IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 12, 2020, accepted April 13, 2020, date of publication April 20, 2020, date of current version May 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2988923

Learning Decorrelated Hashing Codes With Label Relaxation for Multimodal Retrieval

DAYONG TIAN[1], YIWEN WEI[2], (Member, IEEE), AND DEYUN ZHOU[1]
[1]School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China
[2]School of Physics and Optoelectronic Engineering, Xidian University, Xi'an 710071, China

Corresponding author: Dayong Tian (dayong.tian@nwpu.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 61806166 and Grant 61801362, in part by the Natural Science Foundation of Shaanxi Provience under Grant 2020JQ-197, and in part by the Fundamental Research Funds for the Central Universities, NPU under Grant G2018KY0303.

ABSTRACT Due to the correlation among hashing bits, the retrieval performance improvement becomes slower when the hashing code length becomes longer. Existing methods try to regularize the projection matrix as an orthogonal matrix to decorrelate hashing codes. However, the binarization of projected data may completely break the orthogonality. In this paper, we propose a minimum correlation regularization (MCR) for multimodal hashing. Rather than being imposed on projection matrix, MCR is imposed on a differentiable function which approximates the binarization. On the other hand, binary labels could not precisely reflect the distances among data. Hence, we propose a label relaxation scheme to achieve better performance.

INDEX TERMS Multimodality, hashing, binary embedding, minimum correlation regularization.

I. INTRODUCTION

Multimodal hashing which embeds data to binary codes is an efficient tool for retrieving heterogeneous but correlated multimedia data, such as image-text pairs in Facebook and video-tag pairs in Youtube. Unlike real vectors used in traditional retrieval methods [1]–[4], binary codes can greatly reduce the storage requirement and computation costs of nearest neighbors search.

Orthogonality is assumed to be a quality of good hashing codes [5]. However, the orthogonality constraint will lead to an NP-hard problem. Hence, there are two widely used ways to approximate orthogonal code matrix: (1) adopting orthogonal vectors and then thresholding them to generate binary codes [5], [6]; (2) imposing an orthogonality regularization on the objective function [7], [8]. These methods on approximating orthogonality have a theoretical defect that the orthogonality is corrupted by quantization.

Spectral hashing (SH) [5] and iterative quantization (ITQ) [6] are two representative works in way (1). SH selects eigenfunctions corresponding to several smallest eigenvalues and thresholds eigenfunctions at zero. ITQ rotates the principal components and thresholds data projected by those principal components at zero. Obviously, thresholding orthogonal vectors at zero cannot generate orthogonal binary vectors.

As an representative example for way (2), deep multimodal hashing with orthogonal regularization (DMHOR) [7] is illustrated in Fig. 1. Liong *et al.* [9] and Chen *et al.* [10] also use this orthogonal regularization in their deep hashing model.

Deep multimodal hashing with orthogonality regularization (DMHOR) [7] introduces an orthogonality regularization (OR) to deep neural network (DNN). It uses Restricted Boltzmann Machine (RBM) for image and text data. Each layer of RBM can be represented as a nonlinear activation function of a linear transformation of the input. The OR is applied on the weight matrix of each layer. The authors argue that the proposed OR can lead to an orthogonal code matrix when data matrices are orthogonal. This assumption is unreasonable in real application. In this paper, we will briefly analyze the properties of this OR and demonstrate that it is only suitable for some linear hashing models. Deep cross-modal hashing (DCMH) [11] employs different types of DNN for different modalities. For example, convolutional neural network (CNN) is used for images while fully connected neural network is used for text. The orthogonality of hashing codes is neglected.

In this paper, we propose a hashing method named decorrelated multimodal hashing (DMH). First, a sigmoid function is applied on the linear transformations of original data

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

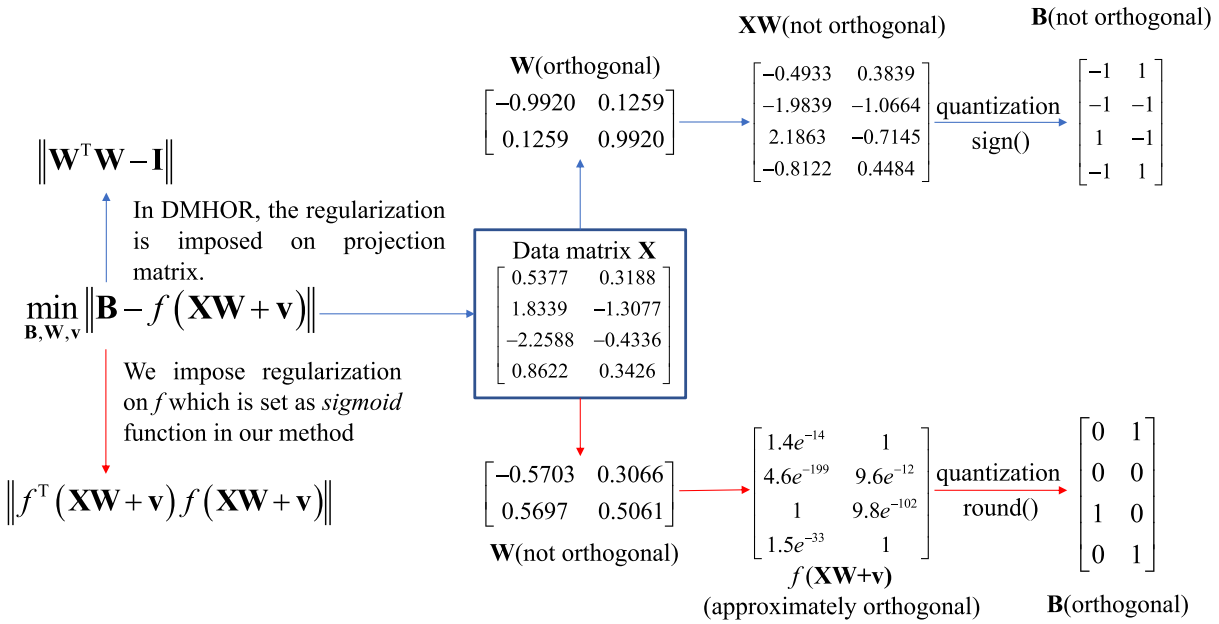
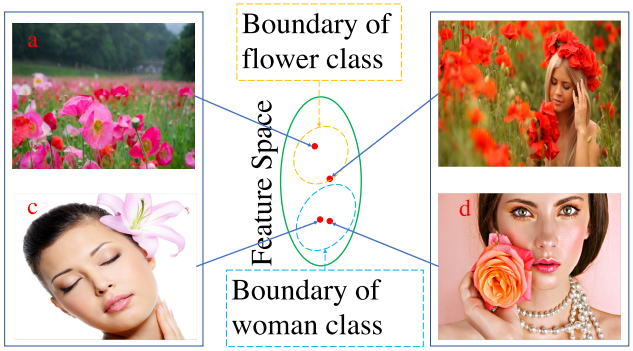


FIGURE 1. Illustration of the difference between our regularization term and that of DMHOR. Red arrows indicate the flowchart of our method, while the blue ones indicate the flowchart of method proposed in DMHOR. W is the projection matrix, X is data matrix, v is a bias, B is the hashing code matrix, I is the identity matrix and *f* is the *sigmoid* function used to approximate binarization.

points to map different modalities into a common code matrix. Then, we devise a minimum correlation regularization (MCR) to improve the retrieval performance on long-bit experiments. Unlike aforementioned orthogonality constraints or regularizations [7] that are usually applied on the linear transformation matrices, the proposed MCR is applied on the sigmoid function. Because the output of sigmoid function approximates a binary code and the hashing code matrix directly depends on the quantization of it, the propose MCR works better on decorrelating hashing codes (Fig. 1).

We do not use the term ''orthogonality'' because the maximum number of mutual orthogonal vectors is equal to the dimension of them and an orthogonal linear transformation does not exist when the rank of a data matrix is less than that of its code matrix. For instance, if an $N \times d$ data matrix is encoded as an $N \times c$ code matrix where N is the number of data and $d < c$, the dimension of the linear transformation matrix W should be $d \times c$. Because we cannot find c d-dimensional column vectors, an orthogonal W does not exist. In Subsection III-B, we will prove that when $d + 1 < c$, the output matrix of sigmoid function cannot be orthogonal and hence the orthogonality of code matrix cannot be even approximated.

Besides the orthogonality regularization, a label relaxation method is proposed for multi-labeled data sets. Labels are generally treated as a special modality in multimodal hashing methods. Therefore, the relaxed labels that can reflect the distances among data will benefit the hashing process. Ji *et al.* [12] proposed a deep multi-level semantic hashing method which is similar to ours. However, their method needs to compute the mutual distances among labels, which makes it intractable for large dataset.



