Unified Binary Generative Adversarial Network for Image Retrieval and Compression

Song, Jingkuan; He, Tao; Gao, Lianli; Xu, Xing; Hanjalic, Alan; Shen, Heng Tao

DOI
10.1007/s11263-020-01305-2

Publication date
2020

Document Version
Final published version

Published in
International Journal of Computer Vision

International Journal of Computer Vision (2020) 128:2243–2264
https://doi.org/10.1007/s11263-020-01305-2

Unified Binary Generative Adversarial Network for Image Retrieval and Compression

Jingkuan Song[1] · Tao He[2] · Lianli Gao[1] · Xing Xu[1] · Alan Hanjalic[3] · Heng Tao Shen[1]

Received: 21 April 2019 / Accepted: 5 February 2020 / Published online: 18 February 2020

Abstract
Binary codes have often been deployed to facilitate large-scale retrieval tasks, but not that often for image compression. In this paper, we propose a unified framework, BGAN+, that restricts the input noise variable of generative adversarial networks to be binary and conditioned on the features of each input image, and simultaneously learns two binary representations per image: one for image retrieval and the other serving as image compression. Compared to related methods that attempt to learn a single binary code serving both purposes, we demonstrate that choosing for two codes leads to more effective representations due to less concessions needed when balancing the requirements. The added value of using a unified framework compared to two separate frameworks lies in the synergy in data representation that is beneficial for both learning processes. When devising this framework, we also address another challenge in learning binary codes, namely that of learning supervision. While the most striking successes in image retrieval using binary codes have mostly involved discriminative models requiring labels, the proposed BGAN+ framework learns the binary codes in an unsupervised fashion, yet more effectively than the state-of-the-art supervised approaches. The proposed BGAN+ framework is evaluated on three benchmark datasets for image retrieval and two datasets on image compression. The experimental results show that BGAN+ outperforms the existing retrieval methods with significant margins and achieves promising performance for image compression, especially for low bit rates.

Keywords Binary codes · Image retrieval · Image compression · Generative adversarial network

1 Introduction

Image retrieval and compression have both been extensively studied, however mostly as two disjointed research topics due to their distinct key techniques and applications. Image retrieval makes use of the representation of visual content to identify relevant images, and image compression searches for ways to achieve efficient image representation to lower the cost of storage and transmission. In this paper we investigate the possibility to address both challenges using a unified framework. This possibility offers itself in the form of binary codes, or *hashes*.

In the context of image retrieval, binary codes have been deployed for approximate nearest-neighbor (ANN) search, which has proven itself as a tractable alternative for the nearest-neighbor search (NN) on large image collections. ANN search is more practical and can achieve orders of magnitude in speed-up compared to exact NN search (Jégou et al. 2011; Wang et al. 2018). Recently, learning-based hashing methods (Wang et al. 2018; Irie et al. 2014; Lin et al. 2014; Song et al. 2013; Shen et al. 2017; Duan et al. 2017) have

Communicated by Li Liu, Matti Pietikäinen, Jie Qin, Jie Chen, Wanli Ouyang, Luc Van Gool.

✉ Lianli Gao
lianli.gao@uestc.edu.cn

✉ Heng Tao Shen
shenhengtao@hotmail.com

Jingkuan Song
jingkuan.song@gmail.com

Tao He
tao.he@monash.edu

Xing Xu
xing.xu@uestc.edu.cn

Alan Hanjalic
a.hanjalic@tudelft.nl

[1] Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, Sichuan, China

[2] Monash University, Clayton, VIC 3800, Australia

[3] Delft University of Technology, Delft, The Netherlands

become the mainstream for scalable image retrieval due to their compact binary representation and efficient Hamming distance calculation. Such approaches embed data points to compact binary codes by hash functions, which can be generally expressed as:

$$\mathbf{b} = \mathbf{h}(\mathbf{x}) \in \{0, 1\}^L \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{D\times 1}$, $\mathbf{h}(.)$ are the hash functions, and $\mathbf{b}$ is a binary vector with code length L.

According to whether labels are leveraged when learning a hashing function, we roughly divide the hashing methods into two categories, supervised and unsupervised. An unsupervised method is aimed at preserving similarity properties of the original data points in the binary codes. Typical techniques preserving the similarity include pairwise similarity (Weiss et al. 2008; Liu et al. 2014), and multi-wise similarity (Norouzi and Fleet 2013; Wang et al. 2013b) or implicit preservation, which means that we do not need to calculate the explicit similarity between the inputs and the compact codes (Heo et al. 2015; Jin et al. 2013). Unsupervised hashing methods show many practical problems, such as how to construct the pairwise data points and how to measure and model different aspects of similarity in training data. Aiming at resolving the problems of unsupervised methods, supervised hashing methods (Lin et al. 2014; Ge et al. 2014; Strecha et al. 2011) have been well studied in recent years. While they usually significantly outperform unsupervised methods, the information that can be used for supervision is also typically scarce.

More recently, deep learning has been introduced in the development of hashing algorithms (Xia et al. 2014b; Lin et al. 2016; Do et al. 2016; Gu et al. 2016; Wang et al. 2017), leading to a new generation of *deep hashing* algorithms. Due to powerful feature representation, remarkable image retrieval performance has been reported using the hashes obtained in this way. However, a number of open issues have still remained open. The most successful deep hashing methods are usually supervised and require labels. The labels are, however, scarce and subjective. Unsupervised approaches, on the other hand, cannot take full advantages of the current deep learning models, and thus yield unsatisfactory performance (Lin et al. 2016). Another issue is a non-smooth sign function used to generate the binary codes, which, despite several ideas being proposed to tackle it (Li et al. 2016; He et al. 2019; Gao et al. 2019; Song et al. 2019), still makes the standard back-propagation infeasible. Meanwhile, GAN-based hash methods (Cao et al. 2017; Zieba et al. 2018; Wang et al. 2017) have been recently proposed. Compared to them, one advantage of our method is that BGAN+ is unsupervised, i.e., without requiring the label information. For example, HashGAN (Cao et al. 2017) is implemented by a conditional GAN, which needs additional supervised information. However, BGAN+ can be easily modified to the supervised version by incorporating the label information. In addition, most of GAN-based hash methods proposed to reconstruct the original image in order to transform original image information into the referring feature. Their reconstructed images are only used in the training stage and they are not required to be similar to their source images. To remedy this defect, our BGAN+ can make the best use of the generated images and reconstruct highly vivid images which are comparable to some classical image compression methods, such as JPEG and JPEG 2000.

The research field of image compression has already developed over many decades. The key challenge here is to find a pair of well-matched encoder and decoder. The encoder is used to transform the original large discrete data into low dimensional codes with minimal entropy (Shannon 2001), while the decoder acts as a translator which decodes the compressed codes into new data that should be identical as the original. In fact, the compression system is heavily associated with the probabilistic structure of the original data so the solution is similar to modeling a probabilistic source. In practice, since the compression codes always have finite entropy, we can not avoid the constructed errors. In this context, lossy compression problem has been studied for many years and generally, we must trade off two costs: the loss from the quantization (distortion) and the entropy of the discretized representation (rate). To be specific, low compression rate results in high entropy loss and high distortion directly leads to low-quality constructed data. However, joint learning of rate and distortion is difficult. Farvardin (1994) has demonstrated that it is intractable to optimize vector quantization without other constraints. Wintz (1972) utilized linear projection to transform the original data into a continuous-valued image representation, and then independently quantized its elements and finally encoded the discrete representation in a lossless fashion. The widespread compression technique, JPEG (Wallace 1991) deploys cosine transform on each pixel block, while another popular technique, JPEG 2000 (Rabbani and Joshi 2002), applies a multi-scale orthogonal wavelet decomposition of the original data. On the other hand, there is another direction of image compression, i.e., to encode images into binary codes and then decode them to original images, such as Toderici et al. (2015), Rippel and Bourdev (2017) and Li et al. (2018). Specifically, Toderici et al. (2015) proposed a RNN-based encoder and decoder to deal with variable compression rates after only once training, and Rippel and Bourdev (2017) proposed a GAN based auto-encoder framework to efficiently compress images into binary codes. However, in this work, our work is focusing on how to combine two tasks, image compression and retrieval into a unified framework instead of separating them. To the best of our knowledge, we are the first to combine the two tasks into a unified network.

Another widely-used technique is multi-tasks learning (Collobert and Weston 2008; Ruder 2017). It has been demonstrated that multi-tasks learning is an efficient strategy in lots of fields, such as speech recognition (Deng et al. 2013), computer vision (Girshick 2015) and natural language processing (Collobert and Weston 2008). Ruder (2017) explains why the multi-tasks learning is effective from five perspectives: data augmentation, attention focusing, eavesdrop, representation bias and regularization. One main feature of multi task learning is that the learned tasks should be related or have some overlap parts with each other, which means that they should share some parameters in the unified network, such as feature learning layers. Take Fast-rcnn (Girshick 2015) as an example. In fact, Faster-rcnn has implemented several tasks in a unified framework, such as region proposal selection, object detection and object recognition, all of which are relevant. Specifically, region proposal selection is to recognize whether the bounding box contains a object and the task of object recognition is to classify the objection in the box. Both of the tasks interact with each other and the former can be also regarded as the base task of the latter. In terms of the implementation of multi tasks, there are two ways: hard parameters sharing and soft parameters sharing, and more details can be found in Deng et al. (2013). In this paper, we focus on solving two tasks, image retrieval and compression in a multi-task learning fashion, due to the fact that the tasks are not independent. More concretely, image retrieval task is a procedure to deal with semantic or high-level feature compression while image compression is to compress the raw pixel image, i.e., low-level feature. As a matter of fact, learning the low-level feature is the basis of learning high-level representation. Thus, in this paper, we propose to use a binary generative adversarial network (BGAN+) to convert images to binary codes for both image retrieval and compression in a multi-task learning fashion and an unsupervised way.

While, ideally, one could try to find a binary code that is usable for both tasks, our preliminary study has shown that optimizing a hash-function learning from both perspectives requires the learning algorithms to make too many concessions towards one of the objectives, making either retrieval or compression less effective than the common state-of-the-art. However, we hypothesize that we can come far in unifying the two binary-code learning processes. In this way, we can produce two different codes that are individually optimized for their different purposes, but in a way that the two learning procedures optimally benefit from each other in terms of learning efficiency and effectiveness. In view of the analysis above, our contribution can be threefold:

- We propose a binary generative adversarial network (BGAN+) to convert images to binary codes for both image retrieval and compression in a multi-task learning fashion. To the best of our knowledge, this is the first attempt to combine the two tasks into a unified network.
- We take the challenge of learning these codes in an unsupervised way in order not to rely on typically scarce training data. Alternatively, BGAN+ can also be easily modified to supervised version to significantly improve the retrieval performance. We also propose several solutions to address the gradient vanishing problem caused by $sign$ function in the hash-learning process, which enables our method to be trained in an end-to-end strategy.
- We conduct extensive experiments to evaluate the performance of the binary codes generated by BGAN+ in terms of image retrieval and compression. A wide range of results show that our BGAN+ outperforms the existing retrieval methods with significant margins and achieves competitive performance for image compression, especially for low bit rates. Besides, substantial ablation studies also show the proposed each part in BGAN+ is effective and able to contribute to the referring task's performance.

The remainder of this paper is organized as follows. We first review the related work in Sect. 2. Then, we provide the details of our proposed model in Sect. 3, followed by the experimental results in Sect. 5. Section 6 concludes the paper.

2 Related Work

In this section, we briefly review the related work, and then specifically discuss the work on hashing for image retrieval, image generation, image compression and multi-tasks learning. Regarding image retrieval using binary codes, supervised methods generally use information to learn hashing codes in three different formats: point-wise, pair-wise and rank orders. Representative point-wise hashing methods include CCA-ITQ (Gong et al. 2013), supervised discrete hashing (SDH) (Shen et al. 2015), deep hashing (Liong et al. 2015), and BinGan (Zieba et al. 2018). Pair-wise hashing can best be illustrated by the methods, such as SPLH (Wang et al. 2010), which utilizes sequential projection learning strategy to generate efficient hashing codes, and KSH (Liu et al. 2012), which uses kernel function to learn hashing function and outperforms other supervised methods, the fast supervised hashing (Lin et al. 2014) and two-step hashing (TSH) (Lin et al. 2013). At the same time, many other methods based on deep learning have been developed, like the convolutional neural network hashing (Xia et al. 2014a), in which it is proposed to automatically learn convolutional image representation instead of the previous work using hand draft features as input. Furthermore, DPSH (Li et al. 2016) directly combines two independent tasks, learning image

representation and hashing function, into a deep end-to-end network. The representative rank-label methods include column generation hashing (Li et al. 2013), ranking-based supervised hashing (Wang et al. 2013a), discretely semantic rank orders (DSeRH) (Liu et al. 2017) and ranking preserving hashing (Zhao et al. 2015). In our work, we use pair-wise similarity as the hashing-learning strategy. Unlike previous work, we do not use the ground truth labels to construct pairwise labels. Instead, we adopt two ways, via hand-crafted feature and deep feature, to create the similarity matrix. In this sense, our proposed method can be treated as the unsupervised method.

