A Fast Optimization Method for General Binary Code Learning

Fumin Shen, Xiang Zhou, Yang Yang, Jingkuan Song, Heng Tao Shen, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Hashing or binary code learning has been recognized to accomplish efficient near neighbor search, and has thus attracted broad interests in recent retrieval, vision, and learning studies. One main challenge of learning to hash arises from the involvement of discrete variables in binary code optimization. While the widely used continuous relaxation may achieve high learning efficiency, the pursued codes are typically less effective due to accumulated quantization error. In this paper, we propose a novel binary code optimization method, dubbed *discrete proximal linearized minimization (DPLM)*, which directly handles the discrete constraints during the learning process. Specifically, the discrete (thus nonsmooth nonconvex) problem is reformulated as minimizing the sum of a smooth loss term with a nonsmooth indicator function. The obtained problem is then efficiently solved by an iterative procedure with each iteration admitting an *analytical* discrete solution, which is thus shown to converge very fast. In addition, the proposed method supports a large family of empirical loss functions, which is particularly instantiated in this paper by both a supervised and an unsupervised hashing losses, together with the bits uncorrelation and balance constraints. In particular, the proposed DPLM with a supervised ℓ_2 loss encodes the whole NUS-WIDE database into 64-b binary codes within 10 s on a standard desktop computer. The proposed approach is extensively evaluated on several large-scale data sets and the generated binary codes are shown to achieve very promising results on both retrieval and classification tasks.**

***Index Terms*— Binary code learning, hashing, discrete optimization.**

Manuscript received April 7, 2016; revised June 28, 2016; accepted September 6, 2016. Date of publication September 22, 2016; date of current version October 7, 2016. This work was supported in part by the National Natural Science Foundation of China under Project 61502081, Project 61673299, Project 61572108, and Project 61632007, in part by the Fundamental Research Funds for the Central Universities under Project ZYGX2015kyqd017 and Project ZYGX2015J055, and in part by the Australian Research Council Project DP-140102164, Project FT-130101457, and Project LE-140100061. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiaochun Cao. *(Corresponding author: Fumin Shen.)*

F. Shen, X. Zhou, and Y. Yang are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: fumin.shen@gmail.com; johinfly@gmail.com; dlyyangg@gmail.com).

J. Song is with Columbia University, New York, NY 10027 USA (e-mail: jingkuan.song@gmail.com).

H. T. Shen is with the School of Information Technology and Electrical Engineering, The University of Queensland, QLD 4072, Australia, and also with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: shenhengtao@hotmail.com).

D. Tao is with the Centre for Artificial Intelligence and the Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: dacheng.tao@uts.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2612883

I. Introduction

BINARY coding (also known as hashing) has recently become a very popular research subject in information retrieval [8], [18], [44], computer vision [24], [45], [48], [50], machine learning [16], [39], *etc*. By encoding high-dimensional feature vectors (*e.g*, of documents, images, videos, or other types of data) to short hash codes, an effective hashing method is expected to accomplish efficient similarity search while preserving the similarities among original data to some extent. As a result, using binary codes to represent and search in massive data is a promising solution to handle large-scale tasks, owing to reduced storage space (typically several hundred binary bits per datum) and the low complexity of pairwise distance computations in a Hamming space.

The hashing techniques can be generally divided into two major categories: data-independent and data-dependent methods. Locality-Sensitive Hashing (LSH) [8] represents one large family of data-independent methods [4], [13], [14], [26], which generate hash functions via random projections. Although LSH is ensured to have high collision probability for similar data items, in practice LSH usually needs long hash bits and multiple hash tables to achieve both high precision and recall. The huge storage overhead may restrict its applications.

The other category, data-dependent or learning based hashing methods have witnessed a rapid development in the most recent years, due to the benefit that they can effectively and efficiently index and organize massive data with very compact binary codes. Different from LSH, data-dependent binary coding methods aim to generate short binary codes using the training data. A number of algorithms in this category have been proposed, including the unsupervised Spectral Hashing [38], [39], Binary Reconstructive Embedding (BRE) [12], PCA Hashing [36], Iterative Quantization (ITQ) [9], Circulant Binary Embedding (CBE) [46], Anchor Graph Hashing (AGH) [21], [23], Isotropic Hashing (IsoHash) [11], Inductive Manifold Hashing [29], Neighborhood Discriminant Hashing (NDH) [32], Binary Projection Bank (BPB) [19] *etc.*, and the supervised Minimal Loss Hashing (MLH) [25], Semi-Supervised Hashing (SSH) [36], Kernel-Based Supervised Hashing (KSH) [22], FastHash [17], Graph Cut Coding (GCC [7]), Supervised Discrete Hashing (SDH) [28] *etc*. The literature is comprehensive reviewed in [37] recently.

The binary constraints imposed on the target hash codes make the associated optimization problem very difficult to solve, which are generally NP-hard. To simplify the optimization, most of the methods in the literature adopt the following

two-step way: first solve a relaxed problem by discarding the discrete constraints, and then quantize the obtained continuous solution to achieve the approximate binary solution. This two-step scheme significantly simplifies the original discrete optimization. Unfortunately, such an approximate solution is typically of low quality and often makes the resulting hash functions less effective. This is possibly due to the accumulated quantization error, which is especially the case when learning long-length codes. Iterative Quantization (ITQ) [9] is an effective approach to decrease the quantization distortion by applying an orthogonal rotation to projected training data. One limitation of ITQ is that it learns orthogonal rotations over pre-computed mappings (*e.g*, PCA or CCA) and the separate learning procedure usually makes ITQ suboptimal.

It would help generate more effective hashes to directly optimize the binary codes without continuous relaxations. However, the importance of discrete optimization in hashing has been less taken into account by most existing hashing methods. Recent efforts in this direction either lead to intractable optimization or are restricted to specific losses thus not easy to generalize. Very recently, binary optimization was studied in the unsupervised discrete graph hashing (DGH [21]) and promising results were obtained compared to previous relaxed methods. One disadvantage of DGH is that it suffers from an expensive optimization due to the involvement of singular value decomposition in each optimization iteration. In the meantime, supervised discrete hashing (SDH [28]) formulated supervised hashing as a linear classification problem with binary codes, where the associated key binary quadratic program (BQP) was efficiently solved by the discrete cyclic coordinate descent (DCC). However, DCC is limited to solving the standard BQP problem and it is still unclear how to apply DCC to other hashing problems with different objectives. For instance, DCC is not ready to optimize with the uncorrelation and balance constraints, which are widely-used in the hashing literature [39].

To overcome these problems, in this work, we propose a fast discrete optimization method for the general binary code learning problem. Our main contributions are summarized as follows:

1) The general binary code learning problem with discrete constraints is rewritten as an unconstrained minimization problem with an objective comprising two parts: a smooth loss function and a nonsmooth indicator function. The smooth function characterizes the learning loss of target binary codes with training data, while the nonsmooth one indicates the binary domain of the optimizing codes. The simple reformulation greatly simplifies the nonconvex nonsmooth binary code optimization problem.
2) We propose a novel discrete optimization method, termed **Discrete Proximal Linearized Minimization (DPLM)**, to learn binary codes in an efficient iterative way. In each optimization iteration, The corresponding subproblem admits an analytical solution by directly investigating the binary code space. As such, a high-quality discrete solution without resort to the continuous relaxation can eventually be obtained in an efficient computing manner, therefore enabling to tackle massive datasets.
3) Different from other discrete optimization solvers in the hashing literature, the proposed method supports a large family of empirical loss functions. In this work, this method is particularly instantiated by the supervised ℓ_2 loss and unsupervised graph hashing loss. The well-known bits uncorrelation and balance constraints are also investigated in the proposed optimization framework.
4) Comprehensive evaluations are conducted on several representative retrieval benchmarks, and the results consistently validate the superiority of the proposed methods over the state-of-the-art in terms of both efficiency and efficacy. In addition, we also show that the binary codes generated by our algorithm perform very well on the image and scene classification problems.

The rest of the paper is organized as follows. Section II elaborates the details of the proposed DPLM method, which is instantiated by both a supervised and an unsupervised hashing objective in Section III, followed by the exploration of bits uncorrelation and balance constraints. In Section IV, we analyze the proposed discrete algorithm with comparison to the relaxed method and other optimization approach. In Section V, we evaluate our algorithm on several real-world large-scale datasets for both retrieval and classification tasks, followed by the conclusion of this work in Section VI.

