashing methods

The traditional hashing approaches are mainly to transform multi-media data into a common representation using the linear projection algorithms, which are binarized to generate their hash codes. Cross-view hashing (CVH) aims at learning a hash function that maps similar samples to similar codes in Hamming space across the views. Xue et al. [13] proposed a supervised matrix factorization hashing, which preserves the correlation between modalities in the original space. Song et al. [14] introduced the inter-media and the intra-media consistency to learn the common representation in Hamming

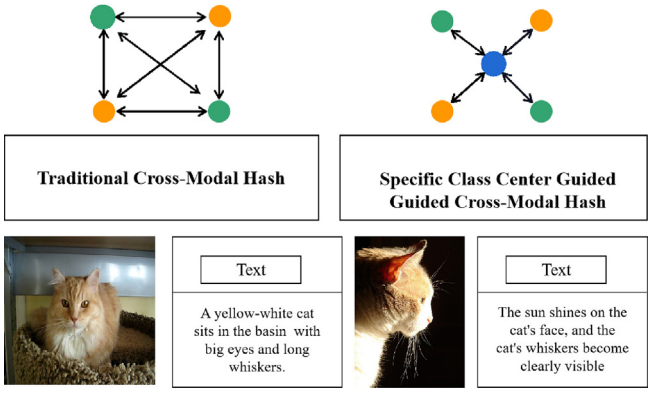

Fig. 1. The difference between our SCCGDH method and traditional hashing methods. The hash codes of both image modality and text modality from the same category are close to the same semantic center from the labels. Green points: hash codes extracted from images. Yellow points: hash codes extracted from texts. Blue points: hash codes learned from labels.

space, in which multi-media data can be consistently connected and represented. Wang et al. [15] proposed to learn hash codes for multimodal data and unimodal hash codes to maintain their common specific properties, simultaneously. Matrix Tri-Factorization Hashing Framework (MTFH) [16] jointly learns different length hash codes and two semantic association matrices to ensure comparability between different modal data. Semantic correlation maximization (SCM) [17] was proposed to naturally embed the label information into the model and guide the optimization without modifying the parameters in the optimization process. In addition, it avoids computing the semantic similarity matrix in the large-scale dataset. Discrete Semantic Alignment Hashing (DSAH)[18] uses collaborative constraints to discover the relationship between the labels and the hash codes to reduce semantic differences. Tang et al.[19] proposed to take full use of the label consistency across different modalities and the manifold geometric structure of each modality, simultaneously. Adaptive Label Correlation Based Asymmetric Discrete Hashing(ALECH)[20] learns the hash codes and the hash functions by utilizing the correlations of labels and preserving pairwise semantic similarity. Multi-hash codes joint learning(MOON)[21] attempts to use labels and cross-modal information to learn multiple length hash codes, simultaneously, and then optimizes the model as a whole.

2.2. Deep Hashing Methods

With the success of deep learning, many deep hashing approaches have emerged in cross-modal retrieval applications in recent years. Deep Cross-Modal Hashing (DCMH) [22] combines hash code learning and feature learning into the same framework for each modality. Autoencoder Hashing (CAH) [3] was proposed to learn discriminative and compact binary codes based on deep autoencoders. Adversarial Cross-Modal Retrieval (ACMR) [23] seeks a common subspace based on adversarial learning. Self-Supervised Adversarial Hashing (SSAH) [12] proposed a self-supervised semantic network, which aims to integrate adversarial learning into cross-modal hashing in a self-supervised manner. Modality-Invariant Asymmetric Networks for Cross-Modal Hashing (MIAN)[24] proposed to learn the internal pairwise similarity of each modality through probabilistic asymmetric learning and use a bottleneck information discrimination module to distinguish two modal information from texts and images, generating more accurate hash codes. Deep Supervised Cross-modal Retrieval (DSCMR) [25] learns to identify features while maintaining modal invariance by minimizing the identification of the label space and also utilizing weight sharing to further eliminate modal differences. Su et al. [26] proposed to construct a new joint semantic matrix that combines the original information of cross-modal data to obtain its intrinsic semantic relevance. Then a reconstruct network is trained to generate hash codes by maximizing the reconstruction of the joint-semantics relations. Hu et al. [27] proposed a triple fusion network to deal with paired data and unpaired data. It explores their association using manifold learning and uses the idea of GAN to train the network. Tu et al. [28] put forward to jointly learn unified hash codes for image-text pairs and a pair of hash functions for unseen query image-text pairs. Semantic ranking structure preserving (SRSP) [29] exploits the potential semantic associations by fully utilizing the label information, and preserves the semantic structure of the common representation using correlation ranking constraints, simultaneously. Graph Convolutional Hashing (GCH) [30] seeks to guide feature encoding through semantics, and thus obtains richer features by using GCH.

3. Formulation

3.1. Symbols and Problems Explanation

This paper uses bold uppercase letters (such as $\boldsymbol{W}$) to indicate matrix $\boldsymbol{W}_i$ represents the i-th column of $\boldsymbol{W}$. $\boldsymbol{W}_{ij}$ denotes the $|(i,j)$ elements of W and W^T denotes the transpose of the matrix W. The element-level symbolic function $sign(\cdot)$ is given as follows:

$$sign(x) = \begin{cases} 1, & if \ x > 0 \\ -1, & if \ x \leqslant 0 \end{cases} \tag{1}$$

Suppose we select n image-text pair samples from the database as training set, in which each pair contains two modality data. $O = \{o_i\}_{i=1}^{n}$ indicates a cross-modal dataset with n instances, and $o_i = (x_i, y_i, l_i)$, where x_i and y_i represent the image sample and text sample of the i-th instance o_i, respectively. $l_i = [l_{i1}, l_{i2}, \ldots, l_{ic}]^T$ is the label of o_i, where c is the class number. If o_i belongs to the j-th class, we set $l_{ij} = 1$, otherwise $l_{ij} = 0$. The pairwise similarity matrix $S \in \{-1, +1\}^{n \times n}$ is utilized to elaborate the semantic similarity between two modality data. Specifically, if two instances o_i and o_j are annotated by multiple label information at the same time, we set $S_{ij} = 1$ when o_i and o_j share one or more labels, otherwise $S_{ij} = -1$.

3.2. Network architecture for hashing

In this subsection, we present the proposed deep hashing framework in detail, which consists of three neural network structures, such as image network, text network and label network. Fig. 2 shows the overall framework of our proposed SCCGDH model.

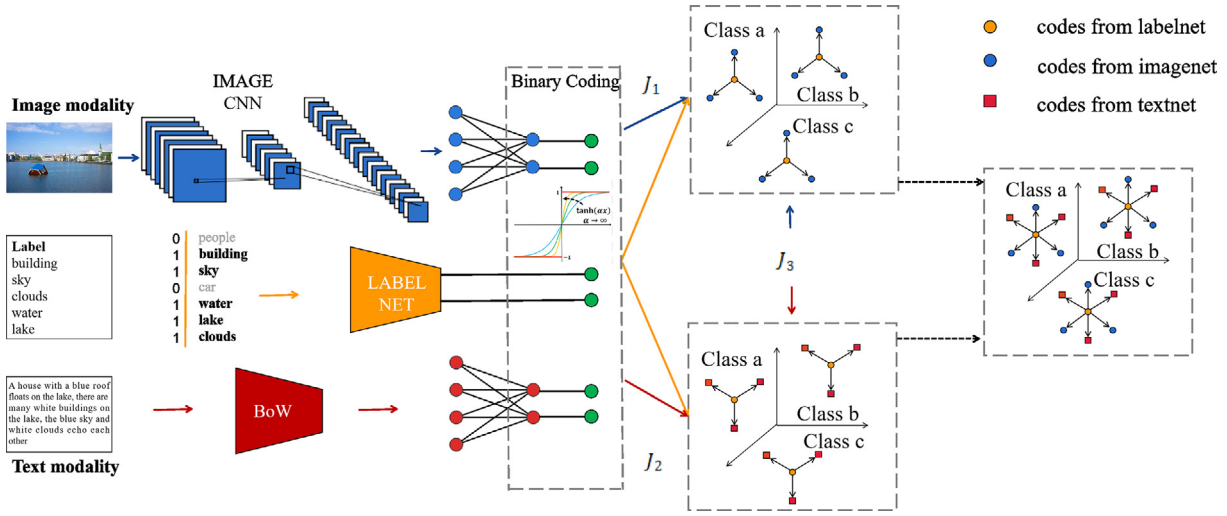

Fig. 2. The overall framework of our proposed SCCGDH approach.