Regarding the research on image generation, generative adversarial networks (GAN) (Goodfellow et al. 2014) has played a critical role recently. GAN usually consists of two networks, a generator and a discriminator network, which are involved in a min–max optimization game. There, the discriminator acting as an adversary to the generator is used to judge whether the generated image from the generator is real or fake, that is, whether it matches the criteria imposed by the input image or not. This is why the goal of the generator is to generate images that resemble the input image in the best possible way so it can 'fool' the discriminator. Theoretically, when the discriminator cannot distinguish the source of the image (original or from the generator), we can consider the overall GAN optimization as converged. Recently, a vast number of image generation methods based on GAN have been explored (Larsen et al. 2016; Ledig et al. 2017).

Lossy image compression has been widely used for data storage and transmission. JPEG (Wallace 1991) and JPEG 2000 (Rabbani and Joshi 2002) are two commonly used methods of lossy compression for digital images. The degree of compression can be adjusted, allowing a selectable trade-off between storage size and image quality. JPEG typically achieves 10:1 compression with little perceptible loss in image quality. After that, more sophisticated compression methods have been proposed, e.g., WebP (Google 2017), JPEG 420, Better Portable Graphics (BPG) (Bellard 2017). Recently, with the wide application of deep learning, there are numerous novel compression methods based on CNN or Recurrent Neural Network (RNN) (Toderici et al. 2016; Li et al. 2018; Ballé et al. 2016). In Toderici et al. (2016) proposed a deep RNN network, which can provide variable compression rates during deployment, and introduced a new hybrid of GRU and ResNet. In Li et al. (2018) explored a content-aware compression method based on the convolutional network, which can generate an importance map of the image content and optimize the compression quality. It can also retain as much detail as possible and in the low bit rate their method outperforms JPEG and JPEG 2000. Baig et al. (2017) proposed multi-stage progressive encoders, whose structure resembles a bottleneck, like VAE (Kingma and Welling 2013). Ballé et al. (2016) proposed an image compression framework, consisting of a nonlinear encoding transformation, a uniform quantizer, and a nonlinear decoding transformation, which only contains three convolutional layers. With the great performance achieved in Gong et al. (2013), the residual block has been proved to be a remarkably efficient way in the aspect of reducing information loss due to deep layers network (Baig et al. 2017). Firstly, the residual block allows the deeper layers to know how to utilize information which could not be generated by the previous stage. Secondly, these connections reduce the distance of the path that information travels, which brings better joint optimization. In Agustsson et al. (2017) proposed to learn compressible representations using deep architectures, which can be trained end-to-end. In Theis et al. (2017) utilized the derivative of a smooth approximation to replace the derivative in the backward pass of back-propagation and optimized the autoencoder network. Outstanding performance was reported.

Multi-task learning has also attracted significant attention. In the application of deep neural network, Long and Wang (2015) proposed an explicit multi-task framework, Multi-linear Relationship Networks (MRN), to discover the task relationships in deep convolutional networks and achieve a promising performance in the multi-task learning datasets. Lu et al. (2017) proposed a compact multi-task deep learning architecture which was initialized with a thin network and dynamically widened during training. Misra et al. (2016) designed a frame that was firstly trained in a separate strategy, and then explored cross-stitch units to predict how to use the knowledge of the other task. On the other hand, multi-task learning also inspires the non-neural models. For example, Argyriou et al. (2007) has proposed to learn a feature representation by sharing across a set of multiple related tasks, based on a 1-norm regularization to control the number of learned features for all the tasks.

3 Proposed Framework

Given N images, $\mathbf{I} = \{\mathbf{I}_i\}_{i=1}^{N}$ without labels, our goal is to learn their compact binary codes $\mathbf{B}$ and $\mathbf{B}^c$ such that: (a) the original image can be reconstructed from $\mathbf{B}^c$, (b) similar images can be retrieved using $\mathbf{B}$, and (c) both $\mathbf{B}^c$ and $\mathbf{B}$ can be computed directly without relaxation.

We illustrate our proposed BGAN+ framework by the scheme in Fig. 1. The framework consists of two components: (1) a binary generative adversarial *compression* network (BGANc), and (2) a binary generative adversarial *retrieval* network (BGANr). Both parts learn their binary codes in an unsupervised fashion. In the BGANc part, $\mathbf{B}^c$ is learned through the interplay between a generator and a discriminator. Specifically, the generator takes an image as input and represents it by a binary code. Then, this code is decoded

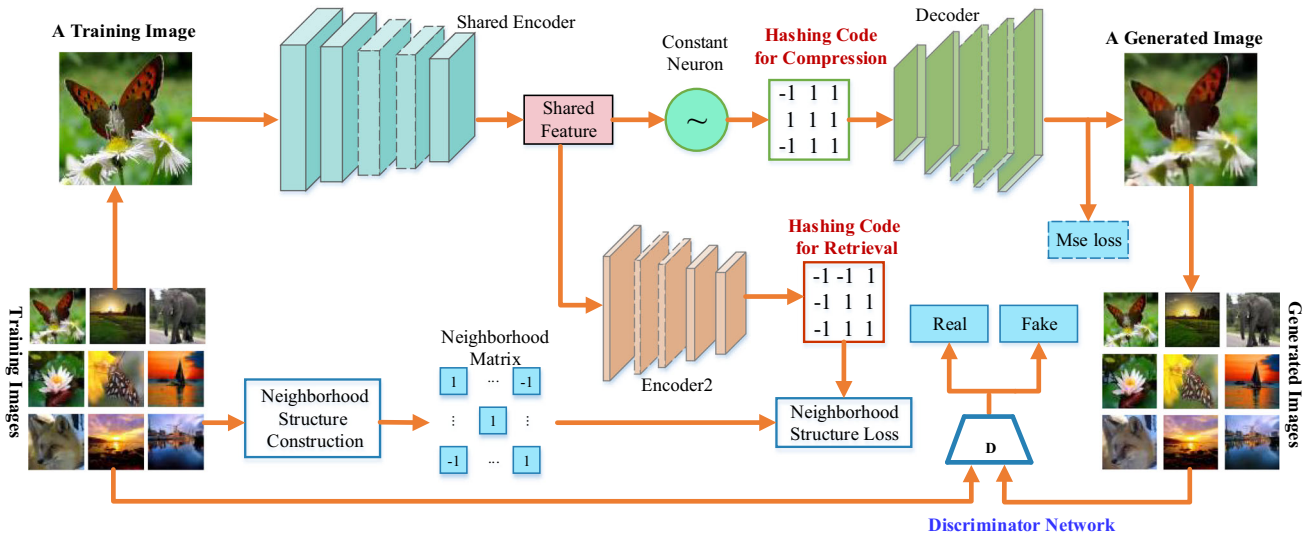


Fig. 1 An overview of our proposed BGAN+ framework for simultaneously learning binary codes for image retrieval and compression. Our framework contains two major networks, i.e., image compression network and image retrieval network. For the image compression network (BGANc), there are four key components: (1) a shared encoder, for learning low-level image representations, (2) a constant neuron layer, for learning the binary codes for image compression, (3) a decoder, to reconstruct the original images, and (4) a discriminator, to distinguish between real and reconstructed images. For the image retrieval network (BGANr), there are three key components: (1) a shared encoder, (2) encoder2, for learning high-level image representations, and (3) a hashing layer, for learning the binary codes for image retrieval. As a preprocessing step, we construct the neighborhood structure of the training images

to reconstruct the image, which enters the verification process in the discriminator to be compared with the original image. The BGANr part learns binary code $\mathbf{B}$ by also taking into account the visual neighborhood structure of the image in order to make sure that the proper notion of image similarity propagates into the similarity of the learned binary codes for retrieval. The two learning processes are coupled by the output of the shared encoder. In this way, the criteria used to learn binary code in one part of the framework helps in learning the binary codes in the other part. Although we learn two separate codes for two different purpose, we hypothesize that this synergetic effect is beneficial for both learning processes and justifies their integration into a single framework, as opposed to creating two binary codes using separate frameworks. In addition, through shared modules, both codes are learned in a more efficient manner than if they are learned independently. Related to the latter, for the learning of the parameters, we train the entire framework at once by a joint training strategy. In the remainder of this section, we provide detailed information about the our proposed BGAN+ framework.

3.1 Binary Generative Adversarial Compression Network (BGANc)

The binary codes $\mathbf{B}^c$ learned from $\mathbf{I} = \{\mathbf{I}_i\}_{i=1}^N$ by BGANc have the task to represent an image such that it can be reconstructed as well as possible back to its original state. We model this goal by the following expression:

$$\ell(\mathbf{I}) = \min_f ||\mathbf{I} - f(\mathbf{I})|| \tag{2}$$

where f denotes the transformation function transforming the original image $\mathbf{I}_i$ into the reconstructed image $\mathbf{I}^R$. The transformation function consists of the elements of the shared encoder, the proposed constant neuron and the decoder. We explain these components in more detail in Sects. 3.1.1, 3.1.2 and 3.1.3, respectively. Then, in Sect. 3.1.4. we come back with an elaborate version of the above expression, taking into account the realizations of the three components.

3.1.1 Shared Encoder

VGG network (Simonyan and Zisserman 2014) utilizing an architecture with 3×3 convolution filters is able to achieve good performance for large scale image classification with 19 weight layers. In this paper, we build our shared encoder with the first five convolution layers of VGG, with the details illustrated in Table 1. Following the architecture of Ledig et al. (2017) and Radford et al. (2015), we avoid two max-pooling operations throughout the shared encoder to allow our network to learn its own spatial downsampling. Specifically, we set the stride of the first four convolutional layers as 2, and thus each convolutional layer has the size (i.e., width and height) of the input feature map. The stride of the last convo-

Table 1 The architecture for feature extraction

Layer	Size of filter	Filters	Others
conv1_1	3×3	64	Stride = 1, padding = 1, relu
conv1_2	3×3	64	Stride = 2, padding = 1, relu
conv2_1	3×3	128	Stride = 1, padding = 1, relu
conv2_2	3×3	128	Stride = 2, padding = 1, relu
conv3_1	3×3	256	Stride = 1, padding = 1, relu
conv3_2	3×3	256	Stride = 1, padding = 1, relu
conv3_3	3×3	256	Stride = 1, padding = 1, relu
Max pooling	2×2		2
conv4_1	3×3	512	Stride = 1, padding = 1, relu
conv4_2	3×3	512	Stride = 1, padding = 1, relu
conv4_3	3×3	512	Stride = 1, padding = 1, relu
Max pooling	2×2		2
conv5_1	3×3	512	Stride = 1, padding = 1, relu
conv5_2	3×3	512	Stride = 1, padding = 1, relu
conv5_3	3×3	512	Stride = 1, padding = 1, relu
Max pooling	2×2		2
FC6	None	4096	relu
FC7	None	4096	relu

lutional layer is set as 1, which can be considered as a fully convolutional layer. Given an image $\mathbf{I}_i$ with the size of $W \times H$ (i.e., W as width and H as height), we can obtain C feature map with the size of $\frac{W}{16} \times \frac{H}{16} \times 3$ through our shared encoder, where C denotes the number of feature map channels.

3.1.2 Constant Neuron

To compress an image into a hash code and then reconstruct the image from the generated hash code, the issue related to binary constraints becomes relevant. The problem of binary constraints has been addressed by relaxing the constraints from $\{0, 1\}$ (Weiss et al. 2008) or by adopting alternating optimization strategies (Gong et al. 2013, 2012). However, they either cause a large optimality gap between non-relaxed and relaxed objectives or substantially weaken the model flexibility, respectively. As a result, in Dai et al. (2017) proposed to define a function $\mathbf{h}$ to re-parameterize Bernoulli distribution over the binary variables to avoid directly working with binary variables. $\mathbf{h}$ is referred to as stochastic neuron:

$$\mathbf{h}(z) = \begin{cases} 1 & if\ z \geq \xi \\ 0 & if\ z < \xi \end{cases} \tag{3}$$

where $\xi \sim \mu\,(0, 1)$. Inspired by the stochastic neuron, in this paper, we propose a *constant neuron*, which is defined as:

$$\mathbf{h}(z) = \begin{cases} 1 & if\ z \geq 0.5 \\ 0 & if\ z < 0.5 \end{cases} \tag{4}$$

Since $\mathbf{h}$ is not smooth, it is still difficult to apply stochastic gradient descent to calculate the gradient of the parameters. To solve this problem, we firstly define $\mathbf{W}_e$, $\mathbf{b}_e$ as the parameters of our shared encoder. As a result, the intermediate compressed hash codes can be formulated as:

$$\mathbf{B}^c = \mathbf{h}(E_n(\mathbf{I}; \mathbf{W}_e, \mathbf{b}_e, \delta)) \tag{5}$$

where δ is the active function of the convolutional layers and E_n is our encoding function. Then, we set the active function of the last convolution layer of shared encoder as $sigmoid$ and the other four layers are set as $ReLu$. Finally, we define our constant neuron as:

$$\mathbf{h}(\mathbf{I}; \Phi) = \frac{sign(E_n(\mathbf{I}; \mathbf{W}_e, \mathbf{b}_e, \delta) - 0.5) + 1}{2} \tag{6}$$

Unfortunately, the $sign$ function is non-smooth and gradient of $sign$ is zero. Following Grubb (2008), we use distributional derivative to estimate the stochastic gradient by:

$$\nabla_\Phi \mathbf{h}(\mathbf{I}; \Phi) = \nabla_\ell(\mathbf{I}; \Theta) E_n(\mathbf{I}; \Phi, \delta) \bullet (1 - E_n(\mathbf{x}; \Phi, \delta))\mathbf{I}^T \tag{7}$$

where $\bullet$ denotes a point-wise product, and where $\Phi = \{\mathbf{W}_e, \mathbf{b}_e\}$ and $\nabla_\ell(\mathbf{I}; \Theta)$ is the gradient from our objective function. More specifically, we utilize chain rules to calculate it. To conduct optimization, we leverage standard stochastic gradient descent algorithm to optimize $\mathbf{B}^c$ by following Nemirovski et al. (2009) and Kingma and Ba (2014).

