Computational Visual Media
https://doi.org/10.1007/s41095-020-0184-6

Research Article

A new dataset of dog breed images and a benchmark for fine-grained classification

Ding-Nan Zou[1,2], Song-Hai Zhang[1] (✉), Tai-Jiang Mu[1], and Min Zhang[3]

Abstract In this paper, we introduce an image dataset for fine-grained classification of dog breeds: the Tsinghua Dogs Dataset. It is currently the largest dataset for fine-grained classification of dogs, including 130 dog breeds and 70,428 real-world images. It has only one dog in each image and provides annotated bounding boxes for the whole body and head. In comparison to previous similar datasets, it contains more breeds and more carefully chosen images for each breed. The diversity within each breed is greater, with between 200 and 7000+ images for each breed. Annotation of the whole body and head makes the dataset not only suitable for the improvement of fine-grained image classification models based on overall features, but also for those locating local informative parts. We show that dataset provides a tough challenge by benchmarking several state-of-the-art deep neural models. The dataset is available for academic purposes at `https://cg.cs.tsinghua.edu.cn/ThuDogs/`.

Keywords fine-grained classification; dog; dataset; benchmark

1 Introduction

Dogs are closely involved in human lives as family members, and are very common as pets. On the other hand, the number of dog-related incidents of injury and uncivilized behavior is increasing. This leads to a need for dog identification using modern visual technology, both for dog recognition and finer-grained classification to breed.

Fine-grained classification is a non-trivial problem, requiring to distinguish different subclasses from subtle inter-class differences. As for other visual tasks, the performance of fine-grained classification has been greatly boosted by the use of deep neural networks [1–4]. However, there are relatively small differences between dogs of different breeds while there can be relatively large differences between those within a breed due to geographic isolation or hybridization. See, for example, Fig. 1: great Dane dogs have multiple colors, while dogs of different breeds, such as Norwich terriers and Australian terriers, may have similar colors. Existing datasets, such as the widely used Stanford Dogs Dataset [5], are not diverse enough to cover such variations, limiting their use for training and testing algorithms.

This paper contributes a new dataset, Tsinghua Dogs, with an emphasis on fine-grained dog classification. It contains 130 breeds of dogs in 70,428 images, with one dog per image, over 65% of which were collected from everyday life. It covers nearly all dog breeds currently found in China. Each breed in our dataset contains at least 200 images, up to a maximum of 7449 images, basically in proportion to their frequency of occurrence in China, so it significantly increases the diversity for each breed over existing datasets. Furthermore, we have annotated bounding boxes of the dog's whole body and head in each image, which can be used for supervising the training of learning algorithms as well as testing them.

We have also benchmarked several classification methods on our dataset, including both general neural networks and fine-grained models which exhibit

1 Department of Computer Science and Technology, BNRist, Tsinghua University, Beijing 100084, China. E-mail: D.-N. Zou, zoudn14@mails.tsinghua.edu.cn; S.-H. Zhang, shz@tsinghua.edu.cn (✉); T.-J. Mu, taijiang@tsinghua.edu.cn.

2 NaJiu Company, Hunan 410022, China.

3 Harvard Medical School, Brigham and Women's Hospital, Boston, MA 02115, USA. E-mail: mzhang@bwh.harvard.edu.

Manuscript received: 2020-05-18; accepted: 2020-06-14

Fig. 1 Dog variations in our dog dataset. (a) Great Danes exhibit large variations in appearance, while (b) Norwich terriers and (c) Australian terriers are quite similar to each other.

good performance on other fine-grained datasets. The results show that the large diversity of our dataset proves to be a tougher challenge, so should be beneficial in the development and testing of algorithms for real-world applications.

Our dataset can be downloaded at `https://cg.cs.tsinghua.edu.cn/ThuDogs/`.

2 Related work

2.1 Fine-grained classification

Fine-grained classification technology is an obvious next step from traditional coarse classification technology [6–9]. Coarse classification is generally intended to distinguish different types of objects such as animals and vehicles, while fine-grained classification usually needs to differentiate subclasses within a class, such as breeds of animals or makes or models of vehicles. CUB200-2011 [10] is a well-known fine-grained classification dataset containing 200 different bird species—see Fig. 2.