(1) Image network architecture
In the image network architecture, we adopt a convolutional neural network (CNN) to extract the feature information of image samples and then generate the hash codes of image modality. Here, the image network modified from Alexnet is used as the hash function to directly learn the hash codes of image modality. Specifically, the image network includes eight layers. The first seven layers are consistent with Alexnet, and the last layer is replaced by a fully connected layer with k hidden units. It is obvious that the output of the image network is hash codes with length k.

(2) Text network architecture
In the text network, we employ the bag-of-words model to extract the feature information of the text modality, and then add two fully connected layers with 4096 and k hidden units at the end of the network, respectively. Therefore, the text network can generate the hash codes of length k from the text modality. The activation functions of two fully–connected layers employ the rectified linear unit (ReLU) and $\tanh(\alpha x)$, respectively.

(3) Label network architecture
In the label network, the multilayer perceptron (MLP) is used to generate hash codes from labels. Specifically, the MLP in the label network includes two full connection layers. The first layer adopts the ReLU as the activation function and the second layer employs $\tanh(\alpha x)$ as the activation function. Obviously, in our SCCGDH method, similar labels can generate similar hash codes using the label network. Therefore, our proposed method is more conducive to dealing with multi-label tasks compared with other methods using the center loss function.

3.3. Loss Function

The binary codes of the image modality, the text modality and the labels are learned from the corresponding hashing network. Specifically, the hash codes generated from the image network are denoted as $U = \{u_1, u_2, \ldots, u_m\}$, where $u_i \in \{1, -1\}^k$. The hash codes generated from the text network are denoted as $T = \{t_1, t_2, \ldots, t_m\}$, where $t_i \in \{1, -1\}^k$ and the hash codes generated from the label network are expressed as $V = \{v_1, v_2, \ldots, v_m\}$, where $v_i \in \{1, -1\}^k$.

To construct the loss function based on semantic similarity, we ensure that the hash codes mapped by similar samples are with a small Hamming distance and the samples with dissimilar semantics are mapped to the hash codes with a large Hamming distance. However, it is difficult to directly optimize the Hamming distance due to the complexity of computation. Consequently, we calculate the inner product of hash codes instead of its Hamming distance. As a result, the loss function can be expressed in the following form:

$$\min_{U,T} J(U,T) = \sum_{i=1}^{m}\sum_{j=1}^{m}\left(u_i^T \mathbf{t}_j - S_{ij}\right)^2$$

$$\begin{aligned} s.t. \quad & U \in \{-1,+1\}^{k\times m}, u_i = h_1(x_i), \forall i \in \{1,2,\ldots,m\} \\ & T \in \{-1,+1\}^{k\times m}, \mathbf{t}_j = h_2(y_j), \forall j \in \{1,2,\ldots,m\} \end{aligned} \tag{2}$$

where $h_1(\cdot)$ and $h_2(\cdot)$ stand for the hash functions of image modality and text modality, respectively. Here, we use image and text modality data as an example and thus the hash codes learned from the two modalities are considered. From Eq. (2), we can see that the variance of the samples from the same modality may be large, while the variance of samples from different modalities may be small. Therefore, it may be inconsistencies between the learned hash codes and the semantic labels, which leads to performance degradation.

In this work, we introduce another hash function $h_3(\cdot)$ considered as label network. the output from the label network is the binary codes, which are defined as the center of each class. Its goal is to guide the hashing learning of image and text modalities. To reduce the intra-class variation of image and text samples from the same category, the hash codes from $h_1(\cdot)$ and $h_2(\cdot)$ are forced to be close to specific–class centers. Therefore, we integrate these three hashing networks into a unified framework and can optimize them at the same time. The objective function is defined as

$$\begin{aligned}
&\min_{U,T,V} J(U,T,V) && = J_1 + J_2 \\
&&& = \sum_{i=1}^{m}\sum_{j=1}^{m}\left(u_i^T v_j - S_{ij}\right)^2 + \sum_{\mathrm{h}=1}^{m}\sum_{j=1}^{m}\left(\mathbf{t}_{\mathrm{h}}^T v_j - S_{hj}\right)^2 \\
&s.t. \quad U \in \{-1,+1\}^{k\times m}, T \in \{-1,+1\}^{k\times m}, \\
&\quad V \in \{-1,+1\}^{k\times m} \\
&\quad u_i = h_1(x_i), \mathbf{t}_{\mathrm{h}} = h_2(y_{\mathrm{h}}), \\
&\quad v_j = h_3(l_j), \forall i,h,j \in \{1,2,\ldots,m\}
\end{aligned} \tag{3}$$

We denotes $h_1(x_i) = \mathrm{sign}(\mathrm{F}_1(x_i;\Theta_1)), h_2(y_{\mathrm{h}}) = \mathrm{sign}(\mathrm{F}_2(y_h;\Theta_2))$ and $h_3(l_j) = \mathrm{sign}\left(\mathrm{F}_3(l_j;\Theta_3)\right)$, respectively, where $\mathrm{F}_1(x_i;\Theta_1) \in \mathbf{R}^{\mathrm{k}}$ is the output of the image network for $x_i, \mathrm{F}_2(y_h;\Theta_2) \in \mathbf{R}^{\mathrm{k}}$ is the output of the text network for y_i, and $\mathrm{F}_3(l_j;\Theta_3) \in \mathbf{R}^{\mathrm{k}}$ is the output of the label network for l_j. Θ_1, Θ_2 and Θ_3 are the to-be-learnt parameters.

