IEEE TRANSACTIONS ON

CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY

A PUBLICATION OF THE IEEE CIRCUITS AND SYSTEMS SOCIETY

CAS
IEEE CIRCUITS AND SYSTEMS SOCIETY

WWW.IEEE-CAS.ORG

MARCH 2022 VOLUME 32 NUMBER 3 ITCTEM (ISSN 1051-8215)

REGULAR PAPERS

Image/Video Processing

(Contents Continued on Page 907)

IEEE

IEEE CIRCUITS AND SYSTEMS SOCIETY

The Circuits and Systems Society is an association of IEEE members with professional interest in the field of circuits and systems theory. For membership and subscription information and pricing, please visit www.ieee.org/membership-catalog.

IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY

Associate Editors

IEEE Officers

IEEE Executive Staff

IEEE Publishing Operations

IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (ISSN 1051-8215) is published monthly by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** To order individual copies for members and nonmembers, please email the IEEE Contact Center at contactcenter@ieee.org. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee of $31.00 is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright ©2022 by The Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage Paid at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE, 445 Hoes Lane, Piscataway, NJ 08854. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

Digital Object Identifier 10.1109/TCSVT.2022.3154898

(Contents Continued from Front Cover)

Image/Video Analysis and Computer Vision

(Contents Continued on Back Cover)

DAHP: Deep Attention-Guided Hashing With Pairwise Labels

Xue Li, Jiong Yu, Yongqiang Wang, Jia-Ying Chen, Peng-Xiao Chang, and Ziyang Li

***Abstract*— To address the problem of inadequate feature extraction and binary code discrete optimization faced by deep hashing methods using a relaxation-quantization strategy, a novel deep attention-guided hashing method with pairwise labels (DAHP) is proposed to enhance global feature fusion, better learn the contextual information of image features to effectively enhance the feature representation, and solve the problem of losing feature information in discrete optimization by optimizing the loss function. First, we introduce a new concept called the anchor hash code generation(AHCG) algorithm, we train the ResNet with the position attention and channel attention mechanisms with the anchor points in Hamming space as supervised information, we fit the binary code representing the picture to the vicinity of each anchor point, and finally, we use the optimized loss function to calculate the pairwise loss and the anchor loss, allowing the hash function to generate hash code with strong discriminative power. The experiments were conducted on four benchmark datasets, and the retrieval accuracy of the proposed method outperformed the retrieval accuracies of the state-of-the-art methods.**

***Index Terms*— Channel attention, deep hashing, image retrieval, position attention.**

I. Introduction

With the advancement of science and technology, the use of image data in networks has increased exponentially, and increasingly more people are keen on using images to express their thoughts in various scenarios such as work, study, and entertainment. The fast retrieval of semantically similar images from large-scale images has become a demand for information exchange in people's daily lives [1]. In addition, as content-based image retrieval technology [2]–[6] gradually replaces earlier text-based image retrieval technology [7]–[10], the limited time and computational resources will shift the challenge to optimize both the memory consumption and retrieval speed [11]. As the most widely used data retrieval technique [12]–[15], the popular hash retrieval methods [16]–[18] map high-dimensional images into fixed-length hashes and keep the similarity in the original space. Hash retrieval methods have obvious advantages in memory consumption and retrieval speed, so they are widely used in related industries such as public security systems, digital libraries and search engines [19], [20]. The current hash retrieval methods have shortcomings such as insufficient utilization of label semantics, slack discrete optimization of hash codes, and insufficient deep feature extraction of images [21]–[23]; these constraints limit the pace of development of large-scale image retrieval. In order to solve the above mentioned problems, this paper proposes a deep attention-guided hashing with pairwise labels (DAHP) image retrieval method, which, compared with the previous methods that simply used similar or dissimilar labels, is able to retrieve images based on position attention and channel attention mechanisms to deeply explore the intraclass connections between rich images and the correlations between semantic features to generate hash codes with strong discriminative power and uses a new loss function to effectively solve the binary code discrete optimization problem. This results in richer semantics expressed by the labels and more pronounced semantic loss during feature learning.

The remainder of this paper is arranged as follows. Section II introduces the application and development of hash methods in image retrieval technology. Section III introduces the deep hash method based on pairwise labels proposed in this paper in detail. Section IV analyzes the experimental results in detail. Finally, Section V summarizes the main innovations and the excellent results obtained in this paper.

II. Related Work

In order to achieve efficient retrieval of large-scale image data, the researchers proposed an approximate nearest neighbor search strategy. The hashing technique is the mainstream method of solving approximate nearest neighbor problem. such as locality sensitive hashing (LSH) [24], which uses random projection to partition the feature space and increase the probability that adjacent points in the original space are partitioned into the same bucket, is constrained by the number of images. Experiments have shown that the effectiveness of such data-independent methods is positively correlated with the length of the hash code. Scholars have proposed a series of data-dependent hash algorithms to generate more compact

Manuscript received January 13, 2021; revised March 3, 2021; accepted March 25, 2021. Date of publication March 31, 2021; date of current version March 9, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61462079, Grant 61562086, Grant 61562078, and Grant 61862060 and in part by the Science and Technology Support Project of Ministry of National Science and Technology of China under Grant 2015BAH02F01. This article was recommended by Associate Editor Q. Wang. *(Corresponding author: Xue Li.)*

Xue Li, Peng-Xiao Chang, and Ziyang Li are with the College of Software, Xinjiang University, Urumqi 830046, China (e-mail: 1547037202@qq.com; changpengxiao@stu.xju.edu.cn; liziyang@stu.xju.edu.cn).

Jiong Yu is with the School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China (e-mail: yujiong@xju.edu.cn).

Yongqiang Wang and Jia-Ying Chen are with the College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China (e-mail: 925968662@qq.com; chenjiaying@stu.xju.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3070129.

Digital Object Identifier 10.1109/TCSVT.2021.3070129

hash codes. These algorithms are divided into two main categories of supervised and unsupervised methods depending on whether the training data have an artificial label or not.

Spectral hashing (SH) [25] and iterative quantization hashing (ITQ) [26] are two of the more classical unsupervised methods. The former obtains the hash function by learning the similarity between pairs of samples in the original space; the latter uses principal component analysis, where the projected binary code is orthogonally transformed to the final obtained principal component, as the hash function.

Supervised discrete hashing (SDH) [27] and kernel supervised hashing (KSH) [28] are two of the more classical methods among supervised methods. The former uses the least squares regression and traditional fixed regression objectives with class labeling information, and the latter is a method that addresses the problem of the linear in distinguishability of data in the kernel space and learns nonlinear hash functions.

Although the above methods have some improvements in image retrieval accuracy, text labeling cannot describe the deep semantic information of images and is subjective. In order to overcome this dilemma, researchers have proposed using deep convolutional neural networks to learn both image feature representations and hash functions [29]–[31]. The AlexNet model proposed in 2012 [32] and the network in network (NIN) model and the deep VGG [33] model proposed in 2014 have successfully validated the extraordinary ability of deep convolutional neural network-based approaches to learn image feature representations. The earliest supervised deep hashing method is convolutional neural network hashing (CNNH) [32], which successfully combines a convolutional neural network and data similarity matrix for image retrieval, but this method is not an end-to-end method, matrix decomposition consumes a large amount of storage memory, and the method is not suitable for large-scale image datasets. To address this problem, researchers have proposed the network in network hashing (NINH) [34] method, which maps the input image from pixels to labels in an end-to-end manner, creating a network that enables simultaneous feature learning and hash coding learning with deep neural networks.

To further improve image retrieval accuracy, researchers have proposed several deep convolutional neural network-based optimizations. Deep binary fast hashing (FastH) [30] introduces new features by using a hidden layer to represent the underlying concept of class labels. Deep semantic ranking hashing (DSRH) [35] learns hash functions by preserving similar semantic information between multilabeled images. Supervised semantics-preserving hashing via deep neural networks (SSHD) [36] implements a deep hash algorithm based on sorting. The robustness, independence, and fast hashing matching algorithm [37] improves the speed dramatically by reducing the complexity using feature selection and binary search tree clustering (BST). The deep hashing network (DHN) [38] learns the associations between samples by constructing pairwise labels. Kanwal [39] uses symmetry, fast scores, shape-based filtering and spatial mapping integrated with a CNN for large scale image retrieval. S. Y. [40] and Z. Y. [41], combined neural networks, PCA and multiscale balancing algorithms to solve the inefficiency and time consumption problems. Deep supervised hashing (DSH) [42] uses two-layer loops without repetition to generate image pairs to learn deep features. To better reduce the negative effects of the quantization loss, deep pairwise-supervised hashing (DPSH) [43] and using deep learning to hash by continuation (HashNet) [44] explicitly penalize the quantification loss after relaxation using regularization terms through different strategies to represent the original relationship of sample pairs in Hamming space. Although the method of adding regularization terms can minimize the loss, it leads to a small gradient descent and slower convergence of the neural network. H. Z. [45] proposed an effective method called deep transfer hashing (DTH) that uses the knowledge from a teacher model as supervised information. H. L. [46] first introduced the MIH mechanism, which divides binary codes into multiple substrings, into their proposed deep architecture. Although these methods reduce the quantization error of the hash code to some extent, error still exists. However, Z. X. [47] proposed a deep balanced discrete hashing method (DBDH), which uses discrete gradient propagation with the straight-through estimator to reduce the quantization error caused by continuous relaxation. The performance of DBDH reaches the level of SOTA. In recent years, attentional mechanisms in neural networks have played a major feature supervisory role, and they have a wide range of applications in different fields. Z. Y and L. J. [48], [49] have already obtained hash codes with strong representation capabilities to improve retrieval accuracy by combining attentional modules and convolutional neural networks. Q. W. [50] explored the attention mechanism and proposed an end-to-end attention recurrent convolutional network (ARCNet) for scene classification. In addition, in the next year, he proposed an adaptive dark-light-dark (ADLD) [51] method for low-level feature extraction, and the proposed ADLD features were used as spatial attention information in a multitask network. C. H. [52] proposed the complementation-reinforced attention network (CRAN) for person reidentification that encouraged each branch to attend to complementary attention regions and enforced orthogonality among the learned features of different regions in the embedding space. L. J. and G. L. [49], [53] proposed learning the rank correlation space by exploiting the local spatial information from a fully convolutional network and the global semantic information from a convolutional neural network simultaneously. However, there is less research on improving the feature representation in image retrieval using an attention mechanism, which allows neural networks to generate hash codes with more discrimination.

