IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 July 2023, accepted 21 August 2023, date of publication 28 August 2023, date of current version 11 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3308931

RESEARCH ARTICLE

Self-Supervised Cluster-Contrast Distillation Hashing Network for Cross-Modal Retrieval

HAOXUAN SUN[1], **YUDONG CAO**[1], **AND GUANGYUAN LIU**[2]
[1]School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou 121001, China
[2]School of Computer Science and Technology, Dalian University of Technology, Dalian 1116024, China

Corresponding author: Yudong Cao (caoyd@lnut.edu.cn)

ABSTRACT Traditional cross-modal hash models enable efficient and fast retrieval between multimodal data by training high-quality hash representations. The key to the cross-modal hashing model is feature extraction. However, the quality of the features largely depends on the semantic similarity between the multi-modal data, and the existing methods do not effectively utilize the semantic information between the data. In this paper, we attempt to explore the semantic information inherent within the data using contrastive learning. Specifically, we propose a end-to-end cluster-level contrastive learning method (SCCDH) for cross-modal hashing. The method utilizes the clustering results to guide feature learning in an appropriately designed contrast framework. In SCCDH, feature-level and hash cluster-level contrastive learning are used to help the model learn discriminative features among multimodal data. In addition, we propose a distillation filtering method to filter out a large amount of noise in the data. Extensive experiments were conducted on the MIRFLICKR-25K, NUS-WIDE, and MS-COCO datasets, and the results demonstrate that the proposed method outperformed several state-of-the-art methods.

INDEX TERMS Hashing, cross-modal, contrastive learning.

I. INTRODUCTION

The goal of cross-modal retrieval is to utilize data from one modality (e.g., images) to retrieve data from another modality (e.g., text) that is semantically relevant to it. The key challenge in cross-modal retrieval is to reduce the heterogeneity of the differences between the modalities, while increasing the differentiation between different kinds of samples within a given modality. Many techniques have been proposed to overcome these difficulties, and relatively satisfactory performance has been achieved. However, the existing methods require large amounts of computer storage and high computational speed because they learn to obtain continuous features, which is difficult to realize due to the ever-increasing volume of large-scale data on the present-day network. Thus, we must develop methods to increase the efficiency of cross-modal retrieval.

Many cross-modal hashing methods [1], [2], [3], [4] have been proposed to reduce heterogeneity differences on large-scale data efficiently. Essentially, these methods project high-dimensional features into a binary Hamming space, thereby improving storage and computational efficiency. Most existing cross-modal hashing methods can be broadly classified as supervised and unsupervised. Supervised methods [5], [6], [7], [8] have achieved better performance by using label semantic information to connect data between different modalities, thereby making it easy to reduce heterogeneity differences. However, data annotation is a costly process, and it may be impractical to obtain sufficient data annotation for the huge volumes of available data. Unlike supervised methods, unsupervised cross-modal hashing methods [5], [9], [10] do not require large amounts of labeled semantic information.

Various computer vision tasks, and cross-modal hashing methods that use deep neural networks can learn both feature representations and hash functions in an end-to-end trainable architecture. In addition, cross-modal hashing based on a deep model can exploit non-linear correlations more effectively and achieve better performance than shallow networks. However, existing deep model based cross-modal

The associate editor coordinating the review of this manuscript and approving it for publication was Shu Xiao.

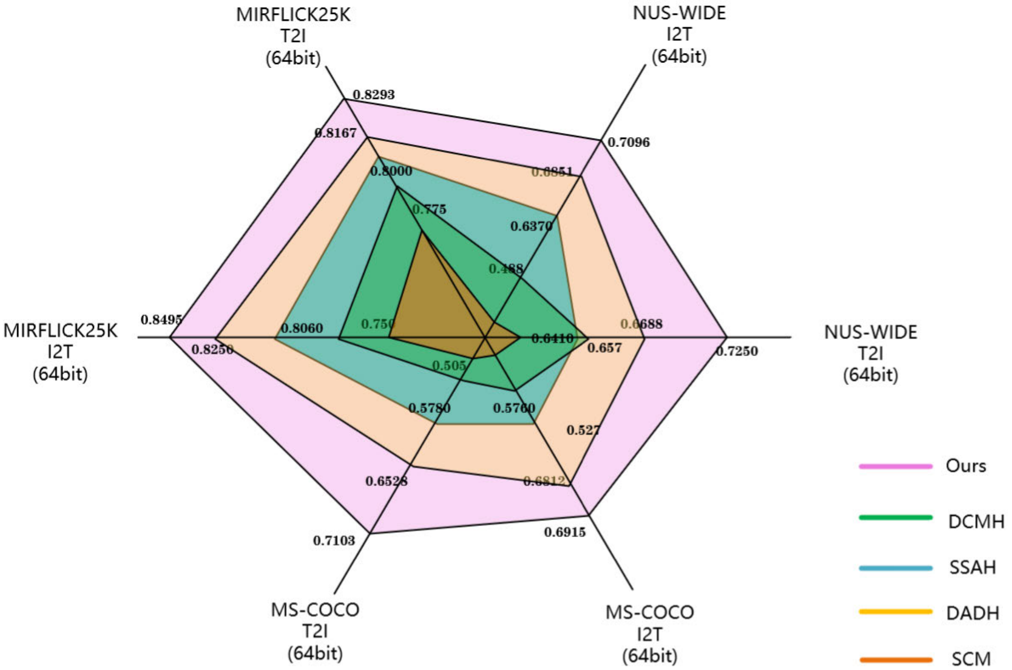


FIGURE 1. The proposed model achieves state-of-the-art performance at 64 bit compared to other customized or base models on three widely used datasets (I2T/T2I: image to text/text to image retrieval).

hashing methods still suffer from following weaknesses. Firstly, high-level representations are usually extracted from deep convolutional networks to represent different modalities. However, simply projecting the extracted features into Hamming space for similarity metric training lacks the exploration of high-level semantic information. Secondly, the model is usually trained using the maximum marginal loss, although this loss function performs well, it introduces a large number of learnable parameters and increases the burden of model training.

