DEEP UNIQUENESS-AWARE HASHING FOR FINE-GRAINED MULTI-LABEL IMAGE RETRIEVAL

Dayan Wu[⋆†] *Zheng Lin*[⋆] *Bo Li*[⋆‡] *Jing Liu*[⋆†] *Weiping Wang*[⋆]

[⋆] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, 100093, China
[†] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, 100049, China

ABSTRACT

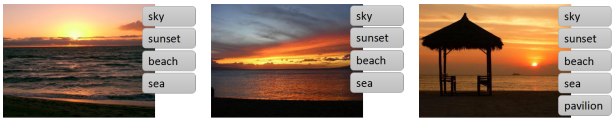
Deep supervised hashing methods for multi-label image retrieval have achieved great success nowadays. However, these methods only take the similarity between the database images and the query images into account, but they ignore the uniqueness of the database images when deciding on their rankings. Here we present a novel Deep Uniqueness-Aware Hashing (DUAH) method for learning hash functions that preserve not only multilevel semantic similarity between multi-label images, but also the unique semantic structure of each image. In our approach, both the pairwise label information and the classification information are fully exploited to maximize the discriminability of the output binary codes within one stream framework. Extensive evaluations conducted on three widely used multi-label image benchmarks demonstrate that DUAH can support fine-grained multi-label image retrieval better.

Index Terms— Deep Learning, Fine-Grained Multi-Label Image Retrieval

1. INTRODUCTION

With the explosive growth of images on the web, much attention has been paid to the nearest neighbor search via hashing methods. Deep supervised hashing methods try to perform simultaneous feature learning and hash-code learning with deep neural networks, which have shown much better performance than traditional hashing methods with hand-crafted features.

However, most of the deep hashing methods aim at preserving binary semantic similarity (i.e. similar or dissimilar) [1, 2, 3, 4, 5, 6, 7, 8], and they are not scalable in multi-label image retrieval [9, 10], which can be observed in our experimental results. Recently, several deep supervised hashing methods [9, 10] for multi-label image retrieval are proposed. More specifically, DSRH in [9] tries to decode the multilevel semantic similarity information with a ranking list, and DMSSPH in [10] is the first pairwise label based deep supervised hashing method for multi-label image retrieval. However, all the previous deep supervised hashing methods for multi-label image retrieval only take the similarity between the database images and the query images into account, but they ignore the uniqueness of the database images when deciding on their rankings. The three images in Fig. 1 are all associated with "sky", "sunset", "beach" and "sea", while the image in (c) is further relevant to "pavilion". For the previous deep hashing methods, the two images in (b) and (c) are considered at the same similarity level when the image in (a) is the query image. However, due to the uniqueness of the image in (c), the image in (a) should be more similar to the image in (b) than to the image in (c). For fine-grained multi-label image retrieval, both the multilevel semantic similarity and the uniqueness of the database images should be taken into consideration when deciding on their rankings.



(a) (b) (c)

Fig. 1. Three exemplary images. Both images in (a) and (b) are associated with four labels: "sky", "sunset", "beach" and "sea". The image in (c) is labeled with "sky", "sunset", "beach", "sea" and "pavilion".

To support fine-grained multi-label image retrieval, we introduce a novel framework, named Deep Uniqueness-Aware Hashing (DUAH). The overview of the proposed framework is presented in Fig. 2. We use a CNN model [11] to learn hash functions directly from images. Meanwhile, a fine-grained multilevel contrastive loss function is elaborately designed to maximize the discriminability of the learned binary codes. To discover the minor semantic differences between the images like Fig. 1a and Fig. 1c, we add a multi-label classification layer *fcc* directly after the hash layer *fch*, aiming at making the learned hash codes ideal for multi-label classification. In general, the contributions of this study can be summarized as follows: 1) We present a novel CNN based framework, named DUAH, for learning hash functions that preserve not only multilevel semantic similarity between images, but also the unique semantic structure of each image. 2) We propose a fine-grained multilevel contrastive loss function to optimize our architecture, which can make the learned hash