II. Fast Binary Optimization for Hashing

Let us first introduce some notations. We denote matrices as boldface uppercase letters like $\mathbf{X}$, vectors as boldface lowercase letters like $\mathbf{x}$ and scalars as x. The $r \times r$ identity matrix is denoted as $\mathbf{I}_r$, and the vector with all ones and zeros as $\mathbf{1}$ and $\mathbf{0}$, respectively. We abbreviate the Frobenius norm $||\cdot||_F$ as $||\cdot||$ in this paper. ∇f denotes the gradient of function $f(\cdot)$. $\mathrm{sgn}(\cdot)$ is the sign function with output $+1$ for positive numbers and -1 otherwise. For binary codes, we use (1, -1) bits for mathematical derivations, and use (1, 0) bits for implementations of all referred binary coding and hashing algorithms.

A. The Binary Code Learning Problem

Suppose we have n samples $\mathbf{x}_i \in \mathbb{R}^d, i = 1, \cdots, n$, stored in matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$. For each sample $\mathbf{x}$, we aim to learn its r-bit binary code $\mathbf{b} \in \{-1, 1\}^r$. We consider the following general binary code learning problem

$$\begin{aligned} &\min_{\mathbf{B}} \ \mathcal{L}(\mathbf{B}) \\ &\text{s.t. } \mathbf{B} \in \{-1, 1\}^{r \times n}. \end{aligned} \tag{1}$$

Here $\mathbf{B}$ is the target binary codes for $\mathbf{X}$ and $\mathcal{L}(\cdot)$ is the smooth loss function. In this work, we aim at a scalable and computationally tractable method which can be applied to a large family of loss functions $\mathcal{L}(\cdot)$.

The binary constraints make problem (1) a mixed-integer optimization problem, which is generally NP-hard. Most previous methods resort to the continuous relaxation by discarding the discrete constraints. As aforementioned, however, this

relaxed solution may cause large error accumulation as the code length increases. This is mainly because the discrete constraints have not been treated adequately during the learning procedure, as shown in [21] and [28].

B. Discrete Proximal Linearized Minimization

In this section, we shown problem (1) can be solved in an efficient way while keeping the discrete variables in the optimization. To simplify the discrete optimization in problem (1), let us first introduce the following *indicator function*

$$\delta_C(\mathbf{B}) = \begin{cases} 0 & \text{if } \mathbf{B} \in C \\ +\infty & \text{otherwise,} \end{cases} \tag{2}$$

where C is a nonempty and closed set. Let $\mathbb{B}$ denotes the binary codes space $\{-1, 1\}^{r\times n}$. The function $\delta_{\mathbb{B}}(\mathbf{B})$ yields infinity as long as one entry of $\mathbf{B}$ does not belong to the binary domain $\{-1, 1\}$. With the indicator function, we are safe to rewrite problem (1) to an unconstrained minimization problem

$$\min_{\mathbf{B}} \mathcal{L}(\mathbf{B}) + \delta_{\mathbb{B}}(\mathbf{B}). \tag{3}$$

The simple reformulation of problem (1) to (3) greatly simplify the optimization therein, as shown below. The objective of (3) consists of two parts: a smooth function and a nonsmooth one. The smooth function $\mathcal{L}(\mathbf{B})$ models the hashing loss which can be chosen freely according to different problem scenarios, while the nonsmooth function $\delta_{\mathbb{B}}(\mathbf{B})$ indicates the domain of the optimizing hash codes.

Solving problem (3) is still nontrivial due to the involvement of nonsmooth indicator. Inspired by the recent advance in nonconvex and nonsmooth optimization [1], [2], we solve problem (3) with the following iterative procedure. Denote Prox_λ^f the *proximal operator* with function f and parameter λ:

$$\text{Prox}_\lambda^f(x) = \arg\min_y f(y) + \frac{\lambda}{2}||y - x||^2. \tag{4}$$

Suppose we have obtained the code solution $\mathbf{B}^{(j)}$ at the j_{th} iteration for problem (3). At the $(j+1)_{\text{th}}$ iteration, $\mathbf{B}$ is updated by

$$\mathbf{B}^{(j+1)} = \text{Prox}_\lambda^\delta\big(\mathbf{B}^{(j)} - \frac{1}{\lambda}\nabla\mathcal{L}(\mathbf{B}^{(j)})\big) \tag{5}$$

$$= \arg\min_{\mathbf{B}} \delta(\mathbf{B}) + \frac{\lambda}{2}||\mathbf{B} - \mathbf{B}^{(j)} + \frac{1}{\lambda}\nabla\mathcal{L}(\mathbf{B}^{(j)})||^2. \tag{6}$$

The optimization procedure with (5) is also known as the forward-backward splitting algorithm [1]. The forward-backward splitting scheme for minimizing the sum of a smooth function $\mathcal{L}(\cdot)$ with a nonsmooth one can simply be viewed as the proximal regularization of $\mathcal{L}(\cdot)$ linearized at a given point $\mathbf{B}^{(j)}$.

By transforming the indicator function back to the binary constraints, solution (6) leads to the following problem

$$\min_{\mathbf{B}} \ ||\mathbf{B} - \mathbf{B}^{(j)} + \frac{1}{\lambda}\nabla\mathcal{L}(\mathbf{B}^{(j)})||^2, \quad \text{s.t. } \mathbf{B} \in \{-1, 1\}^{r\times n}. \tag{7}$$

Remark 1: The derivation of problem (5) *and* (7) *is the key step of our algorithm. By looking at* (7)*, we can see that the problem actually seeks the projection of* $\mathbf{B}^{(j)} - \frac{1}{\lambda}\nabla\mathcal{L}(\mathbf{B}^{(j)})$ *onto the binary code space. Indeed, for the indicator function* $\delta(\mathbf{B})$ *(of the nonempty and closed set* $\mathbb{B}$*), its proximal map* $\text{Prox}_\lambda^\delta(\mathbf{X})$ *reduces to the projection operator:*

$$P_{\mathbb{B}}(\mathbf{X}) = \arg\min\{||\mathbf{B} - \mathbf{X}||^2 : \mathbf{B} \in \mathbb{B}\}. \tag{8}$$

It is clear that, problem (7) has the analytical solution

$$\mathbf{B}^{(j+1)} = \text{sgn}(\mathbf{B}^{(j)} - \frac{1}{\lambda}\nabla\mathcal{L}(\mathbf{B}^{(j)})). \tag{9}$$

We term this optimization method as **Discrete Proximal Linearized Minimization (DPLM)** due to the involvement of discrete variables compared to the linearized proximal method [1].

In the following, we show that for problem (1) the algorithm (9) converges to a critical point. First we introduce the convergence theorem from [1] for the nonconvex gradient projection method.

Theorem 2 [1]: Let $f : \mathbb{R}^n \to \mathbb{R}$ *be a differentiable function whose gradient is L-Lipschitz continuous, and C a nonempty closed subset of* $\mathbb{R}^n$*. Being given* $\epsilon \in (0, \frac{1}{2L})$ *and a sequence of stepsizes* γ_k *such that* $\epsilon < \gamma_k < \frac{1}{L} - \epsilon$*, we consider a sequence* (x^k) *that complies with*

$$x^{k+1} \in P_C(x^k - \gamma_k\nabla f(x^k)), \quad with\ x^0 \in C.$$

If the function $f + \delta_C$ *is a Kurdyka-Lojasiewicz (KL) function and if* (x_k) *is bounded, then the sequence* (x_k) *converges to a point* $x^\star$ *in C.*

Corollary 3: Assume the loss function $\mathcal{L}$ *is a* C^1 *(continuously differentiable) semi-algebraic function whose gradient is L-Lipschitz continuous. By choosing a proper sequence of parameters of* λ*, the sequence* $\mathbf{B}^{(j)}$ *generated by the proposed DPLM algorithm with* (9) *converges to a critical point* $\mathbf{B}^\star$*.*

Proof: This assumption ensures that the objective of (3) $\mathcal{L}(\cdot) + \delta_{\mathbb{B}}(\mathbf{B})$ is a KL function [1]. It is obvious that the sequence $\mathbf{B}^{(j)}$ generated by (9) is bounded in $\mathbb{B}$. Based on Theorem 2, by choosing the parameter λ greater than the Lipschitz constant L, the DPLM algorithm converges to some critical point. ■

Remark 4: The requirement of the presented DPLM method is only mild. The KL assumption of the objective function is very general that $\mathcal{L}$ *being the smooth polynomial is a typical instance. The empirical convergence of DPLM is referred to Section IV-C.*

Till now, we have presented the key optimization method for learning binary codes with a general loss function. The optimization procedure is outlined in Algorithm 1. Despite its simplicity, the proposed method can obtain very high-quality codes for the retrieval and classification tasks, as shown in our experiments.