The main difficulties for fine-grained classification are typically the large number of fine-grained categories, and high intra-class but low inter-class variance. Currently, the best fine-grained classification methods use machine learning techniques, in particular deep neural networks. Research into fine-grained classification considers at least the following issues.

Fig. 2 Birds in the CUB200-2011 dataset [10].

2.1.1 Locating informative parts

In order to distinguish different subclasses, an intuitive approach is to explicitly take advantage of differences between corresponding object parts. Hand-crafted features [11, 12] are extracted from object parts and fed to linear classifiers, such as SVMs. Deep learning methods provide better performance, with the parts located and normalized for pose [13–15].

Since only a few key parts are useful for fine-grained classification, Lam et al. [16] proposed to only search for informative parts in the deep feature map. Chen et al. [17] first decomposed the input image into local parts and found the discriminative regions by reconstructing the image. Ge et al. [18] explored complementary object parts in addition to the dominant one. Du et al. [19] fused parts at various granularities for better performance. Recently, the idea of identifying the most informative parts to provide more robust performance is achieved by exploiting a spatial attention mechanism, such as multi-attention [20], recurrent attention [21], trilinear attention [22], and multi-scale object and part attention [23]. Sun et al. [24] introduced diversification blocks in feature maps to find the most discriminative differences between closely confusing classes.

Bilinear pooling based models can also implicitly learn the informative local parts. Lin et al. [25] explored pairwise relations of local parts using a bilinear pooling of outer products of features from two convolutional extractors. Gao et al. [26] proposed a more compact bilinear pooling. Yu et al. [27] exploited hierarchical bilinear pooling to account for interaction of features between layers.

2.1.2 Learning from image pairs

Learning discriminative cues directly from an image pair is more intuitive since human beings can easily tell fine-grained classes by comparing given image pairs. Metric learning, which is a typical solution for measuring the similarity between image pairs, has also been used for fine-grained classification, e.g., using triplet loss design [28, 29], maximum entropy [30], a multi-stage method [31], multi-attention multi-class constraints [32], and pairwise confusion regularization [33]. These methods are mainly designed to separate images in feature space, but are less capable of discriminating subtle differences between confusing images. Recently, Zhuang et al. [34] suggested finding contrasting cues directly from a pair of images via attentive pairwise interaction. This method achieves state-of-the-art performance on several fine-grained classification datasets.

2.1.3 Data augmentation

Whatever method is used, more meaningful training data always helps to train a more general model [35]. A common approach is to use search engines, crawlers, etc. to search for relevant images and text [36] on the Internet, and to use it to train the fine-grained classification model. However, there is a huge amount of noise in such data [37], and techniques are required to suppress this noise and extract valid information. Hu et al. [38] proposed a weakly supervised data augmentation network (WS-DAN) to augment images guided by attention maps generated by weakly supervised learning. Our dataset ensures data diversity by collecting more samples from real life.

2.2 Datasets for fine-grained classification

To help develop and assess fine-grained classification technology, researchers have released many public fine-grained classification datasets. In addition to the aforementioned CUB200-2011 dataset [10], there are Stanford Cars [39], FGVC Aircraft [40], Oxford 102 Flowers [41], and other datasets.

Stanford Dogs is a public fine-grained classification dataset for dog breeds [5]. It contains 20,580 images of 120 dog breeds, with 150–252 images for each breed. The images in this dataset are clear and obvious; for each dog, its whole body bounding box is annotated.

Other dog datasets have also been provided for classification tasks. For example, ImageNet-1K [42] contains about 116,000 pictures of 117 dog breeds. Some general datasets also contain dog images, but as a single category, without any fine-grained classification information. For example, there are 2079 dog bounding boxes in the VOC dataset training data (2007 and 2012) [43] and 530 images containing dogs in the verification data; in the COCO [44] dataset, there are 5508 bounding boxes of dogs. Our proposed dataset focuses on fine-grained dog classification, and provides sufficient diversity for each breed to test deep neural model generalization.

3 Tsinghua Dogs Dataset

We now introduce how we constructed our Tsinghua Dogs Dataset and present its statistical features.

