DEEP RANK CROSS-MODAL HASHING WITH SEMANTIC CONSISTENT FOR IMAGE-TEXT RETRIEVAL

Xiaoqing Liu[1] *Huanqiang Zeng*[2,†] *Yifan Shi*[2] *Jianqing Zhu*[2] *Kai-Kuang Ma*[3]

[1]School of Information Science and Engineering, Huaqiao University, China
[2]School of Engineering, Huaqiao University, China
[3]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

ABSTRACT

Cross-modal hashing retrieval approaches maps heterogeneous multi-modal data into a common hamming space to achieve efficient and flexible retrieval performance. However, existing cross-modal methods mainly exploit feature-level similarity between multi-modal data, the label-level similarity and relative ranking relationship between adjacent instances have been ignored. To address these problems, we propose a novel **D**eep **R**ank **C**ross-modal **H**ashing(**DRCH**) method that fully explores the intra-modal semantic similarity relationship. Firstly, DRCH preserves semantic similarity by combining both label-level and feature-level information. Secondly, the inherent gap between modalities are narrowed by proposing a ranking alignment loss function. Finally, the compact and efficient hash codes are optimized from the common semantic space. Extensive experiments on two real-world image-text retrieval datasets demonstrate the superiority of DRCH compared with several state-of-the-art(SOTA) methods.

Index Terms— Cross-modal hashing, Rank learning, Heterogeneous gap, Intra-modal similarity

1. INTRODUCTION

With the development of the online social networks, the resulting multi-modal data is growing explosively, most internet users face a challenge to fetch information of interest effectively and efficiently because of the heterogeneous properties of these multi-modal data. Image-text retrieval is one of the most attractive researches in multi-modal data analysis and provides flexible retrieval services between images and text. The core issue of image-text retrieval lies in tackling the semantic gaps between visual and textual data effectively and efficiently [1].

Cross-modal retrieval with hashing methods (CMH) [2, 3, 4, 5, 6] significantly improve the retrieval efficiency on both speed and storage by mapping large-scale and high-dimensional multi-modal media data into a common Hamming space. Traditional cross-modal hashing usually captures the semantic relevance with hand-crafted features in the common representation space [7, 8]. For example, Kumar el al. [9] proposed a cross-view similarity hashing (CVH) aiming at extending spectral hashing from uni-modal to multi-modal scenarios. Xu et al.[10] proposed a semantic correlation maximization approach (SCM) that takes advantage of the supervisory information and maximizes the semantic correlation between different modalities. The semantics-preserving hashing (SePH)[11] generated a unified binary coding scheme by simultaneously calculating the affinity matrix with probability distribution and reducing the Kullbach-Leibler divergence. Zhang et al. [12] directly use semantic tags to transforms heterogeneous data into a modal specific latent semantic representation for guiding hash learning.

In recent years, the content-based CMH combining deep learning technique has been investigated to yield remarkable performance improvement. Existing studies can be roughly divided into two categories: supervised and unsupervised approaches. Unsupervised approaches mainly utilize the underlying data structures and distribution to learn the hash coding function [13, 14, 15], while supervised approaches try to leverage supervised information (i.e. category labels) to further improve the retrieval effectiveness. Deep cross-modal hashing (DCMH) [2] performed an end-to-end learning framework that used a negative log-likelihood loss to preserve semantic information. Inspired by adversarial learning, self-supervised adversarial hashing (SSAH)[3] was proposed to discover the semantic correlation and consistency of data representations between different modalities via an adversarial network. Zhang et al.[16] leveraged adversarial learning and attentional mechanisms to highlight the key information in semantic expression. Shi et al. [17] proposed