a) It cannot break the relaxation-quantization limit and cannot guarantee that the real values requantized after relaxation are still optimal.

b) The loss function employed basically translates the loss of the discrete optimization process directly into the regularization direction, thereby forcing the loss close to the boundary value of the interval and resulting in a small descent gradient of the network and slow convergence of the results.

c) In the process of feature extraction, the convolution operation used in network structure can not take into account all the high-frequency information inevitably, resulting in the loss of some important information or the low

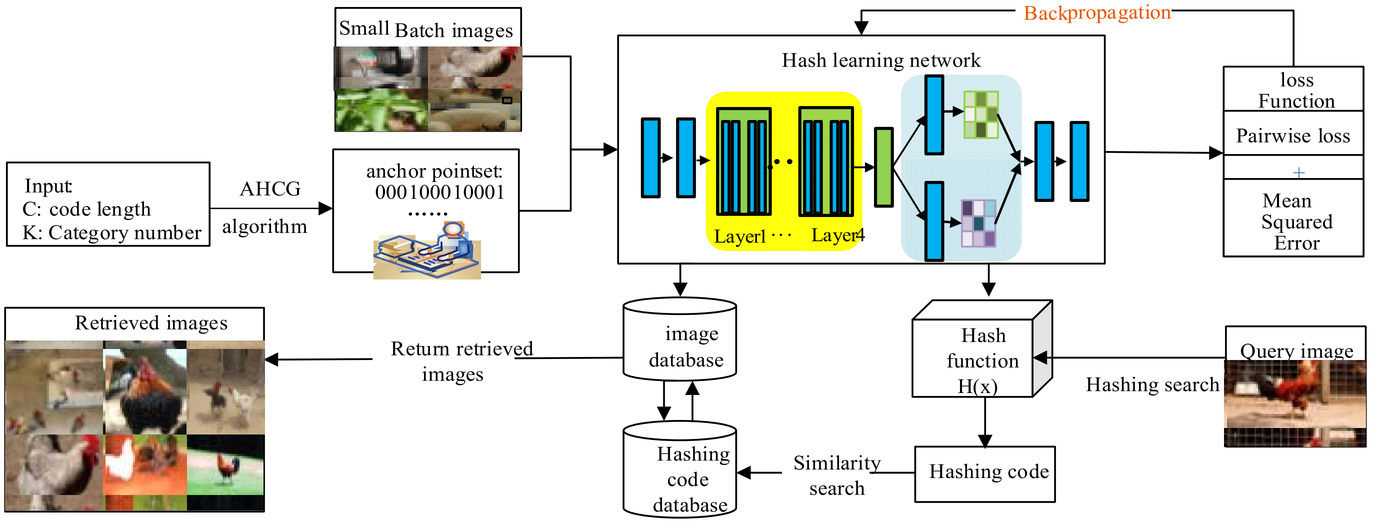


Fig. 1. The image retrieval workflow, which is divided into a training process and a retrieval process, is shown in the figure. A stable DAHP model is obtained in the training process, and the retrieval process is used to output the results.

proportion of some representative features, which makes the final generated hash code have limited discrimination ability.

To solve the above problems, this paper proposes the deep attention-guided hashing with pairwise labels (DAHP) method based on the attention mechanism. The method directly sets some anchor points in Hamming space and constrains the more similar images to the anchor points closer to the anchor point and the less similar images to the anchor points farther away from the anchor point. The optimal hash code is fit by training the neural network model to avoid the inherent drawback of slack quantization. The advantages of DAHP over other related methods are primarily in three aspects:

a) DAHP breaks the original fixed optimization model of relaxation-quantization and introduces anchor algorithm (AHCG) to measure the distance between similar samples directly in Hamming space instead of relaxing into Euclidean space to avoid the semantic loss caused by secondary quantization.

b) DAHP uses the pairwise and mean squared error losses to calculate classification and anchor point errors. When both losses are considered, the image semantic loss is more obvious, the network output is closer to the anchor hash code, and the neural network converges faster, which is more practical and applicable than considering only one loss alone or adding regularization terms.

c) The DAHP method combines position attention and channel attention mechanisms to adaptively integrate local features and global dependencies and is an end-to-end method for simultaneous feature learning and hash code learning. This process can increase the weights of important features, prevent the loss of important features, improve the feature representation, and make the final generated hash code have high discrimination.

The results show that compared with the current advanced image retrieval methods, the proposed method can effectively improve the image retrieval accuracy on four benchmark large-scale image datasets.

III. Deep Attention-Guided Hashing With Pairwise Labels

In this section, we first describe the framework of our proposed method; then, we show the details of our proposed method, including the problem equation, anchor point hash code generation method, and loss function training strategy. Finally, we show the hash function learning process.

A. Retrieval Workflow

To effectively overcome the shortcomings of the relaxation-quantization model, a new deep hash method based on the attention mechanism is proposed. We design a deep residual network based on ResNet18 incorporating position attention and channel attention mechanisms while learning sample features and hash functions to improve the quality of the hash code through continuous iterative fitting. The entire retrieval workflow is shown in Fig. 1. First, the anchor hash code generation(AHCG) algorithm is used to obtain the hash code representing the anchor point for each node in the algorithm. Then, the binary code representing the picture is fit to the vicinity of each anchor point by training a deep residual network. The network adopts the ResNet18 framework, learns the image features from the image label information, and is fine-tuned for the parameters of the network model by using the anchor points in Hamming space as supervised information during the fitting process. The pairwise error and mean squared error loss are used to calculate the pairwise loss and the anchor loss, respectively. Finally, the trained network is the hash function of this method. A small set of images is used as the input to the DAHP model, and the output of the model is quantified using the above method to obtain the hash code

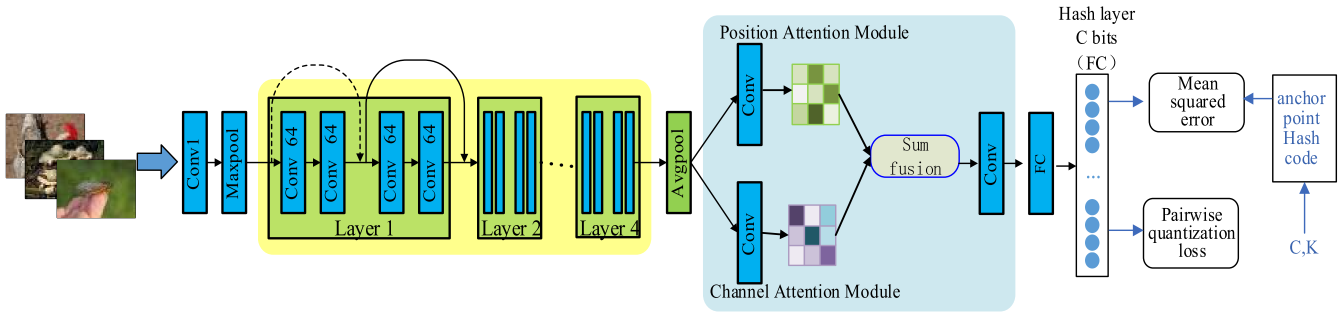


Fig. 2. The network structure of DAHP. The structure can be divided into three parts: the residual network ResNet18, the parallel dual attention mechanism module and two types of loss sum modules.

representing the images, which is then quickly retrieved using common methods such as Hamming distance sorting or hash table lookup.

B. Problem Equationtion

Suppose Ω represents the RGB color image space and assume that k classes of n images are randomly selected from Ω to form a set, which is used as the training set. The strategy of the hash method is to create a transformation $f : \Omega \rightarrow \{0, 1\}^c$ from the image space Ω to a c-bit hash code on the training set X. The transformation process does not disturb the intrarelationship between the images, and the transformed hash code still retains the similarity between the original images. After joining the neural network, this transformation can be decomposed into $f(x_i) = h(\varphi(x_i; \theta))$, where $x_i \in \Omega$ denotes a sample in the image space, and θ denotes all the parameters in the deep convolutional neural network. $\varphi(x_i; \theta) : X \rightarrow feature^d$ denotes the process of extracting d-dimensional image features from the training set X, and $h(x_i) : feature^d \rightarrow \{0, 1\}^c$ denotes the process of fitting high-dimensional features to low-dimensional hash codes. At the end of these two processes, a collection of hash code strings will be obtained, where the more similar the images are, the smaller the Hamming distance between the hash codes, and the higher the quality of the hash code obtained, and vice versa. Briefly, the strategy can be expressed as $h(\varphi(x_i; \theta)) = L(w^T \varphi(x_i; \theta) + v)$, where $w \in R^{4096\times c}$ represents the weight matrix of each neuron in the neural network, $v \in R^c$ represents the partiality vector in the neural network, and the L() function represents the loss function between the output of the hash layer and the anchor point. To simply the understanding, the set of anchor point hash codes is denoted by $M = \{m_i\}_{i=1}^{k}$ below, where m_i denotes the class i anchor point hash code. In this method, the parameters to be learned are wv and θ.