(flower class, woman class)	a	b	c	d
Human	(1,0)	(1,1)	(1,1)	(1,1)
Estimated Labels	(1.0,0.0)	(0.9,0.5)	(0.3,0.9)	(0.3,0.9)

FIGURE 2. Illustration of the proposed relaxation method in multi-labeled data. On image a, both human experts and the proposed method should label it as flower because it only contains flowers. It can be seen that for single-labeled data, there is little difference between human experts and the proposed method. However, when a, b, c and d are treated as data in a multi-labeled data set, human experts label them as "1,0" or "1,1" in traditional way, while the proposed method relaxes labels to be real numbers according to their distances to the classes in the feature space. The closer it is to the class, the larger the label. That is, the generated labels can reflect the distances. As labels are treated as a special modality in the proposed method, it could be better if they can reflect the distances.

The rest of this paper is organized as follows. The related works are reviewed in Section II. In Section III, we, step by step, derive our model from a widely used unimodal hashing method, iterative quantization (ITQ) [6]. The discussions on parameter settings and optimization algorithms are also given in Section III. Experimental results are reported in Section IV. We conclude this paper in Section V

II. RELATED WORKS

Some well-known multimodal hashing models are related to some classical unimodal ones. Hence, in this section, unimodal hashing models will be firstly reviewed and then we will discuss some representative multimodal hashing models and their relations to unimodal ones. For a more comprehensive survey on unimodal hashing methods, please refer to [13].

A. UNIMODAL HASHING

Most existing unimodal hashing models focus on image retrieval tasks. However, they are also feasible for other types of data as long as the data are represented in real vectors. For images, there are lots of popular feature extraction methods [14], [15] available to represent images by real vectors.

The unimodal hashing methods can be divided into two categories according to their dependence on data. Locality-sensitive hashing (LSH) [16] and its kernelized version [17], [18] are well-known data-independent unsupervised unimodal hashing methods. Due to randomized hashing, LSH demands more bits per hashing table [19].

Spectral hashing (SH) [5], one of the most popular and pioneering data-dependent unimodal hashing methods, generate hashing codes by solving a relaxed mathematical problem to avoid computing the affinity matrix that requires calculating and storing pairwise distances of the whole data set [20]. The authors argued that two constraints for a good code matrix are orthogonality and balance, either of which leads to an NP-hard problem. In the following works, balance is generally neglected and orthogonality constraint is relaxed or neglected, too.

Anchor graph hashing (AGH) [21] substitutes the affinity matrix in SH by constructing the a highly sparse one using several anchor points. Discrete graph hashing (DGH) [8] incorporates a relaxed orthogonality constraint into AGH to improve the performance on long-bit experiments.

Methods based on linear transformations, such as principal component analysis (PCA) [22], attract wide interests due to their effectiveness and computation efficiency. ITQ rotates the projection matrix obtained by PCA to minimize the quantization loss. Isotropic hashing (IsoH) [23], harmonious hashing (HH) [23] and ok-means [24] are derived from ITQ. IsoH equalizes the importance of principal components. HH puts an orthogonal constraint on an auxiliary variable for the code matrix. ok-means rotates the data matrix to minimize the quantization loss. ITQ, IsoH and HH depends on principal components whose maximum number is no larger than the minimum dimension of data matrix. Hence, they cannot generate hashing codes longer than the data dimension. Despite of PCA, other linear transformations can be used, such as Linear Discriminant Analysis (LDA) [25]. Unlike these precomputed transformation matrix, neighborhood discriminant hashing [26] calculates the transformation matrix during the iterative minimization procedure.

Inductive manifold hashing [19] embeds some special samples into lower dimensional space and the embeddings of remaining samples are calculated by a linear combination of those special samples. The coefficients of the linear combination are the probabilities that a sample belongs to those special samples.

All aforementioned unimodal hashing models cannot generate balanced code matrix. Spherical hashing (SpH) [27] and global hashing system (GHS) [20] quantize the distance between a data point and a special point. The closer half to a special point is denoted as 1 while the further half is denoted as 0. Therefore, a balanced matrix can be easily generated. Their major difference is on how to find these special points. SpH uses a heuristic algorithm while GHS treats it as a satellite distribution problem of the Global Positioning System (GPS).

Some unsupervised unimodal hashing models can be easily extended to supervised models. For example, substituting PCA by Canonical Correlation Analysis (CCA) [28], the label information can be incorporated. Besides, unsupervised and supervised models, wealy-supervised models [29]–[31] are also promising, since labels can signifcantly improve retrieval accurarcy but manually labelling images is a heavy burden for human experts.

B. MULTIMODAL HASHING

Multimodal hashing models can be classified into unsupervised and supervised ones. Unsupervised multimodal hashing tries to preserve the Euclidean data structure by binary codes. Inter-media hashing [19] learns hashing function by linear regression. IMH models intra-media consistency in a similar way of SH. Like what AGH has done to SH, linear cross-media hashing (LCMH) [32] uses the distances between each data point and each cluster centroid to construct a sparse affinity matrix. Collective matrix factorization hashing (CMFH) [33] can be treated as an extension of NDH. For each modality, CMFH consists of two terms: (1) calculating a transformation matrix for the data matrix to match the code matrix through minimizing quantization loss, and (2) calculating a transformation matrix for the code matrix to match the data matrix through minimizing squared error. Latent semantic sparse hashing [34] is an extension of CMFH and its basic idea is similar to HH that imposes the orthogonality constraint on an auxiliary variable. LSSH imposes the sparse regularization on an auxiliary variable in the latent space. Shen *et al.* [35] proposed a cross-view hashing method for semi-paired data. It jointly learns a correlated representation for each modality and hashing functions. It rotates the hashing code matrix to match the correlated representation matrices. Hence, it can be seen as an extension of ok-means.

By incorporating label information, supervised hashing can achieve higher accuracy. Cross-modality similarity-sensitive hashing (CMSSH) [36] treats hashing as a binary classification problem. Cross-view hashing (CVH) [37] assumes the hashing codes be a linear embedding of the original data points. It substitutes the code matrix by this embedding. The objective function is a weighted summation of that of spectral hashing (SH) [5] on each modality. Multilatent

binary embedding (MLBE) [38] treats hashing codes as the binary latent factors in the proposed probabilistic model and maps data points from multiple modalities to a common Hamming space. Semantics-preserving hashing (SePH) [39] learns the hashing codes by minimizing the KL-divergence of the probability distribution in Hamming space from that in semantic space. CMSSH, MLBE and SePH need to compute the affinities of all data points, which makes it intractable for large data set. Semantic correlation maximization (SCM) [40] circumvents this by learning only one bit each time and the explicit computation of affinity matrix is avoided through several mathematical manipulations. Multimodal discriminative binary embedding (MDBE) models [41] hashing as a minimization problem. There are two main terms in its formulation. One term indicates different modalities and the labels can be embedded to the same latent space, while the other one indicates the embedded modalities can be further embedded as the labels. $l2$-norm is used to regularize the linear embedding matrix. Intra- and Inter-Modality Similarity Preserving Hashing (IISPH) [42] measure the similarity among data within the same modality and across different modalities. SCM, MDBE and IISPH discard the uncorrelation property of the code matrix or embedding matrix, which makes their performance improve slowly as code length increases.

Most hashing methods relax the binary constraint, Xu *et al.* [43] proposes a discrete optimization algorithm to directly learn hashing codes without relaxing the binary constraint. Collective reconstructive embeddings [44], [45] use modality-specific similarity metrics for different modalities. Besides, the above mentioned shallow models. Deep neural networks (DNN) are extensively studied in cross-modal hashing. Among the DNN-based methods, the adversarial learning based models [46]–[48] have achieve appealing results.

III. METHODOLOGY

Terms "view" and "modality" are discriminated in some literatures [41]. Multiple views of data refers to different type of features of one modality, e.g. SIFT [49] and GIST [50] features for images. However, we use these two words interchangeably since our method can be used in either situations as long as the data are represented by real matrices.

First, Let us define the used notations. Suppose that $\mathbf{X}^i$ is the i-th view matrix of the data and $\mathbf{X}^i = \left[\mathbf{x}_1^i, \ldots, \mathbf{x}_n^i\right]^\top$, where $\mathbf{x}_m^i \in \mathbb{R}^{d_i}$, n is the number of data points and $i = 1, \ldots, g$. A binary code corresponding to the m-th data is defined by a row vector $b_m = \{0, 1\}^c$, where c is the code length and the code matrix $\mathbf{B} = \left[\mathbf{b}_1^\top, \ldots, \mathbf{b}_n^\top\right]^\top$. $h^i\left(\mathbf{X}^i\right)$, the hashing function for the i-th view matrix, embeds $\mathbf{X}^i$ into a binary code matrix.

