Deep Supervised Hashing for Fast Image Retrieval

Haomiao Liu[1,2], Ruiping Wang[1], Shiguang Shan[1], Xilin Chen[1]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
haomiao.liu@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

In this document we will give additional experimental results of the corresponding sections in the main paper to support the method we proposed.

1. Evaluation of the Regularizer

This section gives more experimental results of Section 4.2 in the main paper (all results were obtained with 12-bit binary codes). Figure 1 shows the distribution of network outputs on the test set of NUS-WIDE. Table 1 gives retrieval performances (precision within Hamming radius 2) on both datasets. The PR curves of the models are given in Figure 2. These results are consistent with the results in the main paper, thus detailed discussion would not be repeated.

2. Finetuning vs. Training From Scratch

This section gives additional results of Section 4.4 in the main paper. Comparison of the finetuned models and the models trained from scratch are given in Table 2 and Figure 3. Besides, the convergence curves of the 48-bit models trained on NUS-WIDE is provided in Figure 4. Table 2 and Figure 3 are consistent with the results in the main paper, thus no more discussion is presented here. The train/test loss in Figure 4 indicates that the model trained from scratch shows no sign of overfitting, however, the retrieval performance of this model is inferior to the finetuned model. Since the finetuning process is somehow similar to training the "scratch model" with several thousands of additional iterations, increasing the number of training iterations might improve the model trained from scratch (while requiring more computation). The above observations

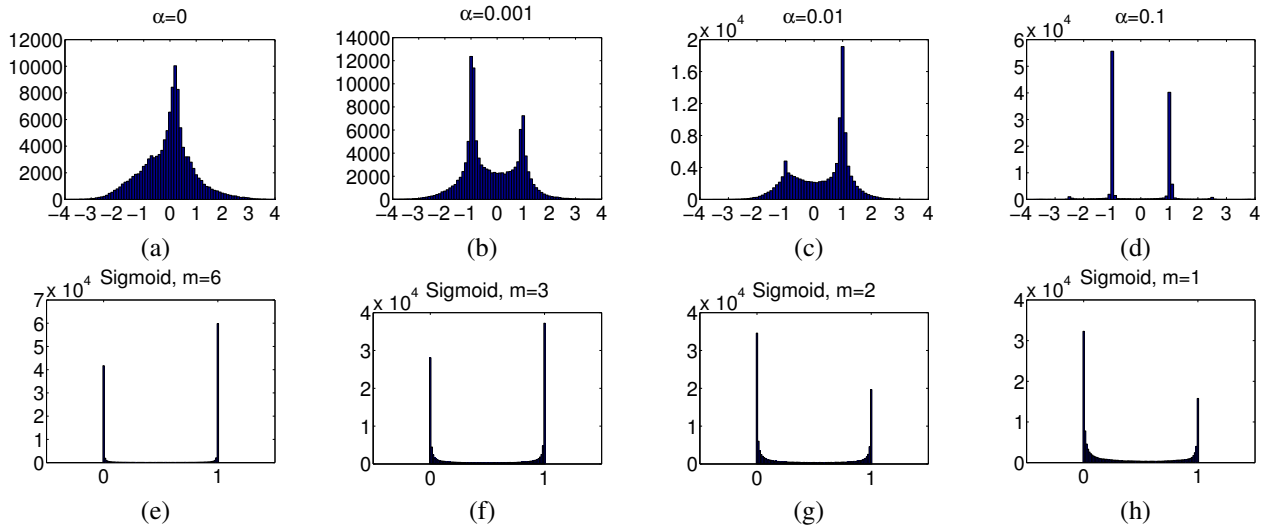
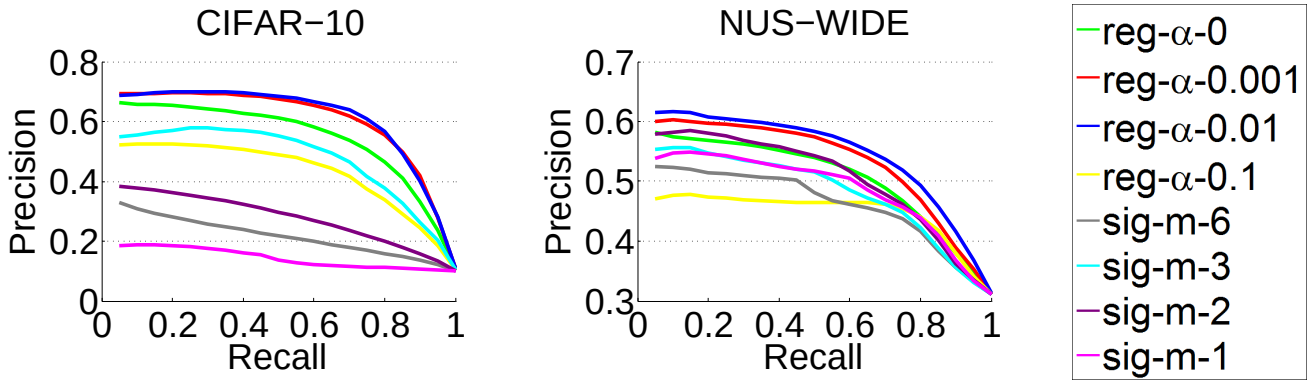


Figure 1. Distribution of network outputs on the test query set of NUS-WIDE. (a)-(d) the models using our proposed regularizer under different settings of α, (e)-(h) the sigmoid relaxed models under different settings of m.