To effectively eliminate the discrepancy between multi-modality data, we try to minimize the distance of hash codes of the image-text pair representations. Therefore, the invariance loss between image modality and text modality is formulated as follows:

$$\min_{U,T} J(U,T) = J_3 = \sum_{\mathrm{i}=1}^{m}\sum_{\mathrm{h}=1}^{m}\eta\left(h_1(x_i) - h_2(y_{\mathrm{h}})\right)^2 \tag{4}$$

where η denotes the hyper-parameter and it aims to control the contribution of loss function J_3 in the total loss function. However, in backpropagation, the corresponding gradient of the sign function is zero for all non-zero inputs, which may lead to the gradient vanishing problem. We refer to the HashNet and employ the scaling tanh function to alleviate this problem. Therefore, the total loss function of our SCCGDH method is given as follows:

$$\begin{aligned}
\min_{\Theta_1,\Theta_2,\Theta_3} J(\Theta_1,\Theta_2,\Theta_3) \quad &= \sum_{i=1}^{m}\sum_{j=1}^{m}\left(h_1(x_i)^T h_3(l_j) - S_{ij}\right)^2 \\
&+ \sum_{\mathrm{h}=1}^{m}\sum_{j=1}^{m}\left(h_2(y_{\mathrm{h}})^T h_3(l_j) - S_{hj}\right)^2 \\
&+ \sum_{\mathrm{i}=1}^{m}\sum_{\mathrm{h}=1}^{m}\eta\left(h_1(x_i) - h_2(y_{\mathrm{h}})\right)^2 \\
s.t. \quad h_1(x_i) &= \tanh\left(\alpha(F_1(x_i;\Theta_1))\right) \in [-1,+1]^{\mathrm{k}\times\mathrm{m}}, \alpha \in \mathbf{R} \\
h_1(y_{\mathrm{h}}) &= \tanh\left(\alpha(F_2(y_{\mathrm{h}};\Theta_2))\right) \in [-1,+1]^{\mathrm{k}\times\mathrm{m}}, \alpha \in \mathbf{R} \\
h_3(l_j) &= \tanh\left(\alpha(F_3(l_j;\Theta_3))\right) \in [-1,+1]^{\mathrm{k}\times\mathrm{m}}, \alpha \in \mathbf{R}
\end{aligned} \tag{5}$$

Due to the non-smooth and non-derivable properties, the sign function cannot use backpropagation to train the deep networks. Similar to HashNet, we design a smooth objective function to alleviate this problem and thus can learn the continuous hash codes instead of the discrete hash codes. Then we integrate the optimization of the hash codes into the neural network. Therefore, we can effectively improve the performances in cross-modal retrieval performances.

As the training phase, the encoding function in formula (1) is replaced by $\tanh(\alpha x)$ function with increasing α. When the values of α approach infinity, the optimization problem converge to the original deep learning to hash problem with sgn(.) activation function. This idea is evolved from the following formula:

$$\lim_{\alpha\to\infty}\tanh(\alpha x)=\operatorname{sgn}(x). \tag{6}$$

Fig. 2 shows this alternative part of the model. Generally, a series of smooth optimization problems are generated from the tightening tanh function and they can converge to the original intractable binarization problem with the increase of the scaling parameter α.

After training model (6), three different hash codes are generated from the image network, text network and label network, respectively. Specifically, the hash functions $h_1(\cdot)$ and $h_2(\cdot)$ can map the images or the text into the Hamming space when the image samples or the text samples are available, and the hash function $h_3(\cdot)$ can project the word vectors into Hamming space when the labels are available. In practical applications, we use the hash function obtained from the label network to hash them when the label of the sample is available. For the image or text query samples without the label information, we can use these samples as the input and their hash codes are generated by the image network or text network.

3.4. Connection

In this subsection, we further illustrate the relationship between our method and the center loss function. Our loss function is designed from the center loss function[31] that is given as the following formula:

$$\mathscr{L}=\frac{1}{2}\sum_{i=1}^{m}\left\|f(x_i)-c_{y_i}\right\|_2^2 \tag{7}$$

where $f(x_i)$ is the representation of the input data, c_{y_i} is the center representation of the y_i-th class. m is the total number of data. By minimizing the center loss function (7), these data can be as close as possible to their corresponding classes, and it can minimize the inter-class differences. In our proposed method, the hash codes extracted from the label network correspond to the class centers using the center loss function, and the hash codes learned from the image and text networks are guided to be close to the corresponding class centers.

Traditional center loss function cannot be applied to multi-label tasks since each sample can only be assigned to one label. In contrast, our proposed method learns similar class centers from label information using the label network and can handle multi-label tasks efficiently.In our proposed method, the hash codes generated by the label network are consistent with the original labels. Meanwhile, we employ them to guide the image and text networks to learn their hash codes, which are encouraged to be close to specific class centers. In this way, the hash codes of different modal data from the same category can be guided to the same class centers. Therefore, our proposed SCCGDH approach can learn more accurate hash codes and can further reduce the heterogeneous gap between multimodal data.

4. Optimization

In this subsection, we introduce the learning procedure of the parameters Θ_1, Θ_2 and Θ_3 by optimizing the image network, text network and label network, respectively. The optimization scheme of the model (5) is summarized in Algorithm 1, and its derivation is represented in detail as follows:

Algorithm 1: SCCGDH

Training Stage
Input: Given a cross-modal dataset $O=\{X,Y,L\}$, where X denotes the image modal data, Y denotes the text modal data and L is corresponding label information; batch size N ($N=16$) and similar matrix S.
Output: Parameters Θ_1 for image network, parameters Θ_2 for text network and parameters Θ_3 for label network; hash codes generated from the three networks.
Procedure:
Initialize parameters Θ_1, Θ_2 and Θ_3.
Repeat
(1) Randomly choose N image, text and label samples to construct a mini-batch for three networks;
(2) Compute the binary codes by forward-propagating them via three networks.
(3) Update the parameters of three networks via Eqs. (8)–(10).
Until *convergence*;
Testing Stage
Input: Image query sample or text query sample; the parameters of three network architecture.
Output: Hash codes of image query sample or text query sample.
Procedure:
(1) The hash codes of the query sample is obtained by forward-propagating it though the image network or text network.
(2) The corresponding hash codes are calculated by Eq. (11) or (12).

4.1. Learning Θ_1 for the image network

During the optimization process, we employ the mini-batch stochastic gradient descent (SGD) method with back-propagation scheme to learn the parameters Θ_1 of image network. In each training process, we randomly select a batch of image samples to iteratively optimize the image network. For simplicity, we define $\tilde{u}_i = F_1(x_i, \Theta_1), \tilde{t}_h = F_2(y_h, \Theta_2), \tilde{v}_j = F_3(l_j, \Theta_3)$. For each image instance x_i in X, the gradient of objective function can be calculated as follows:

$$\frac{\partial J}{\partial \tilde{u}_i} = 2\sum_{i=1}^{m} \tanh\left(\alpha \tilde{l}_j\right) \times \left(\tanh\left(\alpha \tilde{u}_i\right)^T \tanh\left(\alpha \tilde{l}_j\right) - S_{ij}\right)\left(1 - \tanh\left(\alpha \tilde{u}_i\right)^2\right) + 2\sum_{i=1}^{m}\left(\tanh\left(\alpha \tilde{u}_i\right) - \tanh\left(\alpha \tilde{t}_h\right)\right) \tag{8}$$

Then we use the chain rules to calculate $(\partial J/\partial \Theta_1)$ with $(\partial J/\partial \tilde{u}_i)$, and thus update the parameter Θ_1 using BP algorithm.

4.2. Learning Θ_2 for the text network

Similar to the image network, we also use a minibatch SGD with BP algorithm to learn the parameters Θ_2 of the text network. Thus, the gradient of the objective function for $\tilde{t}_h$ is calculated by the following form:

$$\frac{\partial J}{\partial \tilde{t}_h} = 2\sum_{h=1}^{m} \tanh\left(\alpha \tilde{l}_j\right) \times \left(\tanh\left(\alpha \tilde{t}_h\right)^T \tanh\left(\alpha \tilde{l}_j\right) - S_{hj}\right)\left(1 - \tanh\left(\alpha \tilde{t}_h\right)^2\right) + 2\sum_{h=1}^{m}\left(\tanh\left(\alpha \tilde{u}_i\right) - \tanh\left(\alpha \tilde{t}_h\right)\right) \tag{9}$$

Subsequently, we can calculate $(\partial J/\partial \Theta_2)$ through $(\partial J/\partial \tilde{t}_h)$ using the chain derivation rule, and further use BP algorithm to obtain the parameters Θ_2.