A. PROBLEM FORMULATION

ITQ is a successful hashing method for single view data. The formulation of ITQ is

$$\arg\min_{\mathbf{B},\mathbf{R}} E = \|\mathbf{B} - \mathbf{X}\mathbf{W}\mathbf{R}\|_F^2, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the data matrix, $\mathbf{W} \in \mathbb{R}^{d\times c}$ is obtained by principal component analysis (PCA) and $\mathbf{R} \in \mathbb{R}^{c\times c}$ is an orthogonal matrix. An intuitive multi-view extension of ITQ can be

$$\arg\min_{\mathbf{B},\mathbf{R}^i} E = \sum_i \alpha_i \left\|\mathbf{B} - \mathbf{X}^i\mathbf{W}^i\mathbf{R}^i\right\|_F^2, \tag{2}$$

where α_i is a positive real constant. As the maximum number of principal components pre-computed by PCA on the ith view matrix is d_i, Eq. (2) cannot be used when $c > d_i$. We remove $\mathbf{R}^i$ from Eq. (2). Then, we simultaneously calculate $\mathbf{W}^i \in \mathbb{R}^{d_i\times c}$ and $\mathbf{B}$ during the optimization process. This method can be modeled as

$$\arg\min_{\mathbf{B},\mathbf{W}^i} E = \sum_i \alpha_i \left\|\mathbf{B} - \mathbf{X}^i\mathbf{W}^i\right\|_F^2. \tag{3}$$

Because $\mathbf{B}$ is a binary matrix, $h^i(\mathbf{X}^i\mathbf{W}^i) = 1/(1+\exp(-(\beta_i * \mathbf{X}^i\mathbf{W}^i + \mathbf{1}\mathbf{v}^i)))$ is applied to transform the values of $\beta_i * \mathbf{X}^i\mathbf{W}^i + \mathbf{1}\mathbf{v}^i$ into interval (0, 1), where $\mathbf{1}$ is a n-dimensional column vector whose elements are equal to 1. β_i is a constant and $\mathbf{v}^i$ is a bias vector. Hence, Eq. (3) can be modified as following.

$$\arg\min_{\mathbf{B},\mathbf{W}^i,\mathbf{v}^i} E = \sum_i \alpha_i \left\|\mathbf{B} - \frac{1}{1+\exp\left(-(\beta_i\mathbf{X}^i\mathbf{W}^i + \mathbf{1}\mathbf{v}^i)\right)}\right\|_F^2. \tag{4}$$

B. MINIMUM CORRELATION REGULARIZATION

The orthogonality condition for good codes [5] is approximated by an orthogonal $\mathbf{W}$ in ITQ. However, when $c > d_i$, an orthogonal $\mathbf{W}^i$ does not exist. In this case, Wang *et al.* [7] introduces the following regularization to decorrelate code matrix:

$$\mathrm{R} = \left\|\mathbf{W}^{i\top}\mathbf{W}^i - \mathbf{I}\right\|_F^2. \tag{5}$$

First, let us discuss some interesting properties of Eq. (5).

Proposition 1: When $c \le d_i$, the $\mathbf{W}^i$ that minimizes Eq. (5) is an orthogonal matrix.

It is easy to prove **Proposition 1** by the definition of orthogonal matrix.

Proposition 2: Let the $\mathbf{W}^i$ that minimizes Eq. (5) consists of column vectors $\mathbf{w}_p^i$ where $p = 1, \ldots, c$. The angle between any pair of column vectors is equal to each other. *Proof:* Let $\mathbf{V} = \mathbf{W}^{i\top}\mathbf{W}^i$ and let V_{pq} be the element in the pth row and qth column of $\mathbf{V}$. V_{pq} is the inner product of $\mathbf{w}_p^i$ and $\mathbf{w}_q^i$. When $\|\mathbf{w}_p^i\|_F^2 = 1$, the diagonal elements of R will be 0 and the angle between $\mathbf{w}_p^i$ and $\mathbf{w}_q^i$ will be $\arccos(\mathbf{w}_p^{i\top}\mathbf{w}_q^i)$. Eq. (11) can be written as:

$$\mathrm{R} = \sum_{p,q} \mathbf{w}_p^{i\top}\mathbf{w}_q^i, \quad p \ne q. \tag{6}$$

According to the inequality of arithmetic and geometric means, it can be deduced that

$$\frac{\sum_{p,q}\mathbf{w}_p^{i\top}\mathbf{w}_q^i}{c^2 - c} \ge \prod_{p,q} \sqrt[c^2-c]{\mathbf{w}_p^{i\top}\mathbf{w}_q^i}. \tag{7}$$

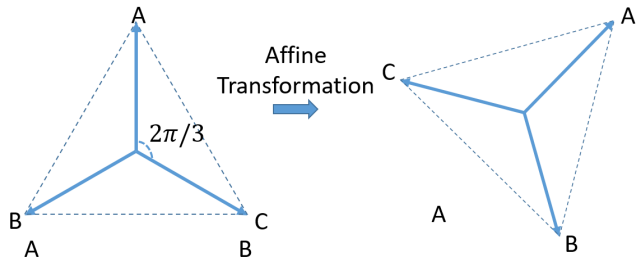


FIGURE 3. Illustration of Proposition 2 and Proposition 3. If $\mathbf{W}^i \in \mathbb{R}^{2\times 3}$, its column vectors will align with the centerlines of an equilateral triangle. The affine transformation will change the relative positions among vectors but the overall structure is kept. In the equilateral triangle, point B is transformed to the clockwise direction of point A.

The equality holds if and only if all ${\mathbf{w}_p^i}^\top \mathbf{w}_q^i$ are equal. That is, the angle between any pair of column vectors is equal when $\mathbf{W}^i$ minimizes Eq. (5). □

Proposition 3: If $\mathbf{W}^i$ minimizes Eq. (5), the affine transformation of $\mathbf{W}^i$, i.e. $\mathbf{W}^i\mathbf{R}$ also minimizes Eq. (5) where $\mathbf{R}$ is an orthogonal matrix.

Proof: As $\mathbf{R}$ is orthogonal, we have

$$\left\|{\mathbf{W}^i}^\top \mathbf{W}^i - \mathbf{I}\right\|_F^2 = \left\|\mathbf{R}^\top \left({\mathbf{W}^i}^\top \mathbf{W}^i - \mathbf{I}\right)\mathbf{R}\right\|_F^2. \tag{8}$$

Eq. (8) can be rewritten as

$$\left\|{\mathbf{W}^i}^\top \mathbf{W}^i - \mathbf{I}\right\|_F^2 = \left\|\mathbf{R}^\top {\mathbf{W}^i}^\top \mathbf{W}^i\mathbf{R} - \mathbf{I}\right\|_F^2. \tag{9}$$

Here, $\mathbf{R}^\top\mathbf{R} = I$ is used in the deduction. Hence, $\mathbf{W}^i\mathbf{R}$ also minimizes Eq. (5). □

In Fig. 3, we illustrate **Proposition 2** and **Proposition 3** in 2-dimensional case. Following the flowchart of ITQ, one can find c d-dimensional vectors distributed like those in Fig. 3 and then transform them by $\mathbf{R}$ to minimize Eq. (2). However, the complexity of theoretically finding such vectors increases dramatically in high dimensional spaces. Wang *et al.* [7] use Eq. (5) as a regularization and argue that Eq. (5) will lead to an orthogonal code matrix when the data matrices are orthogonal. It is easy to find an example demonstrating Eq. (5) can only be used in some linear models. For simplicity, let us consider the following model,

$$\arg\min_{\mathbf{B},\mathbf{W}} E = \|\mathbf{B} - f(\mathbf{X}\mathbf{W})\|_F^2, \tag{10}$$

where $\mathbf{X}$ is an orthogonal data matrix and $f(\cdot)$ is a linear or nonlinear function. Please note Eq. (10) is not a unimodal hashing model, because the binary constraint is not imposed to $\mathbf{B}$. Let us suppose the dimensions of $\mathbf{B}$ and $\mathbf{X}$ are equal. According to **Proposition 1**, Eq. (5) will lead to an orthogonal $\mathbf{W}$. If $f(\mathbf{X}\mathbf{W}) = \mathbf{X}\mathbf{W}$, then $\mathbf{B} = \mathbf{X}\mathbf{W}$ is also an orthogonal matrix. However, if $f(\cdot)$ is a sign function which is nonlinear, we can get a binary code matrix $\mathbf{B} = sign(\mathbf{X}\mathbf{W})$ and Eq. (10) becomes a nonlinear unimodal hashing model. Obviously, an orthogonal $\mathbf{W}$ cannot ensure an orthogonal $\mathbf{B}$.