[‡]Corresponding author:libo@iie.ac.cn

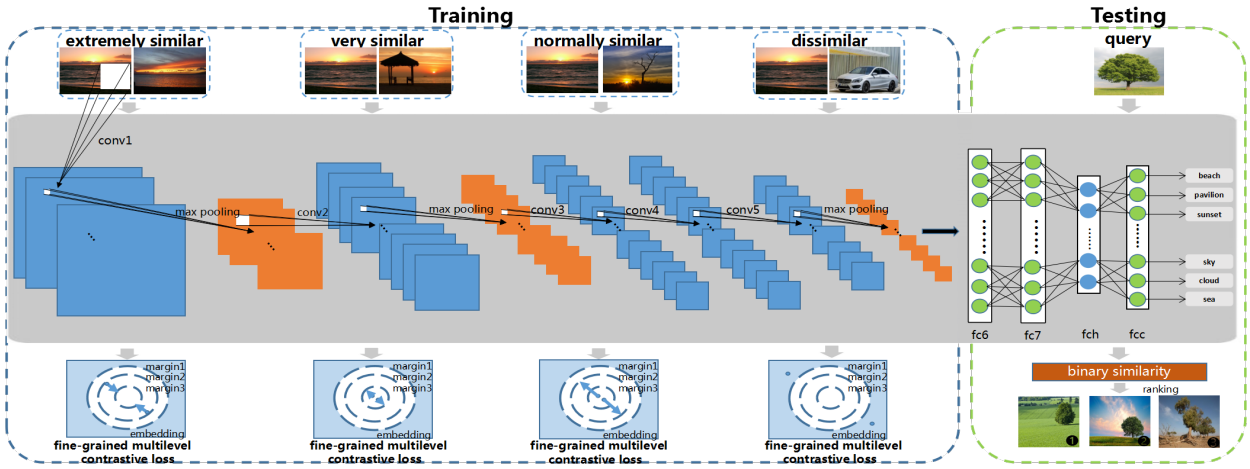


Fig. 2. Overview of the proposed framework. The input to the proposed framework is in the form of two-tuples, i.e., extremely similar images, very similar images, normally similar images and dissimilar images. Through the proposed architecture, the image tuples are first encoded into a pairwise of image feature vectors by five convolution layers and two fully connected layers. Then each image vector in the two-tuple is converted to a hash code by a hash layer *fch*. Note that the hash layer *fch* is simply a fully connected layer without any activation functions. After that, these hash codes are used in a fine-grained multilevel contrastive loss that aims to preserve fine-grained multilevel similarities on images. Besides, a multi-label classification layer *fcc* is directly connected with the layer *fch*, which aims at making the learned hash codes ideal for multi-label classification.

codes more discriminating. 3) A multi-label classification loss function is proposed to learn hash codes together with the fine-grained multilevel contrastive loss function within one stream framework, which can preserve the semantic structure and the uniqueness of images. 4) Extensive experiments on three widely used multi-label datasets demonstrate the advantages of DUAH over several state-of-the-art hashing techniques.

2. THE PROPOSED APPROACH

Assume we are given a data set $D = \{x_n\}_{n=1}^{N}$ where each data $x \in \mathcal{R}^M$ is associated with a set of class labels $l \subseteq L$, we aim to learn a set of hash functions $h(x) = [h_1(x), h_2(x), ..., h_k(x)]$ which generate k-bits binary codes ($k \ll M$).