In addition, due to the analytical solution at each iteration, this method enjoys very fast optimization. We note that the analytical solution does not depend on a specific loss function. In Section III, we will discuss the application of DPLM to different hashing losses, such as supervised ℓ_2 hashing and

Algorithm 1 Discrete Proximal Linearized Minimization

Input: Training data $\mathbf{X}$; code length r; maximum iteration number t; parameters λ.

Output: Binary codes $\mathbf{B} \in \{-1, 1\}^{r\times n}$; hash function $h(\mathbf{x})$.

1) Initialize $\mathbf{B}$ by the sign of random Gaussian matrix;
2) Loop until converge or reach maximum iterations:
   - Calculate the gradient $\nabla\mathcal{L}$;
   - Update $\mathbf{B}$ by $\mathrm{sgn}(\mathbf{B} - \frac{1}{\lambda}\nabla\mathcal{L}(\mathbf{B}))$;
3) Compute hash function $h(\mathbf{x}) = \mathrm{sgn}(\mathbf{P}^\top\mathbf{x})$ with obtained $\mathbf{B}$ by (10).

unsupervised graph hashing. We will also show the well-known bits uncorrelation balance constraints can be easily incorporated in the binary code optimization with the proposed method.

C. Hash Function Learning

The above method describes the learning procedure for generating binary codes $\mathbf{B}$ for training data $\mathbf{X}$. For a novel query $\mathbf{x} \in \mathbb{R}^d$, we need a hash function to efficiently encode $\mathbf{x}$ into binary code. We here adopt the simple linear hash function $h(\mathbf{x}) = \mathrm{sgn}(\mathbf{P}^\top\mathbf{x})$, which is learned by solving a linear regression system with the available training data and codes. That is

$$\min_{\mathbf{P}} ||\mathbf{B} - \mathbf{P}^\top\mathbf{X}||^2, \tag{10}$$

which is clearly solved by $\mathbf{P} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{B}^\top$. This hash function learning scheme has been widely used, such as in [28] and [49].

III. Case Studies of the Hashing Problems

In this section, we investigate different hashing problems with the proposed DPLM method, where both the supervised and unsupervised losses are studied.

A. Supervised Hashing

We adopt the ℓ_2 loss in the supervised setting, where the learned binary codes are assumed to be optimal for linear classification. The learning objective writes,

$$\begin{aligned}\mathcal{L}(\mathbf{B}, \mathbf{W}) &= \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{W}^\top\mathbf{b}_i||^2 + \delta||\mathbf{W}||^2 \\ &= \frac{1}{2}||\mathbf{Y} - \mathbf{W}^\top\mathbf{B}||^2 + \delta||\mathbf{W}||^2. \end{aligned}\tag{11}$$

Here $\mathbf{Y} \in \mathbb{R}^{k\times n}$ stores the labels of of training data $\mathbf{X} \in \mathbb{R}^{d\times n}$, with its (i, j)-th entry $\mathbf{Y}_{ij} = 1$ if the j-th sample $\mathbf{x}_j$ belongs to the i-th of the total k classes and 0 otherwise. Matrix $\mathbf{W}$ is the classification matrix which is jointly learned with the binary codes. δ is the regularization parameter. The above simple objective has been shown to achieve very promising results recently [28].

With the ℓ_2 loss, the binary codes can be easily computed by the DPLM optimization method as shown in Section II-B. Given $\mathbf{W}$, the key step is updating $\mathbf{B}$ by (9) with the following gradient

$$\nabla\mathcal{L}(\mathbf{B}) = \mathbf{W}\mathbf{W}^\top\mathbf{B} - \mathbf{W}\mathbf{Y}. \tag{12}$$

With $\mathbf{B}$ obtained, the classification matrix $\mathbf{W}$ is efficiently computed by $\mathbf{W} = (\mathbf{B}\mathbf{B}^\top)^{-1}\mathbf{B}\mathbf{Y}^\top$. The whole optimization alternatively runs over variable $\mathbf{B}$ and $\mathbf{W}$. In practice, we simply initialize $\mathbf{B}$ by the sign of random Gaussian matrix, and $\mathbf{W}$ and $\mathbf{B}$ are then updated accordingly. The optimization typically converges within 5 iterations.

B. Unsupervised Graph Hashing

For the unsupervised setting, we investigate the well-known graph hashing problem [39], which has been extensively studied in the literature [23], [29], [39]. The unsupervised graph hashing optimizes the following objective

$$\mathcal{L}(\mathbf{B}) = \frac{1}{2}\sum_{i,j=1}^{n} ||\mathbf{b}_i - \mathbf{b}_j||^2\mathbf{A}_{ij} \tag{13}$$

$$= \frac{1}{2}\mathrm{tr}(\mathbf{B}\mathbf{L}\mathbf{B}^\top), \tag{14}$$

where $\mathbf{A}$ is the affinity matrix computed with $\mathbf{A}_{ij} = \exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/\sigma^2)$ and σ is the bandwidth parameter. $\mathbf{L}$ is the associated Laplacian matrix $\mathbf{L} = \mathrm{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A}$.

To tackle this challenging problem, Spectral Hashing [39] additionally assumes that data are sampled from a uniform distribution, which leads to a simple analytical eigenfunction solution of 1-D Laplacians. However, the strong assumption can hardly be true in practice. AGH [23] employs the anchor graph to facilitate constructing affinity matrix and learning hash functions analytically. IMH [29] learns Laplacian eigenmaps on a small data subset and the hash codes are thus inferred with a linear combination of the base points. All these methods apply spectral relaxation to simplify the optimization by discarding the binary constraints.

Different from these methods, we optimize the graph hashing problem by DPLM directly over the binary variables. With the gradient of (13) as

$$\nabla\mathcal{L}(\mathbf{B}) = \mathbf{B}\mathbf{L}, \tag{15}$$

the optimization is performed by updating variable $\mathbf{B}$ in each iteration with

$$\mathbf{B}^{(j+1)} = \mathrm{sgn}(\mathbf{B}^{(j)} - \frac{1}{\lambda}\mathbf{B}^{(j)}\mathbf{L}).$$

Note that the computation of affinity matrix $\mathbf{A}$ dominates the optimization and is $O(dn^2)$ in time complexity. In practice, we adopt the anchor graph to compute $\mathbf{A} = \mathbf{Z}\mathbf{Z}^\top$ with $\mathbf{Z} \in \mathbb{R}^{n\times m}$ as in [23], which is $O(dnm)$ with m anchors. The gradient (15) is thus computed by $\mathbf{B}\mathrm{diag}(\mathbf{A}\mathbf{1}) - \mathbf{B}\mathbf{A}$ which is $O(rmn)$.

C. Bits Uncorrelation and Balance

The bits uncorrelation and balance constraints have been widely used in previous hashing methods. With these two constraints, problem (1) is rewritten as

$$\begin{aligned}\min_{\mathbf{B}}\ & \mathcal{L}(\mathbf{B}) \\ \text{s.t. }& \mathbf{B}\mathbf{B}^\top = n\mathbf{I}_r, \\ & \mathbf{B}\mathbf{1} = \mathbf{0}, \\ & \mathbf{B} \in \{-1, 1\}^{r\times n}. \end{aligned}\tag{16}$$

The first two constraints force the binary codes to be uncorrelated and balanced, respectively. They are two key features of

TABLE I

EVALUATION OF OUR METHOD WITH/WITHOUT THE CONSTRAINT OF BALANCE OR UNCORRELATION ON THE IMAGE RETRIEVAL TASK. −: THIS OPERATION IS NOT APPLIED; ✓: APPLIED. BOTH THE SUPERVISED AND UNSUPERVISED LOSSES ARE TESTED. THE RESULTS ARE REPORTED IN mAP AND PRECISION OF TOP 500 RETRIEVED SAMPLES WITH 64 AND 128 BITS. THE DATABASE OF NUS-WIDE AND CIFAR-10 ARE USED, WHERE THE DESCRIPTIONS CAN BE FOUND IN SECTION V-A AND [28], RESPECTIVELY