Inspired by this example, we propose the following regularization

$$\left\|\frac{f^\top(\mathbf{X},\Theta)f(\mathbf{X},\Theta)}{n} - \mathbf{I}\right\|_F^2, \tag{11}$$

where $f(\mathbf{X},\Theta)$ is the nonlinear embedding function and Θ is the parameter set of f. In our proposed hashing model, i.e., Eq. (4),

$$f(\mathbf{X}^i,\mathbf{W}^i,\mathbf{v}^i,\beta_i) = \frac{1}{1+\exp\left(-(\beta_i\mathbf{X}^i\mathbf{W}^i + \mathbf{1}\mathbf{v}^i)\right)}. \tag{12}$$

Proposition 4: Minimizing Eq. (11) cannot lead to an orthogonal $f(\mathbf{X}^i,\Theta)$ when $d_i+1 < c$. *Proof:* According to the definitions, we have $rank(\mathbf{X}^i) \le d_i$, $rank(\mathbf{W}^i) \le d_i$ and $rank(\mathbf{1}\mathbf{v}^i) \le 1$. Hence, $rank(\mathbf{X}^i\mathbf{W}^i) \le d_i$ and $rank(\mathbf{X}^i\mathbf{W}^i + \mathbf{1}\mathbf{v}^i) \le d_i + 1$.

According to **Theorem 4.2** in [51], $rank(f(\mathbf{X}^i\mathbf{W}^i+\mathbf{1}\mathbf{v}^i)) \le d_i + 1$, and hence

$$rank(f^\top(\mathbf{X}^i\mathbf{W}^i + \mathbf{1}\mathbf{v}^i)f(\mathbf{X}^i\mathbf{W}^i + \mathbf{1}\mathbf{v}^i)) \le d_i + 1. \tag{13}$$

Because $d_i+1 < c, f^\top(\mathbf{X}^i\mathbf{W}^i+\mathbf{1}\mathbf{v}^i)f(\mathbf{X}^i\mathbf{W}^i+\mathbf{1}\mathbf{v}^i)$ cannot be equal to $\mathbf{I}$ in any cases. Hence, Minimizing Eq. (11) cannot lead to an orthogonal f. □

From **Proposition 4**, we can see that an orthogonal f cannot be acquired when $d_i + 1 < c$. In this case, f cannot even approximate an orthogonal matrix. Minimizing f will only minimize the correlation among the column vectors of f. Fig. 3 illustrates this situation.

It is inessential to name Eq. (11) as "minimum correlation regularization" (MCR) or "maximum uncorrelation regularization". Since Eq. (11) will be added into our hashing model which is formulated as a minimization problem, we use the former one to keep literal consistency.

C. DECORRELATED MULTIMODAL HASHING

In our implementation, we found that subtracting identity matrix is somewhat redundant, so MCR can be simplified as:

$$\left\|\frac{f^\top(\mathbf{X},\Theta)f(\mathbf{X},\Theta)}{n}\right\|_F^2. \tag{14}$$

It is unnecessary to worry about the diagonal elements of $f^\top f$ will be zeros during the proposed minimization procedure, because as long as all variables are randomly initialized, it is nearly impossible for gradient descent algorithm to reach a solution that all variables are zero.

Adding MCR to Eq. (4) leads to the following model.

$$\arg\min_{\mathbf{B},\mathbf{W}^i,\mathbf{v}^i} E = \sum_i \alpha_i \left(\left\|\mathbf{B} - \mathbf{C}^i\right\|_F^2 + \gamma_i \left\|{\mathbf{C}^i}^\top \mathbf{C}^i\right\|_F^2\right), \tag{15}$$

where γ_i is a positive real constant, and

$$\mathbf{A}^i = \exp\left(-(\beta_i\mathbf{X}^i\mathbf{W}^i + \mathbf{1}\mathbf{v}^i)\right), \tag{16}$$

$$\mathbf{C}^i = \frac{1}{1+\mathbf{A}^i}. \tag{17}$$

D. OPTIMIZATION

Eq. (15) is minimized by iterative minimization. Take the partial derivative with respect to $\mathbf{B}$, resulting in

$$\frac{\partial E}{\partial \mathbf{B}} = 2\sum_i \alpha_i \mathbf{B} - 2\sum_i \alpha_i \mathbf{C}^i. \tag{18}$$

Setting Eq. (18) as 0, we can derive that

$$\mathbf{B} = \frac{\sum_i \alpha_i \mathbf{C}^i}{\sum_i \alpha_i}. \tag{19}$$

$\mathbf{B}$ is rounded in each iteration to ensure $\mathbf{B} \in \{0, 1\}^{n\times c}$.

R Take the partial derivative with respect to $\mathbf{v}^i$, resulting in

$$\frac{\partial E}{\partial \mathbf{v}^i} = 2\alpha_i \mathbf{1}^\top \left(\mathbf{C}^i - \mathbf{B} + \frac{\gamma_i}{n}\mathbf{C}^i \mathbf{C}^{i\top}\mathbf{C}^i\right) \circ \left(\mathbf{A}^i \circ \mathbf{C}^{i2}\right). \tag{20}$$

In Eq. (20), "$\circ$" means element-wise multiplication. The division and square are also element-wise. The partial derivative with respect to $\mathbf{W}^i$ is

$$\frac{\partial E}{\partial \mathbf{W}^i} = 2\alpha_i \beta_i \mathbf{X}^{i\top} \left(\mathbf{C}^i - \mathbf{B} + \frac{\gamma_i}{n}\mathbf{C}^i \mathbf{C}^{i\top}\mathbf{C}^i\right) \circ \left(\mathbf{A}^i \circ \mathbf{C}^{i2}\right).$$

The prototype of the proposed training method is shown in **Algorithm 1**. In Subsection III-F, the parameter settings and details for efficient implementation are discussed.

Algorithm 1 the Prototype of the Proposed Training Method

Require: α_i, β_i, $\triangle t$, $\mathbf{X}^i$

1: **while** E not converged **do**
2: Update $\mathbf{B}$ using Eq. (18).
3: $\mathbf{v}^i \leftarrow \mathbf{v}^i - \triangle t \cdot \partial E/\partial \mathbf{v}^i$
4: $\mathbf{W}^i \leftarrow \mathbf{W}^i - \triangle t \cdot \partial E/\partial \mathbf{W}^i$
5: **end while**

Ensure: $\mathbf{B}$, $\mathbf{W}^i$, $\mathbf{v}^i$

E. LABEL RELAXATION

As discussed in Section I, relaxing a few labels to real numbers can benefit on learning hashing codes on multi-labeled data sets. Let us denote label matrix as $\mathbf{L}$ which is the g-th modality and without losing generality, the last r rows are extracted from $\mathbf{L}$ to generate a new matrix $\mathbf{L}_R$ and the remaining $n - r$ rows of $\mathbf{L}$ form matrix $\mathbf{L}_T$. $\mathbf{L}_R$ is used for relaxation. The relaxed $\mathbf{L}_R$ is denoted as $\tilde{\mathbf{L}}_R$.

Let us define $A^i_{pq} = exp(\rho(\mathbf{x}^i_p, \mathbf{x}^i_q))$, where $\mathbf{x}^i_p$ is the p-th row of $\mathbf{X}^i$ and ρ is Euclidean distance. Let us define

$$H_{pq} = \frac{1}{g-1}\sum_i^g A^i_{pq} / \sum_r^n A_{pr}, \tag{21}$$

where H_{pq} is used to build matrix $\mathbf{H}$. A^i_{pq} reflects the data structure of i-th modality. H_{pq} integrates the data structure of all modalities by averaging normalized A^i_{pg}. To make $\tilde{\mathbf{L}}_R$ reflects the data structure, the following objective function can be used:

$$\arg\min_{\tilde{\mathbf{L}}_R} O = trace\left(\begin{bmatrix}\tilde{\mathbf{L}}_T \\ \mathbf{L}_R\end{bmatrix}^\top \begin{bmatrix}\mathbf{H}_{TT} & \mathbf{H}_{TR} \\ \mathbf{H}_{RT} & \mathbf{H}_{RR}\end{bmatrix} \begin{bmatrix}\mathbf{L}_T \\ \tilde{\mathbf{L}}_R\end{bmatrix}\right), \tag{22}$$

where $\mathbf{H}$ is partitioned into four blocks according the dimensions of $\tilde{\mathbf{L}}_R$ and $\mathbf{L}_T$. On the other hand, the original labels $\mathbf{L}_R$ also contain useful information. Hence, $\|\tilde{\mathbf{L}}_R - \mathbf{L}_R\|^2_F$ is added to the above objective function:

$$\arg\min_{\tilde{\mathbf{L}}_R} O = trace\left(\begin{bmatrix}\tilde{\mathbf{L}}_T \\ \mathbf{L}_R\end{bmatrix}^\top \begin{bmatrix}\mathbf{H}_{TT} & \mathbf{H}_{TR} \\ \mathbf{H}_{RT} & \mathbf{H}_{RR}\end{bmatrix} \begin{bmatrix}\mathbf{L}_T \\ \tilde{\mathbf{L}}_R\end{bmatrix}\right) + \|\tilde{\mathbf{L}}_R - \mathbf{L}_R\|^2_F. \tag{23}$$

Gradient descent algorithm is used to minimize Eq. (23). The gradient of Eq. (23) with respect to $\tilde{\mathbf{L}}_R$ is

$$\frac{\partial O}{\partial \tilde{\mathbf{L}}_R} = 2\mathbf{H}_{RT}\mathbf{L}_T + 2\mathbf{H}_{RR}\tilde{\mathbf{L}}_R + 2(\tilde{\mathbf{L}}_R - \mathbf{L}_R). \tag{24}$$

After getting $\tilde{\mathbf{L}}_R$, let $\mathbf{L} = \begin{bmatrix}\mathbf{L}_T \\ \tilde{\mathbf{T}}_R\end{bmatrix}$.