The recent rise of contrastive learning has received more and more attention because of its superior performance, and applying it to cross-modal hashing has pointed out a new direction to researchers. However, it is a difficult task. There are two serious challenges. First, contrastive learning is typically trained to optimize continuous features (either positive or negative samples), which runs against the binary values obtained from cross-modal hashing. As a result, it may be impossible to optimize. Alternatively, if optimization is possible, its performance may be reduced significantly. Second, the image-text pairs used for cross-modal hashing are generally captured from the network and are mostly filled with noise. Here, the text may contain words not represented in the image, and the image may contain objects not described in the text. To overcome these difficulties, we propose the self-Supervised cluster-contrast distillation hashing network (SCCDH) for cross-modal retrieval method that utilizes two-stage contrastive learning and momentum distillation to increase the intrinsic differentiation between samples while filtering out a large amount of useless noise to guide the cross-modal hash model to generate higher quality hash codes. In addition, to reduce the training time of the model under large-scale datasets, we use a clustering algorithm to cluster cross-modal hash features before performing contrastive learning, and replace the hash features with clustering centers, which are stored in a dynamic dictionary, thus, greatly reducing the memory occupation of the dynamic dictionary and improving the training speed. In this paper we also employ Transformers as text feature extractors and Vision Transfomers as image feature extractors, thereby exploiting the advantages of both local and global features. The Transformer model compensates for the shortcomings of CNNs by modeling the global representation of the input features at each step through a highly parallel architecture that captures the relative importance between local representations in the feature representation of the same input sequence using a self-attentive mechanism. Our primary contributions are summarized as follows:

- We introduce cluster cross-modal hash contrast, which trains, updates, and performs contrastive loss computation at the cluster level. It uses a unique cluster representation to address the problem of large memory occupation of dynamic dictionary. To the best of our knowledge, this is the first work that combines cluster contrastive learning with cross-modal hash modeling.
- We propose to use feature level contrast to assist cross-modal hash model training, which can effectively utilize the hidden information between different modalities.
- We propose a new cluster-level loss function for cross-modal contrast learning, which can effectively

reduce the distance between data of the same class and increase the distance between data of different classe.

- We also propose a distillation method to filter out a large amount of noise from the data and verify that the distillation method significantly helps the cross-modal hash contrast learning framework.

The rest of the paper is organized as follows. Section II briefly reviews related work. Section III describes the self-supervised cluster-contrast distillation hashing network (SCCDH) for cross-modal Retrieval. Section IV gives the details of the study of our SCCDH-experimental results, ablation studies and comparisons. Section V summarizes the whole paper.

II. RELATED WORK

The most critical difficulty of cross-modal hashing is the difference in heterogeneity, i.e., the similarity between different modalities cannot be measured directly. In this section, we review previous work in terms of supervised cross-modal hashing methods, unsupervised cross-modal hashing methods, deep cross-modal hashing methods, and contrastive learning methods. Supervised cross-modal hashing algorithms learn cross-modal hash functions with the help of label information, and such methods tend to exhibit higher retrieval accuracy than unsupervised cross-modal hashing algorithms. CMSSH [13] represents the hashing process as a binary classification problem with positive and negative pairs, and it preserves intra-class similarity through a classification paradigm with a raising approach. SCM [14] utilizes label information to construct a semantic similarity matrix to learn the maximum correlation between modalities. SePH [15] converts semantic affinity matrices into probability distributions while minimizing KL divergence to generate efficient binary codes. Note that most methods make use of tagging information; however, tagging information requires a lot of manual annotation, which is costly and defeats our original purpose of studying cross-modal hashing. Unsupervised cross-modal hashing algorithms typically map data from different modalities to a common Hamming space to maximize the correlation between them. without using the semantic information of the labels to learn the hash codes. CVH [16] was proposed to learn generic hash codes by exploiting intramodal and intermodal similarities. In addition, CMFH [17] learns hash codes in the public Hamming space through collective matrix decomposition, and UGACH [10] employs generative adversarial networks and association graphs to obtain cross-modal information about similar structures. ASSPH [61] uses semantic reconstruction matrices to exploit correlations in multimodal data. LSSH [18] utilizes sparse coding and matrix decomposition to obtain latent semantic information. Then, the potential semantic features obtained are mapped to a joint abstraction space to learn a uniform binary representation. The rise of deep learning has provided a new direction to study cross-modal hash retrieval. For example, deep cross-modal hashing (DCMH) [21] learns cross-modal similarity information through negative log-likelihood loss, and self-supervised adversarial hashing (SSAH) [22] improves the quality of hash codes through adversarial learning between the label similarity and the generated hash features. (AMSH) [42] enhances the distinction between latent representations and hash codes through adaptive matrices. However, although these methods utilize semantic similarity to learn high-quality hash codes, they do not focus on the underlying structure between the cross-modal features. As a result, these methods do not realize ideal performance.

Since being proposed, contrastive learning has attracted increasing attention from the community. For example, InstDisc [23] is based on the idea that differentiation between samples should come from their intrinsic inherent properties rather than labels. The increasing popularity of contrastive learning has led to a marked improvement in the performance of unsupervised learning methods and has renewed enthusiasm for exploring unsupervised learning. Inspired by the success of contrast learning, a number of contrastive hashing methods [24], [25], [26] have been proposed and these methods have demonstrated good performance in terms of learning unimodal binary hash codes. NSH [25] ranks the similarity between anchor samples using ranking loss and improved contrast loss to maximize the distance between positive and negative samples. In addition, CIMON [24] utilizes a new consistency loss function from the perspective of semantic matching and contrastive learning to optimize hash models by incorporating semantic information into the training process. CIBhash [26] introduces a probabilistic Bernoulli representation layer into the model to connect mutual information and adjusts the contrastive loss to accommodate hash code learning, thereby generating a more general hash code. ConMH [62] constructs positive and negative samples by randomly sampling video frames using masks for contrastive learning. However these existing methods only discuss the application of contrastive learning to unimodal applications. To the best of our knowledge, there has been little exploration of contrastive hash learning in the cross-modal context.

III. THE PROPOSED METHOD

We consider the unsupervised cross-modal text-image retrieval problem. As shown in Figure 2, the proposed SCCDH method includes three main modules, i.e., the feature extraction, hash learning, and double contrastive distillation. Here the feature extraction module extracts the corresponding features from the input multimodal raw data, which are used to represent images and text. The hash learning module attempts to project different modal features into the common hamming space, where features of the same kind are closed, and features of different classes are pushed away. The double contrastive distillation module is used to explore the intrinsic distinctiveness of image and text features in the original feature space and Hamming space. In addition, the double contrastive distillation module filters out a large amount of noise to generate higher quality hash codes. In the following,