4.3. Learning Θ_3 for the label network

Similar to the setting of above two networks, the same optimization scheme is used to learn the parameters Θ_3 of the label network. Therefore, the gradient of the model for $\tilde{v}_j$ is given as follows:

$$\begin{aligned}\frac{\partial J}{\partial \tilde{v}_j} = &\sum_{j=1}^{m} \tanh\left(\alpha \tilde{u}_i\right) \times \left(\tanh\left(\alpha \tilde{v}_j\right)^T \tanh\left(\alpha \tilde{u}_i\right) - S_{ij}\right) \times \left(1 - \tanh\left(\alpha \tilde{v}_j\right)^2\right) + \sum_{j=1}^{m} \tanh\left(\alpha \tilde{t}_h\right) \\ &\times \left(\tanh\left(\alpha \tilde{v}_j\right)^T \tanh\left(\alpha \tilde{t}_h\right) - S_{ij}\right) \times \left(1 - \tanh\left(\alpha \tilde{v}_j\right)^2\right)\end{aligned} \tag{10}$$

Then we employ the chain derivation rule to obtain $(\partial J/\partial \Theta_3)$ through $(\partial J/\partial \tilde{v}_j)$, and learn the parameters Θ_3 by adopting the BP algorithm.

4.4. Out-of-Sample problem

Given a query sample without the label information, we can forward-propagate it by the image or text network to get its hash codes. The mathematical description can be expressed as follows:

$$b_i = \text{sign}(\tanh(\alpha(F_1(x_i); \Theta_1))) \tag{11}$$

$$b_j = \text{sign}(\tanh(\alpha(F_2(y_j); \Theta_2))) \tag{12}$$

5. Experiments

To evaluate the effectiveness of our proposed method, we conducted extensive experiments on three multi-modal datasets, such as MIRFLICKR-25 K [32], MS COCO [33] and NUS-WIDE [34]. We give a brief introduction to the three datasets as well as the evaluation metrics. To make a fair comparison, some state-of-the-art hashing methods are used as the baseline methods.

5.1. Data Sets

MIRFLICKR-25 K dataset consists of 25,000 instances collected from the Flickr website. We randomly selected 20,015 image-text pairs with at least 20 provided labels. Word vectors are extracted from text using the bag-of-words model and thus each text can be represented as a 1386 dimensional vector. 2243 instances were randomly picked out as testing dataset, and the rest is used as database set, in which we picked out 5000 instances to train the networks.

NUS-WIDE dataset is a multi-label dataset from 81 concepts. In the cross-modal retrieval task, we selected the 21 most common concepts with 195834 images with textual tags. We extracted 1000-dimensional features of text modality data using the bag-of-words model as the input. Here, we randomly selected 2085 instances as the testing set and the resting

were as the database set. Similarly, we randomly selected 21,000 instances from the database set as the training set to train the network.

MS COCO dataset contains 123287 instances with one or more of the 80 label concepts. We employed the bag-of-words model to extract a 2000-dimensional feature for each text. In this experiment, 7762 image-text pairs were randomly selected as the testing set, and the rest instances were used as the database set, in which we randomly selected 18000 instances to train our proposed deep network.

5.2. Evaluation metrics

To verity the effectiveness of our proposed SCCGDH method, two widely-used evaluation metrics based on Hamming distance ranking, such as mean average precision (MAP) and TopN-precision, are used to measure the query results with cosine similarity. In our experiments, we reported the results of all hashing approaches for two retrieval tasks: 1) Using the query image to retrieve the text samples (Image2Text) and 2) Using the query text to retrieve the image samples (Text2Image).

(1) MAP is widely used to evaluate the retrieval performance. The definition of the average precision(AP) is given as follows:

$$\mathrm{AP}(x_r) = \frac{1}{N}\sum_{r=1}^{R} P(r)\delta(r)g \tag{13}$$

where N is the number associated with the query instance in the database. $P(r)$ is the precision of the first r instances retrieved. If the return instance is similar to the query instance, we set $\delta(r) = 1$, otherwise $\delta(r) = 0$. R is the number of the retrieved samples from the database. MAP is used to average precision of all query samples. The definition of MAP is represented as follows:

$$\mathrm{MAP} = \frac{1}{Q}\sum_{q=1}^{Q} \mathrm{AP}(x_q) \tag{14}$$

where Q denotes the number of all query samples.

(2) TopN-precision is the average ratio of similar instances among top N return samples for all query samples. In our experiments, we set N = 5000.

5.3. Baseline and Implementation Details

Our proposed SCCGDH method is compared with seven hashing methods. Specifically, they include five shallow hash methods, i.e, IMH, CVH, CCA-ITQ[35], SMFH, STMH[36] and five deep hashing methods, i.e, DJRSH[26], DCMH, SSAH, DCHUC[28] and MIAN. The codes for these methods were provided by the authors. The shallow hashing methods were implemented with MATLAB, and the deep hashing methods are implemented with PyTorch. To ensure a fair comparison, we employed Alexnet to extract deep features as the input of the shallow hashing methods.

In our proposed method, we adopted the modified Alexnet[37] model pre-trained on ImageNet dataset [38] as the image network and used MLP as the label network to learn its corresponding hash codes. The text network includes the bag-of-words model and two full connection layers. The detailed configuration can be seen in III.B. In addition, the SGD optimizer with 0.9 momentum and 0.0005 weights were used to optimize the proposed framework. The learning rates of the image network, text network and label network were set to 0.001, 0.01 and 0.01, respectively. The hyper-parameter η was set to 0.1. In this experiment, we run our proposed SCCGDH approach and other deep hashing approaches on NVIDIA GeForce RTX 3060 GPU.

5.4. Results and Discussion

In this subsection, we conducted some retrieval experiments to evaluate our SCCGDH approach and other comparison methods.

Table 1 shows the retrieval results of all hashing methods in Text2Image and Image2Text tasks on the three datasets. From Table 1, it can be seen that the retrieval accuracy of the proposed SCCGDH approach outperforms other baseline methods with hash lengths varying from 16 to 128 bits. Specifically, compared with the best performance among the baseline methods on 128 bits, the MAP values of our proposed SCCGDH method can be improved by 2.66% and 2.42% in Text2Image and Image2Text tasks on the MIRFLICKR-25 K dataset, respectively. In addition, by setting the length of hash codes to 128 bits, the map values of our SCCGDH method can be improved by 4.02% and 3.27% in Text2Image and Image2Text compared with the best competitor on the MS COCO dataset, respectively. On the NUS-WIDE dataset, the performances of our SCCGDH approach are 2.12% and 3.28% higher than the best performances among other methods in Text2Image and Image2Text tasks when the hash length is 128bit. Moreover, it is obvious to see that our SCCGDH approach achieves the best performances in two retrieval tasks regardless of the hash code length setting.