F. IMPLEMENTATION DETAILS

α_i is the weight for ith view. We set α_i as 10 for the label view and 1 for any other views. β_i is used to re-scale the view matrix. We empirically found that the proposed method achieves the best performance when the values of the re-scaled view matrix are in the interval [0, 255]. For instance, in the NUS-WIDE data set [52], images are represented by 500-dimensional bag-of-visual-words SIFT feature vectors whose values are in [0, 255], texts are represented by 1000-dimensional index vectors whose values are 0 or 1 and labels are 10-dimensional index vectors. Hence, we set β as 1, 255 and 255 for image view matrix, text view matrix and label view matrix, respectively. To improve computation efficiency, β_i is multiplied with $\mathbf{X}_i$ before the iteration starts. All data matrices are zero-centered, except for the label matrix.

We set the maximum iteration times as K. $\triangle t$ linearly decreases from k_s to k_e by K iterations, i.e., in the k-th iteration, $\triangle t = k_s - (k_s - k_e)k/K$.

For large data set, the first term in Eq. (15) is too large, which makes γ_i and $\triangle t$ difficult to be determined. We normalize the gradients so that we can fix γ_i and $\triangle t$ settings for all our experiments. The efficient version of the proposed method is given in **Algorithm** 2.

Algorithm 2 the Proposed Training Method

Require: α_i, β_i, $\triangle t$, $\mathbf{X}^i$, k, k_s, k_e, K

1: **while** E not converged and $k < K$ **do**
2: $\triangle t = k_s - (k_s - k_e)k/K$
3: Update $\mathbf{B}$ using Eq. (18).
4: $\mathbf{v}^i \leftarrow \mathbf{v}^i - \triangle t \cdot \frac{\partial E/\partial \mathbf{v}^i}{\|\partial E/\partial \mathbf{v}^i\|_F}$
5: $\mathbf{W}^i \leftarrow \mathbf{W}^i - \triangle t \cdot \frac{\partial E/\partial \mathbf{W}^i}{\|\partial E/\partial \mathbf{W}^i\|_F}$
6: $k \leftarrow k + 1$
7: **end while**

Ensure: $\mathbf{B}$, $\mathbf{W}^i$, $\mathbf{v}^i$

IV. EXPERIMENTAL RESULTS

In this section, we evaluate the retrieval performance and computational efficiency of the proposed method. First,

TABLE 1. MAP results on Wiki, MIRFlickr and NUS-WIDE data sets.

Task	Data set	Wiki				MIRFlickr				NUS-WIDE			
	Methods	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
I2T	CMSSH	0.3213	0.2927	0.2831	0.2747	0.5966	0.5674	0.5581	0.5701	0.4152	0.3515	0.3510	0.3556
	CVH	0.3227	0.2981	0.2149	0.1772	0.6591	0.6145	0.6133	0.6052	0.4794	0.4195	0.3901	0.3501
	SePH	0.2280	0.2315	0.2387	0.2457	0.6505	0.6447	0.6453	0.6612	0.7185	0.7258	0.7390	0.7491
	MDBE	0.3963	0.4205	0.4173	0.4389	0.6784	0.7050	0.7083	0.7156	0.7623	0.7737	0.7953	0.7987
	DMHOR	0.1920	0.1859	0.1841	0.1853	0.5854	0.5835	0.5827	0.5831	0.3677	0.3639	0.3681	0.3039
	SSAH	0.3579	0.3681	0.3991	0.4012	0.7822	0.7901	0.8004	0.8012	0.6423	0.6364	0.6386	0.6382
	DJSRH	0.3884	0.4029	0.4115	0.4205	0.8103	0.8427	0.8619	0.8758	0.7241	0.7728	0.7980	0.8165
	DMH	**0.4372**	**0.4581**	**0.4654**	**0.4807**	**0.8224**	**0.8459**	**0.8795**	**0.8832**	**0.7653**	**0.7827**	**0.8150**	**0.8246**
T2I	CMSSH	0.3922	0.2917	0.2845	0.2708	0.6613	0.6510	0.6756	0.6471	0.4124	0.3533	0.3540	0.3600
	CVH	0.3180	0.2783	0.1635	0.1531	0.6495	0.6213	0.6179	0.5948	0.4733	0.3505	0.2900	0.2950
	SePH	0.6073	0.6110	0.6259	0.6192	0.6745	0.6824	0.6917	0.7110	0.5573	0.5481	0.5589	0.5569
	MDBE	0.6675	0.6739	0.6841	0.6937	0.7521	0.7793	0.7894	0.7919	0.6281	0.6409	0.6617	0.6644
	DMHOR	0.4298	0.4851	0.4913	0.4905	0.5682	0.5673	0.5565	0.5634	0.3842	0.3645	0.3529	0.3505
	SSAH	0.5235	0.5578	0.5723	0.5940	0.7987	0.7948	0.8025	0.8017	0.6685	0.6615	0.6659	0.6654
	DJSRH	0.6113	0.6354	0.6458	0.6583	0.7859	0.8221	0.8347	0.8465	0.7124	0.7443	0.7712	0.7888
	DMH	**0.6849**	**0.6971**	**0.7093**	**0.7239**	**0.7939**	**0.8317**	**0.8462**	**0.8570**	**0.7317**	**0.7506**	**0.7991**	**0.8037**

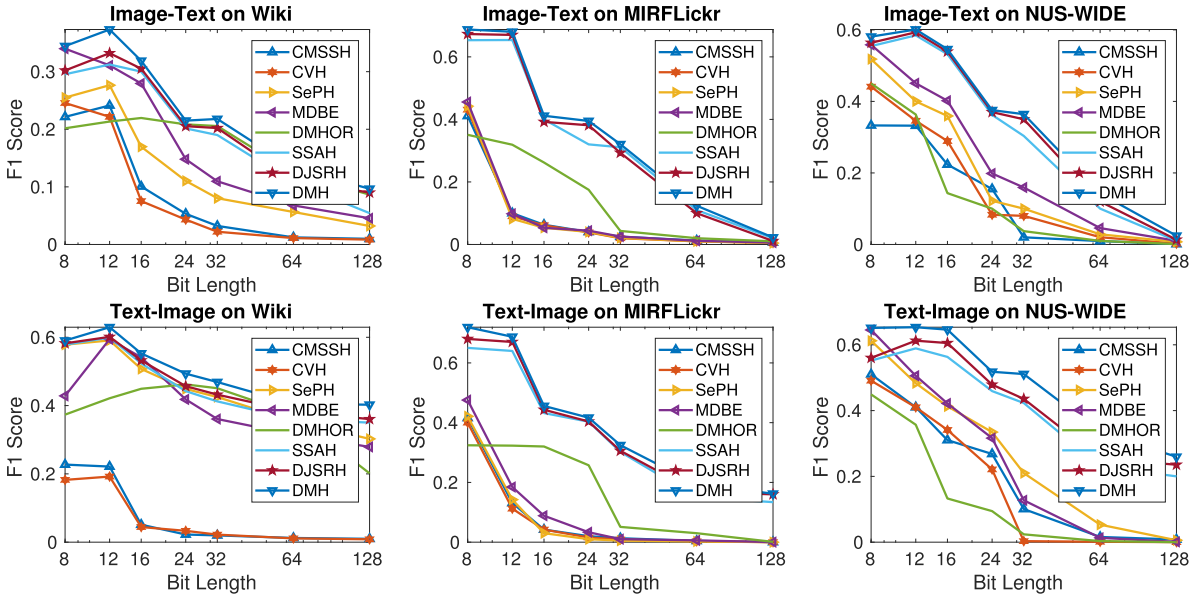


FIGURE 4. F1-score on Wiki, MIRFlickr and NUS-WIDE data sets. The "image-query-text" (Image-Text) and "text-query-image" (Text-Image) results are shown in the first row and second row, respectively.

we introduce the data sets, evaluation metrics and comparison methods. Then, two types of experiments - *Hamming ranking* and *hash lookup* were conducted. Finally, we analyze the convergence and computational efficiency.

A. DATA SETS

Wiki[1] contains 2,866 image and text pairs. Each image is represented by a 4,096-dimensional feature extracted by the Caffe implementation of AlexNet [53] as [41] did and each text is represented by a 10-dimension topics' vector generated by latent Dirichlet allocation (LDA) model. Each pair uniquely belongs to one of the 10 categories. Ground-truth neighbors for a test entry is defined as those in the same category.