2.1. Fine-Grained Multilevel Contrastive Loss

To preserve fine-grained multilevel semantic similarity between images, we further subdivide very similar images into extremely similar images and very similar images. For a pair of images I_1 and I_2 and their associated class label sets l_1 and l_2, we define $n_1 = |l_1|$, $n_2 = |l_1 \cap l_2|$ and $n_3 = |l_2|$. The similarity degrees can be further divided into:

$$y = \begin{cases} 0\ (extremely\ similar), & if\ n_1 = n_2 = n_3 \\ 1\ (very\ similar), & if\ n_1 = n_2 \neq n_3 \\ 2\ (normally\ similar), & if\ n_1 > n_2 > 0 \\ 3\ (dissimilar), & if\ n_2 = 0 \end{cases} \tag{1}$$

Then the fine-grained multilevel contrastive loss function is defined as:

$$\begin{aligned} \mathcal{L}_{fg} = \sum_{i=1}^{N_1} \Big[& \frac{1}{2} I_{y_i=0} \max\left(D_H\left(h(I_{i,1}), h(I_{i,2})\right) - m_1, 0\right) + \\ & \frac{1}{2} I_{y_i=1} \max\left(m_1 - D_H\left(h(I_{i,1}), h(I_{i,2})\right), 0\right) + \\ & \frac{1}{2} I_{y_i=2} \max\left(m_2(n_{i,1} - n_{i,2})/n_{i,1} - D_H\left(h(I_{i,1}), h(I_{i,2})\right), 0\right) + \\ & \frac{1}{2} I_{y_i=3} \max\left(m_2 - D_H\left(h(I_{i,1}), h(I_{i,2})\right), 0\right) \Big] \\ s.t. \quad & h(I_{i,j}) \in \{+1, -1\}^k, j \in \{1, 2\} \end{aligned} \tag{2}$$

where $D_H(\cdot, \cdot)$ denotes the Hamming distance between two binary vectors, N_1 is the number of the training pairs randomly selected from training images, and $m_1, m_2 > 0$ are two margin threshold parameters. The indicator function $I_{condition} = 1$ if condition is true; otherwise $I_{condition} = 0$.

Different from the multilevel contrastive loss function, we don't further penalize extremely similar images if the Hamming distance between their binary codes falls below m_1. In fact, the indefinite contraction of extremely similar images is a damaging behaviour [12]. Besides, to distinguish very similar images from extremely similar images, we punish very similar images if the Hamming distance between their hash codes falls below m_1. We punish normally similar or dissimilar images mapped to close binary codes when their Hamming distance falls below a certain margin threshold determined by their similarity degrees.

However, it is hard to minimize $\mathcal{L}_{fg}$ because it is a dis-

crete optimization. We adopt the relexation method proposed in [1, 10], and $\mathcal{L}_{fg}$ can be transformed to:

$$\begin{aligned}\mathcal{L}_{fg} = \sum_{i=1}^{N_1} \Big[&\frac{1}{2} I_{y_i=0} \max\left(||f(I_{i,1}) - f(I_{i,2})||_2^2 - m_1, 0\right) + \\ &\frac{1}{2} I_{y_i=1} \max\left(m_1 - ||f(I_{i,1}) - f(I_{i,2})||_2^2, 0\right) + \\ &\frac{1}{2} I_{y_i=2} \max\left(m_2(n_{i,1} - n_{i,2})/n_{i,1} - ||f(I_{i,1}) - f(I_{i,2})||_2^2, 0\right) + \\ &\frac{1}{2} I_{y_i=3} \max\left(m_2 - ||f(I_{i,1}) - f(I_{i,2})||_2^2, 0\right) + \\ &\alpha\left(|||f(I_{i,1})| - \mathbf{1}||_1 + |||f(I_{i,2})| - \mathbf{1}||_1\right)\Big]\end{aligned} \tag{3}$$

where $f(I_{i,j})$ is the continuous output vector of the layer *fch*, $\mathbf{1}$ is a vector of all ones, $||\cdot||_1$ is the L1-norm of vector, $|\cdot|$ is the element-wise absolute value operation, and α is a weight parameter that controls the strength of the regularizer. To generate binary codes, we set $h(I_i) = sgn(f(I_i))$.