Models	CIFAR-10	NUS-WIDE
Regularizer-α-0	0.5604	0.5572
Regularizer-α-0.001	0.6195	0.5716
Regularizer-α-0.01	0.5671	0.5853
Regularizer-α-0.1	0.4418	0.4275
Sigmoid-m-6	0.2903	0.3216
Sigmoid-m-3	0.4224	0.4034
Sigmoid-m-2	0.3441	0.5418
Sigmoid-m-1	0.1783	0.5243

Table 1. Retrieval performances (precision within Hamming radius 2) of models under different settings of α, relaxation, and m. The results are obtained with 12-bit binary codes.



(a) (b)

Figure 2. Comparison of precision-recall (PR) curves of the models with different settings of α, relaxation, and m. The results are obtained with 12-bit binary codes. (a) CIFAR-10, (b) NUS-WIDE.

	Code Length	CIFAR-10	NUS-WIDE
Trained From Scratch	12	0.5671	0.5853
	24	0.6774	0.5157
	36	0.6321	0.3902
	48	0.5690	0.3572
Finetuned	24	0.6731	0.5353
	36	0.6377	0.4409
	48	0.5791	0.3805

Table 2. Comparison of retrieval performances (precision within Hamming radius 2) of the models trained from scratch and the finetuned models.

along with the results in the main paper validate that our finetuning scheme is beneficial to training the CNN models in terms of efficiency and avoiding overfitting.

Moreover, we give the detailed performances of the ensemble models in Figure 5 and Table 3. The four 12-bit models with different initializations are denoted as M1, M2, M3, and M4 respectively. In Figure 5, the ensemble models compared are "M3 + M4", "M1 + M3 + M4", and "M1 + M2 + M3 + M4" on CIFAR-10 and "M2 + M3", "M1 + M2 + M3", and "M1 + M2 + M3 + M4" on NUS-WIDE respectively (the choice of these combinations is based on the performances listed in Table 3, the best models in terms of mAP were chosen). We can see that the ensemble binary codes outperforms the finetuned codes in most cases. Nevertheless, the ensemble codes were not adopted in the main paper for efficiency consideration. As discussed in our conclusion part, we have been conducting further study on this point for possible speed up strategy.