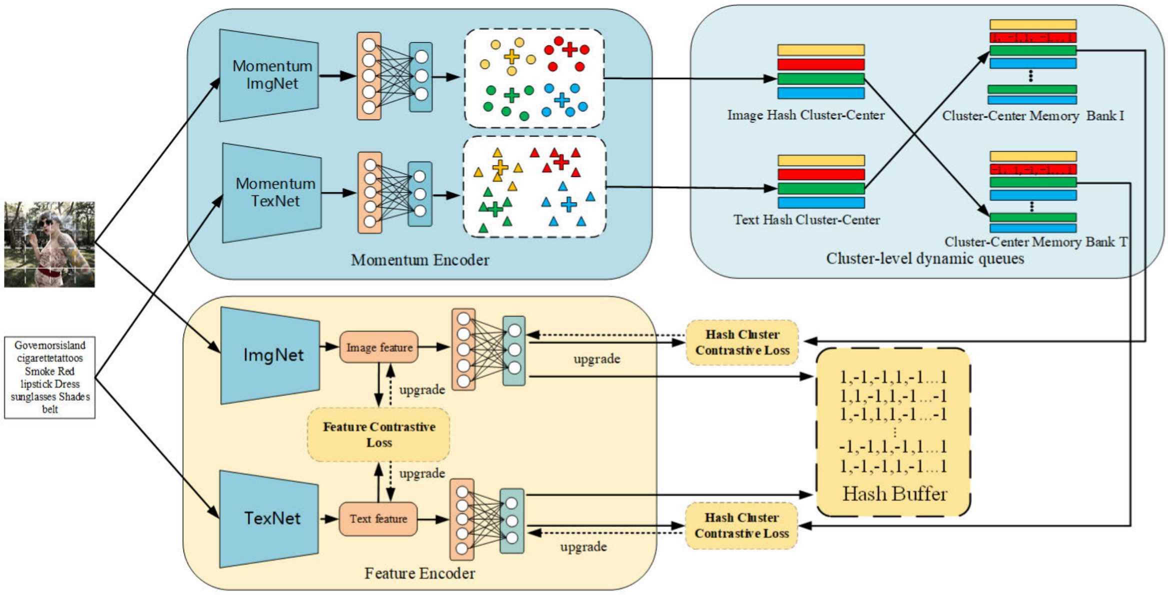


FIGURE 2. Proposed SCCDH method. The dark blue part is the momentum encoding module, the light blue part is the cluster contrast learning module, and the yellow part is the original feature encoding module.

we describe the proposed SCCDH method in detail, including the problem formulation and the hash learning algorithm.

A. PROBLEM FORMULATION

First, we provide relevant definitions for cross-modal hashing problems. Uppercase boldface (e.g., $\mathbf{X}$) and lowercase boldface (e.g., $\mathbf{x}$) denote matrices and vectors, respectively. In addition, let $O = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n}$ denote a cross-modal dataset with n image text pairs. $\mathbf{x_i} \in \mathbb{R}^{d_x \times 1}$ and $\mathbf{y}_i \in \mathbb{R}^{d_y \times 1}$ are the i-th image modal and the text modal instance, and d_x and d_y are the dimension of the image and text features, respectively.

The goal of cross-modal hashing is to learn a uniform binary hash code for both modalities in the Hamming space. Here, $\mathbf{B}^x = \left\{\mathbf{b}_i^x\right\}_{i=1}^{n}$ denotes the hash code of the image modality and $\mathbf{B}^y = \left\{\mathbf{b}_i^y\right\}_{i=1}^{n}$ denotes the hash code of the text modality, $\mathbf{b}_i^* \in \{-1, +1\}^L$, $* \in \{x, y\}$ denotes the hash code of the image/text modality, and L is the length of the hash code.The similarity between hash codes is measured by the Hamming distance, which is expressed as $\mathrm{dis}_H\left(\mathbf{b}_i, \mathbf{b}_j\right) = \frac{1}{2}\left(K - \left\langle \mathbf{b}_i, \mathbf{b}_j \right\rangle\right)$, where $\left\langle \mathbf{b}_i, \mathbf{b}_j \right\rangle$ denotes their inner product. To obtain multimodal data features from an image/text feature extraction network, we designed a dedicated network denoted $f^{x,y}\left(\mathbf{x}, \mathbf{y}; \theta^{x,y}\right)$. Here $\theta^{x,y}$ is the network parameter to learn for the corresponding modality. Then, the binary code $\mathbf{B}^{x,y}$ can then be generated by $f^{x,y}$. Recently, several effective methods have been proposed to compute the gradients of neural networks that contain discrete random variables. In this study, we used the simple straight-through gradient estimator (STE) [19]:

$$\mathbf{B} = \frac{\operatorname{sign}\left(\sigma\left(f^{x,y}\left(x, y; \theta^{x,y}\right)\right) - u\right) + 1}{2}, \tag{1}$$

where u denotes a sample from the uniform distribution [0, 1]. As proposed in the STE, the gradient can then be estimated by applying the backpropagation algorithm on the approximate loss. To learn the hash function, we propose a self-supervised objective function that attempts to eliminate cross-modal discrepancies. Unlike the supervised approach, the proposed SCCDH utilizes unsupervised contrastive learning rather than manual labeling of information to explore the intrinsic differences between image-text pairs. The overall loss function in the proposed SCCDH is:

$$\underset{\Theta^x, \Theta^y}{\arg\min}\left(\alpha \mathcal{L}_S + (1 - \alpha)\mathcal{L}_C\right), \tag{2}$$

where $\alpha (0 < \alpha < 1)$ is a balancing hyperparameter, $\mathcal{L}_c$ is cross-modal contrastive hash loss, $\mathcal{L}_s$ is cross-modal similarity loss.

B. CROSS-MODAL CONTRASTIVE HASH LEARNING

1) CLUSTER-LEVEL CONTRASTIVE DISTILLATION HASH LEARNING

The basic principle of contrastive learning is to select a pair of samples (one positive sample and one negative sample) and then map the selected samples into a common embedding space using two equal or different feature extractors. Then, their distance in the embedding space is compared to identify whether the pair of samples are similar. The distance metric is then optimized to bring similar samples closer together and dissimilar samples further apart such that a better feature representation can be learned. Note that the positive and negative samples are typically stored in a dictionary; thus, this can be considered a dictionary query problem. Inspired by [20], we propose a novel cluster contrastive learning method applicable to cross-modal hashing that exploits the intrinsic differences of samples in the common Hamming space using binary queues of two different modalities. Here, given a query $\mathbf{b}_i^*$, $* \in \{x, y\}$, the proposed method attempts to retrieve the most similar key from the momentum queue of another modality $\{\widetilde{\mathbf{c}}_0^*, \widetilde{\mathbf{c}}_1^*, \widetilde{\mathbf{c}}_2^* \ldots \widetilde{\mathbf{c}}_{n-1}^*\}, * \in \{y, x\}$, The i-th key $\widetilde{\mathbf{c}}_i^*$ in the queue corresponds to the i-th hash clustering center of the text/image. Typically, only one key $\widetilde{\mathbf{c}}_i^*$ (denoted as $\widetilde{\mathbf{c}}^+$) matches the query $\mathbf{b}_i^*$, the rest are all $\widetilde{\mathbf{c}}^-$. This is realized by measuring the distance between the query and the key through contrastive loss and continuously learning to make $\mathbf{b}_i^*$ closer to $\widetilde{\mathbf{c}}^+$ and further away from $\widetilde{\mathbf{c}}^-$ in the Hamming space. The proposed hash contrastive learning architecture is illustrated in Figure 3. A small-batch image-text pair is divided into two parts and entered into the feature and momentum encoders of the corresponding modality. Then, the batches of image-text pairs are split into two groups and input to the corresponding feature and momentum encoders. The binary output of the momentum encoder is subjected to a K-means clustering algorithm, and the resulting cluster centers are stored in a predefined dynamic queue for contrastive learning with the binary output of other modal feature encoders. The dynamic queue is then updated with momentum after the completion of training on the batch of instances. The designed cluster-level contrastive loss is expressed as follows:

$$\begin{aligned}\mathcal{L}_h &= \mathcal{L}_h^{IT} + \mathcal{L}_h^{TI}\\ &= \frac{\exp\left(\left\langle \mathbf{b}_i^x, \widetilde{\mathbf{c}}_i^{y+}\right\rangle/\tau\right)}{\exp\left(\left\langle \mathbf{b}_i^x, \widetilde{\mathbf{c}}_i^{y+}\right\rangle/\tau\right) + \sum_{j=1}^K \exp\left(\left\langle \mathbf{b}_i^x, \widetilde{\mathbf{c}}_j^{y-}\right\rangle/\tau\right)}\\ &\quad + \frac{\exp\left(\left\langle \mathbf{b}_i^y, \widetilde{\mathbf{c}}_i^{x+}\right\rangle/\tau\right)}{\exp\left(\left\langle \mathbf{b}_i^y, \widetilde{\mathbf{c}}_i^{x+}\right\rangle/\tau\right) + \sum_{j=1}^K \exp\left(\left\langle \mathbf{b}_i^y, \widetilde{\mathbf{c}}_j^{x-}\right\rangle/\tau\right)} \qquad (3)\end{aligned}$$

where τ is a temperature hyperparameter. The similarity between the different hash points is measured via the dot product. Here, the first query $\mathbf{b}_i^x$ is from the image modality, and the keys $\widetilde{\mathbf{c}}_{ij}^y$ are from the text modality. Conversely, the second query $\mathbf{b}_i^y$ is from the text modality, and the keys $\widetilde{\mathbf{c}}_{ij}^x$ are from the image modality, where the keys are all from the momentum queue of the corresponding modality. The cluster centroids are calculated by the mean feature vectors of each cluster as:

$$c_k = \frac{1}{|\mathcal{D}_k|} \sum_{B_* \in \mathcal{D}_k} b_i^* \qquad (4)$$

where $\mathcal{D}_k$ denotes the $k-th$ cluster set and $|\cdot|$ indicates the number of instances per cluster. $\mathcal{D}_k$ contains all the feature vectors in the cluster $k-th$. c_k does not perform a momentum update directly, but is calculated according to the following equation:

$$\begin{aligned}c_k &\leftarrow \frac{1}{|\mathcal{D}_k|} \sum_{\substack{b_i^* \in \mathcal{D}_k\\ b_i^* \in \mathcal{Q}}} \left[m b_i^* + (1-m) q^*\right]\\ &= m \frac{1}{|\mathcal{D}_k|} \sum_{\substack{b_i^* \in \mathcal{D}_k\\ b_i^* \in \mathcal{Q}}} b_i^* + (1-m) \frac{1}{|\mathcal{D}_k|} \sum_{\substack{b_i^* \in \mathcal{D}_k\\ b_i^* \in \mathcal{Q}}} q^*\\ &= m c_k + (1-m) q^* \qquad (5)\end{aligned}$$

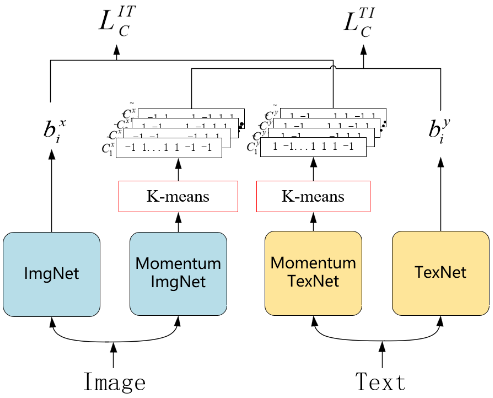


FIGURE 3. ImgNet and TexNet train the image encoder and text encoder using hash contrastive loss by matching the encoded hash query $\mathbf{b}_i^*$ with the encoded key $\widetilde{\mathbf{c}}_i^*$ in the queue.

where $\mathcal{Q}$ denotes the query instance features set, which contains all the feature vectors of query images for one iteration. $\mathbf{q}^*, * \in \{x, y\}$ denotes the hash vector from the original encoder.

In order to remove the large amount of noise from the image-text pairs. The binary code obtained by the momentum encoder is used as the ground truth label, the binary code generated by the original hash network is used as the predicted label, and the gap between the predicted label and the ground truth label is reduced continuously using the improved momentum distillation loss to optimize the momentum distillation network, thereby realizing our goal. The designed momentum distillation loss is expressed as follows:

$$\begin{aligned}\mathcal{L}_d = \frac{1}{2}\left(\mathrm{KL}\left(\boldsymbol{P}^{\mathrm{IT}}(I) \| \boldsymbol{H}\right) + \mathrm{KL}\left(\boldsymbol{Q}^{\mathrm{IT}}(I) \| \boldsymbol{H}\right)\right)\\ + \frac{1}{2}\left(\mathrm{KL}\left(\boldsymbol{P}^{\mathrm{TI}}(I) \| \boldsymbol{M}\right) + \mathrm{KL}\left(\boldsymbol{Q}^{\mathrm{TI}}(T) \| \boldsymbol{M}\right)\right)\\ (6)\end{aligned}$$

where $\boldsymbol{H} = 1/2\left(\boldsymbol{P}^{\mathrm{IT}}(I) + \boldsymbol{Q}^{\mathrm{IT}}(I)\right)$ and $\boldsymbol{M} = 1/2\left(\boldsymbol{P}^{\mathrm{TI}}(T) + \boldsymbol{Q}^{\mathrm{TI}}(T)\right)$, To combine with our contrastive learning, we let $\boldsymbol{Q}^{\mathrm{IT}}(I) = \mathcal{L}_h^{IT}$, $\boldsymbol{Q}^{\mathrm{TI}}(T) = \mathcal{L}_h^{TI}$ denotes the predictive distribution of image and text modalities, respectively. $\boldsymbol{P}^{\mathrm{IT}}(I) = \mathcal{L}_h^{IT}$, $\boldsymbol{P}^{\mathrm{TI}}(T) = \mathcal{L}_h^{TI}$. where $\mathbf{b}_i^x$ in $\mathcal{L}_h^{IT}$ with $\widetilde{\mathbf{c}}_i^x$ and $\mathcal{L}_h^{TI}$ denotes the replacement of $\mathbf{b}_i^y$ in $\mathcal{L}_h^{TI}$ with $\widetilde{\mathbf{c}}_i^y$. Representing the true distribution of image and text modalities. Here the KL-divergence is modified to compress its value to [0, 1], which is more accurate in discriminating similarity and solves the accompanying asymmetry problem.