C. Network Architecture

The proposed network structure is shown in Fig. 2. Using the ResNet18 structure, the last layer of the ResNet18 network is the softmax classification layer, which is classifies results as 0 or 1 according to the sigmoid function. The last layer is now converted to a hash layer that represents the output of the hash code predicted by the neural network. The use of the ReLU as an activation function for RNNs in ResNet18 solves the vanishing gradient problem when the network is deeper and speeds up the training. We incorporate a dual attention mechanism to integrate local features and global dependencies adaptively. Using the traditional ResNet18 as the backbone network, the downsampling operation is removed, and the void convolution is used before the last convolution block. The final feature map is 1/8 the size of the input image. The resulting feature map is then fed into two parallel attention modules to simulate semantic interdependencies in the spatial and channel dimensions, respectively. Finally, the output features of the two attention modules are aggregated to obtain a better pixel-level predicted feature representation, which is represented by two parallel attention mechanisms (position attention and channel attention). The position attention module selectively aggregates features at each location using a weighted sum of the features at all locations, and similar features are correlated with each other, regardless of distance; the channel attention module selectively highlights the presence of interdependent channel mappings by integrating correlated features between all channel mappings. This network design further improves the feature representation by correlating the outputs of the two attention modules, which contributes to more accurate retrieval results. The detailed configuration is shown in Table I. The method contains 18 layers with weights, including 17 convolutional layers (Conv1+Layers 1-4) and 2 fully connected layers (FC1-2), where FC2 is the hash layer.

TABLE I

CONFIGURATION OF FEATURE LEARNING PART OF DAHP

Layer	Configuration
Conv1	{Conv, 64x56x56, k=3, s=1, p=1, ReLU}*4
Maxpool	{Conv, 128x28x28, k=3, s=2, p=1, ReLU}*4
Layer 2	{Conv, 256x14x14, k=3, s=2, p=1, ReLU}*4
Layer 3	{Conv, 512x7x7, k=3, s=2, p=1, ReLU}*4
Layer 4	{Conv, 512x7x7, k=3, s=2, p=1, ReLU}*4
Avgpool	512x1x1
FC1	512x1000
FC2	512x1000

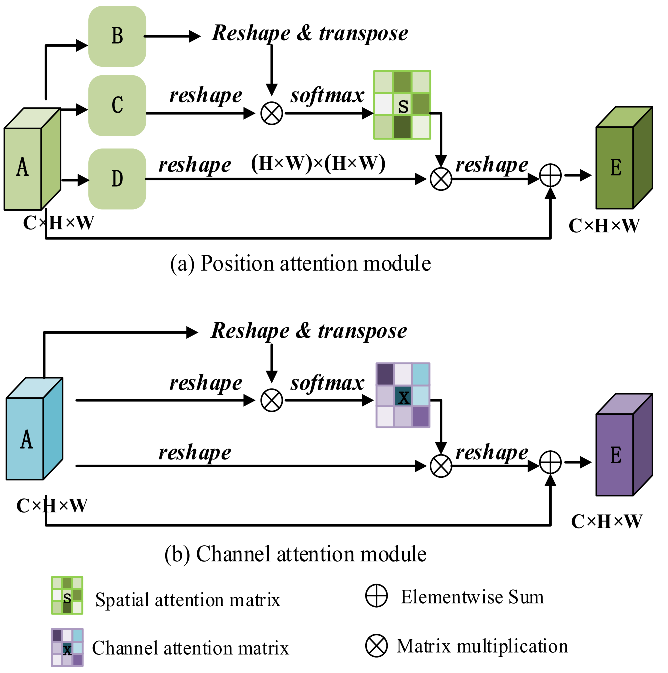


Fig. 3. The dual attention network is composed of two parts: (a) the position attention module and (b) the channel attention module.

The convolutional layer consists of a set of filters which can be regarded as a two-dimensional digital matrix. Each node of the full connection layer is connected to all nodes of the previous layer, which is used to synthesize the features extracted from the previous layer. A neuron in one of the fully connected layers can be regarded as a polynomial. Many neurons are used to fit the data distribution, but sometimes the nonlinear problem cannot be solved with only one fully connected layer. Therefore, the mapping between features and hash codes can be well solved with two fully connected layers.

D. Dual Attention Module

The dual attention network is composed of two parts: a position attention module and a channel attention module. The detailed structure is shown in Fig. 3. Fig. 3(a) represents the detailed structure of the position attention module. Since the characteristics generated by traditional FCNs will limit the learning ability of ResNet18, the position attention module is introduced into the network to establish a rich contextual relationship on local features and encode broader contextual information into local features, thus enhancing their representation ability.

In Fig. 3(a), the original value of E is the location of each feature in each position of the weighted summation. In the first step, the characteristics of figure A (C × H × W) were obtained by the three convolution layers B, C, and D. Then, B, C, and D were reshaped to C×N, where N = H×W. Then, in the second step, the transpose of B (N × C) and C (C × N) was reshaped after the reshape multiplication and then passed through the softmax spatial attention map S (N × N). In the last step, D (C × N) is reshaped and S (N × N) is transposed using matrix multiplication. Then, they are multiplied by the scale coefficient α and reshaped again to return to their original shapes. Finally, the final output of the elementwise summation of A is E.

The calculation of the spatial attention map is as follows:

$$S_{ji} = \frac{\exp(B_i \cdot C_j)}{\sum\limits_{i=1}^{N} \exp(B_i \cdot C_j)}, S \in R^{N\times N} \tag{1}$$

S_{ji} is used to measure the influence of the ith position on the jth position, that is, the degree of correlation between the ith position and the jth position. The deeper the degree of correlation between the two positions, the more similar they are.

$$E_{\rm j} = \alpha \sum_{i=1}^{N} (S_{ji} D_i) + A_j, E \in R^{C\times H\times W} \tag{2}$$

α is initialized to 0 and gradually learns to gain more weight. From the above equation, it can be directly seen that each location is integrated with the information of other locations. Therefore, it has a global context view and can selectively aggregate context based on spatial attention.

Fig. 3(b) shows the detailed structure of the channel attention module. Since the channel graph of each high-level feature can be regarded as a class-specific response, the interdependent feature graph can be highlighted, and the feature representation of specific semantics in the image can be improved by mining the interdependence relationship between channel graphs. Therefore, a channel attention module is created in the network to explicitly model the dependencies between channels. In Fig. 3(b), E is the ultimate characteristic of each channel of all channels and the original characteristics of the weighted summation. In the first step, to reshape A (C×N) and reshape & transpose A (N×C), which will be the two characteristics of diagram multiplication, the channel is obtained the by softmax attention map X (C×C). In the second step, the transpose of X (C × C) and the reshape of A (C×N) are conducted via matrix multiplication, and then they are multiplied by the scale coefficient of β to reshape them back to the original shape. In the final step, the elementwise summation of A obtains the final output E.

The calculation of the channel attention map is as follows:

$$X_{ji} = \frac{\exp(A_i \cdot A_j)}{\sum\limits_{i=1}^{C} \exp(A_i \cdot A_j)}, X \in R^{C\times C} \tag{3}$$

X_{ji} is used to measure the effect of the ith channel on the jth channel, namely, the correlation between the ith and jth positions; the correlation and dependency are positively correlated.

$$E_{\rm j} = \beta \sum_{i=1}^{C} (X_{ji} A_i) + A_j, E \in R^{C\times H\times W} \tag{4}$$

$\boldsymbol{\beta}$ is initialized to 0 and gradually learns to assign more weight to it. The above equation directly shows that the resulting feature E of each channel is the weighted sum of all channel features and original features. The semantic

Algorithm 1 AHCG: Learning Algorithm for Anchor Hash Code Generation

```
Input:
generate_code: c bit, Hamming distance: H, category: k.
Output:
Hamming distance: H, M {anchor hash code_set}.
Initialization: Initialize $c = [12, 24, 36, 48]$,
$H = 6$, $k = 10$, $j = 0^c$.
REPEAT
Establish a set $M = \{0 \leq m_i \leq 2^c |$ binary-coded format $m_i\}$;
for each sampled point $m_i$, perform the following operations:
Calculate   $m_i$^j;
Count       the number of '1's in the result of $m_i$^j
            according to (5);
Compute     Count('1') $\geq$ H, leave $m_i$ in M;
Update      the parameters M $\{m_i\}$;
UNTIL       the end of $2^c$ iterations
```

dependency relationship between feature graphs is modeled, which is helpful to improve feature discrimination.

In order to make full use of the contextual information of image features, the features of these two attention modules are aggregated. In other words, the convolutional layer is used to transform the output of two attention modules, and elementwise summation is performed to realize feature fusion. Finally, the convolution is followed by the final prediction characteristic diagram. The attention module is simple and can be plugged directly into the current FCN. This does not add too many parameters, and it effectively enhances the feature representation.

E. Anchor Hash Code Generation

An anchor hash code is a set of k binary codes that satisfy the conditions. The coding condition requires that the length of the code can be controlled by parameters and that the distance of each code in the set is maximized. The role of the anchor hash is to serve as supervisory information in hash learning so that the original image can be transformed into a binary code close to the anchor hash, thus improving the discriminatory power of the hash code. Given the original image training set $X = \{x_i\}_{i=1}^{n}$, containing n images divided into k classes, set the number of hash codes in the anchor hash code set as k and the code length as c bits, denoted by $M = \{m_i\}_{i=1}^{k}, m_i \in \{0, 1\}^c$. When the Hamming distance between m_i and $m_j (i \neq j \langle k)$ is the maximum, the set of anchor hash codes is a subset of the M set. The anchor hash code is obtained in three steps. First, the code length c and the minimum Hamming distance H are defined. Then, the binary code with the smallest value of c bits is listed. According to algorithm I of AHCG, the binary code with Hamming distance H on the basis of the smallest binary code is placed in set M, and this process iteratively repeats until the number of values in set M is greater than k, $H + 1$. Finally, the above operations are repeated until an appropriate H that satisfies the condition that the number of sets M is less than k is reached. Then, H-1 is the maximum Hamming distance that the problem is seeking, leaving k random binary codes from the corresponding M sets as the set of anchor point hash codes that resolves the problem.