NUS-WIDE									
Balance	Uncorrelation	Supervised				Unsupervised			
		mAP		Precision@500		mAP		Precision@500	
		64 bits	128 bits	64 bits	128 bits	64 bits	128 bits	64 bits	128 bits
−	−	0.6955	0.7167	0.6079	0.6041	0.3719	0.3750	0.4407	0.4466
✓	−	0.7527	0.7600	0.7524	0.7543	0.3903	0.3849	0.4414	0.4506
−	✓	0.734	0.7580	0.7523	**0.7598**	0.4022	0.4087	0.4596	0.4752
✓	✓	**0.7603**	**0.7610**	**0.7545**	0.7554	**0.4055**	**0.4119**	**0.4803**	**0.4920**
CIFAR-10									
Balance	Uncorrelation	Supervised				Unsupervised			
		mAP		Precision@500		mAP		Precision@500	
		64 bits	128 bits	64 bits	128 bits	64 bits	128 bits	64 bits	128 bits
−	−	0.6809	0.7029	0.6129	0.6286	0.1786	0.1920	0.2377	0.2670
✓	−	0.6842	0.7012	0.6133	0.6298	0.1872	0.2112	0.2596	0.3285
−	✓	0.6821	0.7039	0.6140	0.6302	0.1820	0.1927	0.2441	0.2738
✓	✓	**0.6852**	**0.7046**	**0.6155**	**0.6365**	**0.1960**	**0.2117**	**0.2724**	**0.3303**

compact binary code learning [39]. A large family of existing hashing algorithms can be seen as the instances of this general model, such as unsupervised graph hashing [21], [23], [39], and supervised hashing [7]. However, these two constraints often make the hashing problem computationally intractable, especially with the binary constraints. The recent proposed discrete optimization solver [28] discards these constraints for algorithm feasibility.

We rewrite (16) as follows

$$\min_{\mathbf{B}} \ \mathcal{L}(\mathbf{B}) + \frac{\mu}{4}||\mathbf{B}\mathbf{B}^\top||^2 + \frac{\rho}{2}||\mathbf{B}\mathbf{1}||^2 \quad \text{s.t. } \mathbf{B} \in \{-1, 1\}^{r\times n}. \tag{17}$$

Note that $||\mathbf{B}\mathbf{B}^\top - n\mathbf{I}_r||^2 = ||\mathbf{B}\mathbf{B}^\top||^2 + const$. With sufficiently large parameters $\mu > 0$ and $\rho > 0$, problems (16) and (17) will be equivalent. Denoting the objective of (17) as $g(\mathbf{B})$, its gradient is given by

$$\nabla g(\mathbf{B}) = \nabla \mathcal{L}(\mathbf{B}) + \mu\mathbf{B}\mathbf{B}^\top\mathbf{B} + \rho\mathbf{B}\mathbf{1}\mathbf{1}^\top. \tag{18}$$

With this, the binary optimization is conducted by updating variable $\mathbf{B}$ in each iteration with

$$\mathbf{B}^{(j+1)} = \text{sgn}(\mathbf{B}^{(j)} - \frac{1}{\lambda}\nabla g(\mathbf{B}^{(j)})). \tag{19}$$

In Section IV, we will explore the impact of these two binary codes properties for both the supervised and unsupervised hashing problems studied in this work.

D. Complexity Study

In this part, we discuss the computational complexity of our algorithm. For the supervised method in Section III-A, the main step is updating $\mathbf{B}$ by computing its gradient $\mathbf{W}\mathbf{W}^\top\mathbf{B} - \mathbf{W}\mathbf{Y}$, for which the time complexity is $O(r^2k + r^2n + rkn)$, thus making the total time complexity of updating $\mathbf{B}$ be $O(t(r^2k + r^2n + rnk))$, where t is the maximum iteration number during the $\mathbf{B}$ updating step. The complexity of updating $\mathbf{W}$ is $O(r^3 + rkn + r^2k)$. Therefore, the total time complexity of the supervised algorithm is $O(T(tr^2n + rkn)$ with T iterations (updating $\mathbf{W}$ and $\mathbf{B}$).

The unsupervised algorithm in Section III-B comprises two components: the anchor graph construction and binary code learning. As mentioned in III-B, the first part costs $O(dnm)$ time and the second part costs $O(trmn)$ with t iterations.

The time complexity of computing the hash functions with equation (10) is $O(d^3 + 2d^2n)$. As for these algorithms with bit uncorrelation and balance constraints, the additional computation is due to the computing of $\mu\mathbf{B}\mathbf{B}^\top\mathbf{B} + \rho\mathbf{B}\mathbf{1}\mathbf{1}^\top$ in (18) in each iteration of updating $\mathbf{B}$, which is $O(r^2n + 2rn)$ in time.

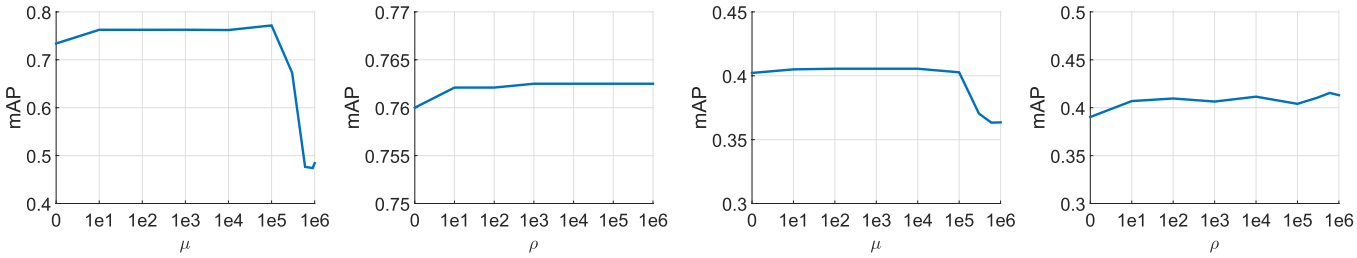
To summarize, the training time complexities for the proposed supervised and unsupervised algorithms are both linear as the data size n. For a novel query, the predicting time with hash function $h(\mathbf{x})$ is $O(dr)$ which is independent of n.

IV. ALGORITHM ANALYSIS

In this section, we evaluate the proposed method from the following aspects: the impact of bits uncorrelation and balance, the optimization performance of DPLM compared to other solvers.

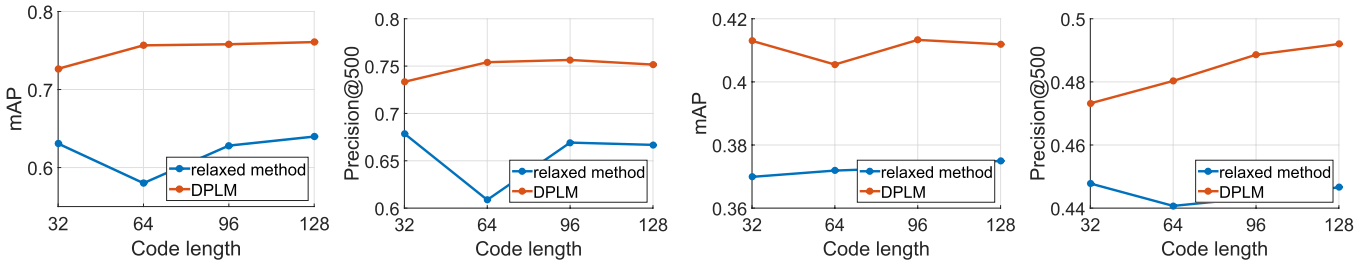
A. The Impact of Bits Uncorrelation and Balance

We first evaluate the impact of the two well-known constraints on binary codes: bits uncorrelation and balance. Both the supervised loss and unsupervised loss are evaluated. The performance of our method with or without each of the constraints is shown in Table I. The database of NUS-WIDE



(a) Supervised hashing (b) Unsupervised hashing

Fig. 1. The function of MAP w.r.t. parameters μ and ρ with (a) the ℓ_2 supervised loss (11) and (b) unsupervised graph hashing loss (13), respectively. The evaluation is performed on NUS-WIDE. 64 bits are used.