The proposed hash contrastive loss can exploit the intrinsic variability of cross-modal hash codes in Hamming space effectively; however, it does not perform well when used independently to optimize models. Thus, after completing the cross-modal feature extraction process, feature contrastive loss is added such that the cross-modal samples are initially distinguished in the original feature space, which further helps the optimization of hash features in the common Hamming space, thereby enabling the hash generation module to produce higher quality hash codes. Inspired by [48], we improved the traditional contrast loss [36]. The difference

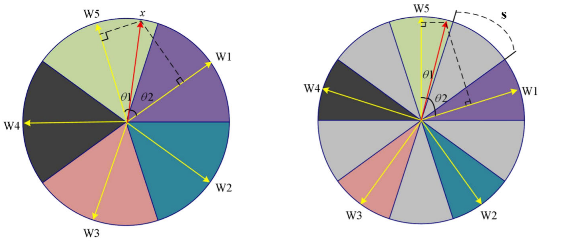


FIGURE 4. W denotes the weight vector, W= 1, Different color areas represent feature space from distinct classes. cos $(\theta) - m$ has a relatively compact feature region compared with cos (θ).

between improved feature contrastive loss and traditional contrastive loss shown in Figure 4. The feature contrastive loss is expressed as follows:

$$\mathcal{L}_f = -\log \frac{\exp\left(\left(\operatorname{sim}\left(f^x, f^y\right) - s\right)/\tau\right)}{\sum_{k=1}^{N} \nVdash_{[k\neq x]} \exp\left(\left(\operatorname{sim}\left(f^x, f^k\right) - s\right)/\tau\right)} \tag{7}$$

where $\operatorname{sim}\left(f^x, f^y\right) = f^{x\top} f^y / \|f^x\| \|f^y\|$, $\nVdash_{[k\neq x]} \in \{0, 1\}$ is an indicator function evaluating to 1 if $k \neq x$. $s \geq 0$ is a fixed parameter introduced to control the magnitude of the cosine margin. Since $\cos(\theta) - m$ is lower than $\cos(\theta)$, the constraint is more stringent for classification. Therefore, the altered loss reinforces the discrimination of learned features by encouraging an extra margin in the cosine space. In summary, the final contrastive distillation loss is expressed as follows:

$$\mathcal{L}_C = \beta(\mathcal{L}_h + \mathcal{L}_d) + (1 - \beta)\mathcal{L}_f, \tag{8}$$

2) CROSS-MODAL SIMILARITY LEARNING

The proposed contrastive distillation learning can explore the intrinsic differentiation of image-text pairs; however, it is not suited to the target downstream task (i.e., cross-modal hash retrieval). To make the model more applicable to downstream tasks, we also propose an improved pairwise similarity loss that attempts to process samples from different modalities and compare the fine-grained intraclass and interclass relationships between them. This method utilizes the data relationships between the cross-modal data as supervised signals to train the model. As a result, the different modal data are mapped into a shared low dimensional space. Here, the distance of negative image-text pairs belonging to the same class is minimized, and the distance of the negative image-text pairs whose distance is greater than the margin is maximized. The pairwise similarity loss is expressed as follows:

$$\mathcal{L}_S = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(S_{pair} W_{ij}^{IT}\right) + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\left(\upsilon - S_{pair}\right) \cdot \left(1 - W_{ij}^{IT}\right)\right), \tag{9}$$

where $W_{ij}^{IT} = 1$ means that $\mathbf{x}_i$ is semantically similar to $\mathbf{y}_i$; otherwise, $W_{ij}^{IT} = 0$. In a multi-label setting, two instances ($\mathbf{x}_i$ and $\mathbf{y}_i$) are annotated by multiple labels. Thus, we define $W_{ij}^{IT} = 1$ if $\mathbf{x}_i$ and $\mathbf{y}_i$ share as least one label; otherwise, $W_{ij}^{IT} = 0$. and $S_{pair}(\mathbf{b}^x, \mathbf{b}^y) = \mathbf{b}^{x\top}\mathbf{b}^y / \|\mathbf{b}^x\| \|\mathbf{b}^y\|$. Here, υ denotes the distance margin.

3) OPTIMIZATION

The process of cross-modal hash learning process minimizes the contrastive distillation loss $\mathcal{L}_C$ and the cross-modal similarity loss $\mathcal{L}_S$, as shown in equation(2), with the following overall loss:

$$\mathcal{L} = \alpha\mathcal{L}_S + (1 - \alpha)\mathcal{L}_C, \tag{10}$$

The proposed SCCDH method can be optimized continuously and iteratively in small batches, and it learns to increase the differentiation of cross-modal samples by minimizing $\mathcal{L}_C$ and encoding them into high-quality binary codes. In addition, $\mathcal{L}_S$ is added to better handle false negative pairs. Note that the complete model of the proposed SCCDH method can be optimized using any stochastic gradient descent optimization algorithm. Our optimization process is given in Algorithm 1.

Algorithm 1 Optimisation Process of SMDCH

```
Input:
    The training image-text pairs $O = \{x_i, y_i\}_{i=1}^{n}$, the length of the hash codes L, batch size $N$, balance parameter $\beta$,$\alpha$. momentum coefficient $\delta$, and learning rate $\lambda$
Output:
    Network parameters $\Theta^x$, $\Theta^y$.
Randomly initialize $\Theta^x$, $\Theta^y$.
while not converge do
    Randomly sample $N$ image-text pairs from $O$ to construct an image-text mini-batch $\{x_i, y_i\}_{i=1}^{N}$.
    Constructing two momentum queues $q_x$, $q_y$.
    The corresponding hash representation is obtained from ImageNet, TextNet and the corresponding momentum encoder.
    Comparison of contrastive distillation loss and cross-modal similarity loss calculated by equations 8 and 9.
    Updating network parameters by minimizing equation 10
    $\Theta_* = \Theta_* - \lambda\left(\alpha\frac{\partial\mathcal{L}_s}{\partial\Theta_*} + (1-\alpha)\frac{\partial\mathcal{L}_c}{\partial\Theta_*}\right)$ $(* \in \{x, y\})$.
    Update image text queue $q_x$, $q_y$.
end while
```

IV. EXPERIMENTS

A. DATASETS

The datasets used in our experiments are summarized as follows.

1) MIRFLICKR-25K

This dataset contains 25,000 image-text pairs, each containing an image and multiple text markers. The markers were annotated manually using multiple tags in 24 unique semantic

TABLE 1. Basic settings for our experimental dataset.

Dataset	Total	Train / Test	Labels
MIRFLICKR-25K	20,015	10,000/ 2,000	24
NUS-WIDE	190,421	10,500/ 2,100	21
MS-COCO	122,218	10,000/ 5,000	80

TABLE 2. Comparison of MAP performance with different metric distances.