In set M, the number of bits with different values of code bits on the corresponding bits between any two code words is defined as the Hamming distance between these two code words, and d() is used to represent the Hamming distance between two hash codes. The mathematical equation is as follows:

$$\mathrm{d}(m_k, m_j) = \sum m_k[i] \oplus m_j[i] \tag{5}$$

where i = {0, 1,…,n-1}, both m_k and m_j are n-bit codes, and $\oplus$ represents the XOR. $M = \{m_i\}_{i=1}^{k}$ and $m_i \in \{0, 1\}^c$ can be obtained according to algorithm I of AHCG.

Any two code words in the set meet the following equation:

$$\mathrm{d}(m_\mathrm{i}, m_j) = \mathrm{d}(m_j, m_k) = \cdots = \mathrm{d}(m_k, m_{n-1}) \tag{6}$$

According to the final hash code set $M = \{m_i\}_{i=1}^{k}$ obtained by algorithm I, AHCG contains multiple binary codes, and it is verified by equation (6) that any subset composed of the same number of elements in set M is selected, and the Hamming distance between elements is consistent. For example, if we want to represent 10 classes of images in the CIFAR-10 dataset with anchor points, we need to construct 10 independent and uniformly distributed hash codes in Hamming space. When setting $c = 12$ and $H = 6$, 16 different 12-bit hash codes can be calculated according to the AHCG algorithm, from which 10 hash codes can be selected to represent each class of images and used as anchor points in Hamming space.

F. Loss Function Optimization

When training deep convolutional neural networks, the role of the anchor hash code is to supervise the network to generate strong discriminative hash codes. Thus, to make the neural network become increasingly more intelligent in training, we need an excellent loss function to guide the network to reduce the error so that each part of the neural network in the error feedback promotes the work to achieve the ideal state. Therefore, it is crucial to design a loss function that can reduce the Hamming distance of similar instance hash codes and increase the Hamming distance of dissimilar instance hash codes.

The loss function is used to express the degree of difference between the prediction and the actual data, which can be used to measure the good or bad prediction performance of the model. The smaller the difference is, the more accurate the prediction results of the model. In the training of the deep convolutional neural network, we use the anchor hash code as supervisory information. The role of the loss function is to let the network learn to obtain the hash code with the anchor point as the center of the distribution in its vicinity as much as possible. Each anchor point and its surrounding hash code form a class of high cohesion, and there is low coupling between classes of the distribution. To achieve this purpose, the function of the loss function is completed. To summarize, there are two constraints as follows: a). We want to keep decreasing the Hamming distance of similar instance hash codes and keep increasing the Hamming distance of dissimilar instance hash codes. b). We want to continuously reduce the distance from the instance hash code to the corresponding anchor point and incorporate the supervisory role of the anchor point. The former is called the pairwise loss, and the latter is called the anchor loss. The specific design is as follows.

Assuming that the hash codes of the images in the training set X are output by the last layer of the neural network and transposed and represented as a binary matrix $B^T = \{b_i\}_{i=1}^n, b_i \in \{0,1\}^c$, for the semantic labeling information given in the training set, the matrix of pairwise label information $S = \{s_{ij}\}$, $s_{ij} \in \{0,1\}$ can be obtained, and the possibility of pairwise labeling matrix S is defined according to the LFH method as:

$$P\left(s_{ij}|B\right) = \begin{cases} \sigma(\pi_{ij}), & s_{ij=1} \\ 1-\sigma(\pi_{ij}), & s_{ij=0} \end{cases} \quad (7)$$

where $\pi_{ij} = \frac{1}{2}b_i^T b_j, \sigma(\pi_{ij}) = \frac{1}{1+e^{-\pi_{ij}}}$.

$\sigma(\pi_{ij})$denotes the degree of similarity between x_i and x_j predicted by the network, $s_{ij} = 0$ implies that the two image samples x_i and x_j are labeled differently, $P(s_{ij}|B)$ denotes how likely it is that x_i and x_j are not similar, $s_{ij} = 1$ implies that x_i and x_j are labeled identically and $P(s_{ij}|B)$ denotes how likely it is that x_i and x_j are similar. Based on the negative log-likelihood of the above function, the loss function is rewritten as follows:

$$L(P(s_{ij}|B)) = -\log P(s_{ij}|B) + \lambda(\frac{1}{n}\sum_{i=1}^{n}(m_i - b_i)^2) \quad (8)$$

where λ denotes the weights.

The loss function consists of two parts. The first part represents the pairwise loss, which uses the label information to measure the Hamming distance between sample pairs. If the pairwise labels are consistent, then the Hamming distance of the sample pair is reduced as much as possible. If the pair labels are not consistent, then the Hamming distance of the sample pair is increased as much as possible. The second part represents the anchor loss, which is used to measure the Hamming distance between the hash codes output by the neural network and the anchor point. First, we determine the anchor position of the sample label and then keep decreasing the Hamming distance between the hash code of the sample and the corresponding anchor hash code so that the spatial position of the sample keeps moving closer to the anchor. The purpose of the loss function is to make the loss value as close as possible to the minimum, that is, to make the output of the neural network as close as possible to the anchor hash code, maintain the original spatial similarity, and finally generate a hash code with high discriminative power. Substituting (7) into (8), the derivation is as follows:

$$\begin{aligned}\min_{B,M} L^1 &= -\log P(s_{ij}|B) + \lambda(\frac{1}{n}\sum_{i=1}^{n}(m_i - b_i)^2) \\ &= -\sum_{s_{ij}\in S}(s_{ij}\pi_{ij} - \log(1+e^{\pi_{ij}}) + \lambda(\frac{1}{n}\sum_{i=1}^{n}(m_i - b_i)^2)\end{aligned} \quad (9)$$

It is not difficult to find that when x_i and x_j are more similar, the larger π_{ij} is, the more likely it is that the network predicts that the labels of the two samples are similar, and the smaller the pairwise loss. The interanchor loss is derived by calculating the average of the sum of squares of the Hamming distances between m_i and b_i, which is called the mean squared error loss and reflects the degree of difference between the predicted value b_i and the true value m_i. The smaller the sum of the pairwise loss and the mean-squared error loss is, the better the trained neural network model predicts the true value, and the higher the accuracy of the model in the retrieval process. Thus, it seems that equation (9) can be satisfied so that the similarity between the hash codes obtained by transforming similar images still remains similar. The similarity is measured in Hamming space using the Hamming distance; the closer the similar hash codes in the Hamming space are, the smaller the Hamming distance is. The goal of the loss function optimization network model is better achieved, and adjusting the parameters has a desirable effect on the pretrained neural network. The final form of the loss function is organized as follows:

$$\begin{aligned}\min_{B,M} L^2 = &-\sum_{s_{ij}\in S}(s_{ij}\pi_{ij} - \log(1+e^{\pi_{ij}})) \\ &+\lambda(\frac{1}{n}\sum_{i=1}^{n}(m_i - b_i)^2)\end{aligned} \quad (10)$$

where the weight is set as $\lambda = 0.05$ using the experimental data on the hyperparameter as the selection criterion, as shown in Table VI.

In existing work, the optimization strategy for the loss function is to convert the binary code from a discrete state to a continuous state by relaxing to a real-valued matrix, which seriously affects the performance of the algorithm. However, the use of the anchor hash code in (10) is a good way to circumvent the discrete optimization problem, and the loss can be calculated directly in Hamming space. In contrast, the loss function proposed in this paper is more conducive to the learning of the hash function, and the final hash code obtained is more discriminating.

G. Hash Function Learning

The abovementioned loss function only works in the sample training process, which makes the neural network output satisfactory results; however, the image retrieval process is different from the tedious training process, and the purpose of retrieval is to efficiently complete the user task of searching images using images. Therefore, it is necessary to make the deep convolutional neural network after training able to achieve the role of the hash function, which can transform the image after functional transformation into a hash code with a high discriminative ability. To obtain the hash function for encoding, θ denotes all the parameters of the seven layers of the feature learning part, x_i denotes the input of the network, b_i denotes the output of the network, $\varphi(x_i;\theta)$ denotes the image features of the output of the 7 full layers, $w \in R^{4096\times c}$ denotes the weight matrix, and v $\in R^c$ is a partiality vector. The two parts are connected to a framework by a fully connected layer. A equation containing weight matrix W and bias vector V is used to conduct data transfer:

$$b_i = w^T \varphi\,(x_i;\theta) + \upsilon \tag{11}$$

Therefore, substituting (11) into (10) can be written as follows:

$$\min_{B,U} L = -\sum_{s_{ij}\in S}(s_{ij}\pi_{ij} - \log(1+e^{\pi_{ij}})) \\ + \lambda(\frac{1}{n}\sum_{i=1}^{n}(m_i - (w^T\varphi(x_i;\theta)+\upsilon)^2) \tag{12}$$

Using control variables for learning, the parameters that need to be constantly adjusted are $w\nu$ and θ. Optimizing one parameter while keeping the other parameters unchanged and alternately optimizing other parameters is a strategy that can effectively adjust the model parameters. First, the derivative of loss function L to b_i is computed, given that

$$\begin{aligned} a_{ij} &= \sigma(\frac{1}{2}b_i^T b_j) = \frac{1}{1+e^{-\frac{1}{2}b_i^T b_j}} : \\ \frac{\partial l}{\partial b_i} &= \frac{1}{2}\sum_{l:s_{ij}\in S}(a_{ij}-s_{ij})b_j \\ &\quad + \frac{1}{2}\sum_{l:s_{ji}\in S}(a_{ji}-s_{ji})b_j + 2\lambda(b_i - m_i) \end{aligned} \tag{13}$$

Then, according to the chain rule, we update the parameters $w\nu$ and θ using b_i:

$$\frac{\partial l}{\partial w} = \varphi(x_i;\theta)(\frac{\partial l}{\partial b_i})^T \tag{14}$$

$$\frac{\partial l}{\partial \upsilon} = \frac{\partial l}{\partial b_i} \tag{15}$$

$$\frac{\partial l}{\varphi(x_i;\theta)} = w\frac{\partial l}{\partial b_i} \tag{16}$$

Finally, each parameter in this neural network can be optimized by the standard backpropagation algorithm. After the training of the neural network is completed, the model is used as a hash function for the DAHP method. In the image retrieval process, the optimized hash function can efficiently generate the sample image with discriminative hash code, perform a similar lookup operation with the hash code in the training image hash code database, and finally sort the output result according to the Hamming distance.