2.2. Multi-Label Classification Loss

Observing that the semantic structure of images can reflect their uniqueness, we propose a multi-label classification loss function to learn the hash codes together with the fine-grained multilevel contrastive loss function within one stream framework. Unlike the two stream framework proposed in [2], the multi-label classification loss function in our framework has a direct impact on the hash functions, because the layer *fcc* is directly connected with the layer *fch*. The multi-label classification loss function can be computed by:

$$\mathcal{L}_{mc} = -\sum_{i=1}^{N_2}\sum_{j=1}^{c}\left[I_{y_j^i=1}\frac{1}{c_i} log(p_j^i) + I_{y_j^i=0} log(1-p_j^i)\right] \tag{4}$$

in which $y^i \in \{0,1\}^c$ is a binary label vector, c is the number of classes, c_i is the number of labels the image I_i is associated with, N_2 is the number of training images, and p_j^i is the predict probability defined as:

$$p_j^i = \frac{e^{W_j^T f(I_i)}}{\sum_{t=1}^{c} e^{W_t^T f(I_i)}} \tag{5}$$

where $W \in \mathcal{R}^{k\times c}$ denotes the weight matrix of the layer *fcc*.

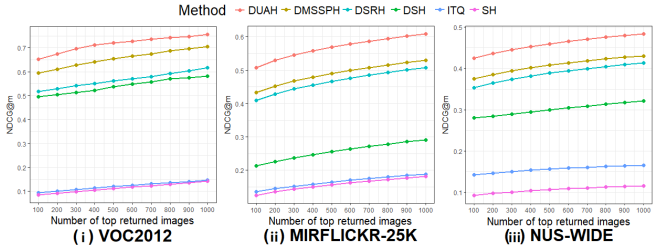
2.3. Learning

We use back-propagation algorithm with mini-batch gradient descent method to train the network. Our goal is to minimize $\mathcal{L} = \mathcal{L}_{fg} + \mathcal{L}_{mc}$. The gradient of $\mathcal{L}_{mc}$ is:

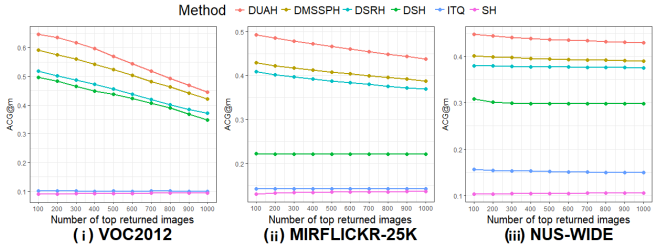
$$\frac{\partial \mathcal{L}_{mc}}{\partial \hat{f}_j^i} = I_{y_j^i=1}((1-S)p_j^i - \frac{1}{c_i}) + I_{y_j^i=0}(2 - S + \frac{p_j^i}{1-p_j^i})p_j^i \tag{6}$$

where $\hat{f}_j^i = W_j^T f(I_i)$ is the j-th output of the layer *fcc*, and S is computed by:

$$S = \sum_{j=1}^{c}(I_{y_j^i=0}\frac{p_j^i}{1-p_j^i}) \tag{7}$$



(a) NDCG curves with 64 bits w.r.t different number of top returned images.



(b) ACG curves with 64 bits w.r.t different number of top returned images.

Fig. 3. NDCG and ACG curves with 64 bits w.r.t different number of top returned images.