3.1.3 Decoder

The decoder of BGANc takes the output of the constant neuron as input to reconstruct the original image. Therefore, the input for the decoder is $\mathbf{B}_i^c$ and the output is an image $\mathbf{I}_i^R$. Transferring such low dimensional features $\mathbf{B}_i^c$ to a high dimensional feature is a challenging task. In previous work, most auto-encoding systems (Larsen et al. 2016; Kingma and Welling 2013) use a fully connected layer as the first layer of a decoder for transforming a low dimensional feature into a high dimensional feature by a non-linear projection, but this substantially reduces the model flexibility: the input size must be fixed. However, cropping or wrapping an image into a fixed size can lead to a loss of image information (He et al. 2015). More importantly, in reality, we need to provide an efficient approach to compress images with an arbitrary size. Previous work (Gong et al. 2013; Ledig et al. 2017; Baig et al. 2017) demonstrated that residual blocks have a significant effect on reducing the information loss as the network deepening. Inspired by this observation, we design our decoder network by integrating deconvolutional layers (Odena et al. 2016), residual blocks with fully convolutional layers, for efficiently reconstructing images from binary codes. Specifically, the decoder consists of four deconvolutional layers (i.e., setting as $3 \times 3 \times 512$, $3 \times 3 \times 256$, $3 \times 3 \times 128$ and $3 \times 3 \times 64$) and three residual blocks, each with two convolutional layers, followed by a convolutional layer ($3 \times 3 \times 64$ and two fully convolutional layers ($1 \times 1 \times 64$, and $1 \times 1 \times 3$). The structure of our proposed decoder is shown in Fig. 2.

3.1.4 BGANc Optimization Objective

Based on the realizations of the three BGANc components as explained above, we can now define the expression Eq. (2) more concretely. What we minimize in Eq. (2) is actually the loss of reconstructing the input image. The definition of the corresponding loss function as the optimization objective is critical for the performance of our generator network. In this subsection, we define two loss functions that contribute to the optimization objective of the BGANc network.

Content Loss The first component is the *content loss*, which directly measures the deviation of the reconstructed image from the original. While both l_1 loss and l_2 loss are applied for image generation task and previous work (Ledig et al. 2017) has proven that l_1 loss performs better than l_2 loss, thus we define our content loss function as below:

$$\ell_{D_e} = \min_{\Omega} \sum_{i=1}^{N} \left\| \mathbf{I}_i - D_e(\mathbf{B}_i^c; \Omega) \right\| \tag{8}$$

where D_e denotes the decoding operation, Ω denotes the parameters of decoder and $\mathbf{B}^c$ is seen as the compressed hashing codes generated by our encoder. Furthermore, we can rewrite Eq. 8 as the following pixel-wise l_1 loss:

$$\ell_{D_e}(\mathbf{I}; \Omega) = \min_{\Omega} \frac{1}{WH} \sum_{i=1}^{N} \sum_{p=1}^{W} \sum_{q=1}^{H} \left\| I_{i,p,q} - I_{i,p,q}^R \right\| \tag{9}$$

Obviously, Eq. 9 is continuous and can be directly optimized by the stochastic gradient descent algorithm.

Adversarial Loss In order to make the optimization of BGANc more robust, we also take the quality of the reconstructed image from another perspective. Following the approach by Goodfellow et al. (2014), we define a "Discriminator" network $\mathbf{D}$ in such a way that it is optimized using criteria conflicting with those of $\mathbf{G}$. $\mathbf{G}$ is the "Generator" network (i.e., the decoder D_c shown in Fig. 1). In this way, $\mathbf{D}$ can act as adversary to $\mathbf{G}$ in the overall min–max optimization process. The goal of this optimization is to improve $\mathbf{G}$ such to be able to generate the images as well as possible. The process being adversarial to image generation is the process of trying to distinguish between the original and reconstructed images. If $\mathbf{G}$ manages to generate the images so well to "fool" $\mathbf{D}$, then it "wins" the min–max game and the overall GAN optimization has converged. In view of this, given a model of the image classifier $\mathbf{D}$ assessing the original ($\mathbf{I}$) and reconstructed ($\mathbf{I}^R$) image, we can formally define the min–max game resulting in the optimal system parameters as follows:

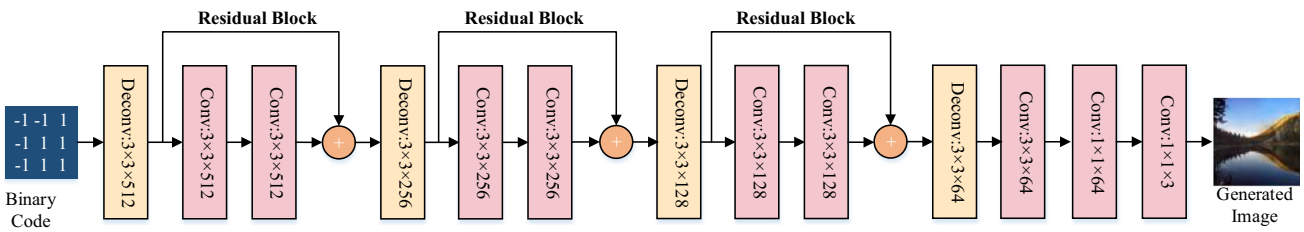


Fig. 2 Configuration of the decoder. $\mathbf{B}_i^c \in \mathbb{R}^{\frac{W}{16} \times \frac{H}{16} \times C}$ is the input code, where C controls the bit-rate, and $\mathbf{I}_i^R \in \mathbb{R}^{W \times H \times 3}$ is the output image. The input image $\mathbf{I}_i$ is firstly compressed to $\mathbf{B}_i^c$. The decoder reconstructs an image $\mathbf{I}_i^R$ from $\mathbf{B}_i^c$ using several Deconv and Conv layers, and ensures that the final output image $\mathbf{I}_i^R$ has the same size as the original image $\mathbf{I}_i$. Skip connection works as the shortcut in residual network

$$\ell_A(\mathbf{I}_i; \Phi, \Omega, \Theta) = \min_{\Phi,\Omega} \max_{\Theta} \log(D(\mathbf{I}_i)) + \log(1 - D(\mathbf{I}_i^R)) \tag{10}$$

where Φ, Ω are, respectively, the parameters of the encoder and decoder network, and Θ is the vector of the parameters of the discriminator.

Here we follow the architecture design summarized by Radford et al. (2015). We use ReLU activation and avoid max-pooling throughout the network. It contains 4 convolutional layers with an increasing number of 5×5 filter kernels (32, 128, 256, and 512). Strided convolutions are used to reduce the image resolution and each time the number of features is doubled. The resulting 512 feature maps are followed by a dense layer with the size of 1024 and a final sigmoid activation function to obtain a probability for sample classification.

We can formulate the objective function of compression network as the weighted sum of the pixel-wise l_1 loss and the adversarial loss as:

$$\ell_C = \ell_{D_e} + \lambda \ell_A \tag{11}$$

where λ is the weighted factor to balance the impact of pixel-wise loss and adversarial loss.

3.2 Binary Generative Adversarial Retrieval Network (BGANr)

The task of our retrieval network BGANr is to generate a hash code $\mathbf{B}_i$ from an image $\mathbf{I}_i$. The structure of BGANr consists of two parts: shared encoder, the encoder2 and the hashing layer (see Fig. 1).

3.2.1 Encoder2

Specifically, our BGANr is based on the VGG network and the specific configuration is defined in Table 1. Theoretically, we can directly use $\mathbf{B}^c$ to retrieve images. However, it is unlikely to acquire excellent results due to the reason that compression network only encodes low-level information without high-level semantic information. To conduct an efficient search, hash code $\mathbf{B}$ must encode both low-level and high-level semantic information. As a result, we design our BGANr by sharing the encoder of $\mathbf{G}$ to extract better low-level information.

3.2.2 Construction of Neighborhood Structure

Moreover, for the training of the system, we first conduct the neighborhood structure of images and then train the network. Neighborhood structure is beneficial to exploiting the manifold structure of the training data, and can improve the performance of image retrieval (Wang et al. 2018). Next, we introduce our approach to construct a similarity matrix by an unsupervised method.

In our unsupervised approach, we propose to exploit the neighborhood structure of the images in feature space as information source steering the process of hash learning. Specifically, we propose a method based on the K-Nearest Neighbor (KNN) concept to create a neighborhood matrix of $\mathbf{S}$. We use two types of features to construct $\mathbf{S}$: non-deep features and deep features. For non-deep features, we use the hand-crafted features provided with the dataset. For deep features, we extract 2048-dimensional features from the pool5-layer based on He et al. (2016). This results in the set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ where $\mathbf{x}_i$ is the feature vector of image $\mathbf{I}_i$.

For the representation of the neighboring structure, our task is to construct a matrix $\boldsymbol{S} = \{s_{ij}\}_{i,j=1}^N$, whose elements indicate the similarity ($s_{ij} = 1$) or dissimilarity ($s_{ij} = -1$) of any two images i and j in terms of their features $\mathbf{x}_i$ and $\mathbf{x}_j$.

We compare images using cosine similarity of the feature vectors. For each image, we select $K1$ images with the highest cosine similarity as its neighbors. Then we can construct an initial similarity matrix $\mathbf{S}_1$:

$$(S_1)_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_j \text{ is K1-NN of } \mathbf{x}_i \\ -1, & \text{otherwise} \end{cases} \tag{12}$$

The precision curve (evaluated using the labels) in Fig. 3 indicates the quality of the constructed neighborhood for different values of $K1$. Due to rapidly decreasing precision with increasing $K1$, creating a large-enough neighborhood by simply increasing $K1$ is not the best option. In order to find a better approach, we borrow ideas from the domain of graph modeling. In an undirected graph, if a node v is connected to a node u and if u is connected to a node w, we can infer that v is also connected to w. Inspired by this, if we treat every training image as a node in an undirected graph, we can expand the neighborhood of an image i by exploring the neighbors of its neighbors. Specifically, if $\mathbf{x}_i$ connects to $\mathbf{x}_j$ and $\mathbf{x}_j$ connects to $\mathbf{x}_k$, we can infer that $\mathbf{x}_i$ has the potential to be also connected to $\mathbf{x}_k$.

In view of the above, we use the initial similarity matrix $\mathbf{S}_1$ to expand the neighborhood structure. Specifically, based on $\mathbf{S}_1$, we calculate the similarity of two images by comparing the corresponding columns in $\mathbf{S}_1$ using the expression $\frac{1}{||(\mathbf{S}_1)_i - (\mathbf{S}_1)_j||^2}$. Then we again construct a ranked list of $K2$ neighbors, based on which we generate the second similarity matrix $\mathbf{S}_2$ as:

$$(\mathbf{S}_2)_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_j \text{ is K2-NN of } \mathbf{x}_i \\ -1, & \text{otherwise} \end{cases} \tag{13}$$

Finally, we construct the neighborhood structure by combining the direct and indirect similarities from the two

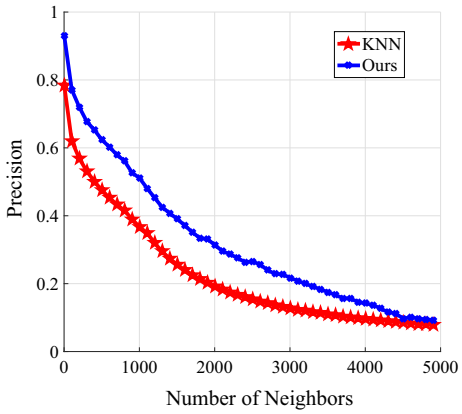


Fig. 3 Precision of constructed labels on cifar-10 dataset with different K, and different methods (deep features are used)

Algorithm 1 Construction of neighborhood structure

Input: Images $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, the number of neighbors K1, the number of neighbors K2 for the neighbors expansion;
Output: Neighborhood matrix $\mathbf{S} = \{s_{ij}\}$;
First ranking: Use cosine similarity to generate the index of K1-NN of each image $L_1, L_2, \ldots, L_N$;
Neighborhood expansion:
for j=1,...,N **do**
  Initialize $num \leftarrow 0$;
  for j=1,...N **do**
    $num_j \leftarrow$ the size of $L_i \cap L_j$;
  end for
  Sort num by descending order and keep the top K2 $\{L_j\}$;
  Set new $L'_i \leftarrow$ union of the top K2 $\{L_j\}$;
end for
for j=1,...,N **do**
  Construct $\mathbf{S}$ with new L'_i based on Eq. 14;
end for
return $\mathbf{S}$;

matrices together. This results in the final similarity matrix $\mathbf{S}$:

$$S_{ij} = \begin{cases} 1, & \text{if } (\mathbf{S}_1)_{ij} = 1 \text{ or } \mathbf{x}_j \text{ is a K1-NN of } \mathbf{x}_i\text{'s K2-NN} \\ -1, & \text{otherwise} \end{cases} \tag{14}$$

The whole algorithm is shown in Algorithm 1. We note here that we could have also omitted this preprocessing step and construct the neighborhood structure directly during the learning of our neural network. We found, however, that the construction of neighborhood structure is time-consuming, and that updating of this structure based on the updating of image features in each epoch does not have a significant impact on the performance. Therefore, we chose to obtain this neighborhood structure as described above and fix it for the rest of the process.