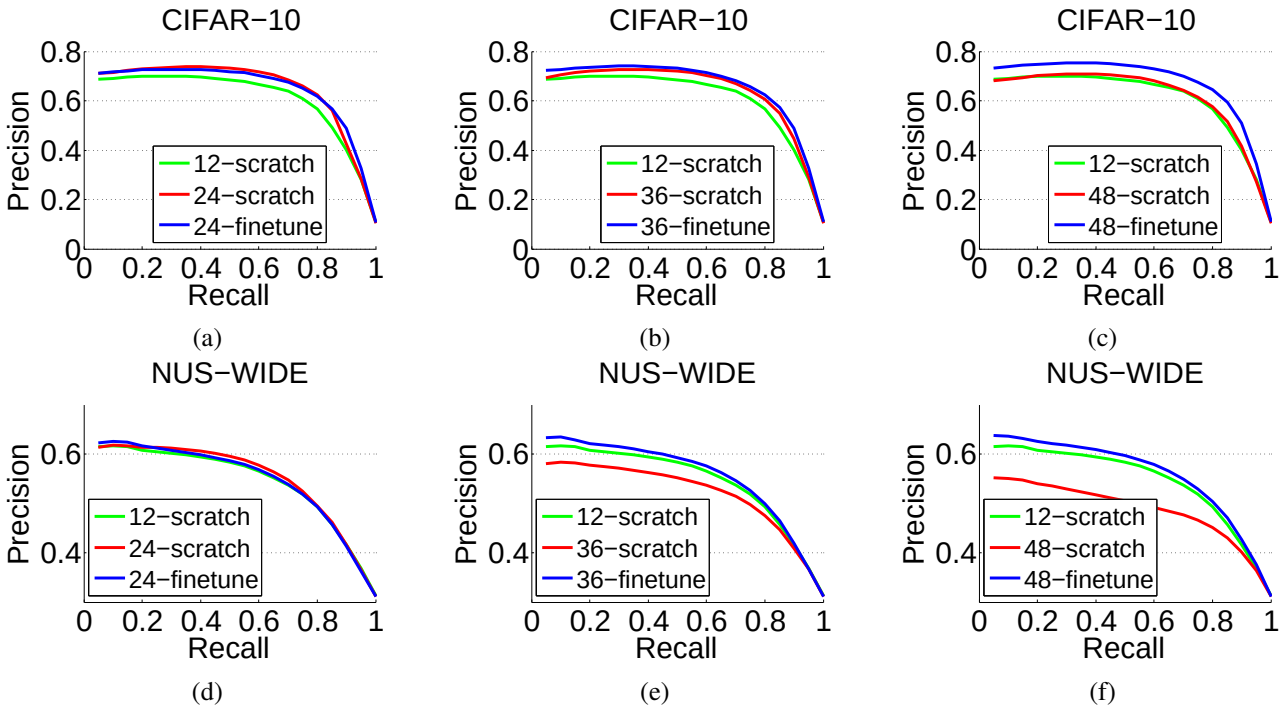


Figure 3. Comparison of the finetuned models against the models trained from scratch under different code lengths (PR curves).

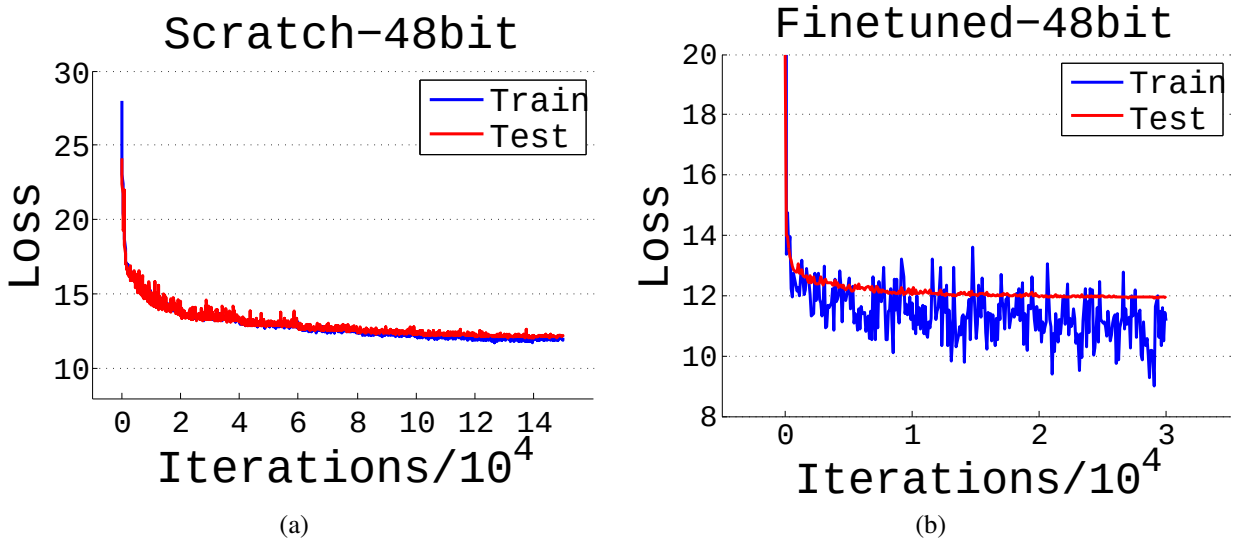


Figure 4. Comparison of (a) the model trained from scratch and (b) the finetuned model in terms of training/test loss. Both models produce 48-bit binary codes, and are trained on the NUS-WIDE dataset.

Model(s)	mAP		Precision within radius 2	
	CIFAR-10	NUS-WIDE	CIFAR-10	NUS-WIDE
M1	0.6157	0.5483	0.5671	0.5853
M2	0.6126	0.5585	0.6202	0.5953
M3	0.6358	0.5588	0.6397	0.5924
M4	0.6289	0.5353	0.6237	0.5756
M1 + M2	0.6685	0.5751	0.6814	0.5720
M1 + M3	0.6826	0.5777	0.6978	0.5722
M1 + M4	0.6755	0.5628	0.6886	0.5560
M2 + M3	0.6806	0.5794	0.6919	0.5776
M2 + M4	0.6740	0.5690	0.6857	0.5716
M3 + M4	0.6850	0.5722	0.6979	0.5752
M1 + M2 + M3	0.6996	0.5870	0.6181	0.5035
M1 + M2 + M4	0.6937	0.5785	0.6154	0.4913
M1 + M3 + M4	0.7022	0.5811	0.6292	0.4931
M2 + M3 + M4	0.7013	0.5832	0.6197	0.5073
M1 + M2 + M3 + M4	0.7106	0.5882	0.5486	0.4525

Table 3. Retrieval performances (mAP and precision within Hamming radius 2) of network ensembles. The four models evaluated are denoted as M1, M2, M3, and M4 respectively. Each one of these four models produces 12-bit binary codes.