IV. Experiments

In this section, four benchmark datasets, CIFAR-10, ImageNet-100, Flickr25K and NUS-WIDE, are used to verify the excellent performance of the DAHP method in image retrieval based on the mean average accuracy metric. In addition, comparison experiments are conducted with eight typical image retrieval methods based on three other commonly used evaluation metrics, and all the experimental results show that the DAHP method proposed in this paper outperforms the current mainstream methods.

TABLE II
Configuration Information

Item	Parameters
Operating System	Windows Server 2012 R2 Standard
CPU	Intel(R) Xeon(R) Gold 5117 CPU @ 2.00 GHz
Memory	256 GB DDR3
GPU	Tesla V100

A. Experiments Environment and Datasets

The experiments use a pretrained model, and PyTorch is used as the environment for adjusting the parameters of the model. A total of 128 training images are initially input to the network in each iteration, pairwise images are used as the input of the hash function via pairing, and the stochastic gradient descent (SGD) optimization model is used. The learning rate of SGD was 0.04, and the decay rate of the learning rate was 10^{-7}. The experimental process is accelerated using a single GPU, and the detailed experimental machine configuration information is shown in Table II.

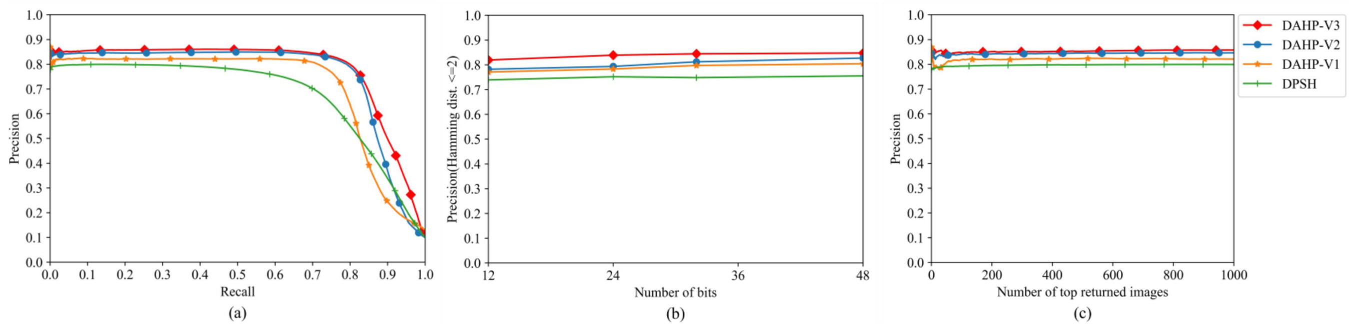
In order to make the experiments fair, the four most commonly used datasets in the field of image retrieval were selected for the experiments.

CIFAR-10: The CIFAR-10 image dataset has a total of 80,000 small images. Of these images, 6,000 images are randomly selected from 10 classes to form a 60,000 32 × 32 color image dataset, and each image belongs to only one class. In the experiment, 100 images (1,000 images in total) are randomly selected from each class as the test set, and 500 images (5,000 images in total) are selected from each class are used as the training set. This paper uses a 512-dimensional GIST descriptor to represent the images of the CIFAR-10 dataset.

ImageNet-100: ImageNet-100 is a widely used single-label large dataset with 100 categories extracted from ImageNet-1000. The dataset contains 128,503 images, and each image belongs to only one category. Of these images, 500 images are used as a test set, and 128,000 images are used as a training set.

Flickr25k: There are 25,000 images in the Flickr25k image dataset, and each image has its own tags and annotation. Tags can be used as text descriptions. In total, 1,386 tags appear in at least 20 pictures, and there are 24 annotations as labels. In the experiment, 500 images of each class are taken as the training set, 100 images of each class are taken as the test set, and other images are taken as the database.

NUS-WIDE: NUS-WIDE is a web image dataset created by the media search laboratory of the National University of Singapore. NUS-WIDE contains 269,648 images from the website and 5,018 different tags. Six types of low-level features are extracted from these images, a including 64-d color histogram, a 144-d color correlation map, a 73-d edge orientation histogram, a 128-d wavelet texture, a 225-d block mode color moment and a 500-d packet based on the SIFT description. The NUS-WIDE dataset is very suitable for large-scale image retrieval research.



(a) (b) (c)

Fig. 4. The performance of ten image retrieval methods on three different evaluation criteria on the CIFAR-10 dataset is divided into three subgraphs: (a), (b) and (c). DAHP performed better than the other methods.

In order to reduce the adverse effects caused by potential errors in the reproduction process, all methods use only these four datasets for comparison.

B. Evaluation Metrics and Benchmarks

In this paper, four commonly used metrics are used as metrics to evaluate the retrieval performance of DAHP: the mean average precision (mAP), the Precision-Recall curves (PR), the precision curves with different numbers of top returned samples (P@N), and precision curves with a Hamming radius of 2 between the query samples and the dataset (P@H $=$ 2). The purpose of the image retrieval algorithm is to return accurate results as quickly and comprehensively as possible with less consumption. The precision is the proportion of correctly categorized images among all retrieval results, AP is the average of the correct rate at different recall points, and P is the average precision of all classes in the dataset. Specifically, the average precision of all classes is summed and then divided by the number of all classes. It is a single-value metric that can reflect the performance of the algorithm on the whole dataset. The mAP is likely to be higher when the retrieved correctly categorized images are more advanced (higher rank), and the accuracy defaults to 0 if the retrieval results do not return correctly categorized images. Usually, the retrieval results are arranged in a sorted manner, and therefore the user cannot see all the images immediately in the user observation process. The correct rate and recall rate are constantly and dynamically changing, and the precision-recall curves denote the correct rate corresponding to the recall rate on a scale from 0%-100%. The results are simple and intuitive, considering both the coverage of the retrieval results and the sorting of the retrieval results. P@N denotes the correct rate at the Nth position, and this index can be controlled by the number of retrieval results is different when observing the change of accuracy. Judging the robustness of the current algorithm is very effective in measuring the robustness and stability of the algorithm in large-scale image retrieval. The precision@H $=$ 2 represents the accuracy rate within a Hamming distance of 2. Specifically, the Hamming distance between the query detection result and the given query image is calculated, and metric measures the percentage of correct results in the images with a Hamming distance less than 2. The precision@H $=$ 2 value of different numbers of hash codes is used to evaluate the impact of the number of hash code bits on the accuracy. In the field of large-scale image hash retrieval, the above four metrics are the most commonly used metrics, and references [27], [30], [34], [43], [46] all use the above four metrics. This paper uses these four metrics to better compare with these algorithms.

Based on the feature extraction methods, the following methods are divided into two categories: traditional handcrafted methods and neural network-based methods. Then, these two categories of methods can be divided into the following five groups:

a) traditional unsupervised hash methods: ITQ [26];

b) traditional supervised hashing methods: SDH [27];

c) deep hash methods for extracting depth features: FastH [30], DHN [38], DSH [42] and DBDH [47];

d) pairwise label depth hashing methods: DPSH [43] and HashNet [44];

e) deep hash methods that introduce the attention mechanism: DAgH [49].

The experiments are conducted on an optimized AlexNet network and two benchmark datasets, and all results are reproduced from existing research results with slightly different performances of reproduced methods. Nevertheless, the performance priority order is consistent with the conclusions of the existing results.

C. Ablation Experiments

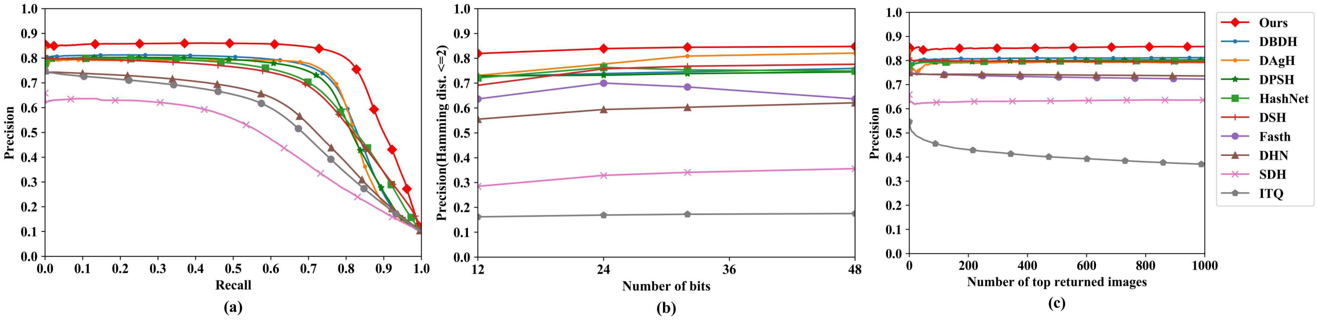
Fig. 4. The performance of three variants of DAHP on three different evaluation criteria on the CIFAR-10 dataset is divided into three subgraphs: (a), (b) and (c). DAHP-v3 performed better than the other methods

This part of the ablation experiment investigates the effects of the dual attention module and the anchor point error loss function module in the DAHP model. The experiments are progressively modified from the baseline DPSH model, which is divided into four different groups to analyze the experimental findings. The overall visual pairing is shown in Table III, where the table columns are the dual attention mechanism and the loss function for calculating the anchor loss with AlexNet as the backbone network and ResNet18 as the backbone network, respectively. The top row of the table indicates the configuration of the experimental models for this group. DPSH represents the baseline model using AlexNet training; the

TABLE III

ABLATION EXPERIMENTS RELATED TO THE DUAL ATTENTION MODULE AND ANCHOR POINT ERROR LOSS FUNCTION MODEL

modules	DPSH	DAHP-V1	DAHP-V2	DAHP-V3
AlexNet	√	×	×	×
ResNet18	×	√	√	√
Dual Attention Module	×	×	√	√
Mean Squared Error	×	×	×	√
(48-bit-CIFAR-10) mAP	0.7455	0.7945	0.8273	0.8477



(a) (b) (c)

Fig. 5. The performance of ten image retrieval methods on three different evaluation criteria on the ImageNet-100 dataset is divided into three subgraphs: (a), (b), and (c). DAHP performed better than the other methods.