3.2.3 Neighborhood Structure Loss

The last section describes how to construct the similarity matrix and in this section we will present our objective function to preserve pair-wise similarity into hashing codes. Like Wang et al. (2018), we define our neighbor loss as below:

$$\ell_N(x; \Lambda, \Phi) = \min_{\Lambda,\Phi} \sum_{s_{ij} \subset \mathbf{S}} \frac{1}{2}\left(\frac{1}{L} b_i^T b_j - s_{ij}\right)^2 \tag{15}$$

where Λ denotes the parameters of the retrieval network and $\mathbf{S}$ is constructed by Algorithm 1 and $s_{ij} \in \mathbf{S}$. Unfortunately, in Eq. 15 b_i is discrete, whose gradient is zero for all nonzero inputs and leads to the failure of training the deep network by disabling the back propagation. A wide range of works have proposed many novel methods to solve this problem. Lin et al. (2016) and Zhang et al. (2014) proposed to use an approximate solution to relax the binary codes, however, which would certainly bring a large quantization error. Therefore, relaxation the binary code is not an efficient way to solve the discrete hashing problem.

In order to address this problem of optimizing binary codes with non-smooth $sign$ activation, we acquire the inspiration from recent works (Cao et al. 2017; Shen et al. 2015). These studies mainly focus on how to convert the difficult optimization problems into several easily optimized subproblems by changing the smoothness of the original function. Specifically, we can gradually reduce the degree of the smoothness of function, which results in a sequence of subproblem optimizations converging to the original optimization problem. Following this idea, if we figure out the similar or approximate smooth function with $sign(.)$, and then gradually make the smooth function non-smooth during the training process, and finally, the results will converge to the desired target.

Motivated by this, we define a function $app(.)$ to approximate $sign(.)$:

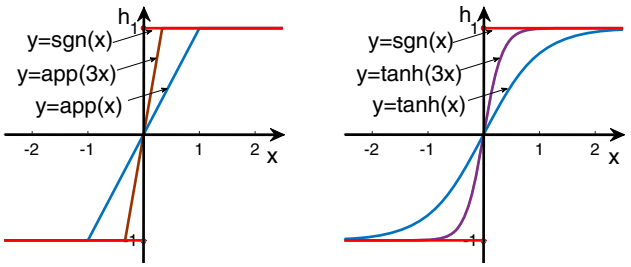
$$app(z) = \begin{cases} +1, & if\ z > 1 \\ z, & if\ 1 \geq z \geq -1 \\ -1, & if\ z < -1 \end{cases} \tag{16}$$

Obviously, there is a relationship between $sign(.)$ and $app(.)$ function:

$$sign(z) = \lim_{\beta \to \infty} app(\beta z) \tag{17}$$

In Fig. 4, we illustrate how $app(.)$ function approximates the original $sign(.)$. In addition, we also introduce an alternative $tanh(.)$:

$$\text{sign}\,(z) = \lim_{\beta \to +\infty} \tanh\,(\beta z) \tag{18}$$



(a) tanh(.) **(b)** app(.)

Fig. 4 Illustrative process of how app(.) and tanh(.) approximate sign(.)

Algorithm 2 Leaning parameters

Input: Images $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$
1: Initialize $\{\Phi, \Omega, \Theta\}$ randomly and Λ with the pre-trained model in Simonyan and Zisserman (2014).
2: **for** j=1,...,t **do**
3: Sample x_i uniformly from $\{x_i\}_{i=1}^N$.
4: Compute the stochastic gradient $\nabla \ell_C$ in 11.
5: Update decoder parameters as
6: $\Omega_{i+1} = \Omega_i - \tau_i \nabla_\Omega \ell_C$
7: Compute the stochastic gradient $\nabla_\Phi \mathbf{h}(x; \Phi)$ in 7.
8: Update encoder parameters as
9: $\Phi_{i+1} = \Phi_i - \tau_i \nabla_\Phi \mathbf{h}(x; \Phi)$
10: Compute the stochastic gradient ∇_{ℓ_A} in 10.
11: Update discriminator parameters as
12: $\Theta_{i+1} = \Theta_i - \tau_i \nabla_\Theta \ell_A(x; \Theta)$
13: Compute the stochastic gradient ∇_{ℓ_N} in 19.
14: Update encoder2 parameters as
15: $\Lambda_{i+1} = \Lambda_i - \tau_i \nabla_\Lambda \ell_N(x; \Lambda)$
16: **end for**

Therefore, as for Eq. 15, we can optimize $\mathbf{Z}$ instead of directly modeling the neighbors structure loss on the binary codes $\mathbf{B}$. Then Eq. 15 can be reformulated as:

$$\ell_N(x; \Lambda, \Phi) = \frac{1}{2} \min_{\Lambda,\Phi} \sum_{s_{ij} \subset \mathbf{S}} \left(\frac{1}{L} z_i^T z_j - s_{ij} \right)^2 + \alpha \|\mathbf{Z} - \mathbf{B}\|^2 \tag{19}$$

where α is the hyper-parameter to balance the terms.

4 Learning

Using the loss functions in Eqs. 9, 10 and 19, we train our network.

The forward propagation is as follows. First, we use a deep convolutional network as the encoder to extract the features and then use the constant neuron layer to embed the real-valued features into binary codes:

$$\mathbf{B}_i^c = h(E_n(\mathbf{I}_i; \Phi)) \tag{20}$$

where $\mathbf{I}_i$ is an input image, Φ is the parameter of the encoder, E_n is the encoder operation and h is the constant neuron layer. Then, $\mathbf{B}_i^c$ is the input for a generator (decoder) D_e to reconstruct an image $\mathbf{I}^R$:

$$\mathbf{I}_i^R = D_e(\mathbf{B}_i^c; \Omega) \tag{21}$$

where Ω stands for the parameters of the generator (decoder) D_e. Finally, a discriminator $\mathbf{D}$ assigns the probability

$$p = \mathbf{D}(\mathbf{I}_i^R; \Theta) \tag{22}$$

if $\mathbf{I}_i^R$ is an actual training sample and probability $1 - p$ if $\mathbf{I}_i^R$ is generated by our model $\mathbf{I}_i^R = D_e(\mathbf{B}_i^c; \Omega)$. Θ stands for the parameters of the discriminator $\mathbf{G}$.

Similarly, for the retrieval network, given $\mathbf{I}_i$, we can obtain its binary codes by:

$$\mathbf{B}_i = sign\left(E_{n2}(E_n(\mathbf{I}_i; \Phi); \Lambda)\right) \tag{23}$$

where E_{n2} is encoder2, and Λ represents its parameters.

As shown in Fig. 1, the compression network and retrieval network shared the top 5 convolutional layers, which allow us to train the entire network through a joint training strategy. In our network, we have parameters of $\{\Phi, \Omega, \Theta, \Lambda\}$ to learn. All parameters can be learned by back-propagation (BP). In particular, we randomly initialize $\{\Phi, \Omega, \Theta\}$ and use the pre-trained model (Simonyan and Zisserman 2014) on ImageNet to initialize Λ. During each iteration, we sample a mini-batch of the images from the training data and use forward propagation to obtain the value of $\{\ell_C, \ell_N\}$, and then apply BP rules to update the associated parameters. The updated formulation is below:

$$\begin{aligned} \Omega_{i+1} &\leftarrow \Omega_i - \tau_i \nabla_\Omega \ell_C \\ \Phi_{i+1} &\leftarrow \Phi_i - \tau_i \nabla_\Phi \mathbf{h}(x; \Phi) \\ \Theta_{i+1} &\leftarrow \Theta_i - \tau_i \nabla_\Theta \ell_A(x; \Theta) \\ \{\Lambda_{i+1}, \Phi_{i+1}\} &\leftarrow \{\Lambda_i, \Phi_{i+1}\} - \tau_i \nabla_{\Lambda,\Phi} \ell_N(x; \Lambda, \Phi) \end{aligned} \tag{24}$$

where ∇ is the gradient and τ_i is the learning rate. Details are shown in Algorithm 2.

For the retrieval hashing codes, we set the $\beta = 1$ at the beginning. For each stage, after the retrieval network converges, we enlarge β for the next stage and use the parameters converging in the last stage to initialize the current stage parameters. By involving $app(\beta z)$ with $\beta \approx \infty$, the retrieval network obtains the same results as using $sign(z)$, which can learn exact binary hash codes as we desire. In the experiment, when we increase β to 10, the network can converge to the expected degree. In addition, we set $\lambda = 0.1$.

5 Experiments

We evaluate our BGAN+ on the task of large-scale image retrieval and image compression. Firstly, we compare BGAN+ with the state of the art methods both in image retrieval and compression. Secondly, we conduct an ablation study to evaluate the effect of each major component.

5.1 Datasets and Settings

To evaluate our method, we conduct our experiments on six public datasets: The Oxford 17 Category Flower (Nilsback and Zisserman 2006), Stanford Dogs-120 (Khosla et al. 2011), CIFAR-10 (Krizhevsky and Hinton 2009), Flickr-25K (Huiskes and Lew 2010), NUS-WIDE (Chua et al. 2009), and Kodak (Franzen 1999). Specifically, the first five datasets are used for retrieval, and NUS-WIDE and Kodak are used for image compression.

The Oxford 17 Category Flower dataset (Nilsback and Zisserman 2006) contains 17 categories and each class consists of 80 images, resulting in a total of 1360 images of flowers.

Brown Brown et al. (2010) consists of three datasets, namely Liberty, Yosemite and Notredame dataset, each of which includes more than 400,000 gray-scale patches, resulting in more than one million patches.

Stanford Dogs-120 Khosla et al. (2011) dataset consists of 20,580 images in 120 mutually categories. Each class contains about 150 images of dogs.

CIFAR-10 dataset consists of 60,000 labeled tiny colored image (32×32). It is a single labeled dataset. Each image has a unique class label belonging to one of the 10 classes.

Flickr-25K contains 25,000 images from Flickr-25K, where each image is labeled with one of the 38 concepts. We resize images of this subset into 256×256.

NUS-WIDE is a Web image dataset containing 269,648 images downloaded from Flickr. Tagging ground-truth for 81 semantic concepts is provided for evaluation. We follow the settings in Zhu et al. (2016) and use the subset of 195,834 images from the 21 most frequent concepts, where each concept consists of at least 5000 images.

Kodak contains 25 uncompressed color images of size 768×512. They are used as a standard test suite for compression testing.

In terms of retrieval, we split the Oxford Flower-17 into the training (40 images per class), validation (20 images per class), and test (20 images per class) sets. The Stanford Dogs-120 is divided into two parts: the train set (100 images per class) and test set (totally 8580 images for all categories). In NUS-WIDE and CIFAR-10, we randomly select 100 images per class as the test query set and 1000 images per class as the training set. In Flickr, we randomly select 1000 images as the test query set and 4000 images for training.

In terms of compression, we randomly select 21,000 images (1000 per class) to train our compression network. After the training, we apply the trained model to evaluate the performance on two testing datasets: (1) randomly select 10,000 from NUS-WIDE dataset as the first testing dataset and (2) the Kodak dataset.

5.1.1 Evaluation Metric

For retrieval task, the hamming ranking is used as the search protocol to evaluate our proposed approaches, and two indicators are reported. (1) Mean Average Precision (**mAP**): for a single query, Average Precision (AP) is the average of the precision value obtained for the set of top-k results, and this value is then averaged over all the queries. (2) **Precision**: we further use the precision-recall curve and precision@K to evaluate the precision of retrieved images.

For compression task, we use MS-SSIM (Wang et al. 2003) to test the quality of the image. The higher MS-SSIM means a better quality.

5.1.2 Compared Methods

For retrieval Task, we compare our BGAN+ with other state-of-the-art hashing algorithms. Specifically, we compare with four non-deep hashing methods [iterative quantization (ITQ) hashing (Gong et al. 2013), spectral hashing (SH) (Heo et al. 2015), Locality Sensitive Hashing (LSH) (Datar et al. 2004), Spherical Hashing (Heo et al. 2015)], PGDH (Yuan et al. 2018), DTH (Wang et al. 2016), and two unsupervised deep hashing methods [DeepBit (Lin et al. 2016) and Deep Hashing (DH) (Liong et al. 2015)]. To make a fair comparison, we also apply the non-deep hashing methods on deep features extracted by the VGG network [VGG-fc7 (Simonyan and Zisserman 2014)].