(a) Supervised hashing (b) Unsupervised hashing

Fig. 2. Comparison of DPLM and relaxed optimization with (a) the ℓ_2 supervised loss (11) and (b) unsupervised graph hashing loss (13), respectively. The evaluation is performed on NUS-WIDE in terms of MAP and Precision@500.

and CIFAR-10 are used for evaluation. As we can see, the constraints play important roles in binary code learning. For the two hashing losses, better results are obtained by imposing both these constraints than discarding them or keeping only one in most cases. The ability to incorporate these two constraints into binary code optimization is one of the advantages of our method over other discrete methods such as DCC [28]. We also show in details the impacts of parameters μ and ρ with both the ℓ_2 supervised loss (11) and unsupervised graph hashing loss (13). The MAP results with varying μ and ρ on NUS-WIDE are are shown in Figure 1.

B. DPLM vs. the Relaxed Method

One may be interested in how the proposed DPLM method performs compared to the relaxed method: first relaxing the original discrete problem to a continuous one and then rounding the resultant solution. In this part, we compare DPLM and the relaxed method for both the supervised ℓ_2 loss and unsupervised graph hashing loss. The mAP and Precision500 results are shown in Figure 2 with code length varying from 32 to 128 bits.

From Figure 2, large performance gains are clearly observed with DPLM over the relaxed approach for both of the two hashing losses. The superior results demonstrate the effectiveness of our optimization method and the importance of discrete optimization for binary code learning or hashing problems. That is, it would be preferred to directly pursue the discrete codes in the binary space without continuous relaxations, provided that scalable and tractable solvers are accessible. In the next part, we will evaluate the convergence speed of DPLM.

C. DPLM vs. DCC

One main contribution of this work is the fast binary optimization method for hashing. In this part, we compare the optimization speed between the proposed Discrete Proximal Linearized Minimization (DPLM) and the recent discrete cyclic coordinate descent (DCC) method [28]. To be fair, we omit the uncorrelation and balance constraints which DCC cannot handle. That is, both DPLM and DCC both minimize the following objective function and are compared according to the obtained optimal solutions:

$$\min_{\mathbf{B},\mathbf{W}} \frac{1}{2}||\mathbf{Y} - \mathbf{W}^\top \mathbf{B}||^2 + \delta||\mathbf{W}||^2$$
$$\text{s.t. } \mathbf{B} \in \{-1, 1\}^{r\times n}.$$

The objective value as a function of optimization time is shown in Figure 3. We can clearly see that both these two methods converge to similar objective value. However, DPLM obtains much faster convergence than DCC. DPLM only costs about 1 second to achieve the convergence. This is mainly because DPLM updates all bits at the same time in each iteration. In contrast, DCC computes the codes in a bit-by-bit manner, where each bit is computed based on the previously updated bits. In the next section, we will extensively compare the two algorithms on both the retrieval and classification tasks.

The advantages of DPLM over DCC is summarized as the following points: 1) DPLM is much more efficient than DCC for the binary optimization problem, as shown in Figure 3; 2) DPLM is developed to solve the general binary code learning problem while DCC can only solve the BQP problem and cannot handle the bit uncorrelation constraints; 3) By adopting the bit balance and uncorrelation constraints,

TABLE II

RESULTS IN TERM OF mAP AND MEAN PRECISION OF THE TOP 500 RETRIEVED NEIGHBORS (PRECISION@500) OF THE COMPARED SUPERVISED METHODS ON THE **NUS-WIDE** DATABASE WITH 64 AND 128 BITS, RESPECTIVELY. THE TRAINING AND TESTING TIME ARE REPORTED IN SECONDS

Method	mAP		Precision@500		Training time (s)		Testing time (s)	
	64 bits	128 bits	64 bits	128 bits	64 bits	128 bits	64 bits	128 bits
Ours	**0.7603**	**0.7610**	**0.7545**	**0.7547**	**8.32**	**13.97**	1.31e-6	**2.47e-6**
SDH	0.6955	0.7167	0.6079	0.6041	34.72	119.8	1.65e-6	2.55e-6
CCA-ITQ	0.6232	0.6239	0.5919	0.5962	8.93	19.71	**1.08e-6**	3.03e-6
KSH	0.6091	0.6129	0.5638	0.5659	2092.3	4384.3	5.84e-6	1.33e-5
FastHash	0.5346	0.5507	0.6013	0.6197	3486.16	7091.99	1.31e-2	2.86e-2
MLH	0.4726	0.4689	0.5540	0.5583	8413.4	13791	2.23e-5	3.79e-5

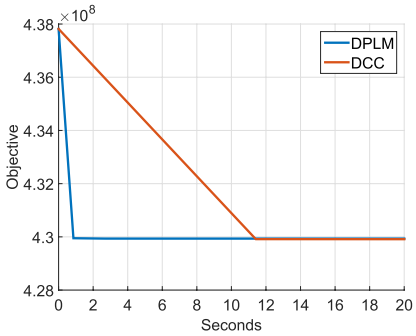


Fig. 3. Objective value as a function of binary code training time for DPLM and DCC on NUS-WIDE. We use 64 bits in this experiment.

DPLM can achieve better performance than DCC, as shown in Table II.

V. EXPERIMENTS

In this section, extensive experiments are conducted to evaluate the proposed hashing methods in both computational efficiency and retrieval or classification performance. We test our method on three large-scale public databases, i.e., SUN397 [40], NUS-WIDE [3] and ImageNet [5]. The detailed descriptions of these databases are introduced in the corresponding subsection. Several state-of-art hashing methods are taken into comparison, including the supervised MLH [25], CCA-ITQ [9], KSH [22], FastHash [17], SDH [28] and the unsupervised SH [39], AGH [23], PCA-ITQ [9], and IMH [29]. For these methods, we employ the implementations and suggested parameters provided by the authors. For our method, since the bits uncorrelation and balance constraints help produce better codes, we impose these two constraints in our algorithm. We empirically set $\lambda = 0.1$, $\mu = 1e3$ and $\rho = 1e2$. Since CCA-ITQ, SDH and our methods can efficiently handle large data during training, we use all the available training data for training. For KSH, MLH and FastHash, we learn these models with 50k training samples due to the large computational costs of these methods.

For the retrieval experiments, we report the compared results in terms of mean average precision (mAP), mean precision of the top 500 retrieved neighbors (Precision@500) and the precision and recall curves. Note that we treat a query a false case if no point is returned when calculating precisions. Ground truths are defined by the category information from the datasets. For computational efficiency, we compare the training and testing time of the evaluated methods. For the compared supervised hashing approaches, we also test the performance of these methods on the classification task, where the metric of classification accuracy is used. If not otherwise specified, the experiments are conducted with MATLAB implementations on a standard PC with an Intel 6-core 3.50GHz CPU and 64G RAM.

A. NUS-WIDE: Retrieval With Multi-Labeled Data

The NUS-WIDE database contains about 270,000 images collected from Flickr. The images in NUS-WIDE are associated with 81 concepts, with each image containing multiple semantic labels. We define the true neighbors of a query as the images sharing at least one labels with the query image. The provided 500-dimensional Bag-of-Words features are used. we collect the 21 most frequent label for test. For each label, 100 images are uniformly sampled for the query set and the remaining images are for the training set. The results in terms of retrieval performance (mAP and Precision@500) and training/testing time efficiency are reported in Table II.

It is clear from Table II that our approach achieves the best results in terms of both mAP and precision among all the compared supervised methods. In particular, with 64 bits, our method outperforms the best of all other methods (obtained by SDH) by more than 6% and 15% in terms of mAP and precision, respectively. The precision-recall and precision curves of these compared methods with 32 to 128 bits are shown in Figure 4. Our method consistently outperforms all other methods by large margins in all situations.

We also evaluate these methods in terms of training and testing efficiencies. We can clearly see from Table II that, our method costs less training time than all other compared methods. Specifically, DPLM only consumes only about 8.3 seconds to train the hashing model on the NUS-WIDE database. CCA-ITQ also has a high computational efficiency, which is much faster than other methods. In terms of testing time (encoding a query image into binary code), our method together with SDH and CCA-ITQ run very fast on the same scale while FastHash suffers from a slow encoding speed.