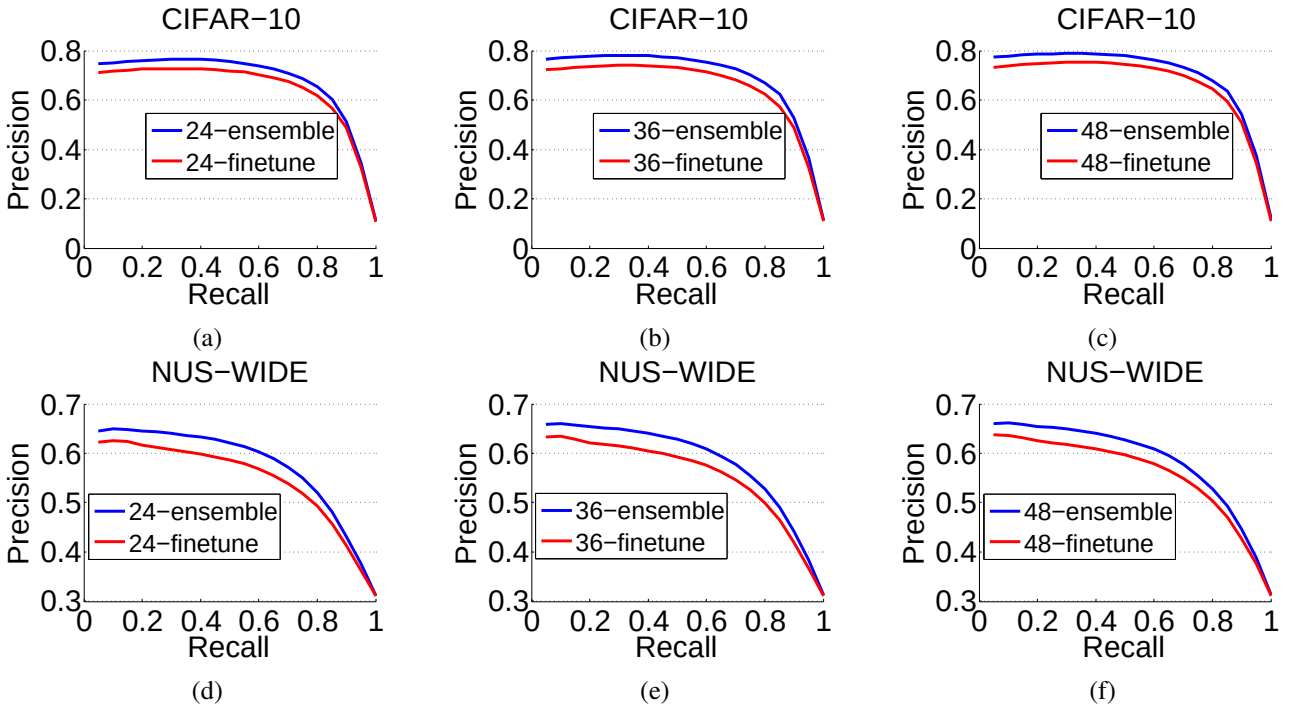
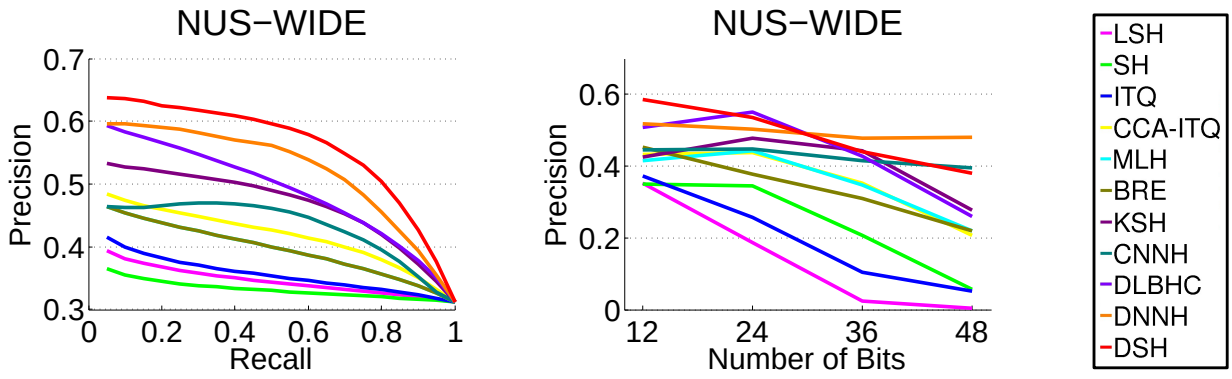


Figure 5. Comparison of the finetuned codes and the ensemble codes on both datasets under different code lengths (a)∼(c) CIFAR-10, (d)∼(f) NUS-WIDE. The models used in the ensembles are: (a) M3 + M4, (b) M1 + M3 + M4, (c) all four models, (d) M2 + M3, (e) M1 + M2 + M3, (f) all four models.

	LSH	CCA-ITQ	KSH	DSH
Hand	0.1492	0.2176	0.4167	-
CNN-cls (b)	0.1783	0.5443	0.5949	-
CNN-ours (a)	0.3154	0.6576	0.6498	0.6755

Table 4. Retrieval mAP on CIFAR-10 with 48-bit binary codes. The conventional hashing methods were trained with different features, including **Hand**: hand-crafted features (i.e. 512-D GIST features, as reported in our main paper), **CNN-cls**: CNN features from the classification model, and **CNN-ours**: CNN features from our 12-bit model.