The subgradients of the first four items and the regularizer part of $\mathcal{L}_{fg}$ are respectively written as:

$$\begin{aligned}\frac{\partial \mathcal{L}_{fg1}}{\partial f_{i,j}} &= (-1)^{j+1}\left(f_{i,1} - f_{i,2}\right) I_{y_i=0 \&\& ||f_{i,1}-f_{i,2}||_2^2 > m_1} \\ \frac{\partial \mathcal{L}_{fg2}}{\partial f_{i,j}} &= (-1)^{j}\left(f_{i,1} - f_{i,2}\right) I_{y_i=1 \&\& ||f_{i,1}-f_{i,2}||_2^2 < m_1} \\ \frac{\partial \mathcal{L}_{fg3}}{\partial f_{i,j}} &= (-1)^{j}\left(f_{i,1} - f_{i,2}\right) I_{y_i=2 \&\& ||f_{i,1}-f_{i,2}||_2^2 < m_2(\frac{n_{i,1}-n_{i,2}}{n_{i,1}})} \\ \frac{\partial \mathcal{L}_{fg4}}{\partial f_{i,j}} &= (-1)^{j}\left(f_{i,1} - f_{i,2}\right) I_{y_i=3 \&\& ||f_{i,1}-f_{i,2}||_2^2 < m_2} \\ \frac{\partial \mathcal{L}_{fg-regularizer}}{\partial f_{i,j}} &= \alpha\delta\left(f_{i,j}\right)\end{aligned} \tag{8}$$

where

$$\delta(x) = \begin{cases} 1 & -1 \le x \le 0 \, or \, x \ge 1 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

is applied element-wisely, and $f_{i,j}$ denotes $f(I_{i,j})$. With the computed gradients (subgradients) over mini-batches, the rest of the back-propagation can be done in standard manner.

3. EXPERIMENTS

3.1. Datasets and Evaluation Metrics

We test the proposed hashing method on three multi-label benchmark datasets, i.e., VOC2012 [13], MIRFLICKR-25K [14] and NUS-WIDE [15]. VOC2012 consists of 22,531 multi-label images in 20 classes. MIRFLICKER-25K consists of 25,000 multi-label images in 38 classes. NUS-WIDE contains 269,648 images collected from Flickr, and we follow the settings in [3, 16, 17] to use the subset of 195,834 images that are associated with the 21 most frequent concepts, where each concept consists of at least 5,000 images.

Table 1. Comparison of NDCG@1000, ACG@1000 and weighted mAP w.r.t. different number of bits on three datasets. Note that the weighted mAP performance is calculated on the top 5,000 returned images for MIRFLICKR-25K and NUS-WIDE.

