A^2-NET: Learning Attribute-Aware Hash Codes for Large-Scale Fine-Grained Image Retrieval (Supplementary Materials)

Xiu-Shen Wei[1,2], **Yang Shen**[1], **Xuhao Sun**[1], **Han-Jia Ye**[2], **Jian Yang**[1]
[1]Nanjing University of Science and Technology
[2]State Key Lab. for Novel Software Technology, Nanjing University

In the supplementary materials, we present further information about the proposed A^2-NET model, including: 1) Additional experimental results of other comparison methods, especially DSaH [4]; 2) More examples of retrieved results on other fine-grained benchmark datasets.

1 Additional experimental results of other comparison methods

Apart from ExchNet, DSaH [4] is another fine-grained hashing method which has achieved good retrieval accuracy. For fair comparisons, we strictly control empirical settings as the same as those of [4] and compare the results of our A^2-NET with its results and three following methods, *i.e.*, DPSH [6], DTQ [7] and HBMP [2].

Specifically, we follow the settings of DSaH [4] and conduct experiments on two fine-grained datasets, *i.e.*, *Stanford Dogs* [5] and *CUB200-2011* [10]. In concretely, *Stanford Dogs* consists of 20,580 images in 120 classes while each class contains about 150 images. The dataset is divided into the train set (100 images per class) and the test set (totally 8,580 images for all categories). *CUB200-2011* contains 11,788 bird images from 200 bird species and is officially split into 5,994 images for training and 5,794 images for test. We use AlexNet as backbone and it is not fine-tuned on each dataset.

As shown in Table 1, our A^2-NET significantly outperforms the other baseline methods on these two datasets by following the same settings of [4]. In particular, compared with DSaH [4], our A^2-NET achieves 10% and 7% improvements on *Stanford Dogs* and *CUB200-2011* in average.

Table 1: Comparisons of retrieval accuracy (% mAP) on two benchmark fine-grained datasets.

Methods	*Stanford Dogs*				*CUB200-2011*			
	12 bits	24 bits	36 bits	48 bits	12 bits	24 bits	36 bits	48 bits
DPSH [6]	17.7	22.1	26.5	31.5	7.2	7.6	8.4	7.9
DTQ [7]	18.5	18.7	18.7	18.8	7.3	11.3	15.4	18.3
HBMP [2]	19.0	23.8	28.7	32.8	8.9	10.9	14.2	16.8
DSaH [4]	24.4	28.7	36.3	40.8	14.2	20.9	23.2	28.5
Ours	**36.6**	**44.8**	**46.9**	**47.4**	**19.2**	**27.2**	**32.5**	**36.7**

2 More examples of retrieved results on other fine-grained datasets

We present more retrieval results on *Aircraft* [8], *Food101* [1], *NABirds* [9] and *VegFru* [3]. As shown in the following figures, our proposed A^2-NET can retrieve well among multiple subordinate categories. There also exist several failure cases, where quite tiny differences (*e.g.*, caused by different views) between the query image and the returned images are demanded by carefully observations.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

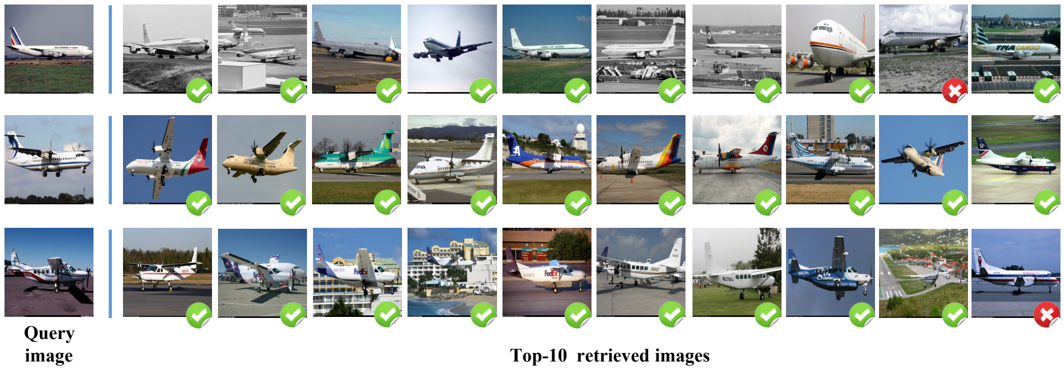

Figure 1: Examples of top-10 retrieved images on *Aircraft* of 48-bit hash codes by our A^2-NET.