From the experimental results on three multi-modal datasets, we can get the following observations:

Table 1
The mAP@50 results of our method and other baseline methods varied with different hash codes lengths on three datasets. (U respresent unsupervised method, S respresent supervised method)

Task	Methods	MIRFLICKR-25 K				MSCOCO				NUS-WIDE			
		16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
Text2Image	IMH(U)	0.6276	0.6471	0.6576	0.6302	0.4313	0.4490	0.4400	0.4256	0.5697	0.6374	0.7028	0.7015
	CVH(S)	0.6151	0.6263	0.6311	0.6842	0.4267	0.4340	0.4262	0.4634	0.6540	0.6578	0.6764	0.6637
	CCQ(S)	0.7509	0.7594	0.7708	0.7598	0.4006	0.4124	0.4106	0.4256	0.5221	0.6006	0.5817	0.5856
	SCM(S)	0.6853	0.6909	0.7433	0.7510	0.4296	0.4383	0.4614	0.4713	0.6556	0.6898	0.7015	0.7125
	STMH(U)	0.6360	0.6457	0.6617	0.6796	0.4147	0.4197	0.4191	0.4188	0.6195	0.6942	0.7449	0.7532
	SMFH(S)	0.5499	0.5482	0.5522	0.5550	0.4452	0.4438	0.4436	0.4435	0.3142	0.3307	0.3441	0.3560
	DJRSH (U)	0.7826	0.8143	0.8262	0.8467	0.6909	0.7606	0.8118	0.8380	0.7531	0.7652	0.7700	0.7793
	DCMH(S)	0.8092	0.8355	0.8293	0.8384	0.5723	0.5916	0.5986	0.6300	0.6835	0.6982	0.7130	0.7219
	SSAH	0.8081	0.8341	0.8413	0.8538	0.5839	0.6203	0.6323	0.6506	0.6894	0.6858	0.6869	0.6952
	DCHUC(S)	0.8177	0.8296	0.8417	0.8465	0.5608	0.5903	0.6284	0.6453	0.6816	0.6726	0.6968	0.7422
	MIAN(S)	0.7445	0.7537	0.7949	0.8083	0.5503	0.6176	0.6523	0.7030	0.6579	0.6735	0.6738	0.7221
	SCCGDH(S)	**0.8204**	**0.8562**	**0.8716**	**0.8804**	**0.7741**	**0.7813**	**0.8315**	**0.8782**	**0.7669**	**0.7734**	**0.7722**	**0.8005**
Image2Text	IMH(U)	0.5919	0.5951	0.5952	0.6049	0.4316	0.4293	0.4292	0.4074	0.5307	0.5957	0.6620	0.6736
	CVH(S)	0.5981	0.5988	0.6190	0.6002	0.3611	0.3497	0.3587	0.3805	0.6202	0.6360	0.6358	0.6350
	CCQ (S)	0.7814	0.7939	0.8043	0.8022	0.3504	0.3931	0.4120	0.5472	0.5955	0.6823	0.7303	0.7122
	SCM (S)	0.6800	0.6903	0.6972	0.7003	0.3697	0.3803	0.4161	0.3886	0.6587	0.6913	0.7072	0.7147
	STMH(U)	0.6212	0.6383	0.6496	0.6643	0.3106	0.4299	0.3910	0.4144	0.5575	0.6272	0.6483	0.6485
	SMFH(S)	0.5369	0.5390	0.5497	0.5489	0.4231	0.4236	0.4233	0.4234	0.2960	0.3146	0.3314	0.3424
	DJRSH(U)	0.8057	0.8420	0.8713	0.8870	0.7137	0.7587	0.7743	0.7897	0.7604	0.8160	0.8259	0.8384
	DCMH (S)	0.7911	0.8128	0.8288	0.8205	0.5549	0.6053	0.6085	0.6299	0.7076	0.7248	0.7272	0.7853
	SSAH(S)	0.8409	0.8607	0.8781	0.8728	0.6203	0.6231	0.6390	0.6460	0.7103	0.7240	0.7794	0.7785
	DCHUC(S)	0.8390	0.8236	0.8299	0.8404	0.6454	0.6138	0.6595	0.6687	0.7432	0.7626	0.7678	0.8202
	MIAN(S)	0.7101	0.7344	0.7301	0.7480	0.5266	0.5278	0.5280	0.5401	0.6105	0.6370	0.6226	0.6315
	SCCGDH(S)	**0.8817**	**0.8918**	**0.8926**	**0.9112**	**0.7317**	**0.7772**	**0.8084**	**0.8224**	**0.8124**	**0.8372**	**0.8401**	**0.8712**

(1) It is clear to see that the deep hashing methods achieve superior retrieval performances than traditional shallow hashing methods in most cases on three datasets. The main reason may be that deep learning methods can extract more essential features than traditional shallow methods, and effectively reduces the semantic gap between different modalities. This phenomenon is consistent with our understanding of deep learning.
(2) Compared with other deep hashing methods, the proposed SCCGDH approach also achieves more promising performances in Image2Text and Text2Image tasks on three datasets. It is worth noting that the former also adopts the similar pairwise loss function to our proposed SCCGDH method. However, the pairwise similar loss in the former is only used between the image modality and text modality. The pairwise similar loss in our proposed SCCGDH method is used to measure the image modality and labels or text modality and labels. Therefore, the hash codes of the image modality and text modality are close to the centers of the labels. Thus, the intraclass variation of the image hash codes and text hash codes from the same class can be significantly reduced.
(3) It can be found that our SCCGDH approach outperforms other baseline approaches on the MS COCO dataset. This is because the MS COCO dataset can provide more label categories than the other two datasets. Moreover, the class-specific centers learned from the label network in our SCCGDH proposed method can effectively utilize multi-label dependencies and thus achieves superior performances in multi-label tasks.
(4) Compared with SSAH, our proposed method shows more superiority on different retrieval tasks. Specifically, both our SCCGDH method and the SSAH method construct a three-stream network framework to extract the hash codes of the images, texts and labels, simultaneously. Among them, SSAH desgins a semantically supervised network to learn and optimize the semantic features and binary features of images and texts, respectively, and uses the hash codes of the labels for adversarial learning to optimize the network. In addition, our proposed SCCGDH method generates the hash codes of the labels through the label network, which is highly similar to the label information. Therefore, we employ them to guide more accuary hash codes of the image modality and the text modality. The experimental results on three benckmark datasets also demonstrate the superiority of our proposed method.
(5) We can see that the supervised learning methods including DCMH, SSAH, DCHUC outperform the unsupervised learning method, such as DJRSH, on the MIRFLICKR-25 dataset, but DJRSH outperforms DCMH, SSAH, DCHUC on the COCO and NUS-WIDE datasets. However, it can be seen that our proposed SCCGDH method achieves the best performances on the three datasets. Therefore, it shows that our proposed SCCGDH model can more effectively utilize label information than other supervised learning methods.

In addition, we conducted some experiments to verify the lookup performances of various hashing methods, when hashing length is set to 128 bits. It is widely known that TopN-precision is also a popular metric of Hamming ranking. Therefore, we reported the TopN-precision of all hashing methods. Fig. 3 shows the TopN-precision curves of different hashing methods on three datasets. We can see that our SCCGDH approach achieves better retrieval performances than other compared hash-