3.1 Data collection

Our data capture system has collected more than 100,000 images of dogs captured and uploaded by users in three Chinese cities. We removed sensitive information from the data and selected more than 46,000 images to build the dataset. As the numbers of images of each breed of dog reflects their actual distribution in these three cities, there is a long tail to this data. Teddy dog pictures are the most frequent (7449 images), while Cassell pictures are the least frequent (4 images). See Fig. 3.

While this reflects the real distribution of dog breeds, to make the dataset friendly to algorithms, we wish to ensure that each breed has no fewer than 200 pictures, to ensure diversity of images for

Fig. 3 Teddy and Cassell Dogs.

each breed. Therefore, we also added data from the Stanford Dogs dataset and by using an image search engine. We integrated 18,000 pictures with only one dog from Stanford Dogs into our dataset. We also crawled and manually selected more than 6000 pictures using Baidu image search to ensure that our dataset contained no fewer than 200 pictures per breed.

We removed duplicate images in the dataset by computing image structural similarity (SSIM) [45]. After collecting the data, we determined the true dog breed in the images through expert review. We also asked the annotators to filter out low-quality images, i.e., any that were seriously blurred, deformed, occluded, or where the dog was too small a part of the image. The final number of images in the Tsinghua Dogs Dataset is 70,428, from a total of 130 breeds, with no less than 200 images per breed. Part of the image of the dataset is shown in Fig. 4.

3.2 Data annotation with active learning

Manual labeling using tools such as Labelme [46] is laborious and inefficient (annotating 300–500 pictures per hour). To reduce the effort of manual labeling, we used an active learning strategy to label the dataset in a semi-automatic manner which can increase the efficiency to 1500–2500 pictures per hour. The approach was:

1. For 2000 randomly chosen pictures from the dataset, label tight bounding boxes around the dog's whole body, and the dog's head (including the ears), as shown in Fig. 5.
2. Train the RetianNet model and automatically generate 2000 new data annotations.
3. Manually correct labellings with low confidence by our own adjustment tool (see Fig. 6) using only keyboard interaction, insert those data into the training data, and return to step 2. Repeat until all the data are labeled.

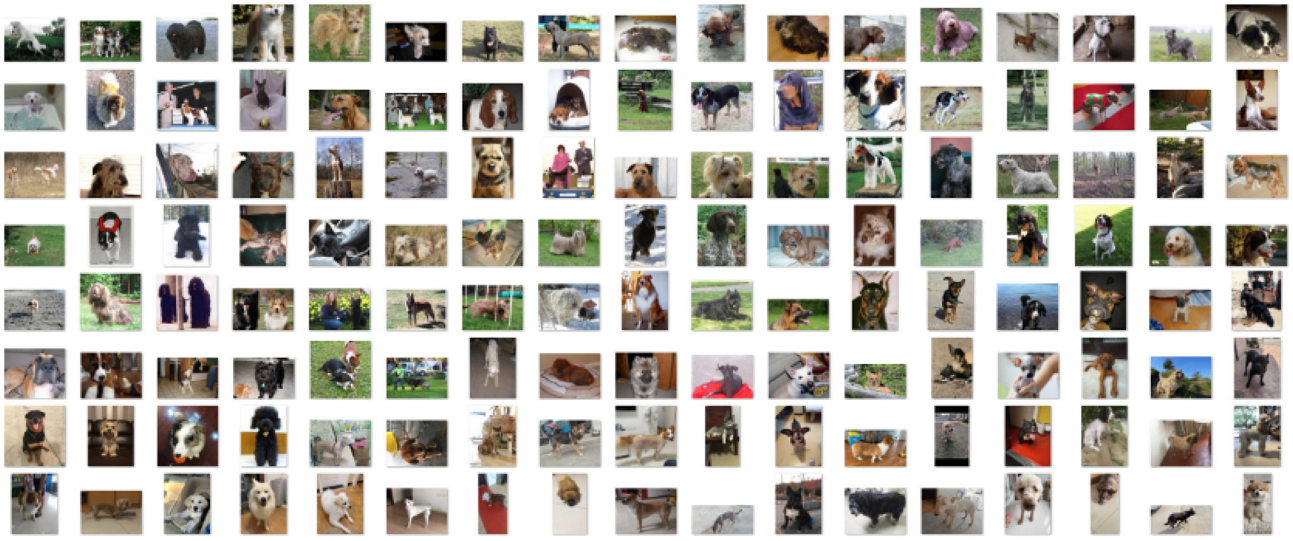

Fig. 4 Snapshots of Tsinghua Dogs Dataset.