Methods	VOC2012				MIRFLICKR-25K				NUS-WIDE			
	24 bits	32 bits	48 bits	64 bits	24 bits	32 bits	48 bits	64 bits	24 bits	32 bits	48 bits	64 bits
NDCG@1000												
DUAH	**0.7375**	**0.7320**	**0.7547**	**0.7561**	**0.5524**	**0.5819**	**0.6020**	**0.6081**	**0.4612**	**0.4638**	**0.4788**	**0.4838**
DMSSPH	0.6662	0.6738	0.7012	0.7064	0.4975	0.5113	0.5208	0.5301	0.4127	0.4189	0.4238	0.4306
DSRH	0.5670	0.5893	0.6110	0.6164	0.4797	0.4959	0.5046	0.5079	0.3696	0.3931	0.4043	0.4131
DSH	0.5720	0.5590	0.5889	0.5824	0.2845	0.2843	0.2720	0.2915	0.3186	0.3240	0.3217	0.3208
ITQ+CNN	0.1411	0.1675	0.1438	0.1475	0.1987	0.2001	0.1888	0.1883	0.1512	0.1629	0.1667	0.1659
SH+CNN	0.1555	0.1689	0.1665	0.1421	0.1842	0.2298	0.1762	0.1812	0.1359	0.1175	0.1413	0.1161
ACG@1000												
DUAH	**0.4364**	**0.4339**	**0.4429**	**0.4446**	**0.3992**	**0.4207**	**0.4339**	**0.4381**	**0.4122**	**0.4136**	**0.4253**	**0.4292**
DMSSPH	0.4002	0.4075	0.4193	0.4218	0.3640	0.3756	0.3812	0.3879	0.3740	0.3791	0.3840	0.3892
DSRH	0.3494	0.3618	0.3706	0.3714	0.3547	0.3630	0.3680	0.3692	0.3395	0.3587	0.3680	0.3751
DSH	0.3483	0.3386	0.3550	0.3488	0.2154	0.2161	0.2083	0.2213	0.2961	0.3005	0.2989	0.2980
ITQ+CNN	0.0956	0.1145	0.0948	0.1001	0.1489	0.1485	0.1418	0.1431	0.1352	0.1453	0.1504	0.1492
SH+CNN	0.1088	0.1104	0.1092	0.0937	0.1405	0.1738	0.1333	0.1365	0.1232	0.1063	0.1275	0.1055
weighted mAP												
DUAH	**0.5593**	**0.5573**	**0.5818**	**0.5868**	**0.3585**	**0.3791**	**0.3900**	**0.3935**	**0.4020**	**0.4025**	**0.4127**	**0.4165**
DMSSPH	0.4933	0.4953	0.5250	0.5307	0.3349	0.3441	0.3492	0.3550	0.3661	0.3711	0.3760	0.3815
DSRH	0.3936	0.4113	0.4302	0.4365	0.3312	0.3363	0.3402	0.3410	0.3355	0.3551	0.3639	0.3708
DSH	0.4372	0.4246	0.4560	0.4542	0.2154	0.2158	0.2076	0.2207	0.2962	0.3006	0.2987	0.2983
ITQ+CNN	0.1104	0.1177	0.1008	0.1034	0.1497	0.1485	0.1430	0.1441	0.1346	0.1442	0.1502	0.1472
SH+CNN	0.1115	0.1149	0.1136	0.0983	0.1424	0.1700	0.1359	0.1387	0.1235	0.1079	0.1271	0.1071

For VOC2012 and MIRFLICKER-25K, 2000 images are randomly selected as testing queries and the remaining images are used as the database for training and retrieval. For NUS-WIDE, we randomly select 2,100 images (100 images per class) for testing queries and the rest is used for training and retrieval. We resize all images into 256×256.

Following the previous works [9, 10], the evaluation metrics are NDCG, ACG and weighted mAP. Note that we use the Jaccard coefficient based similarity measurement ($s_{i,j} = \frac{l_i \cap l_j}{l_i \cup l_j}$) proposed in [5, 10] to measure the fine-grained multi-level semantic similarity between the two images I_i and I_j, which takes both the number of common labels and that of unique labels into account.

3.2. Method Comparison

Comparative methods: We compare our method with SH [18], ITQ [19], DSH [1], DSRH [9] and DMSSPH [10]. DMSSPH and DSRH are two state-of-the-art deep supervised hashing methods for multi-label image retrieval. DSH is one of the state-of-the-art deep supervised hashing methods for single label image retrieval. ITQ and SH are two representative data-dependent hashing methods. Our method is implemented with Caffe [20]. Following [10], we set $m_2 = (\lfloor \frac{k}{2n_1} \rfloor + 1) \times 4n_1$ and $\alpha = 0.01$. For m_1, we empirically set it to 4.

The comparison results are shown in Table 1 and Fig. 3. The weighted mAP results of DUAH indicate a relative increase of **10.8**$\% \sim$ **13.4**$\%$ / **7.0**$\% \sim$ **11.7**$\%$ / **8.3**$\% \sim$ **9.8**$\%$ over the second best baseline on VOC2012 / MIRFLICKER-25K / NUS-WIDE, respectively. The NDCG@1000 values of DUAH indicate a **7.0**$\% \sim$ **10.7**$\%$ / **11.0**$\% \sim$ **15.6**$\%$ / **10.7**$\% \sim$ **13.0**$\%$ relative increase over the second best baseline on VOC2012 / MIRFLICKER-25K / NUS-WIDE, respectively. The ACG@1000 values indicate a **5.4**$\% \sim$ **9.0**$\%$ / **9.7**$\% \sim$ **13.8**$\%$ / **9.1**$\% \sim$ **10.8**$\%$ relative increase on VOC2012 / MIRFLICKER-25K / NUS-WIDE, respectively. Besides, with the decreasing number of top returned images, DUAH keeps obvious advantages in terms of NDCG and ACG, as shown in Fig. 3. Note that DUAH, DMSSPH and DSRH perform much better than DSH in most cases, indicating that the methods aiming at preserving binary semantic similarity can not well apply to multi-label image retrieval.