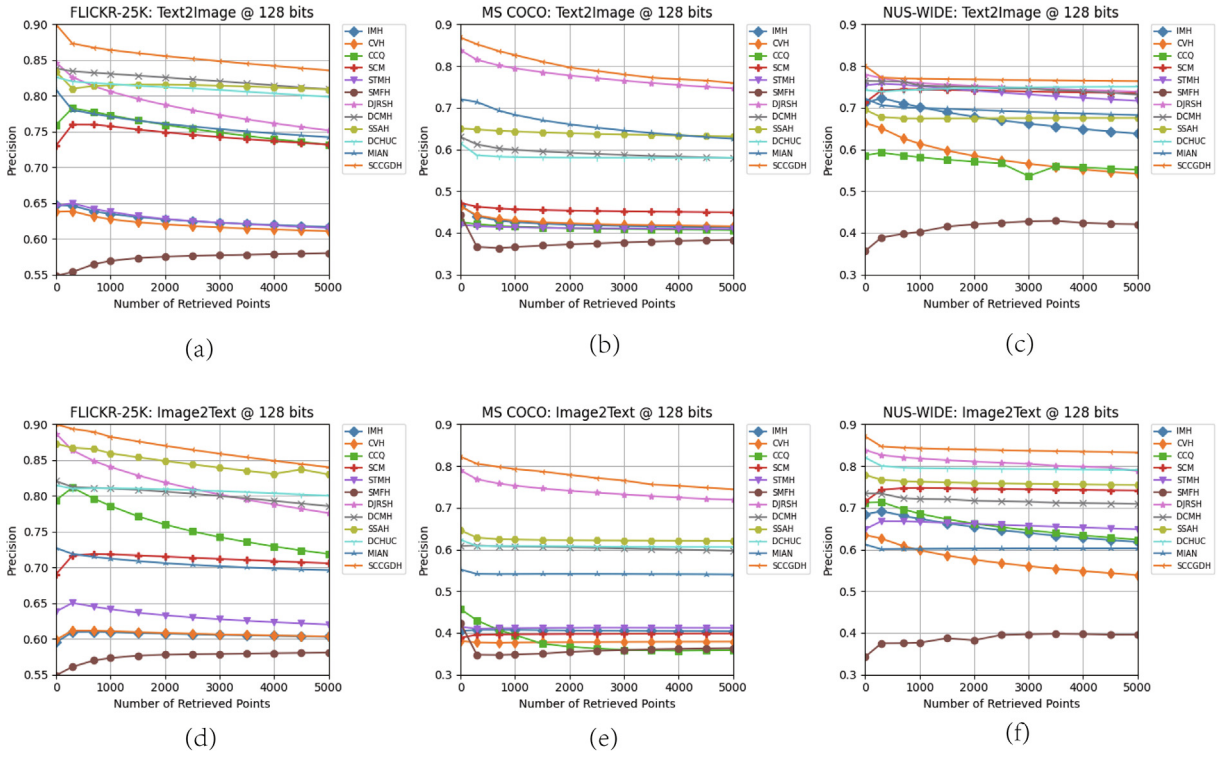

Fig. 3. TopN-precision curves of 128-bit hash code on FLICKR-25 K, MS COCO and NUS-WIDE.

ing methods in TopN-precision. This is consistent with our previously observed experimental results. Therefore, it can be found that our SCCGDH method has obtained higher retrieval performances than other hashing methods on these two metrics, which demonstrates its superiority in cross-modal retrieval applications. Besides, to explore the performance effect of different hash code lengths, the length of hash codes ranged from 16 bits to 128 bits. Fig. 4 and 5 show the MAP values of the proposed SCCGDH method under different hash code lengths. From Fig.4 and 5, we can intuitively see that the performances of our SCCGDH method are generally improved with the increase of hash code length on three benchmark datasets. The main reason is that longer hash codes usually contain more semantic information.

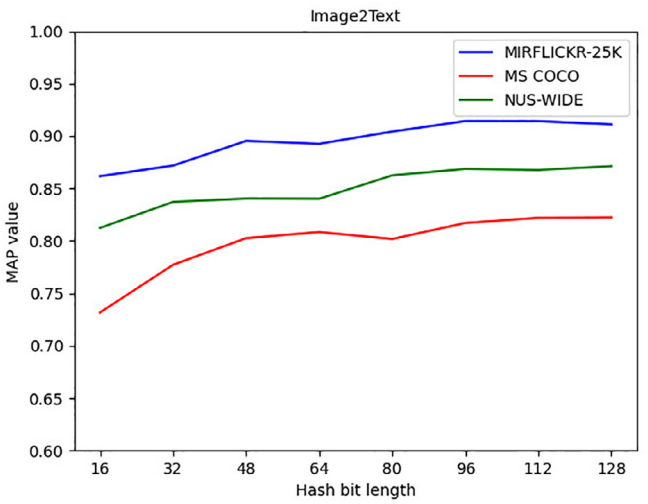

Fig. 4. MAP@50 of our SCCGDH method in image2text task on three datasets.

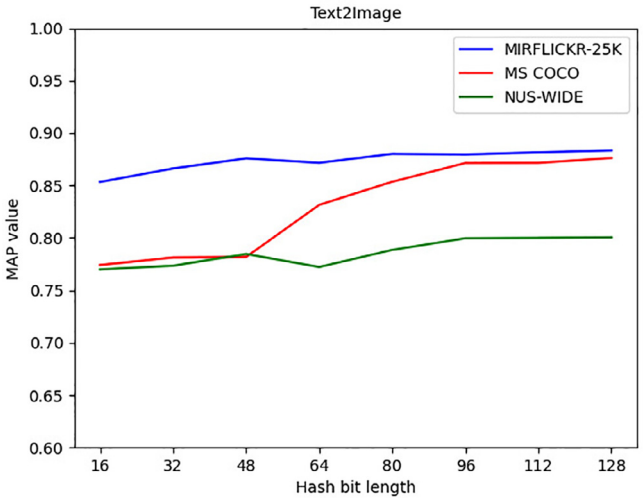

Fig. 5. MAP@50 of our SCCGDH method in Text2Image task on three datasets.

5.5. Convergence analysis

In this subsection, we carried out some experiments to verify the convergence of our proposed method on the MIRFLICKR-25 K when its hash code length is 128 bits. Fig. 6 shows that our proposed method tends to converge after about 25 times. Fig. 7 shows the results of our proposed method varies with the number of iterations in the different retrieval tasks. Therefore, it can be seen that our proposed method shows relatively good convergence in cross-modal retrieval.

5.6. Parameters analysis

In this subsection, we investigate the retrieval performance under different settings of the parameter η in our proposed model on three datasets. Fig. 8 shows the performances of our SCCGDH approach varied with different values of the parameter η. Specifically, the range of the parameter η is set between 10^{-3} and 10 and its hash code length is set to 128 bit. It can be seen that the retrieval results of our proposed SCCGDH method can keep relatively stable in a large range. Therefore, it can be verified that our SCCGDH method is not sensitive to the parameter η.

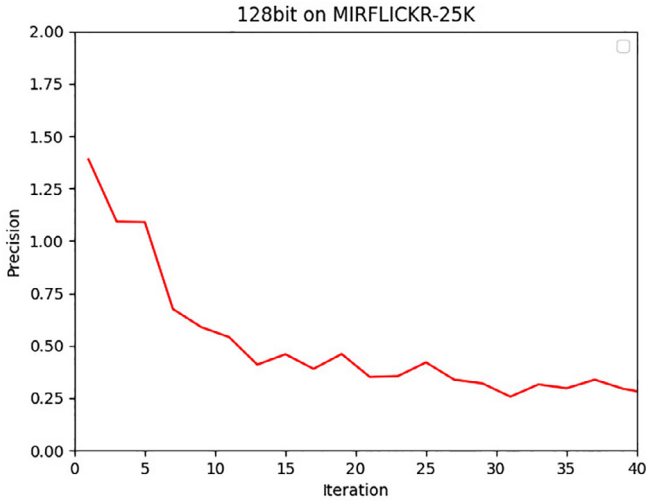

Fig. 6. Convergence curve.

Fig. 7. Accuracy Growth Curve.