Metric distances	I2T	T2I
Euclidean	0.8394	0.8235
Manhattan	0.8401	0.8255
cosine	**0.8495**	**0.8293**

TABLE 3. Comparison of MAP performance for different contrastive learning combinations ("global" denotes the hash contrastive module, "local" denotes the feature contrastive module, and "distillation" denotes distillation module).

Method	I2T	T2I
global	0.8042	0.7984
global+local	0.8320	0.8145
global+distillation	0.8452	0.8273
global+local+distillation	**0.8495**	**0.8293**

classes and can be used for cross-modal hash retrieval. After removing the class information, the dataset includes a total of 20,015 pairs.

2) NUS-WIDE

This dataset is a public web image dataset containing 269,648 images. The NUS-WIDE dataset has been annotated manually with 81 basic truth concepts for search evaluation. In this study, we selected a subset of 190,421 image-text pairs belonging to the 21 most common concepts, excluding data with no label or tag information.

3) MS-COCO

This dataset contains 123,287 images with five annotated sentences per image divided into 80 categories. After removing the image-text pairs without label information, 122,218 image-text pairs were used as experimental data. The details of the experimental setup are shown in table 1.

B. IMPLEMENTATION DETAIL

Most researchers use neural network models as feature encoders for image and text data, where the text data must be converted into BOW form to be input to the text feature encoder. Transformer models have demonstrated excellent performance in computer vision tasks. Thus, in this study, we used vision Transformers and Transformers to replace the neural network as the image and text encoders, respectively. In addition, the original MLP layer of the model was replaced with a new hash head at the back of the encoder for a specific task. To simplify the model, the image-text-momentum encoder used for contrast learning was the same as our original feature encoder. Note that the proposed method was implemented with PyTorch on four NVIDIA GEFORCE RTX 3080 Ti GPUs.

TABLE 4. Comparison of MAP performance for different contrastive learning combinations ("global" denotes the hash contrastive module, "local" denotes the feature contrastive module, and "distillation" denotes distillation module).

Method	I2T	T2I
global	0.8042	0.7984
global+local	0.8320	0.8145
global+distillation	0.8452	0.8273
global+local+distillation	**0.8495**	**0.8293**

TABLE 5. Comparison of MAP performance before and after applying contrastive loss, ("S_{pair}" denotes cross-modal pairwise similarity loss, and "Contra" denotes comparison loss).

Method	I2T	T2I
S_{pair}	0.8001	0.7922
Contra	0.7141	0.7013
S_{pair}+Contra	**0.8495**	**0.8293**

C. EVALUATION AND BASELINES

1) EVALUATION

Hamming sort and hash lookup are two classical protocols used to evaluate the performance of cross-modal retrieval. Two evaluation criteria were used in our experiments, i.e., the mean accuracy performance (MAP) to measure the accuracy of the Hamming distance and precision-recall (PR) curves and top$N$-precision curves (top$N$ Curves) to measure the accuracy of the hash lookup protocol.

2) BASELINES

In this study, we compared the proposed SCCDH method to nine state-of-the-art deep architecture methods (including supervised and unsupervised methods), i.e., SCAHN [4], AGAH [30], CMHH [31], GCH [32], CHN [47], SCM [14], DCMH [21], SSAH [22], DADH [33]. To facilitate a fair comparison, and all hyperparameters were set as provided by the corresponding authors, since the code for some of these models is not released, we implement them as we understood them. For the proposed SCCDH method, all hyperparameters were set as per experimental experience: queue size $q = 4096$, temperature hyperparameters $\tau = 0.6$, and momentum coefficient $m = 0.97$. Note that we used the validation set to select hyperparameters α, β.

D. PERFORMANCE ANALYSIS

Tables 6, 7 and 8 show the results of the proposed SCCDH compared to existing advanced cross-modal hashing methods for MAP on the MIRFLICKR-25K, NUS-WIDE, and MS-COCO datasets. Here, "I2T" means that the image data were used to query the text data in the database, and "T2I" means that the text data were used to query the image data in the database. Compared to the SCAHN [4], AGAH [30], CMHH [31], GCH [32], CHN [47], SCM [14],

TABLE 6. Comparison of MAP scores on MIRFLICKR-25K dataset.

Method	I2T			T2I		
	16bit	32bit	64bit	16bit	32bit	64bit
SCAHN [4]	0.8025	0.8198	0.8268	0.7976	0.8065	0.8200
AGAH [30]	0.7187	0.7445	0.7652	0.7090	0.7303	0.7581
CMHH [31]	0.7246	0.7251	0.7049	0.7674	0.7655	0.7632
GCH [32]	0.7300	0.7480	0.7530	0.7660	0.7790	0.7890
CHN [47]	0.7504	0.7495	0.7461	0.7776	0.7775	0.7798
SCM [15]	0.6354	0.5618	0.5634	0.6340	0.6458	0.6541
DCMH [21]	0.7316	0.7343	0.7446	0.7607	0.7737	0.7805
SSAH [22]	0.7890	0.8005	0.8060	0.7880	0.8001	0.8000
DADH [33]	0.8110	0.8185	0.8250	0.7993	0.8051	0.8167
SCCDH	**0.8290**	**0.8301**	**0.8495**	**0.8110**	**0.8126**	**0.8293**

TABLE 7. Comparison of MAP scores on NUS-WIDE dataset.

Method	I2T			T2I		
	16bit	32bit	64bit	16bit	32bit	64bit
SCAHN [4]	0.6690	0.6763	0.6874	0.6727	0.6798	0.6849
AGAH [30]	0.4082	0.4928	0.4845	0.4171	0.5227	0.5282
CMHH [31]	0.6125	0.5714	0.5618	0.5984	0.5823	0.5724
GCH [32]	0.6560	0.6630	0.7070	0.6690	0.6620	0.6780
CHN [47]	0.5754	0.5966	0.6015	0.5816	0.5967	0.5992
SCM [15]	0.3011	0.3001	0.3101	0.4119	0.4210	0.4378
DCMH [21]	0.5445	0.5497	0.5803	0.5793	0.5922	0.6014
SSAH [22]	0.6160	0.6360	0.6370	0.6400	0.6270	0.6410
DADH [33]	0.6683	0.6733	0.6851	0.6594	0.6655	0.6688
SCCDH	**0.6820**	**0.7054**	**0.7096**	**0.6911**	**0.7144**	**0.7250**

TABLE 8. Comparison of MAP scores on MS-COCO dataset.