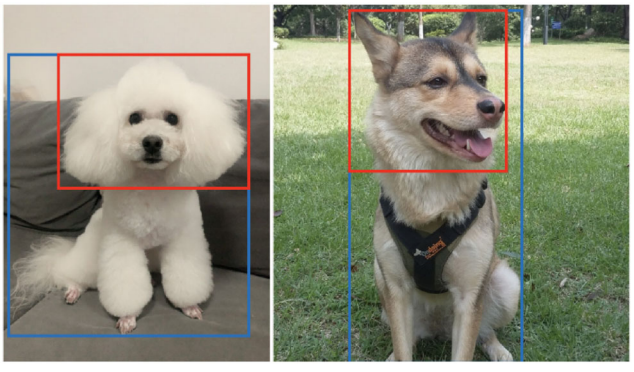

Fig. 5 Bounding boxes for whole dogs (blue) and their heads (red).

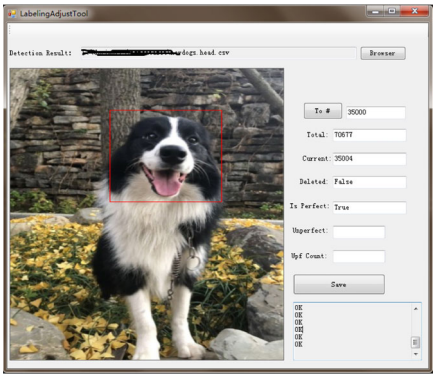

Fig. 6 Adjustment software.

3.3 Statistics

Using the above pipeline, 70,428 dog images were annotated for our Tsinghua Dogs dataset, including about 46,000 images of dogs taken in Chinese cities, 18,000 images from the Stanford Dogs dataset and, 6,000 images downloaded from Baidu, Google and other image search engines. The total number of dog breeds is 130. Each image contains a single dog, annotated with bounding boxes of the dog's head and the whole dog (see Fig. 7).

We compare our dataset, CUB-200, and Stanford Dogs in Table 1.

We now give some statistics for our dataset. The number of dogs for each breed varies from 200 to 7449 (Teddy dogs). Figure 8 shows numbers of the 24 most common dogs in the dataset. Statistics on the fraction of the whole image covered by the bounding

Table 1 Dataset comparison

Dataset	Breeds	Images	Images per breed	Object
CUB-200	200	6033	30	Bird
Stanford Dogs	120	20,580	150–252	Dog
Ours	130	70,428	200–7449	Dog, dog's face