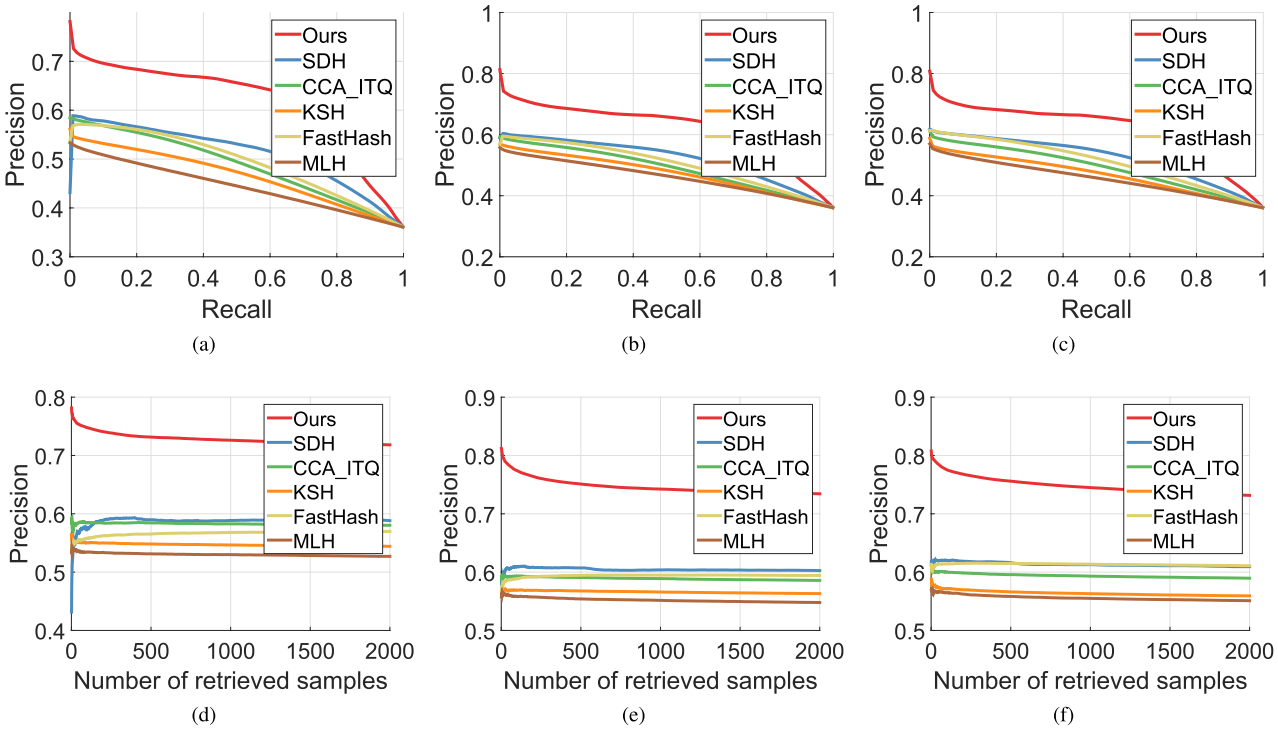


Fig. 4. (Top) Precision-Recall and (Bottom) Precision curves with top 2000 retrieved images of the compared methods on **NUS-WIDE**. (a) 32 bits. (b) 64 bits. (c) 128 bits. (d) 32 bits. (e) 64 bits. (f) 128 bits.

TABLE III

RESULTS IN TERM OF mAP AND MEAN PRECISION OF THE TOP 500 RETRIEVED NEIGHBORS (PRECISION@500) OF THE COMPARED METHODS ON THE **IMAGENET** DATABASE WITH 64 AND 128 BITS, RESPECTIVELY. THE TRAINING AND TESTING TIME ARE REPORTED IN SECONDS. THE EXPERIMENTS ARE CONDUCTED ON A WORKSTATION WITH AN INTEL 6-CORE 2.10GHz CPU AND 188G RAM

Method	mAP		Precision@500		Training time (s)		Testing time (s)	
	64 bits	128 bits	64 bits	128 bits	64 bits	128 bits	64 bits	128 bits
Ours	**0.3235**	**0.4310**	**0.3039**	**0.3996**	**258.82**	**336.84**	**2.80e-6**	**6.93e-6**
SDH	0.3225	0.4261	0.3016	0.3987	1568.66	4015.20	3.01e-6	8.24e-6
CCA-ITQ	0.1086	0.1694	0.1651	0.2461	372.52	468.00	3.42e-6	7.64e-6
KSH	0.0897	0.1351	0.1702	0.2381	6953.48	13897.53	6.18e-5	9.67-05
FastHash	0.1062	0.1827	0.1944	0.2598	3486.16	7091.99	1.59e-2	1.31e-2
MLH	0.0739	0.1111	0.1493	0.2069	20864.12	43494.35	1.18e-5	2.74e-5

B. ImageNet: Retrieval With Large-Scale High Dimensional Features

As a subset of ImageNet [5], the large dataset ILSVRC 2012 contains over 1.2 million images of totally 1,000 categories. We use the provided training set as the retrieval database and 50,000 images from the validation set as the query set. We extract the feature for each image by the convolutional neural networks (CNN) model as a 4096D vector. The results are reported in the Table III.

As in the last section, similar results are observed from Table III that our method obtains the best results. On this large dataset our method is slightly better than SDH, while both of them outperforms other methods by even larger gaps than on the relatively smaller NUS-WIDE database. These results demonstrate the importance of discrete optimization for binary code learning.

In addition, our method demonstrates clearer advantages on ImageNet in training efficiency. For example, our method trains on the whole dataset with only about 5.5 minutes with 128 bits, while SDH costs more than 1 hour and KSH, FastHash and MLH runs even slower. From these experiments, it is clear that *our discrete hashing method can generate more effective binary codes with much less learning time*.

C. SUN397: Scene Classification With Binary Codes

In this part, we test the compared hashing methods on the classification task by feeding the classifier with the generated binary feature with these methods. The LIBLINEAR implementation of linear SVM is used here. The proposed approach is compared with several other supervised hashing methods including SDH, CCA-ITQ, KSH, FastHash and MLH. SUN397 is a widely-used scene classification

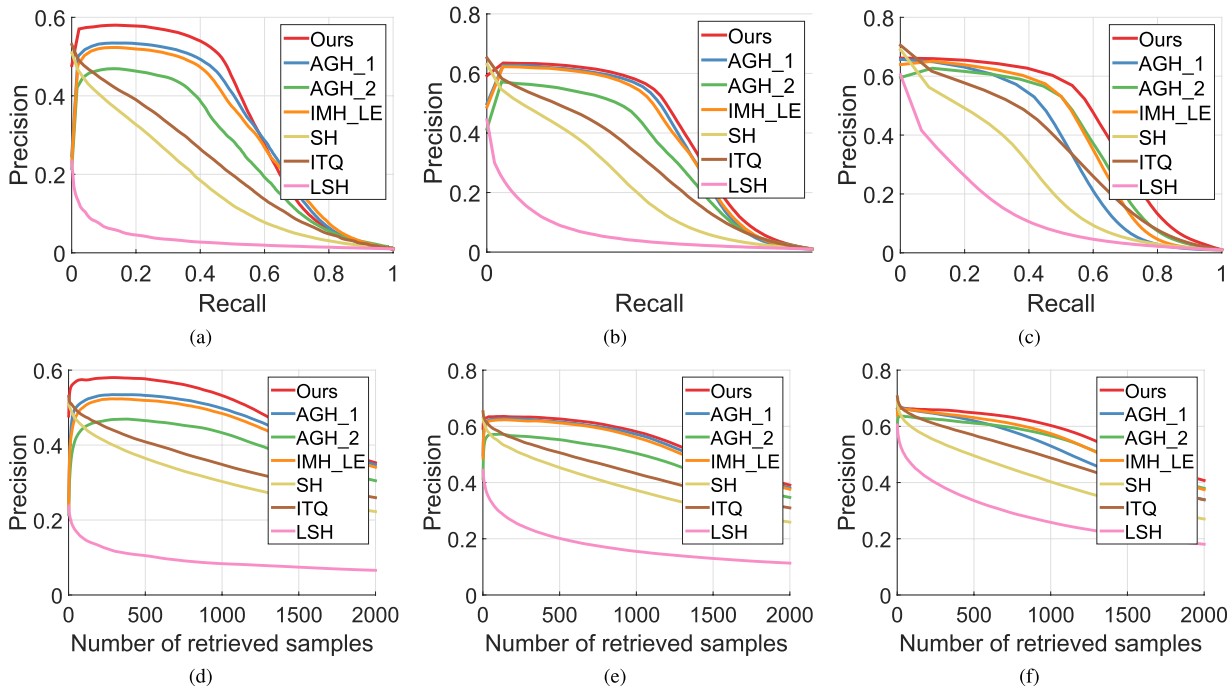


Fig. 5. (Top) Precision-Recall and (Bottom) Precision curves with top 2000 retrieved images of the compared methods on **ImageNet**. (a) 32 bits. (b) 64 bits. (c) 128 bits. (d) 32 bits. (e) 64 bits. (f) 128 bits.