[1]http://www.svcl.ucsd.edu/projects/crossmodal/

MIRFlickr [54] contains 25,000 entries each of which consists of 1 image, several textual tags and labels. Following literature [39], we only keep those textural tags appearing at least 20 times and remove entries which have no label. Hence, 20,015 entries are left. For each entry, the image is represented by a 512-dimensional feature extracted by Resnet-18 [55] and the text is represented by a 500-dimensional feature vector derived from PCA on index vectors of the textural tags. 5% entries are randomly selected for testing and the remaining entries are used as training set. Ground-truth semantic neighbors for a test entry, i.e, a query, are defined as those sharing at least one label.

NUS-WIDE [52] is comprised of 269,648 images and over 5,000 textural tags collected from Flickr. Ground-truth of 81 concepts is provided for the entire data set. Following literatures [33], [39], [40], we select 10 most common

concepts for labels and thus 186,577 entries are left. For each entry, the image is represented as a 512-dimensional feature extracted by Resnet-18 and text is represented as an index vector of the most frequent 1,000 tags. 1% entries are randomly selected for testing and the remaining are used for training. Ground-truth semantic neighbors for a test entry are defined as those sharing at least one label.

For image feature extraction neural networks, AlexNet and Resnet-18, the weights pretrained on ImageNet [56] are used. Fine-tuning is done on Wiki, MIRlickr and NUS-WIDE datasets. For fune-tuning, we resize the images to 244×244, use Adam Optimizer [57] with default settings and run the training process for 10 epoches with batch size 32.

B. EVALUATION METRICS

Hamming ranking and *hash lookup* are two widely used experiments for evaluating retrieval performance. In Hamming ranking experiment, all data points in the training set are ranked depending on their Hamming distances to a given query. The average precision (AP) is defined as

$$AP = \frac{1}{N}\sum_{r=1}^{R} P(r)\delta(r), \tag{25}$$

where N is the number of relevant instances in the retrieved set, $P(r)$ is the precision of the top r retrieved instances, and $\delta(r) = 1$ if the r-th retrieved instance is a true neighbor of the query, and otherwise $\delta(r) = 0$. Mean average precision (MAP) is the mean of APs of all the queries. For the ideal case that all retrieved instance are true neighbors of the queries, MAP is equal to 1, while MAP is equal to 0 for the worst case that all retrieved instance are not the true neighbors. Hence, the closer it is to 1, the better the performance.

In *hash lookup* experiment, the retrieved instances are those whose Hamming distances to a given query are not larger than a given radius, say 2 in our experiment. The performance are evaluated by F1-score which is defined as

$$F1 = 2\frac{precision \cdot recall}{precision + recall}. \tag{26}$$

The F1-scores are averaged for all queries. Similar to MAP, F1 also varies in [0, 1] and the closer it is to 1, the better the performance.

C. BASELINES

The proposed method is compared with seven multimodal hashing methods CMSSH [36], CVH [37], MDBE [41], SCM [40], SePH [39], DMHOR [7], DJSRH [58] and SSAH [48]. DMHOR, DJSRH and SSAH are based on deep neural networks.

CMSSH and SePH requires too much computational cost. Following literatures [39], [40], 10,000 entries are randomly selected for training hashing functions and then we apply these functions to generate hashing codes. We use the codes provided by the authors except for MDBE and DMHOR. We re-implement MDBE and DMHOR, and set parameters following the authors' suggestions. For our method, we use the following parameter settings, $k_s = 0.003$, $k_e = 0.0015$ and $K = 400$. α_i, β_i and γ_i are set as discussed in Subsection III-F.

D. RESULTS

MAP results are shown in Table 1. In Table 1, "I2T" means using images to query texts, while "T2I" means using texts to query images. From Table 1, it can be observed that our method outperforms all compared methods. As the bit length increases, the performance of our method increases faster than baselines, which demonstrates the effectiveness of the proposed minimum correlation regularization. For example, in the "Image-Text" experiment on MIRFlickr, the performance improvement ranges from 3% to 5% as the bit length varies from 16 to 128, compared to the best baseline, i.e., MDBE. The MAP of DMHOR decreases as the code length increases, which demonstrates the inefficiency of the orthogonality proposed in [7] as discussed in Subsection III-B.

F1-score results are shown in Fig. 4. Similar to the MAP results, our method surpasses all baselines by a huge performance improvement, especially on MIRFlickr. On MIRFlickr, the performance improvement ranges from 30% to 3,000%, compared to the best baseline. On NUS-WIDE, it is 5% to 200%. A reasonable explanation is that our method can precisely preserve the inter-class structure and therefore the lookup performance is significantly improved. Because the ranking performance depends on the preservation of the structure of the whole data set regardless of inter-class or intra-class structure, the performance improvement is not as significant as that of the lookup experiment. The size of MIRFlickr is only about 1/10 of NUS-WIDE, so the simple non-linearity introduced in our method works much better on MIRFlickr. To achieve comparable performance improvement on NUS-WIDE data set, more sophisticated non-linear models are expected.

In both experiments, MDBE achieves the best performance among all the baselines. Actually, the main part of MDBE,

$$\|\mathbf{LU} - \mathbf{XW_x}\|_F^2 + \|\mathbf{LU} - \mathbf{YW_y}\|_F^2, \tag{27}$$

is equivalent to Eq. (3) which is an intuitive multimodal extension of ITQ, where L is the label matrix, X is the image view matrix and Y is the text view matrix. W_x, W_y and U are variables. If we treat the label matrix as another view of the data and introduce an auxiliary variable B, it is easy to figure out that Eq. (27) and Eq. (3) are equivalent. By introducing non-linearity and minimum correlation regularization, our method performs much better than MDBE. An illustrative experiment on MIRFlickr data set are shown in Fig. 5.

E. PARAMETER SETTINGS

In Fig. 6, we show the MAP and F1-score of DMH on MIRFlickr data set with various parameter settings. The default setting is $\alpha = 10$, $\beta = 255$ and $\gamma = 0.001$. For label relaxation, we set $r = 1$ for generating relaxed label matrix

FIGURE 5. Some examples of text-query-image retrieval on MIRFlickr data set. The query text and its corresponding image are chosen from the sport class. Top five retrieved images of different methods are shown below the dotted line. Irrelevant images are with red bounding boxes.

L_R. In each figure, only the tested parameter varies and the other two parameters keep their default values.

In the left column of Fig. 6, α varies in $\{1, 5, 10, 15, 20, 25\}$. It can be seen that the highest MAP is usually achieved by $\alpha = 5$ or $\alpha = 10$. The highest F1-score is got when $\alpha = 1$. However, when $\alpha = 1$, DMH performs badly in MAP. Hence, $\alpha = 10$ is selected for our experiments to achieve a balanced performance on these two types of experiments.

In the middle column of Fig. 6, β varies in $2^{\{1,2,4,6,8,9\}} - 1$. In the long-bit experiment ($c > 16$), the performance is relatively robust to β. The highest F1-score is achieved when $\beta = 255$. Hence, $\beta = 255$ is used in our experiments.

In the right column of Fig. 6, γ varies in $10^{\{-5,-4,-3,-2,-1,0,1\}}$. It can be seen that DMH performs best in MAP when $\gamma = 0.001$. When $\gamma > 0.1$, F1-score rockets up, while MAP dumps. A possible explanation is that the regularization overly decorrelates a few columns of the code matrix and leaves other columns highly mutually correlated. The resulting code matrix will be similar to a short-bit code matrix. That is why MAP and F1-scores in all 6 experiments with different lengths of bits are rather close in this situation. Although the global optimum of MCR tends to generate column vectors similar to those illustrated in Fig. 3, the gradient descent algorithm cannot guarantee such solutions since MCR is not convex. Hence, $\gamma = 0.001$ is used in our experiments.

F. CONVERGENCE STUDY

The objective function of our method is minimized by **Algorithm** 2. In **Algorithm** 2, we empirically amend the derivatives of E for easy parameter tuning. The convergence property is experimentally studied in this subsection. Fig. 7 shows the convergence curves. It can be seen that the objective function value decreases fast in the first 100 iterations and then slides relatively slowly except for that of Wiki data set. The preset iteration step is too large for Wiki data set, so the object function value increase incrementally after reaching the smallest value. The convergence curves of experiments on Wiki and MIRFlickr is smooth, while those of experiments on NUS-WIDE jitters because of more sophisticated data structure and therefore more saddle points across which the algorithm jumps.