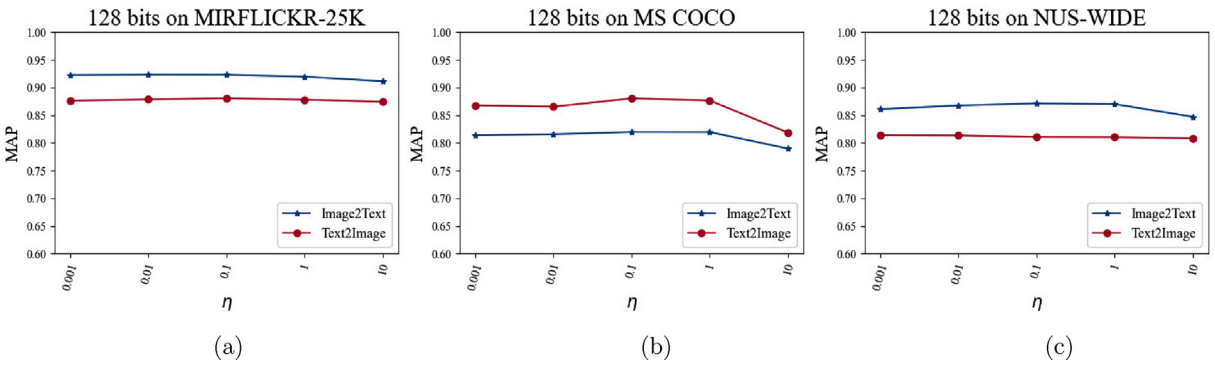

Fig. 8. MAP values with different values of the parameter η on three datasets.

5.7. Impact of different components

The loss function of the proposed SCCGDH method includes three parts. To further explore the role of three parts on the performance of the proposed method, we evaluated each part of SCCGDH, respectively. SCCGDH1 and SCCGDH2 only contain the loss functions J_1 and J_2, respectively, and SCCGDH3 contains the lost functions J_1 and J_2. We can adopt the scheme of the SCCGDH method to optimize these three variants.

Tables 2–4 show the MAP values of our SCCGDH approach and its variants on three multi-modal datasets. We can see that the SCCGDH3 method achieves more performances than both SCCGDH1 and SCCGDH2. It indicates that the retrieval performances can be improved by simultaneously forcing the hash codes of multi-modality data to approximate the class-specific centers from the label network. In addition, it can be observed that the full SCCGDH method outperforms the SCCGDH3 method without eliminating the discrepancy between the two modalities.

Table 2
MAP@50 values of the proposed SCCGDH method and its three variations on the MIRFLICKR-25 K dataset.

Method	Image2Text	Text2Image	Average
Full SCCGDH	**0.9287**	**0.8946**	**0.9117**
SCCGDH1	0.6554	0.6239	0.6397
SCCGDH2	0.5922	0.6315	0.6119
SCCGDH3	0.9165	0.8875	0.9020

Table 3
MAP@50 values of the proposed LDGH method and its three variations on the MS COCO dataset.

Method	Image2Text	Text2Image	Average
Full SCCGDH	**0.8224**	**0.8782**	**0.8503**
SCCGDH1	0.3369	0.3526	0.3448
SCCGDH2	0.2834	0.3802	0.3318
SCCGDH3	0.8210	0.8602	0.8406

Table 4
MAP@50 values of the proposed LDGH method and its three variations on the NUS-WIDE dataset.

Method	Image2Text	Text2Image	Average
Full SCCGDH	**0.8533**	**0.8298**	**0.8416**
SCCGDH1	0.5329	0.4567	0.4948
SCCGDH2	0.4927	0.4705	0.4816
SCCGDH3	0.8526	0.8021	0.8273

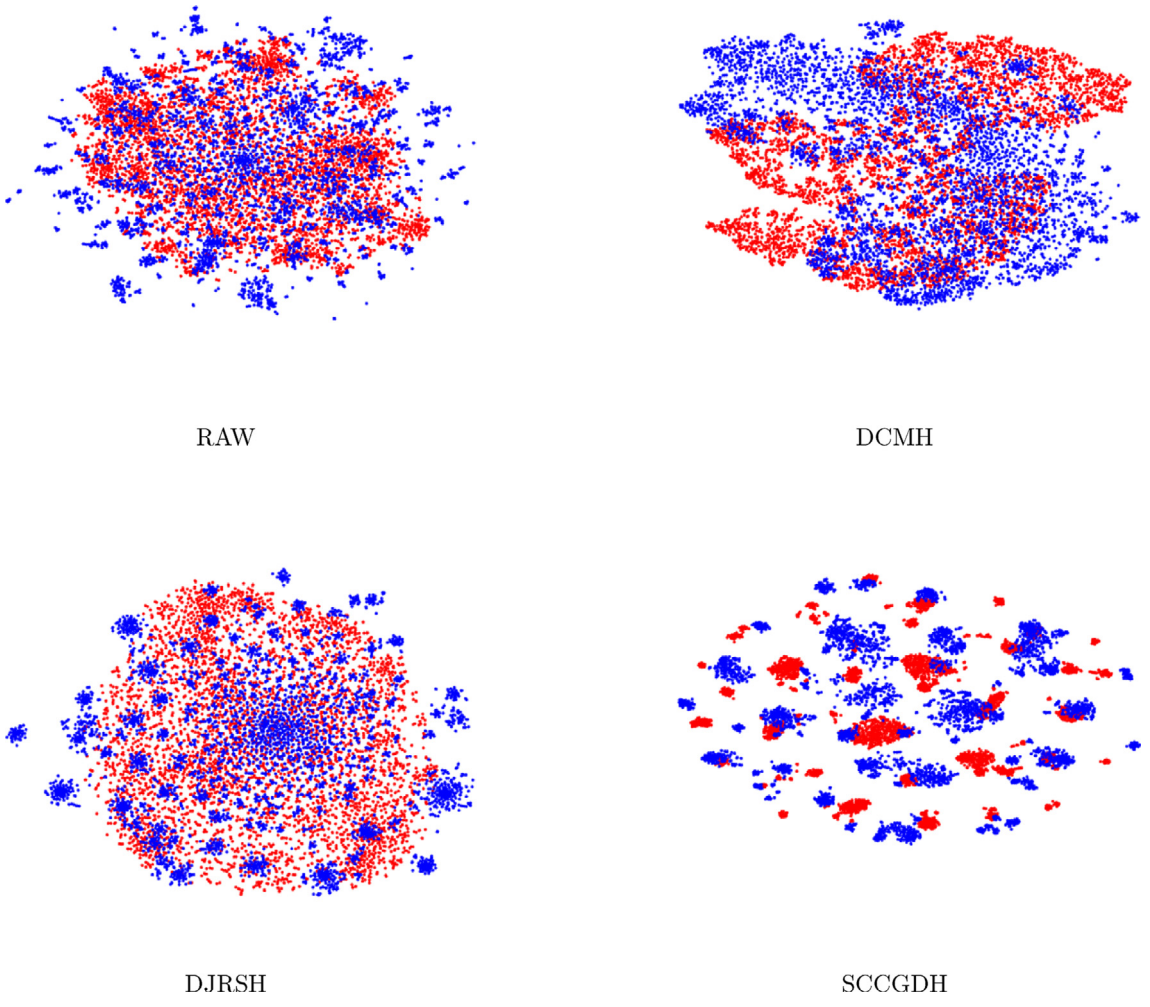

Fig. 9. Visualization results of different deep hashing methods on the MIRFLICKR-25 K dataset.