This work was supported in part by the National Key R&D Program of China under the grant 2021YFE0205400, in part by the National Natural Science Foundation of China under the grants 61871434 and 61976098, in part by the Natural Science Foundation for Outstanding Young Scholars of Fujian Province under the grant 2019J06017, in part by the NTU-WASP Joint Project under Grant M4082184, in part by the Fujian-100 Talented People Program, and in part by the Key Science and Technology Project of Xiamen City under the grant 3502ZCQ20191005 *(†Corresponding author: Huanqiang Zeng, Email: zeng0043@hqu.edu.cn).*

the equally guided discriminative hashing (EGDH) that cooperated hashing-based retrieval with classification by a unified deep learning framework to simultaneously preserve semantic structure and discriminative hash codes.Zuo et al.[18] use ReLU transform to unify the similarity of the learned hash representation and the multi-label semantic similarity corresponding to the original instance to mine the soft semantic information in tags.

However, the state-of-the-art image-text hashing methods have the following limitations: Most existing methods measure the feature-level similarity between instances, while ignoring the label-level similarity in the multi-label scenario. Although label semantic information is explored by adversarial-based methods, the relative ranking relationship between adjacent instances is rarely considered. In order to address these problems, we propose a novel Deep Rank Cross-modal Hashing (DRCH) method, which takes into account the relative ranking relationship from both label-level and feature-level similarity. By designing the additional ranking alignment loss function bridging the heterogeneity gap, DRCH successfully achieves overall better performance on the real-world image-text retrieval tasks. For clarity, the contributions of this paper are summarized as follows:

1. A ranking alignment loss function is designed to capture the relative ranking relationship by preserving semantic similarity between different modalities from both label and feature levels.

2. A novel Deep Rank Cross-modal Hashing (DRCH) method is proposed by taking the ranking alignment loss, the deep semantic consistence loss and the hash quantization loss into consideration. The joint optimization process with an end-to-end way is compatible and convenient for network training and hash learning.

3. Extensive experiments on two real-world image-text retrieval datasets (i.e. MIRFLICKR-25K and NUS-WIDE) demonstrate the superiority of the proposed DRCH compared with several state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 introduces the proposed method in detail. In Section 3, comparative experiments and analysis are conducted. Finally, we conclude our work and discuss future directions in Section 4.

2. PROPOSED METHOD

2.1. Problem Definition

We first define the task of image-text retrieval and the notations involved in this paper. Given a query of one modality (i.e. image or text), the goal of image-text retrieval is to retrieve the most similar instance of another modality. Let

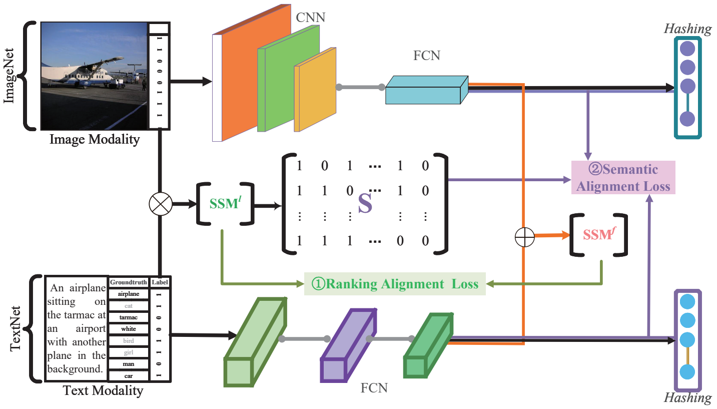

Fig. 1. The overall structure of the proposed DRCH.