TABLE IV

CLASSIFICATION ACCURACY (%) ON **SUN397** WITH THE PRODUCED BINARY CODES BY DIFFERENT HASHING METHODS. THE CODE LENGTH VARIES FROM 32 TO 128 BITS

Method	Accuracy (%)			
	32 bits	64 bits	96 bits	128 bits
Ours	**66.83**	**76.72**	**77.67**	**78.72**
SDH	63.00	75.28	77.44	78.22
CCA-ITQ	60.06	70.50	73.50	73.67
KSH	63.06	68.83	70.44	73.17
FastHash	64.95	71.04	74.17	75.38
MLH	56.56	65.67	68.78	71.44

TABLE V

CLASSIFICATION ACCURACY (%) ON **IMAGENET** WITH THE PRODUCED BINARY CODES BY DIFFERENT HASHING METHODS. THE CODE LENGTH VARIES FROM 32 TO 128 BITS. THE EXPERIMENTS ARE CONDUCTED ON A WORKSTATION WITH AN INTEL 6-CORE 2.10GHZ CPU AND 188G RAM

Method	Accuracy (%)			
	32 bits	64 bits	96 bits	128 bits
Ours	**18.28**	**30.50**	**36.59**	**39.89**
SDH	18.19	30.12	36.03	39.85
CCA-ITQ	10.66	18.31	23.84	27.42
KSH	18.07	27.58	31.62	32.69
FastHash	17.69	28.04	34.36	36.72
MLH	18.00	27.83	33.02	35.70

benchmark, which contains about 108,000 images from 397 scene categories, where each image is represented by a 1,600-dimensional feature vector extracted by PCA from 12,288-dimensional Deep Convolutional Activation Features [10]. In this experiment, 100 images are sampled uniformly randomly from each of the 18 largest scene categories to form a test set of 1,800 images and the rest for training set. The results are reported in Table IV.

As can be clearly seen, the proposed approach obtains the highest classification accuracies on this dataset. Clear advantage of our method is shown over SDH especially with short code length. With long code lengths (128 bits), our method achieves very close results with SDH, while outperforms other methods by more than 5% accuracies.

D. ImageNet: Image Classification With Binary Codes

In this subsection, we test the classification performance of the learned binary codes on the ImageNet benchmark [5]. The same training/test setting is used as in Section V-B. The classification accuracies on this dataset are reported in Table V. Our method performs slightly better than SDH on this dataset (with much lower learning cost however), while much better than all other methods. The results in Table IV and Table V clearly show that the binary codes generated by our methods work very well on the classification problem as well as the retrieval task.

E. Comparison With Unsupervised Methods

In this part, we evaluate our method in the unsupervised setting by performing DPLM with the unsupervised graph hashing loss. Other representative methods of graph hashing including SH, AGH (with one or two layers), IMH with Laplacian Eigenmaps (denoted as IMH_LE) and the well known LSH and ITQ are taken into comparison. We denote AGH with one and two layers by AGH_1 and AGH_2, respectively.

TABLE VI
RESULTS IN TERM OF mAP AND MEAN PRECISION OF THE TOP 500 RETRIEVED NEIGHBORS (PRECISION@500) OF THE COMPARED UNSUPERVISED METHODS ON THE IMAGENET DATABASE WITH 32 TO 128 BITS

Method	mAP				Precision@500			
	32 bits	64 bits	96 bits	128 bits	32 bits	64 bits	96 bits	128 bits
Ours	**0.4985**	**0.5445**	**0.5720**	**0.5784**	**0.5854**	**0.6269**	**0.6450**	**0.6485**
IMH	0.4469	0.5172	0.5261	0.5163	0.5191	0.6121	0.6256	0.6311
AGH-1	0.4587	0.5243	0.5176	0.4537	0.5325	0.6159	0.6346	0.6197
AGH-2	0.3927	0.4645	0.5011	0.5222	0.4653	0.5525	0.5919	0.6114
SH	0.2418	0.3066	0.3243	0.3310	0.3647	0.4532	0.4813	0.4956
ITQ	0.3001	0.3839	0.4244	0.4412	0.4086	0.5063	0.5490	0.5683
LSH	0.0485	0.1010	0.1462	0.1922	0.1050	0.2013	0.2731	0.3359

For AGH, IMH and our method, we use k-means to generate 1,000 cluster centers for anchor or subset points.

The comparison is conducted on the ImageNet dataset, where we form the retrieval and training database by the 100 largest classes with total 128K images from the provided training set, and 50,000 images from the validation set as the query set. The retrieval results of these unsupervised methods are reported in Table VI with code lengths from 32 to 128 bits. Consistent with the supervised experiments, the proposed method outperforms all other methods in both mAP and precision. The advantage of our method is further illustrated by the detailed precision-recall curves and precision curves on top 2,000 retrieved images, as shown in Figure 5.

VI. CONCLUSIONS AND DISCUSSION

This paper investigated discrete optimization in the general binary code learning problem. To tackle the difficult optimization problem over binary variables, we proposed an effective discrete optimization algorithm, dubbed Discrete Proximal Linearized Minimization (DPLM). Profiting from the analytical solution at each iteration, DPLM led very fast optimization. Compared with existing discrete methods, the proposed method supported a large family of empirical loss functions and constraints, which was instantiated by the supervised ℓ_2 loss and unsupervised graph hashing loss. Several large benchmark datasets were used for evaluation and the results clearly demonstrated the superiority of both our supervised and unsupervised approaches over many other state-of-the-art methods, in terms of both retrieval precision and classification accuracy.

A. Deep Learning Based Hashing

Deep learning (DL) has become one of the most effective feature learning approach for vision applications. For image hashing, DL also show its promising performance for the image retrieval task ([6], [15], [51]). We note that, however, in the test phase DL based hashing methods need to forward an image through a deep neural network (usually with many layers of projections), which costs much more time than non-DL hashing algorithms including ours (with only one projection). Therefore, with the same input (*e.g*, raw intensity or GIST feature), the proposed approach can provide more efficient binary code encoding. Another shortcoming of current DL based hashing algorithms is that they usually resort to a continuous relaxation (*e.g*, by the sigmoid function). A reasonable improvement will be incorporating the proposed DPLM binary optimization technique into the deep hash function learning process. This will be a challenging yet valuable research direction deserving further studies.

B. Potential Applications

The proposed DPLM method is developed for general bianry optimization, threrefore another potential application of DPLM will be its deployment with different hashing sennarios. For instance, DLPM could be applied to boost the performance of current hashing algorithms with pairwise supervised information (*e.g*, [7], [22]), multi-model hashing (*e.g*, [30], [31]), where discrete optimization is supposed to produced higher quality hashes. In addition, ranking based loss has been shown very effective in hashing [51] and various vision applications [33], [35], which can also be applied in our discrete optimization based hashing framework.

In addition to hashing, DPLM is also potentialy applied to other binary optimization problems, such as inner product binarizing [27] and collaborative filtering [47]. High-quality binary codes can also potentially boost various visual recognition (*e.g*, [20], [34]) and multimodal learning tasks [41]–[43].

REFERENCES

[1] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods," *Math. Program.*, vol. 137, nos. 1–2, pp. 91–129, 2013.

[2] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, nos. 1–2, pp. 459–494, 2014.

[3] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world Web image database from National University of Singapore," in *Proc. ACM Conf. Image Video Retr.*, 2009, Art. no. 48.

[4] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on *p*-stable distributions," in *Proc. ACM Symp. Comput. Geometry*, 2004, pp. 253–262.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[6] 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.*, Jun. 2015, pp. 2475–2483.

[7] T. Ge, K. He, and J. Sun, "Graph cuts for supervised binary coding," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 250–264.

[8] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proc. Int. Conf. Very Large Databases*, 1999, pp. 518–529.

[9] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, Dec. 2013.

[10] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 392–407.

[11] W. Kong and W.-J. Li, "Isotropic hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1646–1654.

[12] B. Kulis and T. Darrell, "Learning to hash with binary reconstructive embeddings," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1042–1050.