5.8. Visualization

In this subsection, we conducted some experiments to present the visualization results of our proposed SCCGDH method. Fig. 9 shows the visualization results of 128-dimensional hash codes of these methods on the MIRFLICKR-25 K dataset. The red and blue points stand for the hash codes from the image modality and text modality, respectively. From Fig. 9, we can see that the raw images and texts are scattered and have no connection. The DCMH method can carry out a preliminary classification of the text and the images, but it is still unable to classify them more accurately and meticulously. In addition, the DJRSH method cannot effectively classify the image modalities. It is clear to see that our proposed SCCGDH can accurately classify the text and image modalities, and generate more accurate hash codes. The visualization results have further confirmed the superiority of our proposed SCCGDH method.

6. Conclusion

In this work, we design a novel deep hashing learning framework, called Specific Class Center Guided Deep Hashing (SCCGDH), which consists of image network, text network and label network. Specifically, the label network can learn the class-specific centers, and the hash codes of the image network and text network are steered to be close to these centers. Three networks can be trained simultaneously, allowing them to be constrained by each other. Therefore, the proposed SCCGDH method can effectively reduce the intraclass variation of the hash codes between different modalities by introducing class-specific centers. Moreover, the invariance loss between multi–modalities are introduced to eliminate their discrepancy. Experimental results on three multi-modal datasets show that our SCCGDH approach achieves superior performances to other competitors in cross-modal retrieval applications.

CRediT authorship contribution statement

Zhenqiu Shu: Conceptualization, Methodology, Writing – review & editing. **Yibing Bai:** Software, Data curation, Writing – original draft. **Donglin Zhang:** Validation, Writing – review & editing. **Jun Yu:** Validation, Writing – review & editing. **Zhentao Yu:** Project administration, Supervision. **Xiao-Jun Wu:** Supervision, Validation.

Data availability

I have shared our code on Github.

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.

Acknowledgement

This work was supported by the National Natural Science Foundation of China [Grant No. 61603159, 62162033, U21B2027, U1836218], Yunnan Provincial Major Science and Technology Special Plan Projects [Grant No.202002AD080001, 202103AA080015], Yunnan Foundation Research Projects [Grant No. 202101AT070438, 202101BE070001-056].

References

[1] C. Deng, E. Yang, T. Liu, D. Tao, Two-stream deep hashing with class-specific centers for supervised image search, IEEE Transactions on Neural Networks and Learning Systems 31 (6) (2020) 2189–2201.
[2] X. Yao, D. She, S. Zhao, J. Liang, Y.-K. Lai, J. Yang, Attention-aware polarity sensitive embedding for affective image retrieval, in: Proceedings of International Conference on Computer Vision, 2019, pp. 1140–1150.
[3] Y. Cao, M. Long, J. Wang, H. Zhu, Correlation autoencoder hashing for supervised cross-modal search, in: Proceedings of the ACM on International Conference on Multimedia Retrieval, 2016, pp. 197–204.
[4] S. Kumar, R. Udupa, Learning hash functions for cross-view similarity search, in: Twenty-second International Joint Conference on Artificial Intelligence, 2011.
[5] W. Liu, C. Mu, S. Kumar, S.-F. Chang, Discrete graph hashing, Advances in Neural Information Processing Systems 27 (2014) 3419–3427.
[6] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, S.-F. Chang, Supervised hashing with kernels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2074–2081.
[7] G. Ding, Y. Guo, J. Zhou, Collective matrix factorization hashing for multimodal data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2075–2082.
[8] V. Erin Liong, J. Lu, Y.-P. Tan, J. Zhou, Cross-modal deep variational hashing, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4077–4085.
[9] Y. Shen, L. Liu, L. Shao, J. Song, Deep binaries: Encoding semantic-rich cues for efficient textual-visual cross retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4097–4106.
[10] 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, Vol. 31, 2017, pp. 1618–1628.

[11] Y. Cao, M. Long, J. Wang, Q. Yang, P.S. Yu, Deep visual-semantic hashing for cross-modal retrieval, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1445–1454.
[12] 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.
[13] F. Xue, W. Wang, W. Zhou, T. Zeng, T. Yang, Cross-modal retrieval via label category supervised matrix factorization hashing, Pattern Recognition Letters 138 (2020) 469–475.
[14] J. Song, Y. Yang, Y. Yang, Z. Huang, H.T. Shen, Inter-media hashing for large-scale retrieval from heterogeneous data sources, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2013, pp. 785–796.
[15] D. Wang, Q. Wang, L. He, X. Gao, Y. Tian, Joint and individual matrix factorization hashing for large-scale cross-modal retrieval, Pattern Recognition 107 (2020) 107479.
[16] X. Liu, Z. Hu, H. Ling, Y.-M. Cheung, 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.
[17] D. Zhang, W.-J. Li, Large-scale supervised multimodal hashing with semantic correlation maximization, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 28, 2014, pp. 2177–2183.
[18] T. Yao, X. Kong, H. Fu, Q. Tian, Discrete semantic alignment hashing for cross-media retrieval, IEEE Transactions on Cybernetics 50 (12) (2019) 4896–4907.
[19] J. Tang, K. Wang, L. Shao, Supervised matrix factorization hashing for cross-modal retrieval, IEEE Transactions on Image Processing 25 (7) (2016) 3157–3166.
[20] H. Li, C. Zhang, X. Jia, Y. Gao, C. Chen, Adaptive label correlation based asymmetric discrete hashing for cross-modal retrieval, IEEE Transactions on Knowledge and Data Engineering (2021).
[21] D. Zhang, X.-J. Wu, H.-F. Yin, J. Kittler, Moon: Multi-hash codes joint learning for cross-media retrieval, Pattern Recognition Letters 151 (2021) 19–25.
[22] Q.-Y. Jiang, W.-J. Li, Deep cross-modal hashing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3232–3240.
[23] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 154–162.
[24] L.Z.G.L.Z. Zhang, Y. Luo, H.T. Shen, Modality-invariant asymmetric networks for cross-modal hashing, IEEE Transactions on Knowledge and Data Engineering (2022).
[25] L. Zhen, P. Hu, X. Wang, D. Peng, Deep supervised cross-modal retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10394–10403.
[26] S. Su, Z. Zhong, C. Zhang, Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3027–3035.
[27] Z. Hu, X. Liu, X. Wang, Y.-M. Cheung, N. Wang, Y. Chen, Triplet fusion network hashing for unpaired cross-modal retrieval, in: Proceedings of the on International Conference on Multimedia Retrieval, 2019, pp. 141–149.
[28] R.-C. Tu, X.-L. Mao, B. Ma, Y. Hu, T. Yan, W. Wei, H. Huang, Deep cross-modal hashing with hashing functions and unified hash codes jointly learning, IEEE Transactions on Knowledge and Data Engineering (2020) 560–572.
[29] H. Liu, Y. Feng, M. Zhou, B. Qiang, Semantic ranking structure preserving for cross-modal retrieval, Applied Intelligence 51 (3) (2021) 1802–1812.
[30] R. Xu, C. Li, J. Yan, C. Deng, X. Liu, Graph convolutional network hashing for cross-modal retrieval., in: Ijcai, Vol. 2019, 2019, pp. 982–988.
[31] J.W.Z. Cao, M. Long, P.S. Yu, Hashnet: Deep learning to hash by continuation, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5609–5618.
[32] 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.
[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, European Conference on Computer Vision, Springer (2014) 740–755.
[34] 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.
[35] Y. Gong, S. Lazebnik, A. Gordo, F. Perronnin, Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (12) (2012) 2916–2929.
[36] D. Wang, X. Gao, X. Wang, L. He, Semantic topic multimodal hashing for cross-media retrieval, in: Twenty-fourth International Joint Conference on Artificial Intelligence, 2015, pp. 3890–3896.
[37] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems 25 (2012).
[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al, Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.