$Z = \{z_i\}_{i=1}^n$ denotes the cross-modal dataset with n instances where $z_i = (x_i, y_i, l_i)$. Each instance z_i is represented by both visual raw features $x_i \in \mathbb{R}^{1\times d_X}$ and textual raw features $y_i \in \mathbb{R}^{1\times d_Y}$. The task lies in the multi-label scenario, thus $l_i = \{l_{i1}, l_{i2}, \cdots, l_{ik}\}$ denotes the multi-label annotation assigned to z_i, where k is the number of classes. Specifically, if z_i belongs to the j^{th} class, $l_{ij} = 1$, otherwise $l_{ij} = 0$. The image-feature matrix, text-feature matrix and label matrix are defined as $\mathbf{X} = \{x_i\}_{i=1}^n \in \mathbb{R}^{n\times d_X}$, $\mathbf{Y} = \{y_i\}_{i=1}^n \in \mathbb{R}^{n\times d_Y}$ and $\mathbf{L} = \{l_i\}_{i=1}^n$ for all instances, respectively. The label-level semantic similarity matrix (SSM^l), pairwise multi-label similarity matrix S and feature-level semantic similarity matrix (SSM^f) are defined as follows:

$$\text{SSM}_{ij}^l = l_i \cdot l_j^T \tag{1}$$

$$S_{ij} = \begin{cases} 1 & \text{SSM}_{ij}^l > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$\text{SSM}_{ij}^f = \begin{cases} x_i \cdot y_j^T & \text{Querying images} \\ y_i \cdot x_j^T & \text{Quering texts} \end{cases} \tag{3}$$

It can be seen that SSM^l and SSM^f measure the semantic similarity between different instances and modalities from label and feature levels, respectively. Besides, S with binary values describes the semantic relationship between instances.

2.2. Overall Structure

In this section, we first describe the overall structure of DRCH, followed by introducing the ranking alignment, the deep semantic consistence and the hash code learning.

The architecture of DRCH is shown in Fig.1, which mainly includes two modules: ranking alignment loss module and semantic alignment loss module. Firstly, SSM^l and SSM^f are computed by label inner product and feature inner product, respectively. Then the ranking alignment loss is constructed by taking both SSM^l and SSM^f into consideration, while the semantic alignment loss is related to deep semantic consistence and hash quantization. Finally, DRCH provides a unified objective function combining the above loss functions to optimizes the networks.

Two deep neural networks are applied to learn the latent feature spaces for the image and text modalities, respectively. Specifically, the commonly-used deep convolutional network can be employed to capture image features. In the experiments, we choose VGG-16[19] network composed of 13 convolutional layers (*Conv*), 5 pooling layers and 2 fully-connected layers (*ImgFcn*), where the backbone network is pre-trained on ILSVRC[20] for being effective in capturing common semantic information of images, and *ImgFcn* outputs the well-learned image features. In order to extract better features from text modality, we embed textual data into bag-of-words (BOW) representations and then feed them to a two-layer fully-connected network (*TxtFcn*).

In practice, the hash codes with length C (i.e. $B^{x,y} \in \{-1,1\}^C$) are generated from the two modalities. Binary codes $B^{x,y}$ are generated by applying a sign function to $H^{x,y}$ (i.e. $H^* = \mathcal{H}^*(*;\theta^*), * \in \{\mathbf{X},\mathbf{Y}\}$), which can be defined as follows:

$$B^{x,y} = sign(H^{x,y}) \in \{-1,1\}^C \tag{4}$$

where $\mathcal{H}^{x,y}$ is a hash function, and $\theta^{x,y}$ denotes the network parameters to be learned.

2.3. Ranking and Semantic Alignments

We design a ranking alignment loss in the training process to explore the semantic structure information in intra-modal data and inter-modal data. In other words, the ranking relationship between instances should be preserved from both label and feature levels. Given a query q_k and retrieved set $\{r_i^{q_k}\}_{k=1}^m$ with m instances, the ranking alignment loss function is defined as follows:

$$\begin{aligned}\mathcal{L}_k^{rank} = -\sum_{i,j\in m, i\neq j}&\{t_{ij}\times\log(\sigma(\mathrm{SSM}_{ki}^f - \mathrm{SSM}_{kj}^f))\\ &+(1-t_{ij})\times\log(1-\sigma(\mathrm{SSM}_{ki}^f-\mathrm{SSM}_{kj}^f))\}\end{aligned} \tag{5}$$