Method	I2T			T2I		
	16bit	32bit	64bit	16bit	32bit	64bit
SCAHN [4]	0.6095	0.6502	0.6435	0.6035	0.6403	0.6435
AGAH [30]	0.5123	0.5721	0.6087	0.5384	0.5559	0.5821
CMHH [31]	0.5314	0.5116	0.5087	0.5589	0.5495	0.5519
GCH [32]	0.5950	0.6530	0.6790	0.6260	0.6240	0.6220
CHN [47]	0.5763	0.5822	0.5805	0.5198	0.5320	0.5409
SCM [15]	0.3601	0.3574	0.3562	0.4118	0.4183	0.4345
DCMH [21]	0.5225	0.5438	0.5419	0.4883	0.4942	0.5145
SSAH [22]	0.5520	0.5770	0.5760	0.5520	0.5780	0.5780
DADH [33]	0.6388	0.6668	0.6812	0.6027	0.6334	0.6528
SCCDH	**0.6521**	**0.6700**	**0.6915**	**0.6570**	**0.6811**	**0.7103**

DCMH [21], SSAH [22], DADH [33] baseline methods, the proposed SCCDH achieved a significant increase in MAP for I2T/T2I on the MIRFLICKR-25K dataset. On the NUS-WIDE and MS-COCO datasets, which contain more content and more complex relationships, the proposed SCCDH outperformed the compared state-of-the-art methods because, during the learning process, the proposed self-supervised cluster-contrast distillation network better captures the inherent discriminative properties in the different modal data. As a result, the proposed SCCDH method can generate more discriminative hash codes. Figure 5,6 and 7 shows the PR curves and top-N curves for the proposed SCCDH with a hash code length of 64 bits against four state-of-the-art methods on the three datasets. As can be seen, the proposed method outperformed the baseline methods.

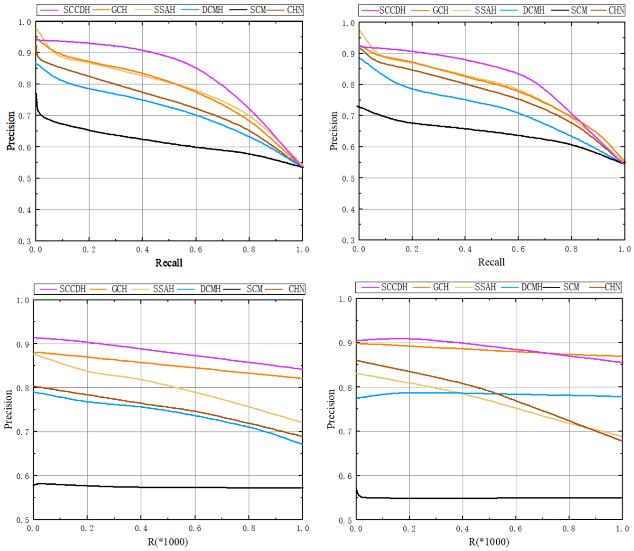


FIGURE 5. PR curves and top*N* curves on MirFlickr25k, from left to right are I2T and T2I (the code length is 64 bits).

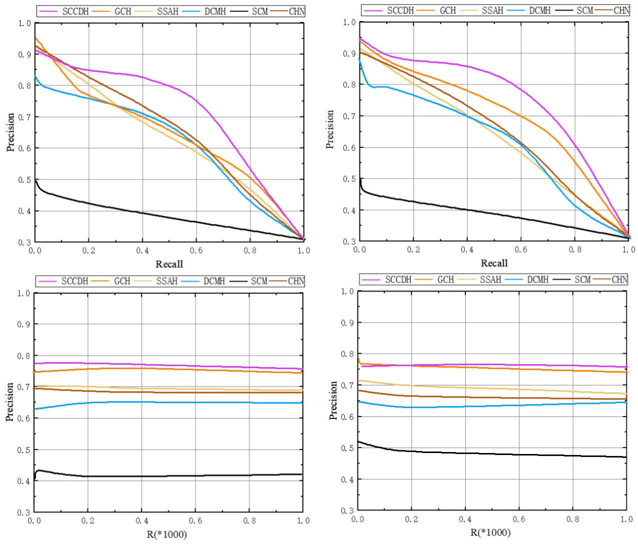


FIGURE 6. PR curves and top*N* curves on NUS-WIDE, from left to right are I2T and T2I (the code length is 64 bits).

E. ABLATION STUDY

We also considered the effect of different modules on the performance of our model. Here, we designed three ablation experiments to demonstrate the superiority of the proposed model. (1) We demonstrate the superiority of the pairwise similarity loss by replacing different distance metric functions. (2) We highlight the importance of the double contrastive network by changing the structure of the contrastive network. (3) By removing the contrastive learning network, we prove that the proposed contrastive learning network is essential.

As shown in Table 2, replacing the cosine distance in the pairwise similarity loss with the Euclidean and Manhattan

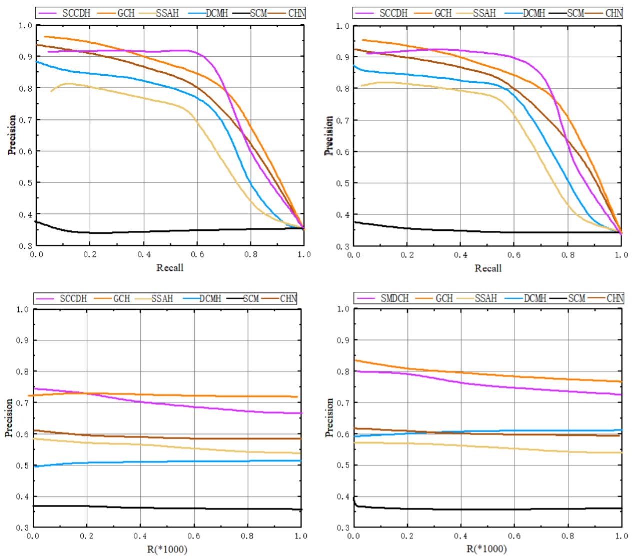


FIGURE 7. PR curves and top*N* curves on MS-COCO, from left to right are I2T and T2I (the code length is 64 bits).

distances resulted in a significant decrease in their MAP results.

To demonstrate the superiority of the proposed contrastive distillation structure, we replaced the contrastive model with three variants. As shown in Table 3, the performance improved significantly with ''global+local'' compared to simply ''global.'' However, our expectations were not reached, and when the distillation module was added, the performance improved further due to the removal of a large amount of noise.

Contrastive learning can tap into the intrinsic distinctiveness of the data. We removed the double contrastive distillation module and achieved a MAP result of 0.8 using only ''S_{pair}''. However, the performance dropped significantly using only ''Contra''. It may have been caused by the fact that contrastive learning is not applicable to a specific cross-modal retrieval task; thus, the combination of the two elements will improve results, and our experiments confirmed this estimation.

To demonstrate the superiority of the proposed contrastive distillation structure, we replaced the contrastive model with three variants. As shown in Table 3, the performance improved significantly with ''global+local'' compared to simply ''global.'' However, our expectations were not reached, and when the distillation module was added, the performance improved further due to the removal of a large amount of noise.