Figure 2: Examples of top-10 retrieved images on *Food101* of 48-bit hash codes by our A^2-NET.

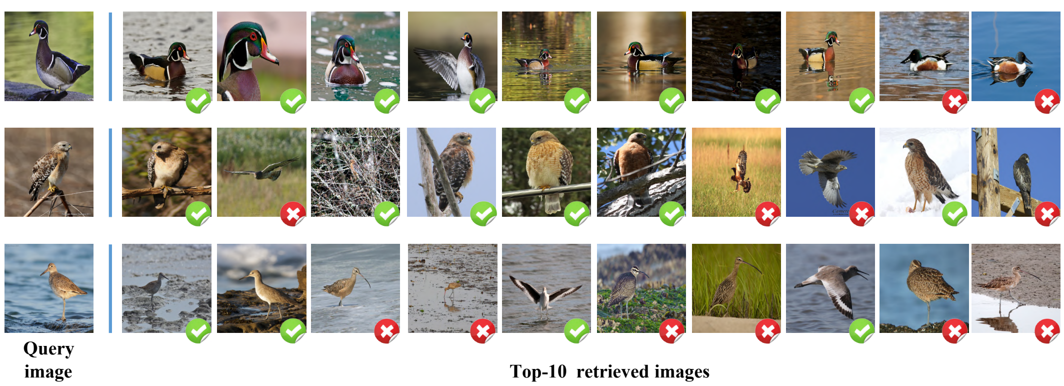

Figure 3: Examples of top-10 retrieved images on *NABirds* of 48-bit hash codes by our A^2-NET.

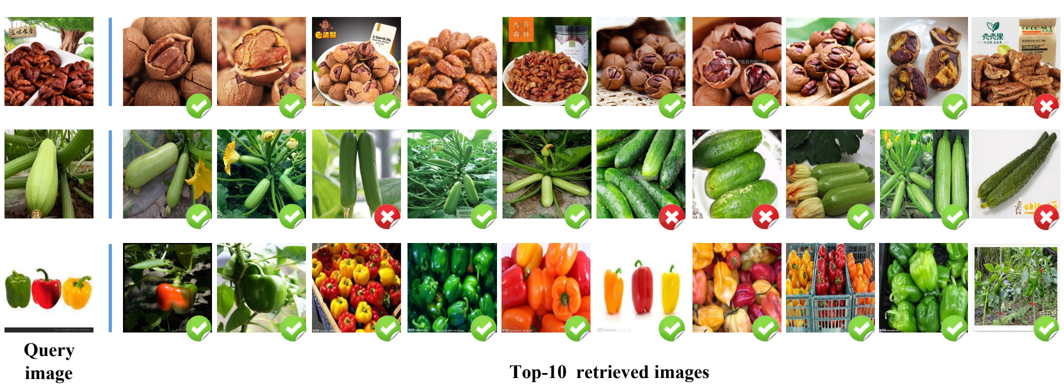

Figure 4: Examples of top-10 retrieved images on *VegFru* of 48-bit hash codes by our A^2-NET.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Proc. Eur. Conf. Comp. Vis.*, pages 446–461, 2014.

[2] Fatih Cakir, Kun He, and Stan Sclaroff. Hashing with binary matrix pursuit. In *Proc. Eur. Conf. Comp. Vis.*, pages 332–348, 2018.

[3] Saihui Hou, Yushan Feng, and Zilei Wang. VegFru: A domain-specific dataset for fine-grained visual categorization. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 541–549, 2017.

[4] Sheng Jin, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, Lei Zhang, and Xiansheng Hua. Deep saliency hashing for fine-grained retrieval. *IEEE Trans. Image Process.*, 29:5336–5351, 2020.

[5] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization*, volume 2. Citeseer, 2011.

[6] Wu-Jun Li, Sheng Wang, and Wang-Cheng Kang. Feature learning based deep supervised hashing with pairwise labels. In *Proc. Int. Joint Conf. Artificial Intell.*, pages 1711–1717, 2015.

[7] Bin Liu, Yue Cao, Mingsheng Long, Jianmin Wang, and Jingdong Wang. Deep triplet quantization. In *Proc. ACM Int. Conf. Multimedia*, pages 755–763, 2018.

[8] Subhransu Maji, Esa Rahtu, Juho Kannala, Mathhew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[9] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 595–604, 2015.

[10] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. *Tech. Report CNS-TR-2011-001*, 2011.