For non-deep hashing algorithms, we use the features provided with the dataset. By constructing the neighborhood structure using the labels, our method can be easily modified as a supervised hashing method, named as (BGAN+_s). Therefore, we also compare with some supervised hashing methods, e.g., iterative quantization hashing (ITQ-CCA) (Gong et al. 2013), KSH (Liu et al. 2012), minimal loss hashing (MLH) (Norouzi and Blei 2011), CNNH (Xia et al. 2014b) and Deep Hashing Network (DHN) (Zhu et al. 2016). For compression task, we compare with four widely used image compression approaches: JPEG (Wallace 1991), JPEG 2000 (Rabbani and Joshi 2002), Theis et al. (2017) and JPEG 420.

5.1.3 Implementation Details

When constructing the neighborhood structure, we use two different types of features: non-deep features provided with

Table 2 mAP on CIFAR-10 using different optimization methods

Methods	mAP		
	24-bit	32-bit	48-bit
Two-step solution	0.344	0.362	0.373
sign $(z) = \lim_{\beta \to +\infty} \tanh(\beta z)$	**0.387**	**0.398**	**0.413**
sign $(z) = \lim_{\beta \to +\infty} app(\beta z)$	0.363	0.371	0.389

Bold values indicate the best result

the dataset, and 2048-dimensional deep features extracted using ResNet. We denote them as **BGAN+_non** and **BGAN+** respectively. The average number of the neighbors for each image is 400, 1021, 1168 for the three datasets: CIFAR-10, NUS-WIDE, and Flickr-25K. By default, we set $\lambda = 0.1$ and the learning rate as 0.001.

5.2 Results on Image Retrieval

5.2.1 The Effect of Binary Optimization

As discussed above, both BGAN and BGAN+ can learn binary hash codes directly while previous hashing methods first learn continuous representations and then generate hash codes using a sign function (denoted as two-step solution). The previous study on BGAN has verified that the two-step solution is sub-optimal, and binary optimization can achieve better performance. In this section, we further study the effect of binary codes optimization on the performance of hash codes to verify the robustness of our binary optimization approach. The performance results of BGAN+ on the CIFAR-10 are shown in Table 2. As shown in Table 2, our binary optimization can improve the performance of the learned binary codes. Specifically, the first *app* solution (Eq. 17) outperforms two-step solution by 1.9%, 0.9%, and 1.6% for 24, 32, and 48-bit hash codes, while the second solution *tanh* (Eq. 18) improves it by 4.3%, 3.6%, and 4.0%. This verifies our argument on BGAN+ that two-step solution is sub-optimal, and binary optimization can achieve better performance. These experimental results show the robustness of our proposed binary optimization approach.

5.2.2 Comparison with State-of-the-Art Methods for Fine-Grained Image Retrieval

In this section, we compare our BGAN+ with state-of-the-art methods for fine-grained image retrieval on two datasets. The mAP results are shown in Table 3.

It shows that our method (BGAN+) significantly outperforms the other unsupervised hashing methods (SH and ITQ+CCA) in both datasets. In Oxford 17 Category Flower dataset, BGAN+ outperforms the best counterpart (SH) by 11.7%, 12.3%, 14.7% and 15.2% for 12, 24, 32 and 48 bits, respectively. On the other hand, both SH and ITQ+CCA have unsatisfactory performance in Stanford Dogs-120 dataset. Their mAP is 0.008 for different bits, which is almost random. This indicates that hashing methods for general image retrieval may not work well on the task of fine-grained image retrieval. Compared with several supervised hashing methods, e.g., SDH (Shen et al. 2015), KSH (Liu et al. 2012), FastH (Lai et al. 2015), DQN (Cao et al. 2016), DQN (Cao et al. 2016), our BGAN+, as an unsupervised method, achieves even better performance in Oxford 17 Category Flower dataset. BGAN+ outperforms the best counter-part (DSH) by 14.0%, 9.8%, 9.8% and 5.8% for 12, 24, 32 and 48 bits, respectively. However, in the Stanford Dogs-120 dataset, FastH has better performance in general, and it is better than BGAN+ by 3.1%, 14.9% and 15.8% for 24, 32 and 48 bits. Nevertheless, FastH is a supervised hashing method while our BGAN+ is unsupervised.

5.2.3 Comparison with State-of-the-Art Methods for General Image Retrieval

In this section, we evaluate our hashing method performance on three datasets. The mAP results are shown in Tables 4 and 5 and Precision-Recall curves are shown in Fig. 5. From Tables 4 and 5, we can obtain the following conclusions.

Table 3 mAP results for fine-grained image retrieval using different number of bits on Oxford Flower-17 and Stanford Dogs-120. Note that our method is unsupervised while FashH is a supervised method

Bits	Oxford Flower-17				Stanford Dogs-120			
	12	24	32	48	12	24	32	48
SH (Weiss et al. 2008)	0.589	0.589	0.588	0.587	0.008	0.008	0.008	0.008
ITQ-CCA (Gong et al. 2013)	0.585	0.587	0.587	0.586	0.008	0.008	0.008	0.008
SDH (Shen et al. 2015)	0.108	0.140	0.117	0.145	0.009	0.018	0.090	0.037
KSH (Liu et al. 2012)	0.243	0.501	0.253	0.355	0.014	0.123	0.136	0.193
FastH (Lai et al. 2015)	0.402	0.524	0.528	0.536	0.044	**0.223**	**0.364**	**0.393**
DQN (Cao et al. 2016)	0.476	0.537	0.562	0.573	0.009	0.013	0.035	0.053
DSH (Liu et al. 2016)	0.566	0.614	0.637	0.680	0.012	0.012	0.012	0.012
BGAN+	**0.706**	**0.712**	**0.735**	**0.738**	**0.163**	0.192	0.215	0.235

Bold values indicate the best result

Table 4 mAP for different unsupervised hashing methods using different bits on two datasets. The first four methods are non-deep hashing methods, and the second the four methods are based on deep networks

	Cifar-10				NUS-WIDE			
Bits	12	24	32	48	12	24	32	48
ITQ (Gong et al. 2013)	0.162	0.169	0.172	0.175	0.452	0.468	0.472	0.477
SH (Weiss et al. 2008)	0.131	0.135	0.133	0.130	0.433	0.426	0.426	0.423
LSH (Datar et al. 2004)	0.121	0.126	0.120	0.120	0.403	0.421	0.426	0.441
Spherical (Heo et al. 2015)	0.138	0.141	0.146	0.150	0.413	0.413	0.424	0.431
ITQ+VGG	0.196	0.246	**0.289**	**0.301**	0.435	0.435	0.548	0.435
SH+VGG	0.174	0.205	0.220	0.232	0.433	0.426	0.426	0.423
LSH+VGG	0.101	0.128	0.132	0.169	0.401	0.442	0.480	0.471
Spherical+VGG	**0.212**	**0.247**	0.256	0.281	**0.549**	**0.614**	**0.653**	**0.678**
DeepBit (Lin et al. 2016)	0.185	0.218	0.248	0.263	0.383	0.401	0.403	0.412
DH (Liong et al. 2015)	0.160	0.164	0.166	0.168	0.422	0.448	0.480	0.493
BGAN_non (Song et al. 2018)	0.361	0.369	0.375	0.395	0.518	0.541	0.545	0.568
BGAN (Song et al. 2018)	0.401	0.512	0.531	0.558	0.675	0.690	0.714	0.728
BGAN+_non	0.375	0.387	0.398	0.413	0.544	0.552	0.561	0.579
BGAN+	**0.531**	**0.543**	**0.564**	**0.586**	**0.682**	**0.719**	**0.723**	**0.736**

Bold values indicate the best result for non-deep and deep methods

Table 5 mAP for different unsupervised hashing methods using a different number of bits on Flickr. The first four methods are non-deep hashing methods, and the second four methods are based on deep networks

	Flickr			
Bits	12	24	32	48
ITQ (Gong et al. 2013)	0.544	0.555	0.560	0.570
SH (Weiss et al. 2008)	0.531	0.533	0.531	0.529
LSH (Datar et al. 2004)	0.499	0.513	0.521	0.548
Spherical (Heo et al. 2015)	0.552	0.547	0.546	0.545
ITQ+VGG	0.553	0.548	0.545	0.560
SH+VGG	0.550	0.544	0.541	0.545
LSH+VGG	0.543	0.549	0.555	0.551
Spherical+VGG	0.569	0.559	0.583	0.572
DeepBit (Lin et al. 2016)	0.501	0.505	0.511	0.513
DH (Liong et al. 2015)	0.553	0.548	0.543	0.556
BGAN_non (Song et al. 2018)	0.591	0.601	0.607	0.626
BGAN (Song et al. 2018)	0.683	0.702	0.703	0.703
BGAN+_non	0.599	0.612	0.618	0.636
BGAN+	**0.715**	**0.719**	**0.723**	**0.736**

Bold values indicate the best result

First, our method (BGAN+) significantly outperforms the other deep or non-deep hashing methods in all datasets. In CIFAR-10, the improvement of BGAN+ over other methods is more significant, compared with NUS-WIDE and Flickr dataset. Specifically, it outperforms the best counterpart (Spherical+VGG) by 31.9%, 29.6%, 30.8% and 30.5% for 12, 24, 32 and 48-bit codes. One possible reason is that CIFAR-10 contains simple images, and the constructed neighborhood structure is more accurate than in the other two datasets. BGAN+ improves the state of the art method by 13.3%, 10.5%, 7.0% and 5.8% in the NUS-WIDE dataset, and 14.6%, 16.0%, 14.0% and 16.4% in Flicker dataset.

Second, comparing with BGAN (or BGAN+), the performance of BGAN_non (or BGAN+_non) is worse. This indicates that the similarity graph plays an important role in the learning of hashing codes, and the non-deep features are not as good as deep features.

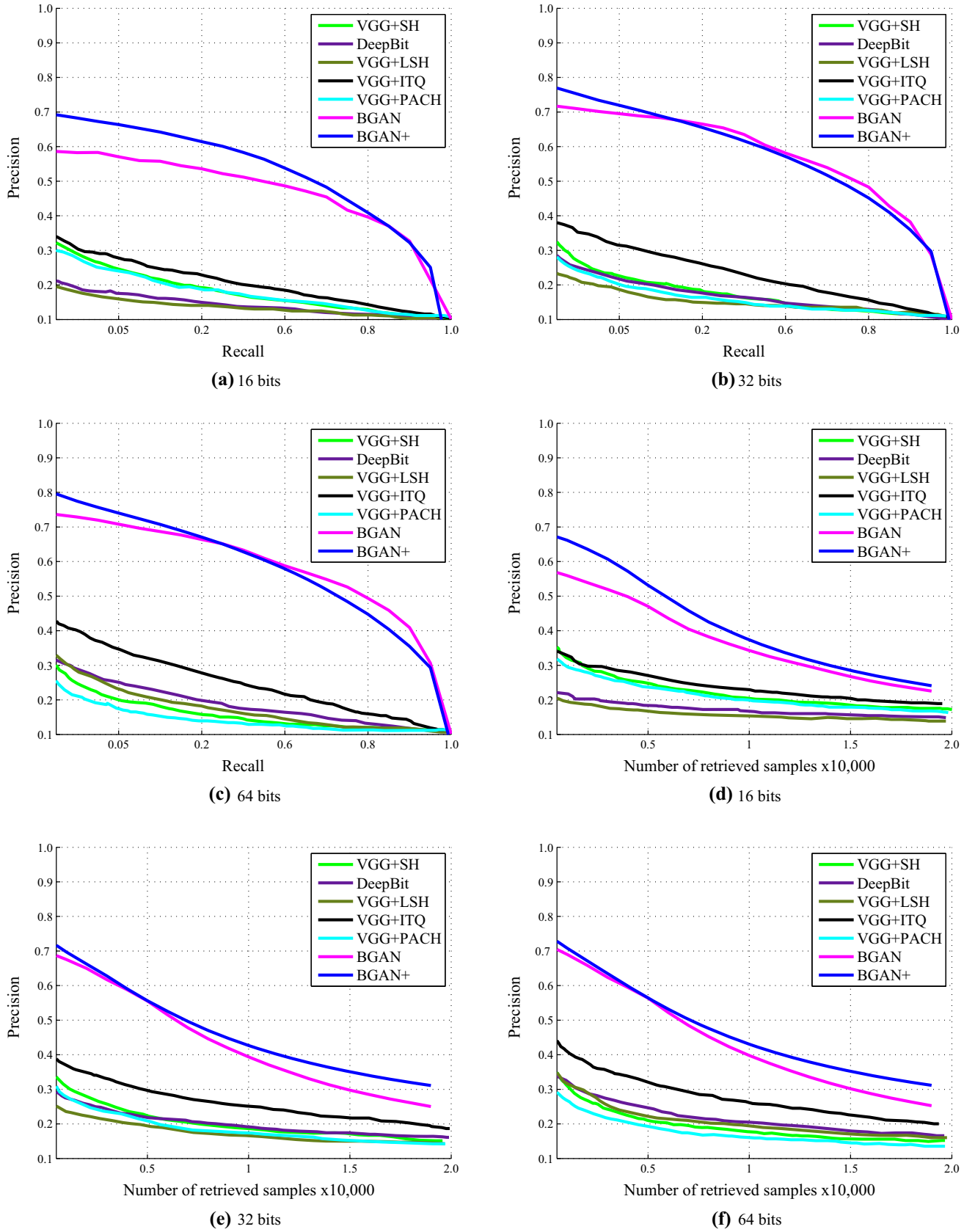
Third, from Tables 4 and 5, we observe that Spherical+VGG is a strong competitor in terms of mAP. On the other hand, the performance of deep hashing methods (i.e., DeepBit and DH) is not superior. A possible reason is that the deep hashing methods use only 3 fully connected layers to extract the features, which is not powerful enough.