DAHP-V1 model, DAHP-V2 model, and DAHP-V3 model are three variants of the DAHP model. $\surd$ indicates the addition of the module to the baseline model, and $\times$ indicates the absence of the module in the baseline model. Fig. 4 shows the performances of the four models on the remaining three metrics, and the ablation study is discussed in detail below.

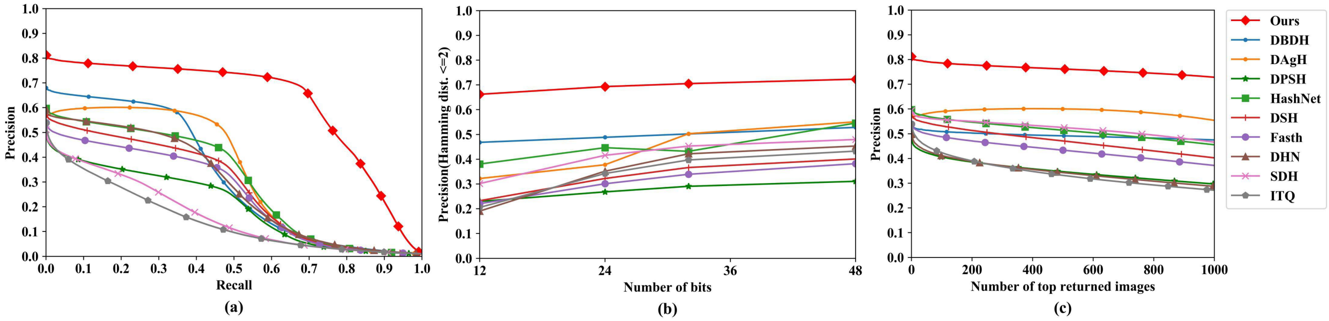
The DAHP-V1 model replaces only the backbone network based on the baseline DPSH model, changing from the original AlexNet to ResNet18; the experimental results show that the mAP is up to 5.16% higher than the original value. In addition, the visualization results of the other three metrics show slight performance improvements. The DAHP-V2 model adds a dual attention module to the DAHP-V1 model, and the experimental results show that the mAP increases by up to 4.13% compared to the original value. The visual results of the other three indicators show that the performance is once again improved. The DAHP-V3 model adds a mean squared error loss module to the DAHP-V2 model to calculate the anchor point error, and the experimental results show that the mAP is improved by 2.47% compared with the original value.

The ablation experiment fully demonstrates that ResNet18 has an advantage over AlexNet in feature extraction depth and that it helps to generate more discriminative hash codes. The dual attention mechanism module and the proposed anchor point error loss function module have a significant additive effect, which effectively circumvents the negative impact of the traditional relaxation-quantization step and confirms that the method of using anchor points as supervisory information in Hamming space can effectively reduce the Hamming distance of similar instance hash codes and increase the Hamming distance of dissimilar instance hash codes.

D. Comparison With Existing Methods

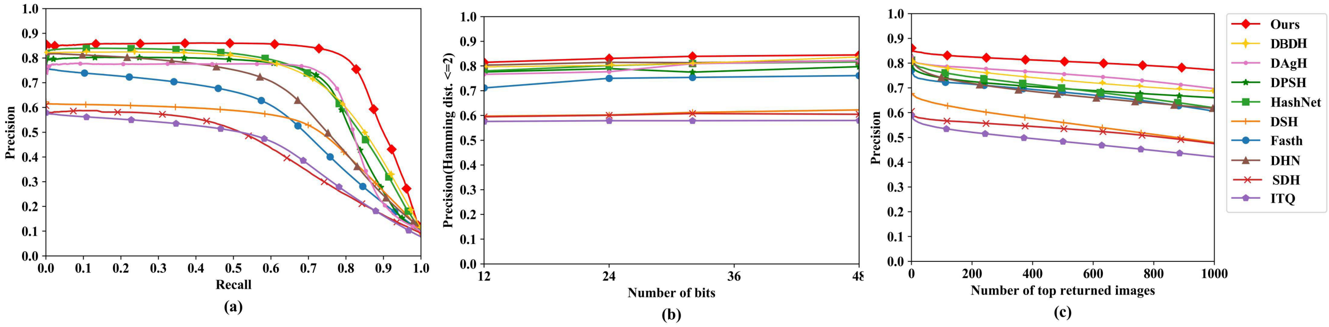
Specifically, all images were first resized to 224 × 224 pixels, after which original image pixels and target hash code were used as model inputs. In order to reduce the risk of overfitting, the initialization of the model is consistent with the work in literature [43]. The first seven layers of the pretrained ResNet18 initialized DAHP framework were used for hash learning. Table IV shows the mAP results of the ten image retrieval methods on CIFAR-10 and ImageNet-100. Table V shows the mAP results of the ten image retrieval methods on Flickr25K and NUS-WIDE.

As shown in Figs. 5(a), 6(a), 7(a) and 8(a), the precision-recall curve (PR) is an important metric to evaluate the image retrieval performance. The performance of DAHP is better than the model with which it is compared. Figs. 5(b), 6(b), 7(b) and 8(b) show the precisions within a Hamming radius of 2 ($P@H = 2$). Figs. 5(c), 6(c), 7(c) and 8(c) show the results of mAP@ALL on CIFAR-10 and NUS-WIDE and the precision curve (P@N) for the first 1000 search results on ImageNet-100 and Flickr25K, respectively. The DAHP model achieves the best results among all the search methods that were compared. The experimental results show that DAHP has better retrieval performance relative to the previous hash algorithm in all evaluation metrics. The DAHP algorithm has achieved SOTA performance.



(a) (b) (c)

Fig. 6. The performance of ten image retrieval methods on three different evaluation criteria on the Flickr25K dataset is divided into three subgraphs: (a), (b) and (c). DAHP performed better than the other methods.



(a) (b) (c)

Fig. 7. The performance of ten image retrieval methods on three different evaluation criteria on the NUSWIDE dataset is divided into three subgraphs: (a), (b) and (c). DAHP performed better than the other methods.

TABLE IV
MAP Results of the Ten Image Retrieval Methods on CIFAR-10 and ImageNet-100

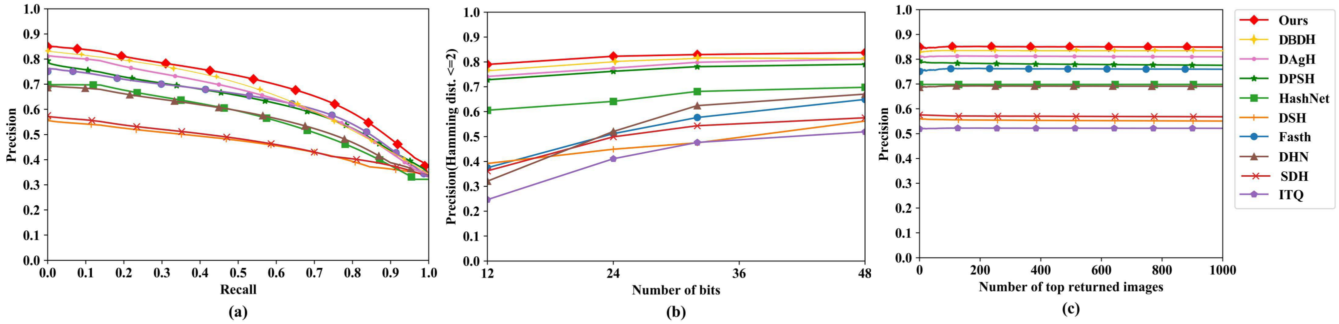
Approaches	CIFAR-10 (mAP@ALL)				ImageNet-100 (mAP@1000)			
	12-bits	24-bits	32-bits	48-bits	12-bits	24-bits	32-bits	48-bits
Ours	0.8194	0.8390	0.8445	0.8477	0.6621	0.6931	0.7052	0.7231
DBDH	0.7266	0.7356	0.7427	0.7483	0.4682	0.4885	0.5021	0.5285
DAgH	0.731	0.777	0.809	0.821	0.322	0.377	0.503	0.551
DPSH	0.7292	0.7320	0.7385	0.7455	0.2296	0.2682	0.2907	0.3108
HashNet	0.7210	0.7632	0.7535	0.7489	0.3806	0.4466	0.4319	0.5460
DSH	0.6914	0.7577	0.7683	0.7761	0.2325	0.3223	0.3665	0.4010
FastH	0.6365	0.7004	0.6849	0.6367	0.2205	0.301	0.3392	0.382
DHN	0.5550	0.5940	0.6030	0.6210	0.1906	0.3517	0.4219	0.4532
SDH	0.2850	0.3290	0.3410	0.3560	0.3018	0.4158	0.453	0.4797
ITQ	0.1620	0.1692	0.1725	0.1754	0.2055	0.3424	0.3971	0.4327

On the CIFAR-10 dataset, ImageNet-100 dataset, Flickr25K dataset and NUS-WIDE dataset, the mAP of 48 bits of the DAHP method improved by 3.25%, 31.23%, 3.07% and 3.75%, respectively, compared with the highest results of other methods. On the CIFAR-10 dataset, the mAP of 12 bits, 24 bits, 32 bits, 48 bits of the DAHP method improved by 12.76%, 14.05%, 13.70% and 9.94%, respectively, compared with the state-of-the-art method DBDH. The experimental results on imagenet-100 dataset, flickr25k dataset and nus-wide dataset also show that the performance of DAHP model