(a) (b)

Figure 6. Comparison of retrieval performances of our DSH method and the other hashing methods on NUS-WIDE. (a) PR curves (48-bit). (b) Mean precision within Hamming radius 2.

3. Comparison with the State-of-the-art

This section corresponds to Section 4.5 in the main paper. The comparison of state-of-the-art hashing methods and our proposed DSH method on NUS-WIDE are shown in Figure 6 (PR curves and precision within Hamming radius 2). The results are similar to the ones on CIFAR-10, thus the discussions are not repeated here.

We also tested the performance of 48-bit codes generated by LSH, CCA-ITQ, and KSH on CIFAR-10 using two kinds of CNN features, *i.e.* the 500-D network activations of the first FC layer, L2 normalized, extracted from (a) our 12-bit model, and (b) a model with the same preceding layers but trained for classification task (obtained by replacing the output layer of our model with a 10-way softmax loss layer), which achieves 74.54% accuracy on the test set. Results are shown in Table 4. The performances of conventional methods improve significantly with CNN features (even comparable to our method), and the features from our model are superior to the ones from the classification model, validating again our motivation for learning binary codes in an end-to-end manner.

Some real failed/successful retrieval cases (3 failed cases and 3 successful cases for each dataset, the true matches are bounded by red frames) on both datasets are provided in Figure 7 to 18 (obtained by 48-bit binary codes, failed: 7, 8, 9, 13, 14, 15, successful: 10, 11, 12, 16, 17, 18). For space limitation, only the top-10 feedbacks are shown here. The number of true matches of each method is listed beside the retrieval results, and the labels of query images are also provided. We can observe that in the failed cases, although the top feedbacks are inconsistent with the query image, the feedbacks themselves are similar to each other. This observation suggests that our method is able to preserve the similarity of images, and can be further improved in terms of recognizing the semantic meanings of query images.

Figure 7. A **failed** retrieval case on CIFAR-10, only the top-10 feedbacks are shown due to space limitation. Results were obtained with 48-bit binary codes. Images with red frames are true matches.