Fourth, when we run the non-deep hashing method on deep features, the performance is usually improved compared with the hand-craft features. The performance gap is larger in CIFAR-10 and NUS-WIDE datasets than in Flicker dataset.

Fifth, with the increase of hash code length, the performance of most hashing method is improved accordingly. More specifically, the mAP improvements using deep features are generally more significant than that of non-deep features in CIFAR-10 dataset and NUS-Wide dataset. An exception is SH, which has no improvement with the increase of code length.

Sixth, compared with BGAN (or BGAN_non), BGAN+ (or BGAN+_non) achieves better performance. In particular, the increase of BGAN+ of 12-bit on the CIFAR-10 dataset is 13.0%. In addition, BGAN+ improves BGAN by 3.2%, 1.7%, 2% and 3.3% of 12, 24, 32 and 48-bit on the Flick dataset.

From Fig. 5, we have the following observations. In terms of the precision-recall curve, the results indicate that BGAN



(a) 16 bits **(b)** 32 bits **(c)** 64 bits **(d)** 16 bits **(e)** 32 bits **(f)** 64 bits

Fig. 5 Precision for different unsupervised hashing methods using different bits on CIFAR-10 dataset

and BGAN+ significantly outperform existing approaches. In general, BGAN+ performs better than BGAN, especially when the hash code is 16-bit. In addition, the bottom row of Fig. 5 shows the precision curves when we set a different number of retrieved samples (times of 10,000) and then train the model with 16, 32 and 64-bit, separately. When code length is 16-bit, BGAN+ achieves better performance. When code length is set as 32 or 64-bit and the number of retrieved samples is 5000, BGAN and BGAN+ obtain the same value of precision. In addition, when the number of retrieved samples gradually increases from 5000, the gap between BGAN and BGAN+ enlarges gradually.

Additionally, we also conduct the experiment to evaluate the quality of binary representation on the Brown dataset (Brown et al. 2010). The experimental setting follows DeepBit (Lin et al. 2016) and Table 6 shows the results of image matching by binary descriptor in terms of 95% error rates. Our method shows a slightly lower performance compared with DeepBit and DBD-MQ (Duan et al. 2017). The possible reason is that each patch in Brown dataset only has 3–5 matched images. So it is hard to construct pair-wise labels by Algorithm 1, i.e., most of matched patches are treated as the mismatched cases, and vice versa. In fact, we also compute the accuracy of the constructed similarity, only about 3% close to random constructing results. In contrast, on the real-world dataset, such as NUS-WIDE and CIFAR-10, BGAN+ can show a great advantage due to the fact that each image has sufficient pairs, even when there are many miss-matched pairs.

We also compare with supervised hashing methods, and present the mAP results on CIFAR-10 dataset in Table 7. It is observed that our BGAN+ reaches the highest mAP scores across hashing code length ranging from 12-bit to 48-bit. Compared with the best deep supervised hashing method DHN, BGAN+ has an increase of 17.6%, 15.0%, 14.4% and 13.6% over 12, 24, 32 and 48-bit. In generally, BGAN+ increases the BGAN by around 2.0% and the improvement is contributed by our compression network. This indicates that the performance improvement of BGAN is not only due to the constructed neighborhood structure, but also the other components. Similarly, HashGAN (Cao et al. 2018) also adapts GAN to learn binary codes in a supervised manner, but BGAN+ still outperforms HashGAN averagely about 11.05% on each bit length. It is worth noting that PGDH (Yuan et al. 2018) shows competitive results and is also a relaxation-free method by a reinforcement learning strategy. However, our method still significantly outperforms it by about 14% in terms of each bits. However, our method is mainly designed for unsupervised learning of hashing codes, and it has a large room to be improved for the task of supervised hashing.

Table 6 95% error rates (ERR) compared with the binary descriptors on Brown dataset (%), where SSC, DBD-MQ, Brisk, and DeepBit are unsupervised binary representation while LDAHash is supervised

Train	Test	Real-valued	Binary				
		SIFT (Lowe 2004)	SSC (Shakhnarovich 2005)	LDAH (Strecha et al. 2011)	DeepBit (Lin et al. 2016)	DBD-MQ (Duan et al. 2017)	BGAN+
Yosemite	Notredame	28.09	72.20	51.58	29.60	**27.20**	32.14
Yosemite	Liberty	36.27	71.59	49.66	34.41	**33.11**	36.11
Notredame	Yosemite	29.15	76.00	52.95	63.68	**57.24**	60.24
Notredame	Liberty	36.27	70.35	49.66	32.06	**31.10**	40.51
Liberty	Notredame	28.16	72.95	51.34	26.66	**25.78**	30.26
Liberty	Yosemite	28.15	77.99	52.95	57.61	57.15	**54.64**
Average		31.17	73.51	51.40	40.67	**38.59**	42.12

Bold values indicate the best result

Table 7 mAP for different supervised hashing methods using different number of bits on CIFAR-10

Method	CIFAR-10			
	12 bits	24 bits	32 bits	48 bits
ITQ-CCA (Gong et al. 2013)	0.435	0.435	0.435	0.435
KSH (Liu et al. 2012)	0.556	0.572	0.581	0.588
MLH (Norouzi and Blei 2011)	0.500	0.514	0.520	0.522
DNNH (Lai et al. 2015)	0.674	0.697	0.713	0.715
CNNH (Xia et al. 2014b)	0.611	0.618	0.625	0.608
DHN (Zhu et al. 2016)	0.708	0.735	0.748	0.758
HashGAN (Cao et al. 2018)	0.668	0.731	0.735	0.749
DTH (Wang et al. 2016)	0.710	0.750	0.765	0.774
PGDH (Yuan et al. 2018)	0.736	0.741	0.747	0.762
BGAN_s	0.866	0.874	0.876	0.877
BGAN+_s	**0.884**	**0.889**	**0.892**	**0.894**

Bold values indicate the best result

Table 8 MS-SSIM on NUS-WIDE at different bit-rate

Methods	MS-SSIM		
	0.15 bit/px	0.25 bit/px	0.5 bit/px
JPEG (Wallace 1991)	0.875	0.894	0.922
JPEG 2000 (Rabbani and Joshi 2002)	0.925	0.937	0.945
BGAN+	**0.927**	**0.939**	**0.948**

Bold values indicate the best result

5.3 Results on Image Compression

JPEG is an image compression standard approach, while JPEG 200 is an improvement on JPEG. They are both widely used for image compression. Additionally, we also compare our method with other deep neural network methods, such as RNN-based (Toderici et al. 2017). However, due to the code access problem, we do not compare our method with RAIC (Rippel and Bourdev 2017) and CWIC (Li et al. 2018). To evaluate the performance of our compression network, we use NUS-WIDE dataset to train our compression model and then evaluate it on two datasets: NUS-WIDE and Kodak dataset.

The experimental results obtained from the NUS-WIDE dataset are shown in Table 8, which demonstrates that our BGAN+ obtains the best performance in terms of MS-SSIM. Compared with JPEG, BGAN+ has an increase of 5.2%, 4.5% and 2.6% for 0.15, 0.25 and 0.5 bit/px. Specifically, the improvement gap becomes narrow with the increase of bit-rate. The lower the bit-rate is, the harder the compression operation is. Furthermore, our BGAN+ performs slightly better than JPEG2000.

In order to test the robustness of our compression network, we further run the trained model on the image compression benchmark dataset Kodak and the experimental results are shown in Table 9. From Table 9, we can find BGAN+ performs the best, except at the 0.5 bit/px where BGAN+ is slightly lower than RNN-based (Toderici et al. 2017), about 0.003. It is worth noting that RNN-based also compresses images into binary codes but on the low compression rate. BGAN+ outweighs RNN-based obviously, about 4.9% and 1.1%. In terms of the non-deep learning method, our method still outperforms them on the low compression rate, such as JPEG and JPEG 2000, by about 10.4%, 8.0% and 0.4% for 0.15, 0.25 and 0.5 bit/px compared with JPEG, respectively. As for JPEG 2000, the improvement is relatively lower, about 0.3% and 0.2% on 0.15 bit/px and 0.25 bit/px, respectively. Generally speaking, our method is more competitive at a lower bit rate because the reconstructed image is generated from the binary codes, which are highly compact codes.

To evaluate the ability of image reconstruction using BGAN+ and to compare with the previous BGAN proposed in Song et al. (2018), we demonstrate some qualitative results on CIFAR-10 dataset in Fig. 6. From Fig. 6, we can see that the images reconstructed from BGAN with 256-bit hash code are blurry compared with the ground-truth images. Compared with the images reconstructed from BGAN with 256-bit, BGAN+ can generate images of high quality with only 128-bit hash code. This indicates the effectiveness of our BGAN+ for image compression. With longer hash code (i.e., 256-bit), it can produce even better quality images, which is as good as the ground-truth images from the human visual aspect.

In addition, more visual examples are provided in Fig. 7 and Fig. 8. All the images are randomly selected from the

Table 9 MS-SSIM on Kodak at different bit-rate

Methods	MS-SSIM		
	0.15 bit/px	0.25 bit/px	0.5 bit/px
JPEG (Wallace 1991)	0.802	0.844	0.945
JPEG 2000 (Rabbani and Joshi 2002)	0.903	0.922	0.951
Theis et al. (2017)	0.901	0.920	0.948
RNN-based (Toderici et al. 2017)	0.857	0.913	**0.952**
JPEG 420	0.824	0.891	0.950
BGAN+	**0.906**	**0.924**	0.949

Bold values indicate the best result

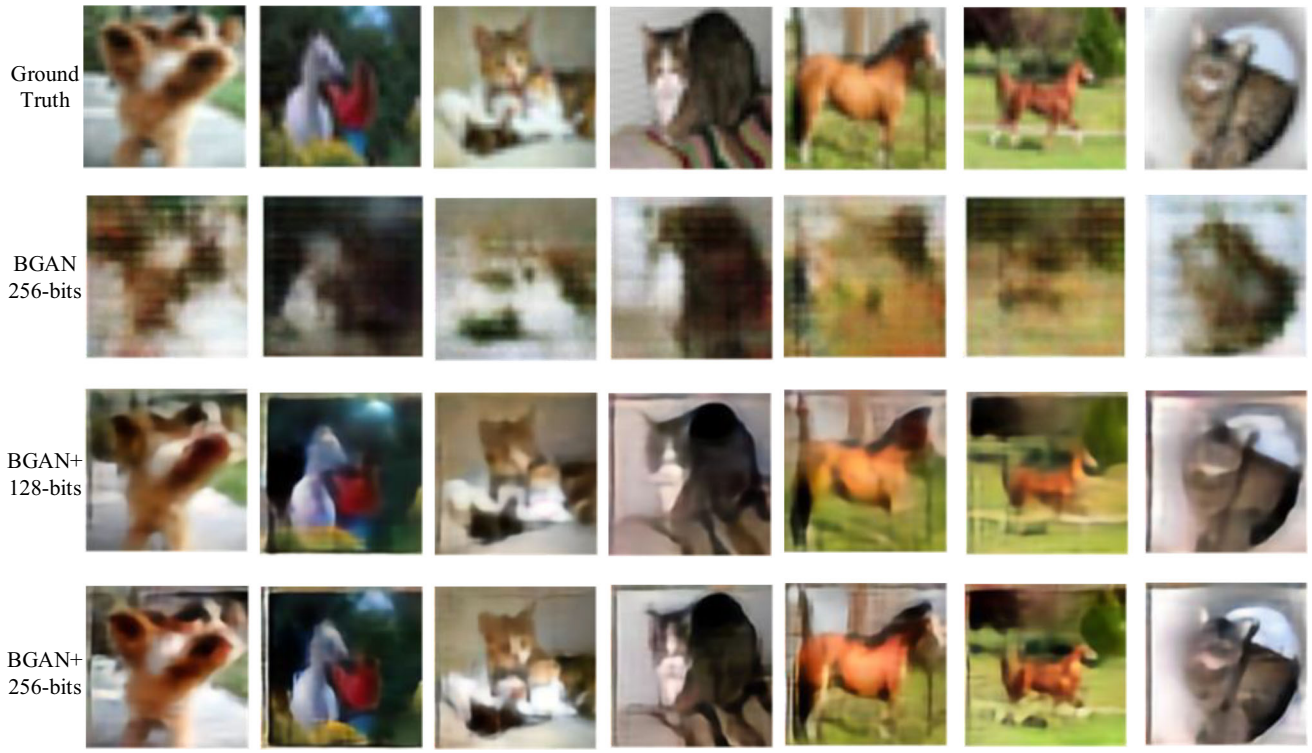


Fig. 6 Reconstructed images on CIFAR-10 using binary codes

NUS-WIDE and Kodak dataset respectively. In Fig. 7, each image is compressed by JPEG, JPEG 2000 and our BGAN+, and their corresponding MS-SSIM values are provided. The higher the MS-SSIM is, the better the compression results are. All the examples indicate that our BGAN+ performs the best. While for some examples (i.e., the third row), JPEG performs better than JPEG 2000. In terms of human visual visualization, the images generated by the JPEG are usually blurring, while BGAN+ and JPEG provide images with high-resolution. Figure 8 shows that our BGAN+ can reconstruct images with photo-realistic details.