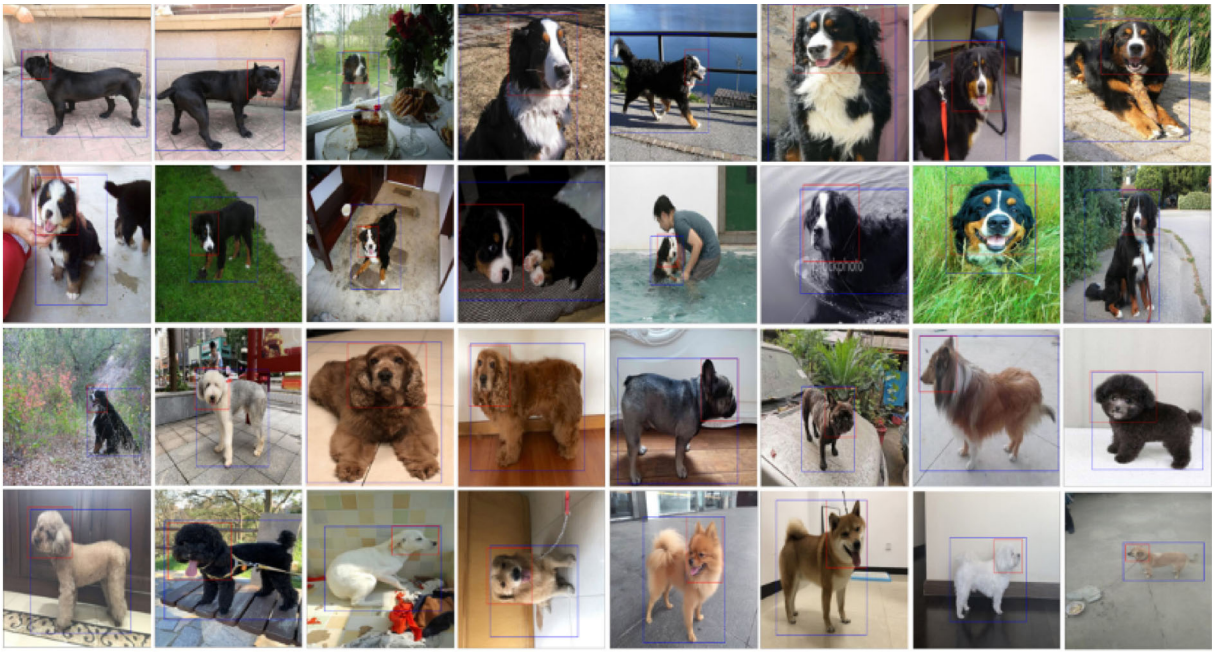

Fig. 7 Labeled Images.

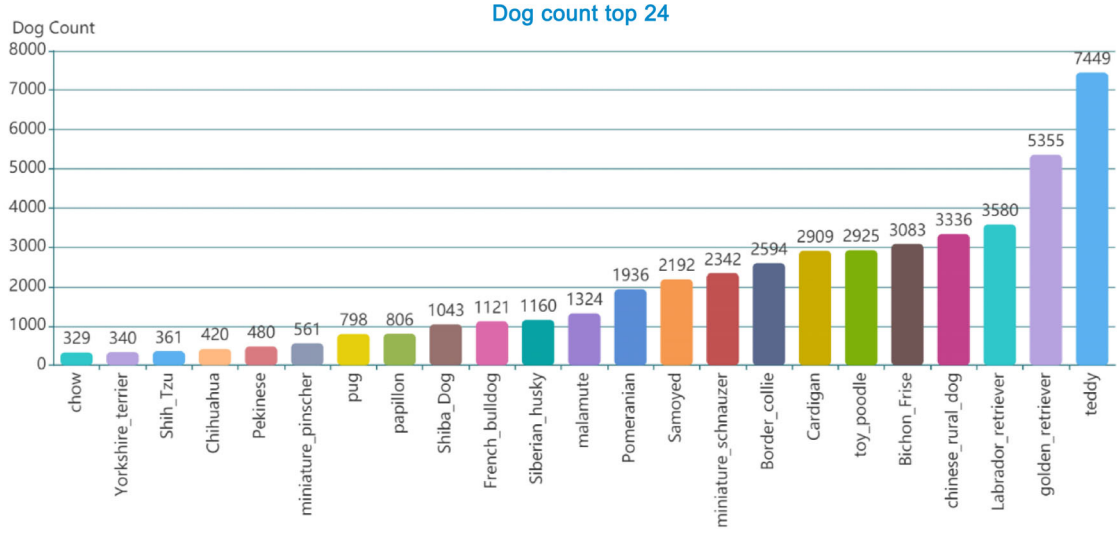

Fig. 8 Top 24 breeds of dogs by number of images.

box of the dog's head are given in Fig. 9, while the fraction of the whole image covered by the dog's body's bounding box is indicated in Fig. 10.

The images do not have a fixed resolution. Very few pictures have a length or width less than 100 pixels, with a minimum of 60, and most images have a relatively high resolution. Image resolution statistics are shown in Fig. 11. At least half of our images have higher resolution than those in the Stanford Dogs dataset.

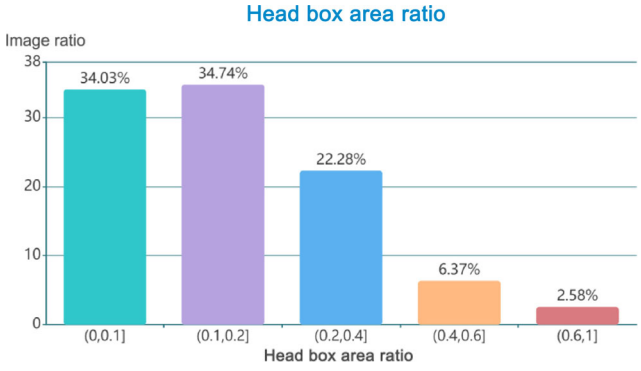

Fig. 9 Fraction of the image covered by the dog's head bounding box.