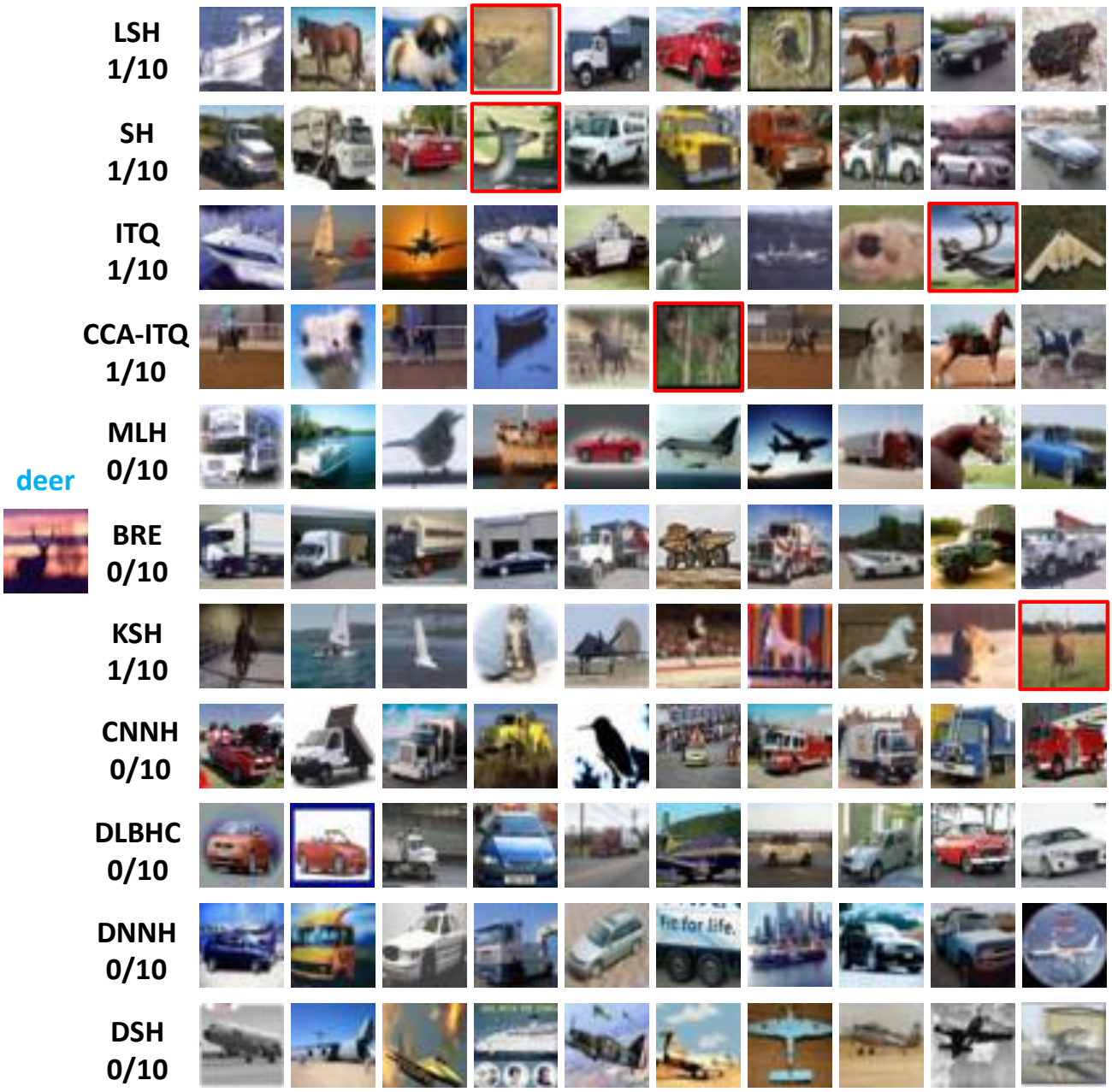


Figure 8. A **failed** retrieval case on CIFAR-10, only the top-10 feedbacks are shown due to space limitation. Results were obtained with 48-bit binary codes. Images with red frames are true matches.

Figure 9. A **failed** retrieval case on CIFAR-10, only the top-10 feedbacks are shown due to space limitation. Results were obtained with 48-bit binary codes. Images with red frames are true matches.

Figure 10. A **successful** retrieval case on CIFAR-10, only the top-10 feedbacks are shown due to space limitation. Results were obtained with 48-bit binary codes. Images with red frames are true matches.

Figure 11. A **successful** retrieval case on CIFAR-10, only the top-10 feedbacks are shown due to space limitation. Results were obtained with 48-bit binary codes. Images with red frames are true matches.

Figure 12. A **successful** retrieval case on CIFAR-10, only the top-10 feedbacks are shown due to space limitation. Results were obtained with 48-bit binary codes. Images with red frames are true matches.

Figure 13. A **failed** retrieval case on NUS-WIDE, only the top-10 feedbacks are shown due to space limitation. Results were obtained with 48-bit binary codes. Images with red frames are true matches.