4. CONCLUSION

In this paper, we have proposed a CNN-based hashing method for fine-grained multi-label image retrieval, called DUAH. DUAH takes into account both the multilevel semantic simiarity and the uniqueness of the database images when deciding on their rankings. We carefully design a fine-grained multilevel contrastive loss function which pays particular attention to the uniqueness of the images that are very similar to the query image, aiming at improving the accuracies of top returned images. To preserve the unique semantic structure of each image, we propose a multi-label classification loss function to learn the hash codes together with the fine-grained multilevel contrastive loss within one stream framework. Extensive experiments demonstrate that DUAH performs much better in fine-grained multi-label image retrieval.

5. REFERENCES

[1] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen, "Deep supervised hashing for fast image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2064–2072.

[2] Ting Yao, Fuchen Long, Tao Mei, and Yong Rui, "Deep semantic-preserving and ranking-based hashing for image retrieval.," in *IJCAI*, 2016, pp. 3931–3937.

[3] Hanjiang Lai, Yan Pan, Ye Liu, and Shuicheng Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3270–3278.

[4] Wu-Jun Li, Sheng Wang, and Wang-Cheng Kang, "Feature learning based deep supervised hashing with pairwise labels," *arXiv preprint arXiv:1511.03855*, 2015.

[5] Ruimao Zhang, Liang Lin, Rui Zhang, Wangmeng Zuo, and Lei Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4766–4779, 2015.

[6] Han Zhu, Mingsheng Long, Jianmin Wang, and Yue Cao, "Deep hashing network for efficient similarity retrieval.," in *AAAI*, 2016, pp. 2415–2421.

[7] Qi Li, Zhenan Sun, Ran He, and Tieniu Tan, "Deep supervised discrete hashing," *arXiv preprint arXiv:1705.10999*, 2017.

[8] Jian Zhang, Yuxin Peng, and Junchao Zhang, "Ssdh: semi-supervised deep hashing for large scale image retrieval," *arXiv preprint arXiv:1607.08477*, 2016.

[9] Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1556–1564.

[10] Dayan Wu, Zheng Lin, Bo Li, Mingzhen Ye, and Weiping Wang, "Deep supervised hashing for multi-label and large-scale image retrieval," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 2017, pp. 150–158.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] Jie Lin, Olivier Morère, Antoine Veillard, Ling-Yu Duan, Hanlin Goh, and Vijay Chandrasekhar, "Deephash for image instance retrieval: Getting regularization, depth and fine-tuning right," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 2017, pp. 133–141.

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[14] Mark J Huiskes and Michael S Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 2008, pp. 39–43.

[15] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng, "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*. ACM, 2009, p. 48.

[16] Yue Cao, Mingsheng Long, Jianmin Wang, Han Zhu, and Qingfu Wen, "Deep quantization network for efficient image retrieval.," in *AAAI*, 2016, pp. 3457–3463.

[17] Yue Cao, Mingsheng Long, Jianmin Wang, and Shichen Liu, "Deep visual-semantic quantization for efficient image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1328–1337.

[18] Yair Weiss, Antonio Torralba, and Rob Fergus, "Spectral hashing," in *Advances in neural information processing systems*, 2009, pp. 1753–1760.

[19] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2916–2929, 2013.

[20] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.