t_{ij} represents the partially ordered relationship between SSM_{ki}^l and SSM_{kj}^l, which is defined as follows:

$$t_{ij} = \begin{cases} 1 & \mathrm{SSM}_{ki}^l > \mathrm{SSM}_{kj}^l \\ 0.5 & \mathrm{SSM}_{ki}^l = \mathrm{SSM}_{kj}^l \\ 0 & \mathrm{SSM}_{ki}^l < \mathrm{SSM}_{kj}^l \end{cases} \tag{6}$$

where $\sigma(u)$ is the sigmoid function as follows:

$$\sigma(u) = \frac{1}{1+e^{-u}} \tag{7}$$

In addition, the semantic alignment loss function is defined as follows:

$$\begin{aligned}\mathcal{L}^{sim} = &-\sum_{i,j=1}^{n}(S_{ij}\Theta_{ij} - \log(1+e^{\Theta_{ij}}))\\ &+\eta(\|B^x-\mathbf{X}\|_F^2 + \|B^y-\mathbf{Y}\|_F^2)\\ &s.t. B^* \in \{-1,1\}\end{aligned} \tag{8}$$

where $\Theta_{ij} = \frac{1}{2}\mathbf{X}_{*i}^T\mathbf{Y}_{j*}$. The first term $(S_{ij}\Theta_{ij} - \log(1+e^{\Theta_{ij}}))$ reflects the deep semantic consistence that similar instances should keep the adjacency relationship in the latent feature space. The second item $\|B^x-\mathbf{X}\|_F^2 + \|B^y-\mathbf{Y}\|_F^2$ is the hash quantization loss, which contributes to generating more compact hash codes for both modalities. In order to reduce the information loss caused by hash process, we add an L2-norm as a regularization term as follows:

$$\mathcal{L}^{regular} = \|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2 \tag{9}$$

In general, the final objective function for ImageNet and TextNet is formulated as follows:

$$\min_{B^{x,y},\theta^{x,y}} \mathcal{L} = \alpha\sum_k \mathcal{L}_k^{rank} + \beta\mathcal{L}^{sim} + \gamma\mathcal{L}^{regular} \tag{10}$$

where $B^{x,y}$ is the hash code generated by ImageNet or TextNet, and $\theta^{x,y}$ is the corresponding network parameters. Empirically, we choose $\alpha = 0.7, \beta = 0.3$ and $\gamma = 0.2$ and apply random gradient descent (SGD) algorithm to alternately optimize the networks in different modalities.

3. EXPERIMENTS

3.1. Experimental Setting

In this section, we test the performance of DRCH and several comparison approaches on two real-world image-text retrieval datasets including MIRFLICKR-25K[21] and NUS-WIDE[22]. MIRFLICKR-25K contains 25,000 instances collected from Flickr, while NUS-WIDE has over 269,000 images with over 5000 user-provided textual tags.

Mean Average Precision (MAP) is commonly used to measure the performance of cross-modal retrieval methods by jointly considering the ranking information and precision. The larger the MAP value, the better the retrieval performance. We repeat every method 20 times and report the average MAP scores for two different image-text retrieval tasks: retrieving texts by querying images (I → T) and retrieving images by querying texts (T → I). For MIRFLICKR, the learning rate is within 1.5e-5 and the batch size is 256. For NUS-WIDE, the learning rate is within 5e-6 and the batch size is 128.

3.2. Comparative Experiments

In order to verify the effectiveness of DRCH, we compare several baselines and state-of-the-art image-text retrieval hashing methods including CVH [9], SCM [10], SePH [11], DCMH [2], SSAH [3], MLCAH [6] and EGDH [17]. Among them, CVH, SCM and SePH are based on hand-craft features, while DCMH, SSAH, EGDH and MLCAH are based on deep learning techniques. Table 1 and Table 2 provide the experimental results of two retrieval tasks on MIRFLICKR-25K, while Table 3 and Table 4 provide the experimental results on NUS-WIDE. Each dataset is splited in the same way as DCMH, the performance of the above comparison methods is directly cited from the original papers.