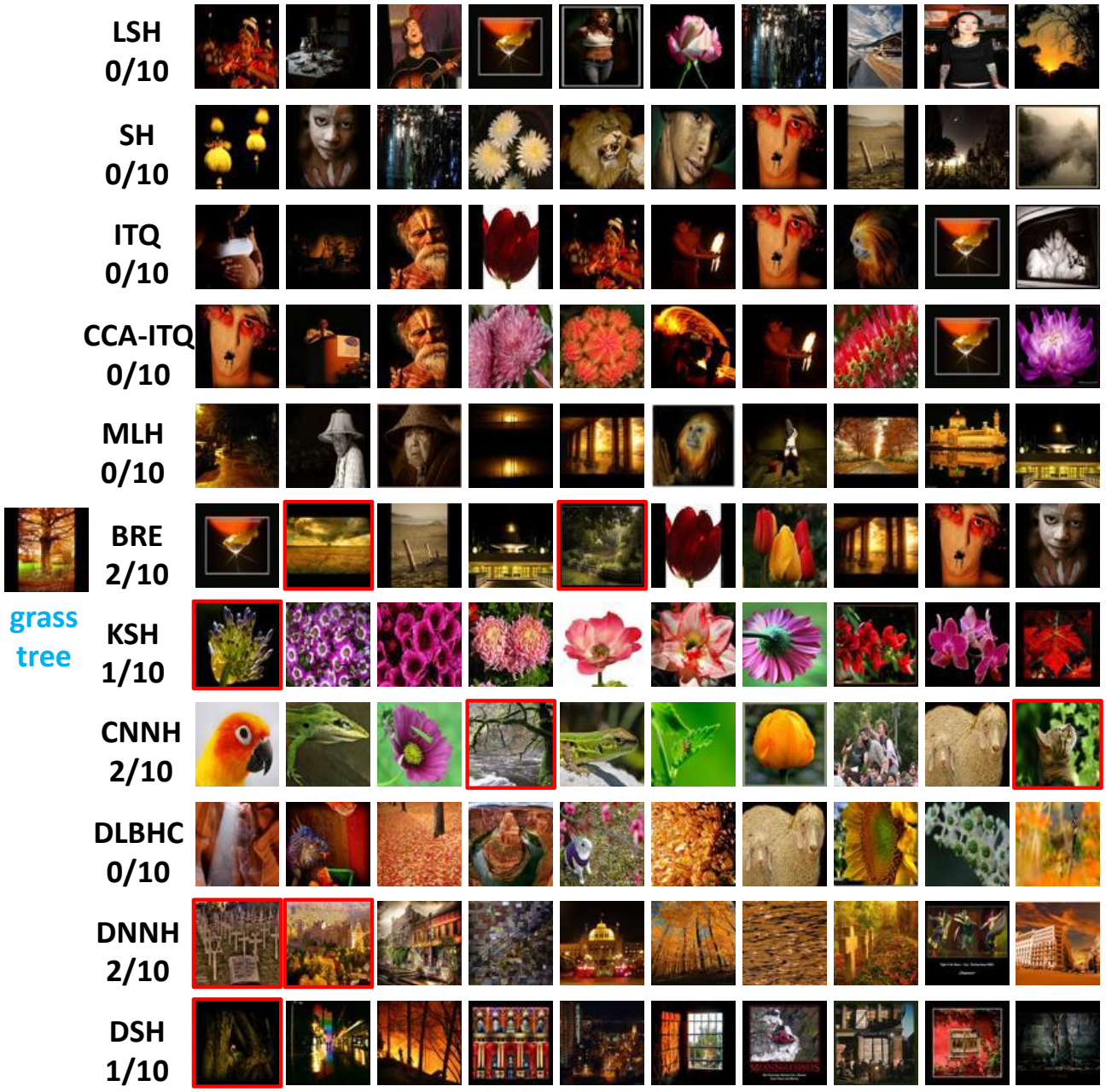


Figure 14. A **failed** retrieval case on NUS-WIDE, only the top-10 feedbacks are shown due to space limitation. Results were obtained with 48-bit binary codes. Images with red frames are true matches.

Figure 15. A **failed** retrieval case on NUS-WIDE, only the top-10 feedbacks are shown due to space limitation. Results were obtained with 48-bit binary codes. Images with red frames are true matches.

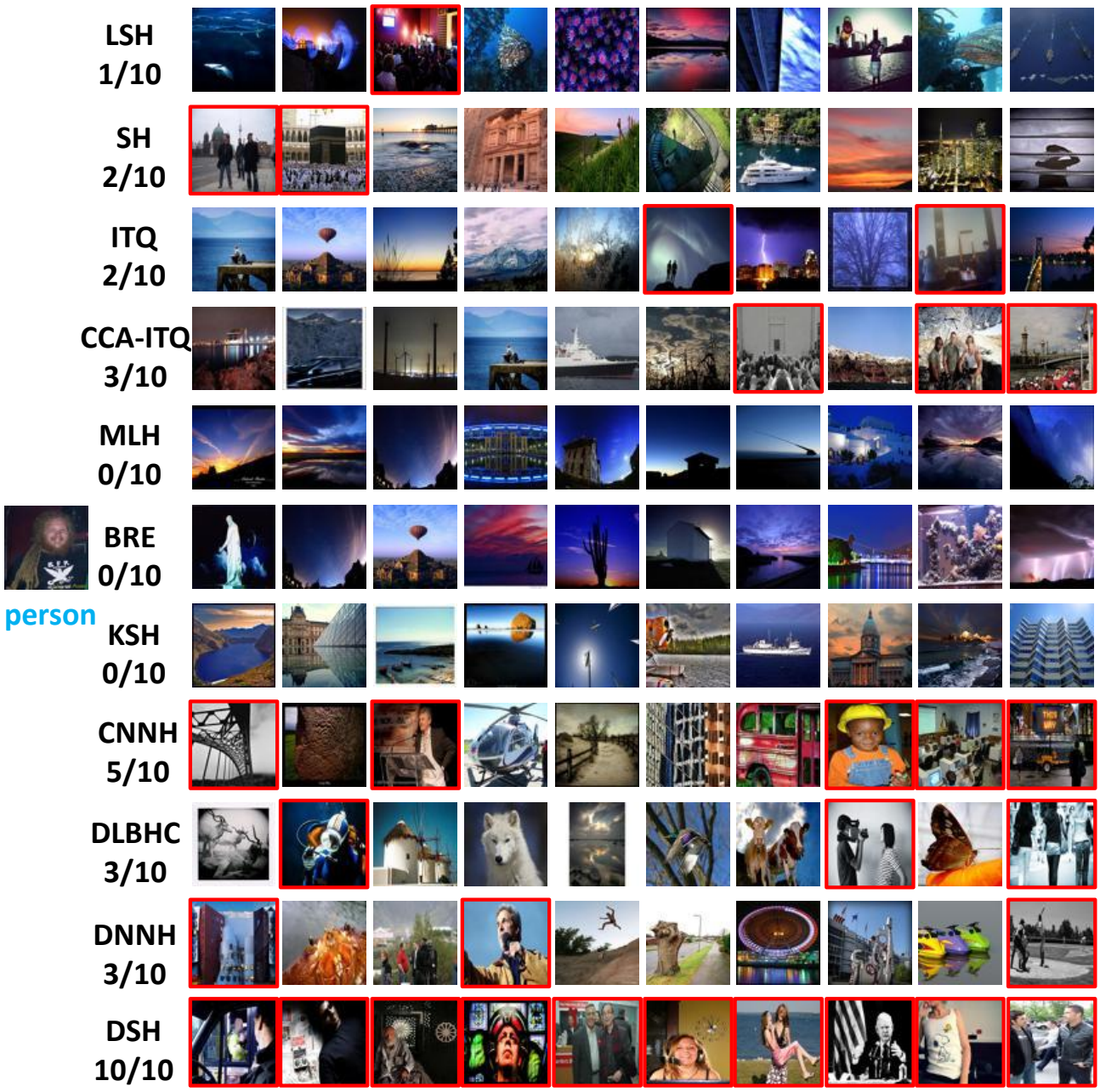


Figure 16. A **successful** retrieval case on NUS-WIDE, only the top-10 feedbacks are shown due to space limitation. Results were obtained with 48-bit binary codes. Images with red frames are true matches.