Contrastive learning can tap into the intrinsic distinctiveness of the data. We removed the double contrastive distillation module and achieved a MAP result of 0.8 using only ''S_{pair}''. However, the performance dropped significantly using only ''Contra''. It may have been caused by the fact that contrastive learning is not applicable to a specific cross-modal retrieval task; thus, the combination of the two elements will improve results, and our experiments confirmed this estimation.

V. CONCLUSION

We propose a novel cross-modal hash retrieval method (SCCDH) that utilizes double contrast learning to exploit the intrinsic distinctiveness between cross-modal samples in the original feature space and the Hamming space. The proposed method also uses the momentum encoder in contrastive learning as a teacher network to filter out large amounts of noise in the samples. The results of extensive experiments have shown that the proposed SCCDH achieves state-of-the-art retrieval performance on three popular public datasets. However, as the proposed SCCDH model uses a Transformer as the backbone network, the training time increases, and the results obtained on the NUS-WIDE and MS-COCO datasets were not satisfactory and we are focused on exploring the semantic distinctiveness between modalities, thus neglecting to explore the intrinsic variability between samples within modalities. Thus, we assume that the potential differentiations in these dataset were not utilized fully. We will investigate reducing the complexity of the backbone network and explore a loss function that is more suitable for cross-modal hash retrieval to solve these problems in the future.

VI. CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this paper.

REFERENCES

[1] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, ''Inter-media hashing for large-scale retrieval from heterogeneous data sources,'' in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2013, pp. 785–796.

[2] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, ''Triplet-based deep hashing network for cross-modal retrieval,'' *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3893–3903, Aug. 2018.

[3] P. Hu, X. Wang, L. Zhen, and D. Peng, ''Separated variational hashing networks for cross-modal retrieval,'' in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1721–1729.

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

[5] H. Liu, R. Ji, Y. Wu, F. Huang, and B. Zhang, ''Cross-modality binary code learning via fusion similarity hashing,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6345–6353.

[6] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang, ''Supervised hashing with kernels,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2074–2081.

[7] K. Li, G.-J. Qi, J. Ye, and K. A. Hua, ''Linear subspace ranking hashing for cross-modal retrieval,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1825–1838, Sep. 2017.

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

[9] C. Li, C. Deng, L. Wang, D. Xie, and X. Liu, ''Coupled cycleGAN: Unsupervised hashing network for cross-modal retrieval,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 176–183.

[10] J. Zhang, Y. Peng, and M. Yuan, ''Unsupervised generative adversarial cross-modal hashing,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–8.

[11] S. R. Dubey, S. K. Singh, and W.-T. Chu, ''Vision transformer hashing for image retrieval,'' in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

[12] P. Hu, H. Zhu, J. Lin, D. Peng, Y.-P. Zhao, and X. Peng, ''Unsupervised contrastive cross-modal hashing,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3877–3889, Mar. 2023.

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

[14] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. 28th AAAI Conf. Artif. Intell.*, vol. 28, no. 1, 2014, pp. 1–7.

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

[16] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1–6.

[17] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5427–5440, Nov. 2016.

[18] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2014, pp. 415–424.

[19] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," 2013, *arXiv:1308.3432*.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

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

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

[23] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.

[24] X. Luo, D. Wu, Z. Ma, C. Chen, M. Deng, J. Ma, Z. Jin, J. Huang, and X.-S. Hua, "CIMON: Towards high-quality hash codes," 2020, *arXiv:2010.07804*.

[25] J. Yu, Y. Shen, M. Wang, H. Zhang, and P. H. S. Torr, "Learning to hash naturally sorts," 2022, *arXiv:2201.13322*.

[26] Z. Qiu, Q. Su, Z. Ou, J. Yu, and C. Chen, "Unsupervised hashing with contrastive information bottleneck," 2021, *arXiv:2105.06138*.

[27] E. Yang, C. Deng, T. Liu, W. Liu, and D. Tao, "Semantic structure-based unsupervised deep hashing," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1064–1070.

[28] E. Yang, T. Liu, C. Deng, W. Liu, and D. Tao, "DistillHash: Unsupervised deep hashing by distilling data pairs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2941–2950.

[29] R.-C. Tu, X.-L. Mao, and W. Wei, "MLS3RDUH: Deep unsupervised hashing via manifold based local semantic similarity structure reconstructing," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 3466–3472.

[30] W. Gu, X. Gu, J. Gu, B. Li, Z. Xiong, and W. Wang, "Adversary guided asymmetric hashing for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2019, pp. 159–167.

[31] Y. Cao, B. Liu, M. Long, and J. Wang, "Cross-modal Hamming hashing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 202–218.

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

[33] C. Bai, C. Zeng, Q. Ma, J. Zhang, and S. Chen, "Deep adversarial discrete hashing for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 525–531.

[34] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[35] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9620–9629.

[36] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn. (PMLR)*, 2020, pp. 1597–1607.

[37] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22243–22255.

[38] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15745–15753.

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

[40] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[41] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "Image as a foreign language: BEiT pretraining for all vision and vision-language tasks," 2022, *arXiv:2208.10442*.

[42] K. Luo, C. Zhang, H. Li, X. Jia, and C. Chen, "Adaptive marginalized semantic hashing for unpaired cross-modal retrieval," *IEEE Trans. Multimedia*, early access, Feb. 15, 2023, doi: 10.1109/TMM.2023.3245400.

[43] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," 2021, *arXiv:2104.08821*.

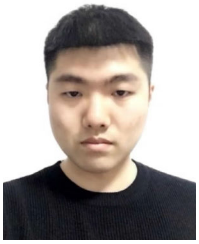
[44] B. Shan, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang, "ERNIE-ViL 2.0: Multi-view contrastive learning for image-text pre-training," 2022, *arXiv:2209.15270*.

[45] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive Captioners are image-text foundation models," 2022, *arXiv:2205.01917*.

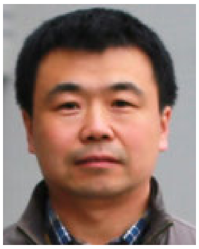
[46] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," 2022, *arXiv:2210.10163*.

[47] Y. Cao, M. Long, J. Wang, and P. S. Yu, "Correlation hashing network for efficient cross-modal retrieval," 2016, *arXiv:1602.06697*.

[48] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.



HAOXUAN SUN is currently pursuing the master's degree majoring in information and communication engineering with the Liaoning University of Technology, China. His current research interests include deep learning and cross-modal hashing.



YUDONG CAO received the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2011. He is currently an Associate Professor with the School of Electronic and Information Engineering, Liaoning University of Technology, Jinzhou, China. His current research interests include pattern recognition and machine learning.

GUANGYUAN LIU received the M.S. degree in electrical and computer engineering from the Dalian University of Technology, Dalian, China. His current research interests include deep learning and optimization.