TABLE V

MAP RESULTS OF THE TEN IMAGE RETRIEVAL METHODS ON FLICKR25K AND NUS-WIDE

Approaches	Flickr25K (mAP@ALL)				NUS-WIDE (mAP@1000)			
	12-bits	24-bits	32-bits	48-bits	12-bits	24-bits	32-bits	48-bits
Ours	0.8149	0.8308	0.8392	0.8469	0.7929	0.8251	0.8386	0.8417
DBDH	0.7859	0.7970	0.8015	0.8135	0.7989	0.8121	0.8352	0.8405
DAgH	0.7665	0.777	0.8095	0.8215	0.7415	0.7751	0.7985	0.8112
DPSH	0.7761	0.7896	0.7754	0.7975	0.7283	0.7624	0.7810	0.7905
HashNet	0.7802	0.8020	0.8104	0.8166	0.6059	0.6415	0.6814	0.6981
DSH	0.5974	0.6022	0.6125	0.6226	0.3920	0.4492	0.4749	0.5623
FastH	0.7115	0.7504	0.7541	0.7614	0.3748	0.5117	0.5767	0.6494
DHN	0.8035	0.8150	0.8136	0.8217	0.3208	0.5205	0.6244	0.6707
SDH	0.595	0.601	0.608	0.605	0.3622	0.4989	0.5436	0.5756
ITQ	0.576	0.579	0.579	0.580	0.2466	0.4109	0.4765	0.5192



(a) (b) (c)

Fig. 8. The performance of ten image retrieval methods on three different evaluation criteria on the NUSWIDE dataset is divided into three subgraphs: (a), (b) and (c). DAHP performed better than the other methods.

is better than that of the DBDH. From the results of the other three evaluate metric, the performance of DAHP algorithm is much better than that of DBDH. Through comparative analysis with existing methods, we found that the proposed method is more suitable for high-precision retrieval scenarios in image datasets with more than 100,000 images than other mainstream methods. In addition, DAHP has the advantages of higher accuracy rates, more stable performance and faster retrieval efficiency on large-scale datasets, which can efficiently complete the image retrieval tasks under the trend of the growth of the volume of image data. In summary, our method could better meet the practical needs of current large-scale image retrieval.

E. Hyperparameter Experiments

In equation (10), the weight of the anchor loss is expressed, and the relationship between the anchor loss and map is explored. As shown in Table VI, the value range is set as [0.01-0.1]. It is not difficult to find that the value of the mAP first increases and then decreases. When the value is in the range of [0.05-0.08], the mAP reaches the maximum, so the value is set = 0.05 during the experiment.

TABLE VI

MAP RESULTS OF THE TEN IMAGE RETRIEVAL METHODS ON FLICKR25K AND NUS-WIDE

λ	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
m A P	0.7571	0.7757	0.7857	0.8224	0.8477	0.8445	0.8421	0.8412	0.7521	0.6446

V. CONCLUSION

Large-scale image retrieval is widely used in daily life. Deep hashing is a popular supporting technology at present. However, the traditional relaxation-quantization method seriously hampers the depth of the hash and lowers its advantages, which causes difficulties. In this paper, we proposed a new attention guided-based image retrieval method with pairwise labels, called DAHP. We replaced the backbone network of the baseline model with ResNet18. Then, we proposed two new modules, namely, the dual attention module and the mean square error loss module. The dual attention module is composed of the parallel structure of position attachment and channel attention mechanisms, and the mean squared error loss module is used to calculate the anchor error. This method was experimentally validated and shown to outperform the original relaxation-quantization fixed optimization model. The experimental results based on actual datasets showed that the DAHP algorithm improved the retrieval performance greatly compared with the baseline and that its performance was better than other methods in image retrieval applications. In addition, the next step is to test the DAHP method on datasets in specific domains, such as medical images, to provide machine-aided diagnosis to test the generality of the DAHP method.

References

[1] H.-M. Liu, R.-P. Wang, S.-G. Shan, and X.-X. Chen, "Learning to hash with discrete optimization," *Chin. J. Comput.*, vol. 42, no. 5, pp. 223–224, May 2019

[2] A. Grigorova, F. G. B. De Natale, C. Dagli, and T. S. Huang, "Content-based image retrieval by feature adaptation and relevance feedback," *IEEE Trans. Multimedia*, vol. 9, no. 6, pp. 1183–1192, Oct. 2007.

[3] R. Datta, D. Joshi, J. Li, and J.-Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, pp. 35–94, May 2008.

[4] J.-H. Su, W.-J. Huang, P. S. Yu, and V. S. Tseng, "Efficient relevance feedback for content-based image retrieval by mining user navigation patterns," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 3, pp. 360–372, Mar. 2011.

[5] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Object bank: An object-level image representation for high-level visual recognition," *Int. J. Comput. Vis.*, vol. 107, no. 1, pp. 20–39, Mar. 2014.

[6] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May 2018.

[7] J.-T. Horng and C.-C. Yeh, "Applying genetic algorithms to query optimization in document retrieval," *Inf. Process. Manage.*, vol. 36, no. 5, pp. 737–759, Sep. 2000.

[8] H. D. Cheng and X. J. Shi, "A simple and effective histogram equalization approach to image enhancement," *Digit. Signal Process.*, vol. 14, no. 2, pp. 158–170, Mar. 2004.

[9] M. Boughanem, C. Chrisment, and L. Tamine, "On using genetic algorithms for multimodal relevance optimization in information retrieval," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 53, no. 11, pp. 934–942, 2002.

[10] X. Yan, F. Zhu, J. Han, and P. Yu, "Feature-based similarity search in graph structures," *ACM Trans. Database Syst.*, vol. 31, no. 4, pp. 1458–1483, Jul. 2006.

[11] M. Wang and J.-F. Jing, "Image retrieval based on hash coding and convolutional neural network," *Comput. Eng. Appl.*, vol. 55, no. 23, pp. 194–199, Nov. 2019.

[12] R. Pasunuri and V. C. Venkaiah, "An optimal proximity method for nearest neighbor search in high dimensional data," in *Proc. 2nd Int. Conf. Contemp. Comput. Informat. (ICI)*, Dec. 2016, pp. 479–483.

[13] W. Wang, R. Wang, S. Shan, and X. Chen, "Probabilistic nearest neighbor search for robust classification of face image sets," in *Proc. 11th IEEE Int. Conf. Workshops Automat. Face Gesture Recognit. (FG)*, May 2015, pp. 1–7.

[14] Y. Cao, H. Qi, J. Gui, S.-A. Li, and K.-Q. Li, "General distributed hash learning on image descriptors for K-nearest neighbor search," *IEEE Signal Process. Lett.*, vol. 26, no. 5, pp. 750–754, May 2019.

[15] T. Reitmaier and A. Calma, "Resp-kNN: A semi-supervised classifier for sparsely labeled data in the field of organic computing," in *Proc. Doctoral Diss. Colloq.*, May 2014, pp. 85–99.

[16] Y. Li, Z. Miao, J. Wang, and Y. Zhang, "Deep binary constraint hashing for fast image retrieval," *Electron. Lett.*, vol. 54, no. 1, pp. 25–27, Jan. 2018.

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

[18] X. Zhu, Z. Huang, H. Cheng, J. Cui, and H. T. Shen, "Sparse hashing for fast multimedia search," *ACM Trans. Inf. Syst.*, vol. 31, no. 2, pp. 1–24, May 2013.

[19] H.-L. Liu, B.-A. Li, X.-Q. Lv, and Y. Huang, "Research on image retrieval algorithm based on deep convolutional neural network," *J. Appl. Res. Comput.*, vol. 34, no. 12, pp. 302–305, Jun. 2017.

[20] L. Liu, M. Yu, and L. Shao, "Latent structure preserving hashing," *Int. J. Comput. Vis.*, vol. 122, no. 3, pp. 439–457, May 2017.

[21] M. Zareapoor, J. Yang, D. K. Jain, P. Shamsolmoali, N. Jain, and S. Kant, "Deep semantic preserving hashing for large scale image retrieval," *Multimedia Tools Appl.*, vol. 78, no. 17, pp. 23831–23846, Sep. 2019.

[22] W. Ying, J. Sang, and J. Yu, "Locality-constrained discrete graph hashing," *Neurocomputing*, vol. 398, pp. 556–573, Jul. 2019.

[23] Z. Chen and J. Zhou, "Collaborative multiview hashing," *Pattern Recognit.*, vol. 52, no. 12, pp. 95–100, Feb. 2017, doi: 10.1016/j.patcog.2017.02.026.

[24] Y. Cao, Y. Liu, J. Xu, and D. Wang, "Image spam filtering with improved LSH algorithm," *Appl. Res. Comput.*, vol. 33, no. 6, pp. 1693–1696, Mar. 2016.

[25] L.-C. Xia, J.-G. Jiang, and M.-B. Qi, "A large-scale image retrieval method based on improved spectral hashing," *J. Hefei Univ. Technol., Natural Sci.*, vol. 38, no. 3, pp. 1049–1054, Aug. 2016.

[26] J.-J. Zhen, Z.-L. Ying, Y.-H. Zhao, and S.-A. Huang, "Research of deep learning and iterative quantization in image retrieval," *J. Signal Process.*, vol. 35, no. 5, pp. 919–925, May 2019.

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

[28] S.-C. Ke, Y.-W. Zhao, B.-C. Li, and T.-Q. Peng, "Image retrieval based on convolutional neural network and kernel-based supervised hashing," *Acta Electronica Sinica*, vol. 45, no. 1, pp. 157–163, Jan. 2017.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015, doi: 10.1109/TPAMI.2015.2389824.

[30] K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen, "Deep learning of binary hash codes for fast image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 27–35.

[31] J. Wan *et al.*, "Deep learning for content-based image retrieval: A comprehensive study," in *Proc. ACM Int. Conf. Multimedia*, Nov. 2014, pp. 157–166.