From the experimental results, we have the following observations: 1) DRCH achieves overall better performance than the comparison methods on the two retrieval tasks of both datasets. Specifically, DRCH outperforms the second-best methods with an improvement of 4%, 2.6%, 0.4% and 0.5% in terms of average MAP score. 2) DRCH with different lengths of hash codes have more competitive and robust retrieval performance. 3) Compared with DCMH that only optimizes the semantic consistence loss and the hash quantization loss, DRCH considers the additional ranking alignment loss and significantly improves the performance.

Table 1. THE MEAN AVERAGE PRECISION COMPARISON IN THE TASK OF IMAGE AS QUERIES(I $\rightarrow$ T) ON DATASET MIRFLICKR-25K. THE LENGTH OF HASH CODE ARE 16BIT, 32BIT AND 64BIT RESPECTIVELY.

Method	16bit	32bit	64bit	Avg
CVH[9]	0.557	0.554	0.554	0.555
SCM[10]	0.671	0.682	0.685	0.679
SePH[11]	0.657	0.660	0.671	0.659
DCMH[2]	0.735	0.737	0.750	0.741
SSAH[3]	0.782	0.790	0.800	0.791
EDGH[17]	0.782	0.790	0.800	0.791
MLCAH[6]	0.782	0.790	0.800	0.791
DRCH	**0.803**	**0.822**	**0.868**	**0.831**

Table 2. THE MEAN AVERAGE PRECISION COMPARISON IN THE TASK OF IMAGE AS QUERIES(T $\rightarrow$ I) ON DATASET MIRFLICKR-25K. THE LENGTH OF HASH CODE ARE 16BIT, 32BIT AND 64BIT RESPECTIVELY.

Method	16bit	32bit	64bit	Avg
CVH[9]	0.557	0.554	0.554	0.555
SCM[10]	0.671	0.707	0.713	0.706
SePH[11]	0.648	0.652	0.654	0.651
DCMH[2]	0.763	0.764	0.775	0.767
SSAH[3]	0.791	0.795	0.803	0.796
EDGH[17]	0.779	0.794	0.799	0.791
MLCAH[6]	0.794	0.805	**0.805**	0.801
DRCH	**0.846**	**0.836**	0.798	**0.827**

Table 3. THE MEAN AVERAGE PRECISION COMPARISON IN THE TASK OF IMAGE AS QUERIES(I $\rightarrow$ T) ON DATASET NUS-WIDE. THE LENGTH OF HASH CODE ARE 16BIT, 32BIT AND 64BIT RESPECTIVELY.

Method	16bit	32bit	64bit	Avg
CVH[9]	0.400	0.392	0.386	0.392
SCM[10]	0.533	0.548	0.557	0.546
SePH[11]	0.478	0.487	0.489	0.485
DCMH[2]	0.478	0.486	0.488	0.484
SSAH[3]	0.642	0.636	0.639	0.639
EDGH[17]	-	-	-	-
MLCAH[6]	0.644	0.641	**0.643**	0.643
DRCH	**0.655**	**0.647**	0.640	**0.647**

Table 4. THE MEAN AVERAGE PRECISION COMPARISON IN THE TASK OF IMAGE AS QUERIES(T $\rightarrow$ I) ON DATASET NUS-WIDE. THE LENGTH OF HASH CODE ARE 16BIT, 32BIT AND 64BIT RESPECTIVELY.