G. COMPUTATION EFFICIENCY

Training and testing time on 32-bit are given in Table 2. The training time is the mean time of 10 runs. The testing time is the average time cost for one query. All experiments were

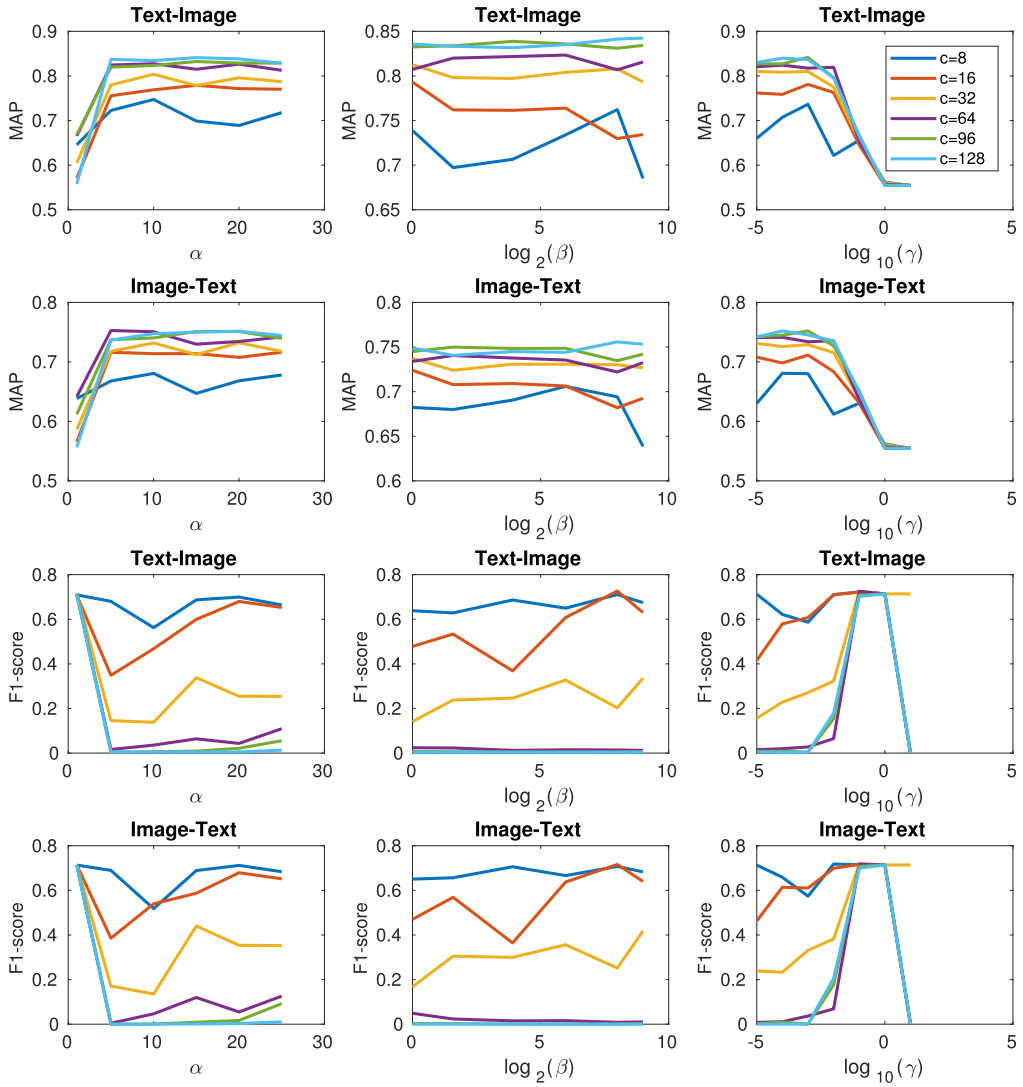


FIGURE 6. MAP and F1-score of DMH on MIRFlickr data set. The first two rows are MAP and the last two rows are F1-score.

performed on MATLAB R2015b installed on a GNU/Linux Server with 2.30 GHz 16-core CPU and 768 GB RAM. The three compared deep models, i.e. DMHOR [7], SSAH [48] and DJSRH [58], were trained on a NVIDIA GeForce 1080TI GPU. It is meaningless to compare running time of methods implemented on different platforms. Hence, the running time of these three deep models are not reported in Table 2. From Table 2, it can be seen that the training time of our method is moderate among all methods. Its testing time is close to that of MDBE, because the encoding procedure for a new query of these two methods are similar.

TABLE 2. Training and Testing Time on MIRFlickr and NUS-WIDE data sets in seconds. The testing time is multiplied with 10^{-5}.

	MIRFlickr		NUS-WIDE	
Method	Training	Testing	Training	Testing
CMSSH	69.7	1.016	705.2	1.270
CVH	0.9	0.910	3.6	1.087
SePH	4711.2	4.244	5082.3	5.550
MDBE	25.0	0.431	241.8	0.572
DMH	29.8	0.432	398.0	0.572

H. COMPARISON OF REGULARIZATIONS

In order to prove the efficiency of the proposed regularization, we imposed four different types of regularization on our method, i.e., 1) no regularization, 2) regularization proposed in [7], 3) Eq. (11) and 4) Eq. (14). The MAP on Wiki data set are shown in Fig. 8. From Fig. 8, we can see that the proposed regularization can improve the performance on experiments of long codes (>64 bits). It is difficult to judge the effects of the regularization proposed in [7], since the performance improvement was not guaranteed on all experiments. The

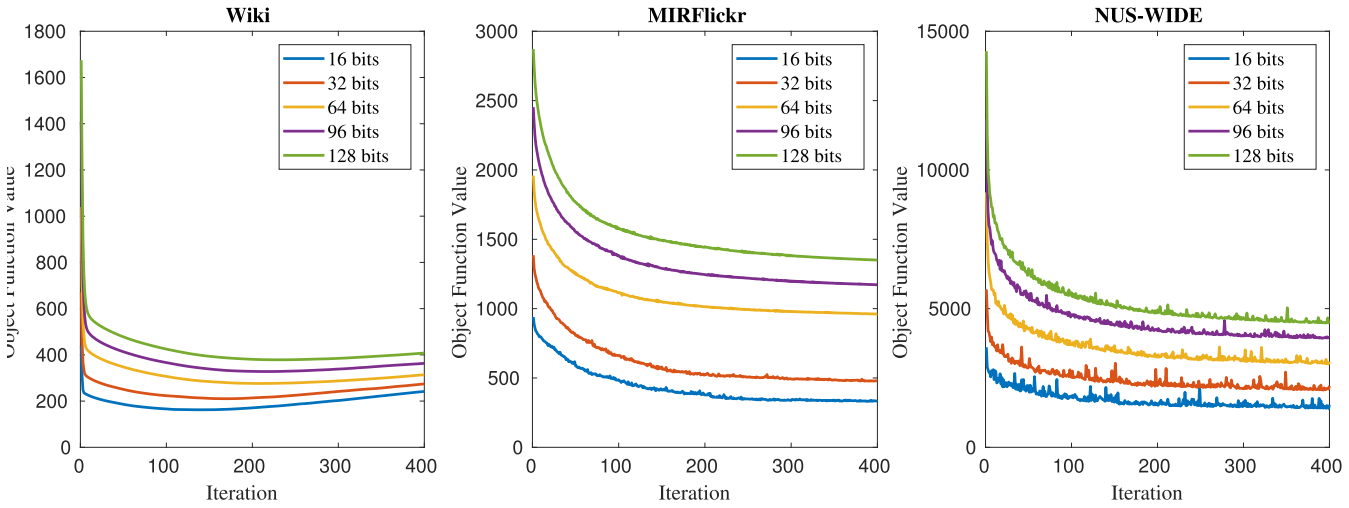


FIGURE 7. Convergence curves on Wiki, MIRFlickr and NUS-WIDE data sets.

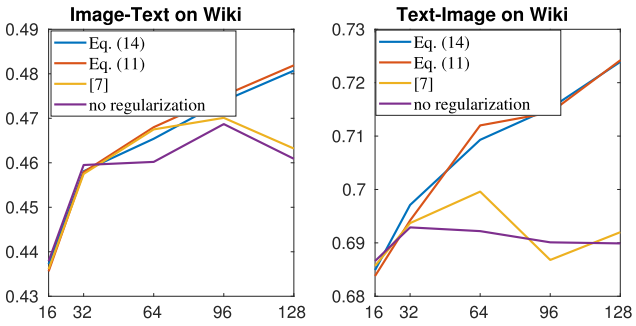


FIGURE 8. MAP of different regularizations on Wiki data set.

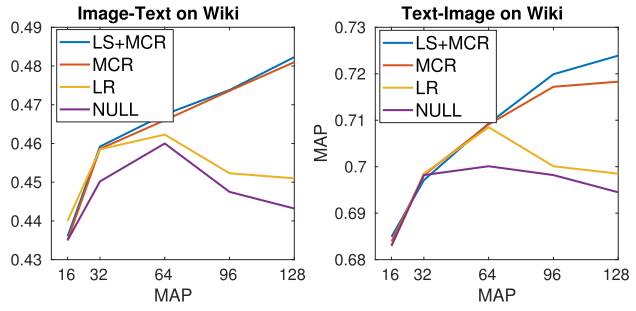


FIGURE 9. Ablation study on Wiki data set.

performance of Eq. (11) and Eq. (14) is close. Hence, it is preferred to use Eq. (14) due to its low computational cost.