[32] W.-C. Kang, W.-J. Li, and Z.-H. Zhou, "Column sampling based discrete supervised hashing," in *Proc. 30th AAAI Conf. Artif. Intell.*, Jul. 2016, pp. 1230–1236.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

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

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

[36] H.-F. Yang, K. Lin, and C.-S. Chen, "Supervised learning of semantics-preserving hash via deep convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 437–451, Feb. 2018.

[37] J. Jeong, I. Won, H. Yang, B. Lee, and D. Jeong, "Deformable object matching algorithm using fast agglomerative binary search tree clustering," *Symmetry*, vol. 9, no. 2, p. 25, Feb. 2017.

[38] H. Zhu, M.-S. Long, J.-M. Wang, and Y. Gao, "Deep hashing network for efficient similarity retrieval," in *Proc. 30th AAAI Conf. Artif. Intell.*, Dec. 2016, pp. 448–456.

[39] K. Kanwal, K. T. Ahmad, R. Khan, A. T. Abbasi, and J. Li, "Deep learning using symmetry, FAST scores, shape-based filtering and spatial mapping integrated with CNN for large scale image retrieval," *Symmetry*, vol. 12, no. 4, pp. 6–12, Apr. 2020.

[40] Y.-J. Su, H. Yu, C. Lei, Q. Deng, and Y.-G. Li, "PCA hashing for image data retrieval," *Appl. Res. Comput.*, vol. 35, no. 10, pp. 273–276, Oct. 2017.

[41] Y.-C. Zhang, Z.-C. Huang, and Y.-X. Chen, "Multi-scale balanced deep hashing method for image retrieval," *Appl. Res. Comput.*, vol. 36, no. 2, pp. 307–315, Oct. 2019.

[42] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2064–2072.

[43] W.-J. Li, S. Wang, and W.-C. Kang, "Feature learning based deep supervised hashing with pairwise labels," in *Proc. Int. Joint Conf. Artif. Intell.*, Nov. 2016, pp. 3270–3278.

[44] Z. Cao, M. Long, J. Wang, and P. S. Yu, "HashNet: Deep learning to hash by continuation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5609–5618.

[45] H. Zhai, S. Lai, H. Jin, X. Qian, and T. Mei, "Deep transfer hashing for image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 742–753, Feb. 2021, doi: 10.1109/TCSVT.2020.2991171.

[46] H. Lai, Y. Pan, S. Liu, Z. Weng, and J. Yin, "Improved search in Hamming space using deep multi-index hashing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2844–2855, Sep. 2019, doi: 10.1109/TCSVT.2018.2869921.

[47] X. Zheng, Y. Zhang, and X. Lu, "Deep balanced discrete hashing for image retrieval," *Neurocomputing*, vol. 403, pp. 224–236, Aug. 2020, doi: 10.1016/j.neucom.2020.04.037.

[48] Z. Yang, O. I. Raymond, W. Sun, and J. Long, "Deep attention-guided hashing," *IEEE Access*, vol. 7, pp. 11209–11221, Aug. 2019.

[49] L. Jin, X. Shu, K. Li, Z. Li, G.-J. Qi, and J. Tang, "Deep ordinal hashing with spatial attention," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2173–2186, May 2019, doi: 10.1109/TIP.2018.2883522.

[50] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019, doi: 10.1109/TGRS.2018.2864987.

[51] Q. Wang, T. Han, Z. Qin, J. Gao, and X. Li, "Multitask attention network for lane detection and fitting," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 8, 2020, doi: 10.1109/TNNLS.2020.3039675.

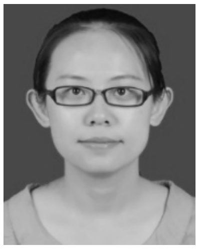
[52] C. Han, R. Zheng, C. Gao, and N. Sang, "Complementation-reinforced attention network for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3433–3445, Oct. 2020, doi: 10.1109/TCSVT.2019.2957467.

[53] L.-W. Ge, J. Zhang, Y. Xia, P. Chen, B. Wang, and C.-H. Zheng, "Deep spatial attention hashing network for image retrieval," *J. Vis. Commun. Image Represent.*, vol. 63, pp. 2173–2186, Aug. 2019.

Xue Li received the bachelor's degree from the School of Software Engineering, Xinjiang University, Urumqi, China, in 2018, where she is currently pursuing the master's degree. Her research interests include image retrieval and learning to hash.

Jiong Yu received the Ph.D. degree from the School of Computer Science and Technology, Beijing University of Technology, China, in 2009. He worked as a Senior Visiting Scholar with the Information Computing Center, National Research Institute, Canada. He is currently a Professor and a Ph.D. Supervisor of Computer Science with the School of Information Science and Engineering, Xinjiang University. His main research interests include grid computing, parallel computing, and deep learning.

Yongqiang Wang received the bachelor's degree in landscape major from Xinjiang Agricultural University, Urumqi, China, in 2018. He is currently pursuing the master's degree with the School of Information Science and Engineering, Xinjiang University, China. His research interests include image super-resolution and image retrieval.



Jia-Ying Chen received the B.E. degree from Northwest A&F University in 2011 and the M.E. degree from Xinjiang University, China, in 2014, where she is currently pursuing the Ph.D. degree with the School of Information Science and Engineering. Her current research interests include machine learning and recommender systems.

Peng-Xiao Chang received the bachelor's degree from the School of Software Engineering, Xinjiang University, Urumqi, China, in 2019, where he is currently pursuing the master's degree. His research interests include image retrieval and deep learning.

Ziyang Li received the master's degree from the School of Software Engineering, Xinjiang University, Urumqi, China, in 2018, where he is currently pursuing the Ph.D. degree. His research interests include stream computing, in-memory computing, and distributed computing.

Digital Object Identifier 10.1109/TCSVT.2022.3154905

(Contents Continued from Page 907)

Image/Video Coding and Compression

Image/Video Communications and Networks

Image/Video Storage and Retrieval

Image/Video Hardware/Software Systems

Image/Video Applications

LETTERS

DAHP：使用成对标签的深度注意力引导哈希

李雪[1]，于炯[1]，王永强[1]，陈嘉颖[1]，常朋肖[2]，李梓杨[2]

（1.新疆大学信息科学与工程学院，新疆乌鲁木齐市；2.新疆大学软件学院，新疆乌鲁木齐市）

摘要：为了解决提取特征不足以及松弛-量化策略的深度哈希面临的哈希码离散优化问题，我们提出一种新的使用成对标签的深度注意引导哈希方法(DAHP)来更好的增强全局特征融合，并通过学习图像特征的上下文信息，来提升特征表示能力，以及优化了损失函数解决了离散优化导致的特征信息丢失问题。首先，我们引入了一个新的概念，称为锚定哈希码生成(AHCG)算法，作用是在汉明空间中生成锚点哈希码，作为模型训练的监督信息。然后，我们训练了带有位置注意和通道注意机制的ResNet网络，网络将表示图片的哈希码匹配到每个锚点附近，最后，我们使用优化后的损失函数来计算成对损失和锚点损失，使得模型具备生成强辨别力的哈希码的能力。我们分别在四个基准数据集上做了对比实验和消融实验，检索精度优于SOTA方法。

关键字： 通道注意力，深度哈希，图像检索，位置注意力

检索工具	SCI（科学引文索引）	查证单位	新疆大学图书馆
版　本	网络版	地　址	新疆大学图书馆信息咨询部
收录作者	于炯	查 证 人	黄晖 0991-8582535
文章作者单位	新疆大学信息科学与工程学院	查证日期	2022/6/2

标题: DAHP: Deep Attention-Guided Hashing With Pairwise Labels

作者: Li, X (Li, Xue); Yu, J (Yu, Jiong); Wang, YQ (Wang, Yongqiang); Chen, JY (Chen, Jia-Ying); Chang, PX (Chang, Peng-Xiao); Li, ZY (Li, Ziyang)

来源出版物: IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY 卷: 32 期: 3 页: 933-946 DOI: 10.1109/TCSVT.2021.3070129 出版年: MAR 2022

Web of Science 核心合集中的 "被引频次": 1

被引频次合计: 1

使用次数 (最近 180 天): 4

使用次数 (2013 年至今): 4

引用的参考文献数: 53

入藏号: WOS:000766700400006

语言: English

文献类型: Article

地址: [Li, Xue; Chang, Peng-Xiao; Li, Ziyang] Xinjiang Univ, Coll Software, Urumqi 830046, Peoples R China.

[Yu, Jiong] Xinjiang Univ, Sch Informat Sci & Engn, Urumqi 830046, Peoples R China.

[Wang, Yongqiang] Xinjiang Univ, Coll Informat Sci & Engn, Urumqi 830046, Peoples R China.

通讯作者地址: Li, X (通讯作者)，Xinjiang Univ, Coll Software, Urumqi 830046, Peoples R China.

电子邮件地址: 1547037202@qq.com; yujiong@xju.edu.cn; 925968662@qq.com; chenjiaying@stu.xju.edu.cn; changpengxiao@stu.xju.edu.cn; liziyang@stu.xju.edu.cn

出版商: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC

出版商地址: 445 HOES LANE, PISCATAWAY, NJ 08855-4141 USA

Web of Science Index: Science Citation Index Expanded (SCI-EXPANDED)

Web of Science 类别: Engineering, Electrical & Electronic

研究方向: Engineering　　ISSN: 1051-8215

IDS 号: ZP8XH　　eISSN: 1558-2205

29 字符的来源出版物名称缩写: IEEE T CIRC SYST VID

ISO 来源出版物缩写: IEEE Trans. Circuits Syst. Video Technol.

来源出版物页码计数: 14

基金资助致谢:

基金资助机构	授权号
National Natural Science Foundation of China	61462079 61562086 61562078 61862060
Science and Technology Support Project of Ministry of National Science and Technology of China	2015BAH02F01

This work was supported in part by the National Natural Science Foundation of China under Grant 61462079, Grant 61562086, Grant 61562078, and Grant 61862060 and in part by the Science and Technology Support Project of Ministry of National Science and Technology of China under Grant 2015BAH02F01.

输出日期: 2022-06-02