Method	16bit	32bit	64bit	Avg
CVH[9]	0.372	0.366	0.363	0.367
SCM[10]	0.463	0.462	0.471	0.465
SePH[11]	0.449	0.454	0.458	0.454
DCMH[2]	0.638	0.651	0.657	0.649
SSAH[3]	0.669	0.662	0.666	0.666
EDGH[17]	-	-	-	-
MLCAH[6]	0.662	0.673	0.687	0.674
DRCH	**0.675**	**0.677**	**0.694**	**0.679**

4. CONCLUSION

In this work, we propose a novel Deep Rank Cross-modal Hashing (DRCH) method, which takes into account the relative ranking relationship from both label-level and feature-level similarity, and bridges the heterogeneous gap between different modalities by designing an additional ranking alignment loss. Finally, DRCH jointly optimizes the ranking alignment loss, the deep semantic consistence loss and the hash quantization loss to maintain the semantic relevance and capture the feature distribution consistency between different modalities. The extensive experiments demonstrate that DRCH achieves overall better retrieval performance on two benchmark datasets compared with multiple state-of-the-art methods.

5. REFERENCES

[1] M. B. Kokare and M. S. Shirdhonkar, "Document image retrieval: An overview," *International Journal of Computer Applications*, vol. 1, no. 7, pp. 128–130, 2010.

[2] Q. Y. Jiang and W. J. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3232–3240.

[3] C. Li, C. Deng, N. Li, W. Liu, X. B. Gao, and D. C. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4242–4251.

[4] X. Lu, L. Zhu, Z. Y. Cheng, J. J. Li, X. S. Nie, and H. X. Zhang, "Flexible online multi-modal hashing for large-scale multimedia retrieval," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1129–1137.

[5] D. Xie, C. Deng, C. Li, X. L. Liu, and D. C. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 29, pp. 3626–3637, 2020.

[6] X. H. Ma, T. Z. Zhang, and C. S. Xu, "Multi-level correlation adversarial hashing for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3101–3114, 2020.

[7] D. X. Wang, P. Cui, M. D. Ou, and W. W. Zhu, "Learning compact hash codes for multimodal representations using orthogonal deep structure," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1404–1416, 2015.

[8] P. C. Zhang, W. Zhang, W. J. Li, and M. Y. Guo, "Supervised hashing with latent factor models," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 173–182.

[9] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Twenty-second international joint conference on artificial intelligence*, 2011.

[10] D. Q. Zhang and W. J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proceedings of the AAAI conference on artificial intelligence*, 2014, vol. 28.

[11] Z. J. Lin, G. G Ding, M. Q. Hu, and J. M. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3864–3872.

[12] D. L. Zhang, X. J. Wu, and J. Yu, "Label consistent flexible matrix factorization hashing for efficient cross-modal retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 3, pp. 1–18, 2021.

[13] S. Liu, S. S. Qian, Y Guan, J. W. Zhan, and L. Ying, "Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1379–1388.

[14] M. Y. Li, Q. Q. Li, L. R. Tang, S. Peng, Y. Ma, and D. G. Yang, "Deep unsupervised hashing for large-scale cross-modal retrieval using knowledge distillation model," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[15] M. Y. Li and H. Y. Wang, "Unsupervised deep cross-modal hashing by knowledge distillation for large-scale cross-modal retrieval," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 2021, pp. 183–191.

[16] X. Zhang, H. J. Lai, and J. S. Feng, "Attention-aware deep adversarial hashing for cross-modal retrieval," in *Proceedings of the European conference on computer vision*, 2018, pp. 591–606.

[17] Y. F. Shi, X. G. You, F. Zheng, S. Wang, and Q. M. Peng, "Equally-guided discriminative hashing for cross-modal retrieval," in *International Joint Conference on Artificial Intelligence*, 2019, pp. 4767–4773.

[18] X. Zou, S. Wu, E. M. Bakker, and X. Z. Wang, "Multi-label enhancement based self-supervised deep cross-modal hashing," *Neurocomputing*, vol. 467, pp. 138–162, 2022.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2009.

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

[22] T.-S. Chua, J. H. Tang, R. C. Hong, H. J. Li, Z. P. Luo, and Y. T. 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*, 2009, pp. 1–9.