5.4 Evaluation of Individual Component

To verify the effects of individual components (i.e., retrieval network and compression network) in our framework and show contribution of each part they made to the performance boost, we evaluate two variants of our approaches. Instead of using multi-task learning (MTL), we assume tasks are independent and learn retrieval network and compression network separately. The resulting retrieval model (ℓ_N) is acquired based on single task learning (STL) by utilizing ℓ_N loss only to train the retrieval network. In this way, STL trains its retrieval model separately without sharing the first five conv layers with compression network. For the retrieval task, the experiments are conducted on the CIFAR-10 dataset and the experimental results are shown in Table 10. The results by STL is worse than by MTL with a decrease of 3.2%, 3.1% and 1.9% over 24, 32 and 48-bit, respectively.

In the second experiment conducted on the NUS-WIDE dataset, we compare BGAN+ trained by MTL with BGAN+

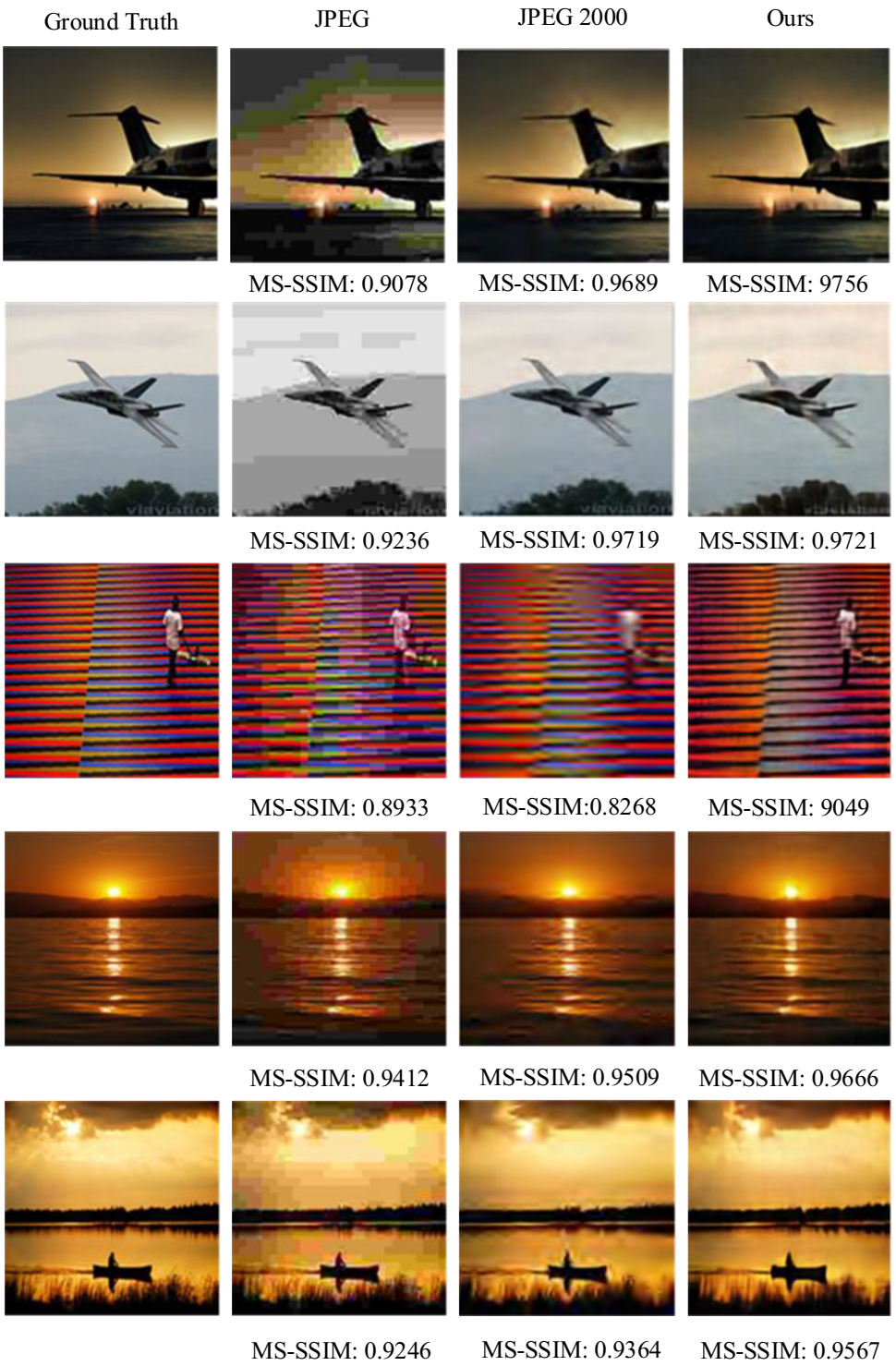


Fig. 7 Samples from NUS-WIDE dataset for visualization (0.15 bit/px)

trained by STL for compressing images and the results are shown in Table 11. The results show that MTL-BGAN+ outperforms STL-BGAN+ by 1.3%, 1.0% and 1.6% for 0.15, 0.25 and 0.5 bit/px respectively. These two experiments indicate that multi-task learning framework using the retrieval and compression network is beneficial for both image retrieval and compression tasks. This is due to the reason that learning-related tasks simultaneously can successfully exploit shared features among tasks and increase the discriminative ability of the learned models. As we can see from the experimental results, our method is able to

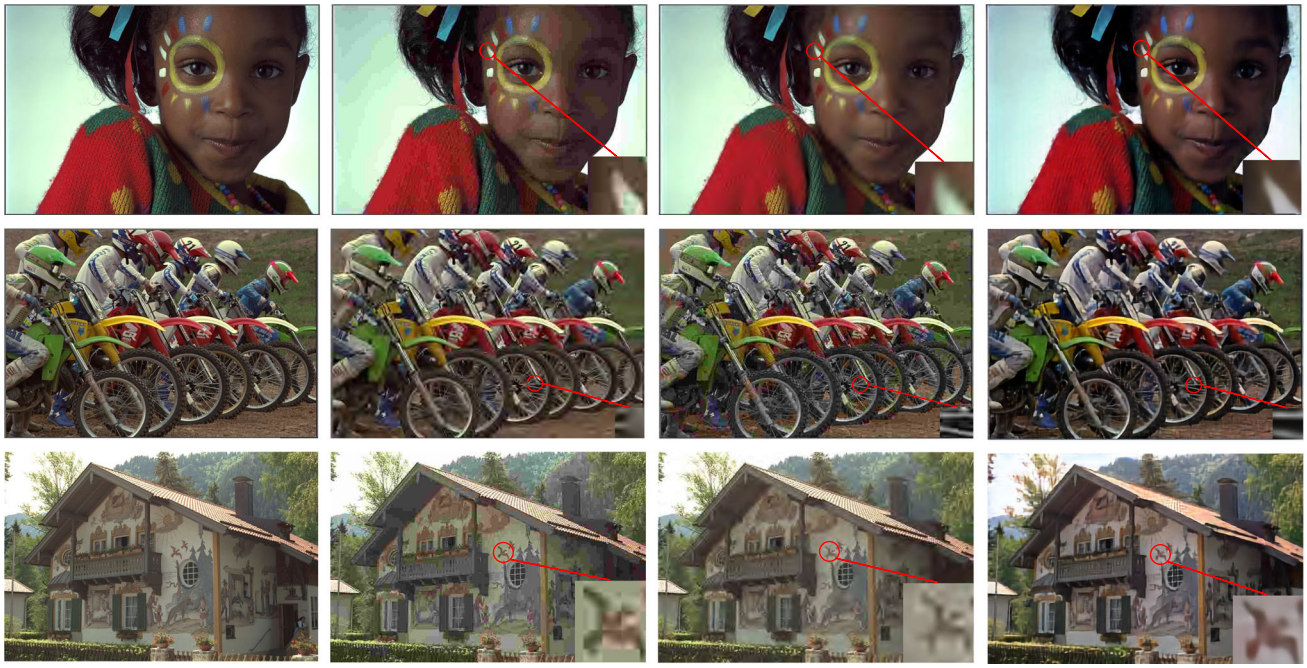


Fig. 8 Samples from Kodak dataset for visualization. Ground truth, JPEG, JPEG 2000 and Ours BGAN+ are from left to right (0.15 bit/pix)

Table 10 The mAP of BGAN+ on CIFAR-10 using different combinations of components (Without/with joint learning using compression net)

	mAP		
Components	24-bit	32-bit	48-bit
STL-BGAN+	0.511	0.533	0.567
MTL-BGAN+	**0.543**	**0.564**	**0.586**

Bold values indicate the best result

Table 11 MS-SSIM on NUS-WIDE at different bit-rate

	MS-SSIM		
Methods	0.15 bit/px	0.25 bit/px	0.5 bit/px
STL-BGAN	0.914	0.929	0.932
MTL-BGAN+	**0.927**	**0.939**	**0.948**

Bold values indicate the best result

Table 12 MS-SSIM on NUS-WIDE based on different dimensions with/without the layer of constant neuron

	MS-SSIM		
Methods	7k	12k	25k
Real value	**0.943**	**0.950**	**0.962**
BGAN+	0.926	0.937	0.948

Bold values indicate the best result

simultaneously generate binary codes for image retrieval and compression.

Table 13 MS-SSIM on NUS-WIDE under different compression rate by different reconstruction strategies

	MS-SSIM		
Methods	0.15 bit/px	0.25 bit/px	0.5 bit/px
Real value	0.801	0.824	0.856
Two steps	0.643	0.669	0.695
BGAN+	**0.927**	**0.939**	**0.948**

Bold values indicate the best result

Furthermore, we also test the efficiency of the constant neuron layer and conduct three different experiments as showing in Tables 12 and 13. To be specific, we test the quality of reconstructed images from real value vectors and binary codes, which is optimized by the two steps relaxation optimizing strategy. From Table 12, we can see that our results from binary codes do not degrade much compared with the real value compression codes. Obviously, in the condition of the same dimension, BGAN+ just declines 0.017, 0.013 and 0.014 on four dimensions 7k, 12k and 25k, respectively. It is worth noting that those results are based on the same compression dimension but not the same compression rate. It is reasonable that the compression strategy of binary codes is worse than the real value because the binary codes lose more information compared to the real value with the same dimension. Table 13 shows the results from the real value vector under the same compression rate. Here, we should note that the dimension of compression codes depends on the size of the input image, and set the input

image size as 224 × 224. However, when based on the same compression rate, our method shows a significant superiority to the real value codes and improves the compression performance about 0.126, 0.115 and 0.092 on the 0.15, 0.25 and 0.5 bit/px. The last experiment is to evaluate different strategies to generate binary codes. From Table 13, we can see that our constant neuron has a huge benefit to the binary learning compared to the two steps relaxation method and has about 0.283, 0.27 and 0.254 improvement on the three compression rates. The reason for this scenario is that two steps approximate strategy can lead to huge quantization errors. To sum up, through the three experiments, we can gain the following conclusions: (1) compared with the reconstruction from the real value vector, BGAN+ has the comparable performance but less storage cost. (2) The constant neuron layer can directly optimize binary codes and avoid large quantization errors, which is the key unit to ensure the high quality reconstructed image and low storage space.

6 Conclusion

In this paper, we propose a unified binary generative adversarial networks (BGAN+) to simultaneously convert images to binary codes for both image compression and retrieval in a multi-task fashion and an unsupervised way. By restricting the input noise variable of generative adversarial networks (GAN) to be binary and conditioned on the features of each input image, BGAN+ can simultaneously learn two binary representations per image: one for image retrieval and one for image compression. To equip the binary representation with the ability of accurate image retrieval and compression, we design a novel loss function. We also propose several solutions to address the gradient vanishing problem caused by $sign$ function. Extensive experiments are conducted for image retrieval and compression. The results show that our BGAN+ outperforms the existing retrieval methods with significant margins and achieves competitive performance for image compression, especially for low bit-rates. And the multi-task strategy is beneficial for both tasks. As far as we know, this is the first work of using binary codes for simultaneous image retrieval and image compression.

Acknowledgements This work is supported by the Fundamental Research Funds for the Central Universities (Grant No. ZYGX2019J073), the National Natural Science Foundation of China (Grant No. 61772116, No. 61872064, No.61632007, No. 61602049), The Open Project of Zhejiang Lab (Grant No.2019KD0AB05).

References

Agustsson, E., Mentzer, F., Tschannen, M., Cavigelli, L., Timofte, R., Benini, L., et al. (2017). Soft-to-hard vector quantization for end-to-end learned compression of images and neural networks. CoRR arXiv:1704.00648.