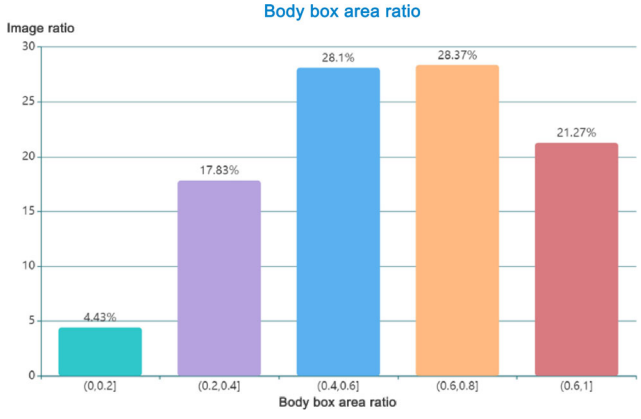

Fig. 10 Fraction of the image covered by the dog's body bounding box.

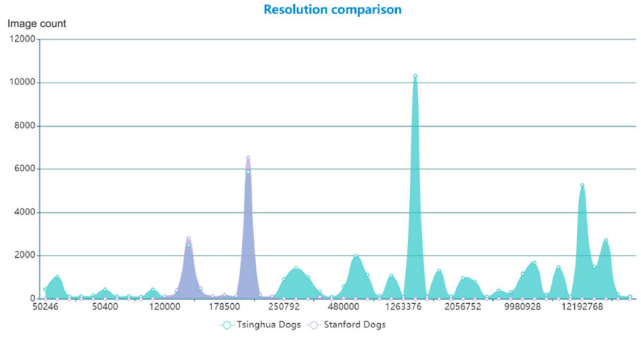

Fig. 11 Image resolutions in the Stanford Dogs and Tsinghua Dogs datasets (pixels).

4 Benchmarking using Tsinghua Dogs

Although most deep neural fine-grained classification models can be retrained on dog datasets, as has been done for the Stanford Dogs [5][①], these methods are usually not optimized for images of dogs. Furthermore, we argue that the diversity within current dog datasets does not provide an adequate test. In this section, we first discuss training procedures for several models we have benchmarked on our dataset. We show benchmarking results using our dataset and analyze how the additional diversity in our dataset improves the robustness of fine-grained classification models.

4.1 Training

Using our new dataset, we trained three state-of-the-art fine-grained classification deep neural networks: PMG [19] (ranked top on several datasets), TBMSL-Net [23] (ranked 1st on FGVC Aircraft [40]), and WS-DAN [38] (ranked 1st on Stanford Dogs)[②], as well as a general classification backbone network, Inception V3 [47]. All models were trained using the Pytorch framework. We compared the accuracy achieved for fine-grained classification using our dataset, Standford Dogs [5], and the CUB dataset [10].

We took PMG[③] as our base model, and used ResNet50 as the backbone network. Parameter settings strictly adhered to those in the original paper. The PMG model starts from the bottom stage network and trains the network stage-by-stage. Each stage is trained with images spliced from image patches of the size specified in the original paper. The experiment used a learning rate of 0.002 for the newly added stage, with a cosine annealing schedule to reduce the learning rate. We trained 200 rounds on each dataset. The input was a 448×448 image cropped from the center after scaling the original image to 550×550. The batch size was 16.

The backbone network of TBMSL-Net[④] is also ResNet50. TBMSL-Net can automatically learn the location and the key parts of an object in an input image. Its final fine-grained classification score is

① https://paperswithcode.com/sota/fine-grained-image-classification-on-stanford-1
② on May 17, 2020.
③ https://github.com/RuoyiDu/PMG-Progressive-Multi-Granularity-Training
④ https://github.com/ZF1044404254/TBMSL-Net

given by combining whole graph features. We used the same algorithm window and other parameter settings for CUB. Both the object and the original image were resized to 448 × 448, but the image of the key part of the object is resized to 224 × 224. The optimizer used was stochastic gradient descent (SGD). The momentum was 0.9, and the weight was 0.0001. The initial learning rate was 0.001; it was multiplied by 0.1 after 60 epochs. We trained 200 rounds in total.