I. ABLATION STUDY

To evaluate the effects of minimum correlation regularization (MCR) and label relaxation (LR) on our proposed method, we evaluated our methods on Wiki dataset in four settings: (1) with both MCR and LR, (2) with only MCR, (3) with only LR and (4) with neither MCR nor LR. The four settings are denoted as "MCR+LR", "MCR", "LR" and "NULL" in Fig. 9. From Fig. 9, we can conclude that the MCR is important for long-bit experiments. In short-bit experiments, the MAP improved by LR and MCR are subtle. However, for code length longer than 64 bits, the benefits from MCR and LR become significant. LR stably improves the MAP on our methods with or without MCR.

V. CONCLUSION

This paper proposed an effective multimodal hashing method which is modeled as a quantization error problem and the minimum correlation regularization is devised to improve the retrieval performance on long codes. Experiments on MIRFlickr and NUS-WIDE data sets show that the proposed method surpasses the compared methods distinctively. Future works include testing more nonlinear embedding functions and refining optimization procedure for high computational efficiency.

REFERENCES

[1] G. Cao, A. Iosifidis, K. Chen, and M. Gabbouj, "Generalized multi-view embedding for visual recognition and cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2542–2555, Sep. 2018.
[2] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.
[3] X. Zhang, S. Wang, Z. Li, and S. Ma, "Landmark image retrieval by jointing feature refinement and multimodal classifier learning," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1682–1695, Jun. 2018.
[4] J. Xie, G. Dai, F. Zhu, L. Shao, and Y. Fang, "Deep nonlinear metric learning for 3-D shape retrieval," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 412–422, Jan. 2018.
[5] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1753–1760.
[6] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *Proc. CVPR*, Jun. 2011, pp. 817–824.
[7] D. Wang, P. Cui, M. Ou, and W. Zhu, "Deep multimodal hashing with orthogonal regularization," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 2291–2297.
[8] W. Liu, C. Mu, S. Kumar, and S.-F. Chang, "Discrete graph hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3419–3427.

[9] V. E. Liong, J. Lu, G. Wang, P. Moulin, and J. Zhou, "Deep hashing for compact binary codes learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2475–2483.

[10] Z. Chen, J. Lu, J. Feng, and J. Zhou, "Nonlinear discrete hashing," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 123–135, Jan. 2017.

[11] 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.

[12] Z. Ji, W. Yao, W. Wei, H. Song, and H. Pi, "Deep multi-level semantic hashing for cross-modal retrieval," *IEEE Access*, vol. 7, pp. 23667–23674, 2019.

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

[14] Z. Li and J. Tang, "Unsupervised feature selection via nonnegative spectral analysis and redundancy control," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5343–5355, Dec. 2015.

[15] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2085–2098, Oct. 2015.

[16] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," *Commun. ACM*, vol. 51, no. 1, pp. 117–122, Jan. 2008.

[17] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1092–1104, Jun. 2012.

[18] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," in *Proc. 24th Annu. ACM Symp. Theory Comput. STOC*, 2002, pp. 380–388.

[19] F. Shen, C. Shen, Q. Shi, A. van den Hengel, and Z. Tang, "Inductive hashing on manifolds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1562–1569.

[20] D. Tian and D. Tao, "Global hashing system for fast image search," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 79–89, Jan. 2017.

[21] W. Liu, J. Wang, and S.-F. Chang, "Hashing with graphs," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 1–8.

[22] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *J. Educ. Psychol.*, vol. 24, no. 7, pp. 498–520, Sep. 1933.

[23] B. Xu, J. Bu, Y. Lin, C. Chen, X. He, and D. Cai, "Harmonious hashing," in *Int. Joint Conf. Artif. Intell.*, 2013, pp. 1820–1826.

[24] M. Norouzi and D. J. Fleet, "Cartesian K-Means," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3017–3024.

[25] C. Strecha, A. M. Bronstein, M. M. Bronstein, and P. Fua, "LDAHash: Improved matching with smaller descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 66–78, Jan. 2012.

[26] J. Tang, Z. Li, M. Wang, and R. Zhao, "Neighborhood discriminant hashing for large-scale image retrieval," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2827–2840, Sep. 2015.

[27] J.-P. Heo, Y. Lee, J. He, S.-F. Chang, and S.-E. Yoon, "Spherical hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2957–2964.

[28] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, p. 321, Dec. 1936.

[29] Z. Li and J. Tang, "Weakly supervised deep metric learning for community-contributed image retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1989–1999, Nov. 2015.

[30] Z. Li and J. Tang, "Weakly supervised deep matrix factorization for social image understanding," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 276–288, Jan. 2017.

[31] J. Tang and Z. Li, "Weakly supervised multimodal hashing for scalable social image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2730–2741, Oct. 2018.

[32] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, "Linear cross-modal hashing for efficient multimedia search," in *Proc. 21st ACM Int. Conf. Multimedia MM*, 2013, pp. 143–152.

[33] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2083–2090.

[34] 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. SIGIR*, 2014, pp. 415–424.

[35] X. Shen, F. Shen, Q.-S. Sun, Y. Yang, Y.-H. Yuan, and H. T. Shen, "Semi-paired discrete hashing: Learning latent hash codes for Semi-paired cross-view retrieval," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4275–4288, Dec. 2017.

[36] 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.

[37] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2011, pp. 1360–1365.

[38] Y. Zhen and D.-Y. Yeung, "A probabilistic model for multimodal hash function learning," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining KDD*, 2012, pp. 940–948.

[39] Z. Lin, G. Ding, J. Han, and J. Wang, "Cross-view retrieval via probability-based semantics-preserving hashing," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4342–4355, Dec. 2017.

[40] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2177–2183.

[41] D. Wang, X. Gao, X. Wang, L. He, and B. Yuan, "Multimodal discriminative binary embedding for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4540–4554, Oct. 2016.

[42] Z. Chen, F. Zhong, G. Min, Y. Leng, and Y. Ying, "Supervised Intra- and inter-modality similarity preserving hashing for cross-modal retrieval," *IEEE Access*, vol. 6, pp. 27796–27808, 2018.

[43] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2494–2507, May 2017.

[44] M. Hu, Y. Yang, F. Shen, N. Xie, R. Hong, and H. T. Shen, "Collective reconstructive embeddings for cross-modal hashing," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2770–2784, Jun. 2019.

[45] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5585–5599, Nov. 2018.

[46] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, "Deep adversarial metric learning for cross-modal retrieval," *World Wide Web*, vol. 22, no. 2, pp. 657–672, Mar. 2019.

[47] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Trans. Cybern.*, early access, Jul. 24, 2019, doi: 10.1109/TCYB.2019.2928180.

[48] 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.

[49] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1150–1157.

[50] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[51] C. L. Wu and R. J. Adler, "Nonlinear matrix algebra and engineering applications—Part 1: Theory and linear form matrix," *J. Comput. Appl. Math.*, vol. 1, no. 1, pp. 25–37, Mar. 1975.

[52] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world Web image database from national university of Singapore," in *Proc. ACM Int. Conf. Image Video Retr. CIVR*, 2009, p. 48.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[54] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retr. MIR*, 2008, pp. 39–43.

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

[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

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

DAYONG TIAN received the B.S. and M.E. degrees from Xidian University, Xi'an, China, in 2010 and 2014, respectively, and the Ph.D. degree from the University of Technology, Sydney, NSW, Australia, in 2017. He is currently an Assistant Professor with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and machine learning, and in particular, on image restoration, image retrieval, and face recognition.

YIWEN WEI (Member, IEEE) received the Ph.D. degree in radio science from the School of Physics and Optoelectronic Engineering, Xidian University, Xi'an, China, in 2016. From 2016 to 2018, she worked as a Research Scientist with the Temasek Laboratories, National University of Singapore, Singapore. She is currently an Assistant Professor with the School of Physics and Optoelectronic Engineering Science, Xidian University, China. Her research interests include electromagnetic wave propagation and scattering in complex systems, computational electromagnetic, remote sensing, and parameters retrieval, and in particular applying machine learning methods on complex electromagnetic problems.

DEYUN ZHOU received the B.E., M.E., and Ph.D. degrees from Northwestern Polytechnical University (NWPU), in 1985, 1988, and 1991, respectively. He has been a Professor at NWPU, since 1997, where he has also been the Dean of the School of Electronics and Information, since 2012. His research interests include self-adaptive control, intelligent control theory, complex systems modeling, multiobjective optimization, information fusion, and aerial electronic systems.