Argyriou, A., Evgeniou, T., & Pontil, M. (2007). Multi-task feature learning. In *Advances in neural information processing systems* (pp. 41–48).

Baig, M. H., Koltun, V., & Torresani, L. (2017). Learning to inpaint for image compression. In *NIPS* (pp. 1246–1255).

Ballé, J., Laparra, V., & Simoncelli, E. P. (2016). End-to-end optimized image compression. CoRR arXiv:1611.01704.

Bellard, M. (2017). *BPG image format*. http://bellard.org/bpg/. Retrieved January 30, 2017 (1, 2).

Brown, M., Hua, G., & Winder, S. (2010). Discriminative learning of local image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(1), 43–57.

Cao, Y., Liu, B., Long, M., & Wang, J. (2018). Hashgan: Deep learning to hash with pair conditional wasserstein GAN. In *CVPR* (pp. 1287–1296).

Cao, Y., Long, M., Wang, J., Zhu, H., & Wen, Q. (2016). Deep quantization network for efficient image retrieval. In *AAAI* (pp. 3457–3463).

Cao, Z., Long, M., Wang, J., & Yu, P. S. (2017). Hashnet: Deep learning to hash by continuation. In *ICCV* (pp. 5609–5618).

Chua, T. S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). Nus-wide: A real-world web image database from National University of Singapore. In *Proceedings of the ACM international conference on image and video retrieval* (p. 48). ACM.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML* (pp. 160–167). ACM.

Dai, B., Guo, R., Kumar, S., He, N., & Song, L. (2017). Stochastic generative hashing. In *ICML* (pp. 913–922).

Datar, M., Immorlica, N., Indyk, P., & Mirrokni, V. S. (2004). Locality-sensitive hashing scheme based on p-stable distributions. In *Symposium on computational geometry*.

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8599–8603). IEEE.

Do, T. T., Doan, A. D., & Cheung, N. M. (2016). Learning to hash with binary deep neural network. In *ECCV* (pp. 219–234). Berlin: Springer.

Duan, Y., Lu, J., Wang, Z., Feng, J., & Zhou, J. (2017). Learning deep binary descriptor with multi-quantization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1183–1192).

Farvardin, N. (1994). Review of 'vector quantization and signal compression' (gersho, a., and gray, r.m.; 1992). *IEEE Transactions Information Theory*, *40*(1), 287.

Franzen, R. (1999). Kodak lossless true color image suite (Vol. 4). http://r0k.us/graphics/kodak.

Gao, L., Li, X., Song, J., & Shen, H. T. (2019). Hierarchical LSTMs with adaptive attention for visual captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,. https://doi.org/10.1109/TPAMI.2019.2894139.

Ge, T., He, K., & Sun, J. (2014). Graph cuts for supervised binary coding. In *ECCV* (pp. 250–264).

Girshick, R. (2015). Fast r-cnn. In *Proceedings of The IEEE international conference on computer vision* (pp. 1440–1448).

Gong, Y., Kumar, S., Verma, V., & Lazebnik, S. (2012). Angular quantization-based binary codes for fast similarity search. In *NIPS* (pp. 1205–1213).

Gong, Y., Lazebnik, S., Gordo, A., & Perronnin, F. (2013). Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2916–2929.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *NIPS* (pp. 2672–2680).

Google. (2017). *Webp: Compression techniques*. http://developers.google.com/speed/webp/docs/compression. Retrieved January 30, 2017 (1, 2, 5).

Grubb, G. (2008). *Distributions and operators* (Vol. 252). Berlin: Springer.

Gu, Y., Ma, C., & Yang, J. (2016). Supervised recurrent hashing for large scale video retrieval. In *ACM multimedia* (pp. 272–276). ACM.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1904–1916.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).

He, T., Li, Y., Gao, L., Zhang, D., & Song, J. (2019). One network for multi-domains: Domain adaptive hashing with intersectant generative adversarial networks. In *IJCAI* (pp. 2477–2483).

Heo, J., Lee, Y., He, J., Chang, S., & Yoon, S. (2015). Spherical hashing: Binary code embedding with hyperspheres. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(11), 2304–2316.

Huiskes, M. J., & Lew, M. S. (2010). New trends and ideas in visual concept detection: The MIR Flickr retrieval evaluation initiative. In *MIR '10: Proceedings of the 2010 ACM international conference on multimedia information retrieval* (pp. 527–536).

Irie, G., Li, Z., Wu, X., & Chang, S. (2014). Locally linear hashing for extracting non-linear manifolds. In *CVPR* (pp. 2123–2130).

Jégou, H., Douze, M., & Schmid, C. (2011). Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(1), 117–128.

Jin, Z., Hu, Y., Lin, Y., Zhang, D., Lin, S., Cai, D., et al. (2013). Complementary projection hashing. In *ICCV* (pp. 257–264).

Khosla, A., Jayadevaprakash, N., Yao, B., & Fei-Fei, L. (2011). Novel dataset for fine-grained image categorization. In *First workshop on fine-grained visual categorization, IEEE conference on computer vision and pattern recognition*. Colorado Springs, CO.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. CoRR arXiv:1412.6980.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. CoRR arXiv:1312.6114.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. Master's thesis, University of Toronto.

Lai, H., Pan, Y., Liu, Y., & Yan, S. (2015). Simultaneous feature learning and hash coding with deep neural networks. In *CVPR* (pp. 3270–3278).

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *ICML* (pp. 1558–1566).

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR* (pp. 105–114).

Li, M., Zuo, W., Gu, S., Zhao, D., & Zhang, D. (2018). Learning convolutional networks for content-weighted image compression. In *CVPR*.

Li, W., Wang, S., & Kang, W. (2016). Feature learning based deep supervised hashing with pairwise labels. In *IJCAI* (pp. 1711–1717).

Li, X., Lin, G., Shen, C., van den Hengel, A., & Dick, A. R. (2013). Learning hash functions using column generation. In *ICML* (pp. 142–150).

Lin, G., Shen, C., Shi, Q., van den Hengel, A., & Suter, D. (2014). Fast supervised hashing with decision trees for high-dimensional data. In *CVPR* (pp. 1971–1978).

Lin, G., Shen, C., Suter, D., & Van Den Hengel, A. (2013). A general two-step approach to learning-based hashing. In *ICCV* (pp. 2552–2559). IEEE.

Lin, K., Lu, J., Chen, C., & Zhou, J. (2016). Learning compact binary descriptors with unsupervised deep neural networks. In *CVPR* (pp. 1183–1192).

Liong, V. E., Lu, J., Wang, G., Moulin, P., & Zhou, J. (2015). Deep hashing for compact binary codes learning. In *CVPR* (pp. 2475–2483).

Liu, H., Wang, R., Shan, S., & Chen, X. (2016). Deep supervised hashing for fast image retrieval. In *CVPR* (pp. 2064–2072).

Liu, L., Shao, L., Shen, F., & Yu, M. (2017). Discretely coding semantic rank orders for supervised image hashing. In *CVPR* (pp. 5140–5149).

Liu, W., Wang, J., Ji, R., Jiang, Y., & Chang, S. (2012). Supervised hashing with kernels. In *CVPR* (pp. 2074–2081).

Liu, X., He, J., Deng, C., & Lang, B. (2014). Collaborative hashing. In *CVPR* (pp. 2147–2154).

Long, M., & Wang, J. (2015). Learning multiple tasks with deep relationship networks (Vol. 2). arXiv preprint arXiv:1506.02117.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., & Feris, R. (2017). Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5334–5343).

Misra, I., Shrivastava, A., Gupta, A., & Hebert, M. (2016). Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3994–4003).

Nemirovski, A., Juditsky, A., Lan, G., & Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, *19*(4), 1574–1609.

Nilsback, M. & Zisserman, A. (2006). A visual vocabulary for flower classification. In *CVPR* (pp. 1447–1454).

Norouzi, M., & Blei, D. M. (2011). Minimal loss hashing for compact binary codes. In *ICML* (pp. 353–360).

Norouzi, M., & Fleet, D. J. (2013). Cartesian k-means. In *CVPR* (pp. 3017–3024).

Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts. *Distill*, *1*(10), e3.

Rabbani, M., & Joshi, R. L. (2002). An overview of the JPEG 2000 still image compression standard. *Signal Processing: Image Communication*, *17*(1), 3–48.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Rippel, O., & Bourdev, L. (2017). Real-time adaptive image compression. In *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 2922–2930). JMLR.org.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.

Shakhnarovich, G. (2005). Learning task-specific similarity. Ph.D. thesis, Massachusetts Institute of Technology.

Shannon, C. E. (2001). A mathematical theory of communication. *Mobile Computing and Communications Review*, *5*(1), 3–55.

Shen, F., Mu, Y., Yang, Y., Liu, W., Liu, L., Song, J., et al. (2017). Classification by retrieval: Binarizing data and classifiers. In *SIGIR* (pp. 595–604).

Shen, F., Shen, C., Liu, W., & Shen, H. T. (2015). Supervised discrete hashing. In *CVPR* (pp. 37–45).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. CoRR arXiv:1409.1556.

Song, J., He, T., Gao, L., Xu, X., Hanjalic, A., & Shen, H. T. (2018). Binary generative adversarial networks for image retrieval. In *AAAI*.

Song, J., Guo, Y., Gao, L., Li, X., Hanjalic, A., & Shen, H. T. (2019). From deterministic to generative: Multimodal stochastic RNNS for video captioning. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(10), 3047–3058.

Song, J., Yang, Y., Yang, Y., Huang, Z., & Shen, H. T. (2013). Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *SIGMOD* (pp. 785–796).

Strecha, C., Bronstein, A., Bronstein, M., & Fua, P. (2011). Ldahash: Improved matching with smaller descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(1), 66–78.

Theis, L., Shi, W., Cunningham, A., & Huszár, F. (2017). Lossy image compression with compressive autoencoders. In *ICLR*.

Toderici, G., O'Malley, S. M., Hwang, S. J., Vincent, D., Minnen, D., Baluja, S., et al. (2015). Variable rate image compression with recurrent neural networks. arXiv preprint arXiv:1511.06085.

Toderici, G., Vincent, D., Johnston, N., Hwang, S. J., Minnen, D., Shor, J., et al. (2016). Full resolution image compression with recurrent neural networks. CoRR arXiv:1608.05148.

Toderici, G., Vincent, D., Johnston, N., Hwang, S.J., Minnen, D., Shor, J., et al. (2017). Full resolution image compression with recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5306–5314).

Wallace, G. K. (1991). The JPEG still picture compression standard. *Communications of the ACM*, *34*(4), 30–44.

Wang, B., Yang, Y., Xu, X., Hanjalic, A., & Shen, H. T. (2017). Adversarial cross-modal retrieval. In *Proceedings of the 2017 ACM on multimedia conference* (pp. 154–162).

Wang, J., Kumar, S., & Chang, S. (2010). Sequential projection learning for hashing with compact codes. In *ICML* (pp. 1127–1134).

Wang, J., Liu, W., Sun, A. X., & Jiang, Y. (2013a). Learning hash codes with listwise supervision. In *ICCV* (pp. 3032–3039).

Wang, J., Wang, J., Yu, N., & Li, S. (2013b). Order preserving hashing for approximate nearest neighbor search. In *ACM multimedia* (pp. 133–142). ACM.

Wang, J., Zhang, T., Song, J., Sebe, N., & Shen, H. T. (2018). A survey on learning to hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 769–790.

Wang, X., Shi, Y., & Kitani, K. M. (2016). Deep supervised hashing with triplet labels. In *Asian conference on computer vision* (pp. 70–84). Berlin: Springer.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *Conference record of the thirty-seventh asilomar conference on signals, systems and computers, 2004* (Vol. 2, pp. 1398–1402).

Weiss, Y., Torralba, A., & Fergus, R. (2008). Spectral hashing. In *NIPS* (pp. 1753–1760).

Wintz, P. A. (1972). Transform picture coding. *Proceedings of the IEEE*, *60*(7), 809–820.

Xia, R., Pan, Y., Lai, H., Liu, C., & Yan, S. (2014a). Supervised hashing for image retrieval via image representation learning. In *AAAI* (pp. 2156–2162).

Xia, R., Pan, Y., Lai, H., Liu, C., & Yan, S. (2014b). Supervised hashing for image retrieval via image representation learning. In *AAAI* (Vol. 1, p. 2).

Yuan, X., Ren, L., Lu, J., & Zhou, J. (2018). Relaxation-free deep hashing via policy gradient. In *The European conference on computer vision (ECCV)*.

Zhang, P., Zhang, W., Li, W., & Guo, M. (2014). Supervised hashing with latent factor models. In *SIGIR* (pp. 173–182).

Zhao, F., Huang, Y., Wang, L., & Tan, T. (2015). Deep semantic ranking based hashing for multi-label image retrieval. In *CVPR* (pp. 1556–1564). IEEE.

Zhu, H., Long, M., Wang, J., & Cao, Y. (2016). Deep hashing network for efficient similarity retrieval. In *AAAI* (pp. 2415–2421).

Zieba, M., Semberecki, P., El-Gaaly, T., & Trzcinski, T. (2018). Bingan: Learning compact binary descriptors with a regularized GAN. In *Advances in neural information processing systems* (pp. 3608–3618).