Figure 17. A **successful** retrieval case on NUS-WIDE, only the top-10 feedbacks are shown due to space limitation. Results were obtained with 48-bit binary codes. Images with red frames are true matches.

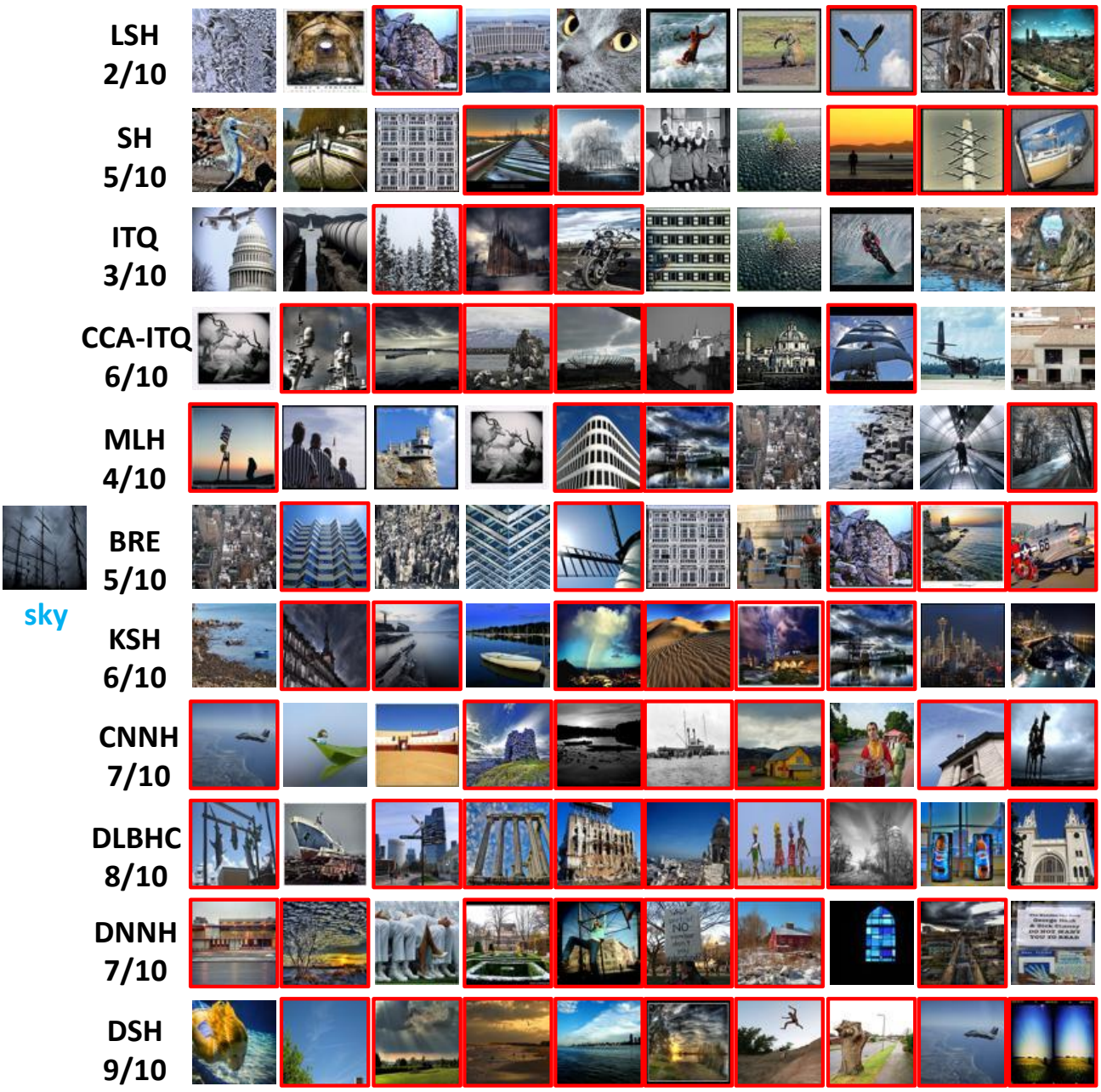


Figure 18. A **successful** retrieval case on NUS-WIDE, only the top-10 feedbacks are shown due to space limitation. Results were obtained with 48-bit binary codes. Images with red frames are true matches.