[13] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing for scalable image search," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2130–2137.

[14] B. Kulis, P. Jain, and K. Grauman, "Fast similarity search for learned metrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2143–2157, Dec. 2009.

[15] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3270–3278.

[16] X. Li, G. Lin, C. Shen, A. van den Hengel, and A. Dick, "Learning hash functions using column generation," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 142–150.

[17] G. Lin, C. Shen, Q. Shi, A. van den Hengel, and D. Suter, "Fast supervised hashing with decision trees for high-dimensional data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1971–1978.

[18] L. Liu, M. Yu, and L. Shao, "Multiview alignment hashing for efficient image search," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 956–966, Mar. 2015.

[19] L. Liu, M. Yu, and L. Shao, "Projection bank: From high-dimensional data to medium-length binary codes," in *Proc. IEEE Int. Conf. Comput. Vis.*, Feb. 2015, pp. 2821–2829.

[20] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 447–461, Mar. 2016.

[21] W. Liu, C. Mu, S. Kumar, and S.-F. Chang, "Discrete graph hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3419–3427.

[22] 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.

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

[24] J. Lu, V. E. Liong, and J. Zhou, "Cost-sensitive local binary feature learning for facial age estimation," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5356–5368, Dec. 2015.

[25] M. Norouzi and D. M. Blei, "Minimal loss hashing for compact binary codes," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 353–360.

[26] M. Raginsky and S. Lazebnik, "Locality-sensitive binary codes from shift-invariant kernels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1509–1517.

[27] F. Shen, W. Liu, S. Zhang, Y. Yang, and H. Tao Shen, "Learning binary codes for maximum inner product search," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4148–4156.

[28] F. Shen, C. Shen, W. Liu, and H. T. Shen, "Supervised discrete hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 37–45.

[29] F. Shen, C. Shen, Q. Shi, A. V. D. Hengel, Z. Tang, and H. T. Shen, "Hashing on nonlinear manifolds," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1839–1851, Jun. 2015.

[30] J. Song, Y. Yang, Z. Huang, H. T. Shen, and J. Luo, "Effective multiple feature hashing for large-scale near-duplicate video retrieval," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1997–2008, Dec. 2013.

[31] 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 Conf. Manage. Data*, 2013, pp. 785–796.

[32] 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.

[33] D. Tao, J. Cheng, M. Song, and X. Lin, "Manifold ranking-based matrix factorization for saliency detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1122–1134, Jun. 2016.

[34] D. Tao, Y. Guo, M. Song, Y. Li, Z. Yu, and Y. Y. Tang, "Person re-identification by dual-regularized KISS metric learning," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2726–2738, Jun. 2016.

[35] D. Tao, L. Jin, Y. Yuan, and Y. Xue, "Ensemble manifold rank preserving for acceleration-based human activity recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1392–1404, Jun. 2016.

[36] J. Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for large-scale search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2393–2406, Dec. 2012.

[37] J. Wang, H. T. Shen, J. Song, and J. Ji. (2014). "Hashing for similarity search: A survey." [Online]. Available: https://arxiv.org/abs/1408.2927

[38] Y. Weiss, R. Fergus, and A. Torralba, "Multidimensional spectral hashing," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 340–353.

[39] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1753–1760.

[40] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3485–3492.

[41] C. Xu, T. Liu, D. Tao, and C. Xu, "Local rademacher complexity for multi-label learning," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1495–1507, Mar. 2016.

[42] C. Xu, D. Tao, and C. Xu, "Large-margin multi-view information bottleneck," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1559–1572, Aug. 2014.

[43] C. Xu, D. Tao, and C. Xu, "Multi-view intact space learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2531–2544, Dec. 2015.

[44] Y. Yang, F. Shen, H. T. Shen, H. Li, and X. Li, "Robust discrete spectral hashing for large-scale image semantic indexing," *IEEE Trans. Big Data*, vol. 1, no. 4, pp. 162–171, Apr. 2015.

[45] Y. Yang, Z.-J. Zha, Y. Gao, X. Zhu, and T.-S. Chua, "Exploiting Web images for semantic video indexing via robust sample-specific loss," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1677–1689, Oct. 2014.

[46] F. Yu, S. Kumar, Y. Gong, and S.-F. Chang, "Circulant binary embedding," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 946–954.

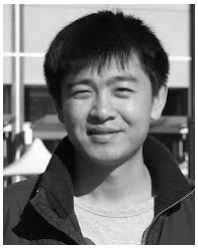
[47] H. Zhang, F. Shen, W. Liu, X. He, H. Luan, and T.-S. Chua, "Discrete collaborative filtering," in *Proc. ACM Conf. Inf. Retr.*, 2016, pp. 325–334.

[48] L. Zhang, H. Lu, D. Du, and L. Liu, "Sparse hashing tracking," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 840–849, Feb. 2016.

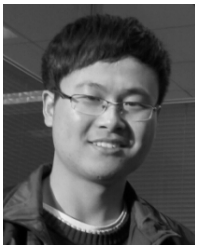
[49] P. Zhang, W. Zhang, W.-J. Li, and M. Guo, "Supervised hashing with latent factor models," in *Proc. ACM Conf. Inf. Retr.*, 2014, pp. 173–182.

[50] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.

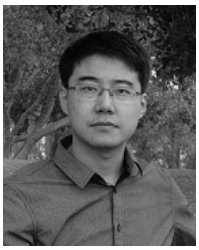
[51] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1556–1564.



Fumin Shen received the B.S. from Shandong University in 2007 and the Ph.D. degree from the Nanjing University of Science and Technology, China, in 2014.He is currently an Associate Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, China. His major research interests include computer vision and machine learning, including face recognition, image analysis, hashing methods, and robust statistics with its applications in computer vision.



Xiang Zhou is currently pursuing the master's degree with the University of Electronic Science and Technology of China. His major research interests include computer vision and machine learning.



Yang Yang received the bachelor's degree from Jilin University in 2006, the master's degree from Peking University in 2009, and the Ph.D. degree from The University of Queensland, Australia, in 2012, under the supervision of Prof. H. T. Shen and Prof. X. Zhou. He was a Research Fellow with the National University of Singapore from 2012 to 2014. He is currently with the University of Electronic Science and Technology of China.

Jingkuan Song received the B.S. degree in computer science from the University of Electronic Science and Technology of China and the Ph.D. degree in information technology from The University of Queensland, Australia, in 2014. He is currently a Post-Doctoral Research Scientist with Columbia University. He joined a University of Trento as a Research Fellow sponsored by Prof. Nicu Sebe from 2014–2016. His research interest includes large-scale multimedia retrieval, image/video segmentation, and image/video annotation using hashing, graph learning, and deep learning techniques.

Heng Tao Shen received the B.Sc. degree (Hons.) and the Ph.D. degree from the Department of Computer Science, National University of Singapore, in 2000 and 2004, respectively. He joined The University of Queensland as a Lecturer, a Senior Lecturer, and a Reader, where he became a Professor in 2011. He is currently a Professor of Computer Science and an ARC Future Fellow with the School of Information Technology and Electrical Engineering, The University of Queensland. He is also a Visiting Professor with Nagoya University and the National University of Singapore. His research interests mainly include multimedia/ mobile/Web search, and big data management on spatial, temporal, multimedia, and social media databases. He has extensively published and served on program committees in most prestigious international publication venues of interests. He received the Chris Wallace Award for outstanding Research Contribution in 2010 conferred by the Computing Research and Education Association, Australasia. He is also an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. He will serve as the PC Co-Chair of the ACM Multimedia in 2015.

Dacheng Tao (F'15) is currently a Professor of Computer Science and the Director of the Centre for Artificial Intelligence, and the Faculty of Engineering and Information Technology with the University of Technology Sydney. He mainly applies statistics and mathematics to Artificial Intelligence and Data Science. His research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and over 200 publications at prestigious journals and prominent conferences, such as the IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, ICML, CVPR, ICCV, ECCV, AISTATS, ICDM, and the ACM SIGKDD, with several best paper awards, such as the Best Theory/Algorithm Paper Runner Up Award in the IEEE ICDM07, the Best Student Paper Award in the IEEE ICDM13, and the 2014 ICDM 10-Year Highest Impact Paper Award. He received the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award, and the 2015 UTS Vice-Chancellors Medal for Exceptional Research. He is a fellow of the OSA, IAPR, and SPIE.