WS-DAN[①] improves the performance of image classification through two mechanisms: one extracts significant features from the image to make the image appearance more effective; the other focuses on the location of the target so that the model can observe the target more "closely" to improve performance. The size of the input images was 512 × 448; 80 epochs were used. SGD optimization was used with a momentum of 0.9 and weight decay of 0.00001. The initial learning rate was 0.001 with exponential decay of 0.9 after every 2 epochs.

For Inception V3[②] we used the training settings from Ref. [42]. Each image size is resized to 224 × 224, and 200 rounds of training were completed. The initial learning rate was 0.05, and it was adjusted as follows: if the accuracy on the validation set did not increase after 10 rounds, then the learning rate was multiplied by 0.1. The optimization function was again SGD (momentum = 0.9) with a batch size of 64. The penultimate layer used a dropout of 0.4.

We split the training data of CUB 200-2011 [10] into training and validation sets according to `train_test_split.txt`. The number of images in the training and validation sets is 5994 and 5794, respectively. Standford Dogs has 12,000 training images and 8580 validation images. Our dataset also provides labels for training and validation (randomly selecting 40 images for each breed), with 65,228 and 5200 cases respectively.

4.2 Results and analysis

Various deep neural classification models have reported their performance on Stanford Dogs. Inception V3 achieved an accuracy of 88.9%, while WS_DAN ranked 1st with an accuracy of 92.2%[③]. However, these models are not optimized for fine-grained classification of dogs, and performance would degrade in real-world applications.

4.2.1 Results

Although PMG [19] reported its classification accuracy on three fine-grained classification datasets, CUB 200-2011 [10], Stanford Cars [39], and FGVC-Aircraft [40], the model has not been tested on Stanford Dogs. To ensure a fair and effective comparison, we first tested the PMG model on CUB to verify that our trained PMG model gave results consistent with the original paper. Then we trained the PMG model on Stanford Dogs and our dataset with the same training parameters. The performance of the PMG model on these datasets is shown in Table 2. As can be seen clearly from the comparison, the accuracy of the PMG model on our dataset is lower than on Stanford Dogs by about 3%, demonstrating that our dataset presents a greater challenge for fine-grained dog classification.

We also benchmarked the accuracy of the deep neural networks described above on our Tsinghua Dogs dataset: see Table 3. Notice that the accuracy of Inception V3 drops by more than 10% from Stanford Dogs to Tsinghua Dogs, while WS_DAN decreases by over 5%. These results imply that current state-of-the-art fine-grained models still have considerable room for improvement.

Table 2 Performance of PMG [19] on different datasets

Dataset	Information	Accuracy reported in Ref. [19]	Accuracy in our test
CUB 200-2011	200 species of birds 11,788 pictures	88.9% (single) 89.6% (combined)	88.609% (single) 89.454% (combined)
Stanford Dogs	120 breeds of dogs 20,580 pictures	—	84.674% (single) 86.515% (combined)
Tsinghua Dogs Dataset	130 breeds of dogs 70,428 pictures	—	81.98% (single) 83.52% (combined)

Table 3 Fine-grained classification accuracy of PMG [19], TBMSL-Net [23], WS-DAN [38], and Inception V3 [47] on our dataset

Model	Backbone	Batchsize	Epochs	Accuracy
Inception V3	—	64	200	77.66%
WS-DAN	Inception	12	80	86.404%
PMG	ResNet50	16	200	83.52%
TBMSL-Net	ResNet50	6	200	83.7%

① https://github.com/wvinzh/WS_DAN_PyTorch
② https://github.com/liuzhuang13/DenseNet
③ on May 17, 2020

4.2.2 *Further analysis*

To better understand how the diversity of our new dog dataset improves the robustness of classification, we now consider a qualitative analysis of the classification results. We trained two WS_DAN models on Stanford Dogs and Tsinghua Dogs respectively, and tested them on a subset of the validation set of Tsinghua Dogs. This subset consists of the 20 breeds with the largest number of images left when those appeared in the training set of Stanford Dogs are excluded, resulting in 657 images.

The model trained on Tsinghua Dogs achieved an accuracy of 82.65%, while the one trained on Stanford Dogs only achieved 58.14%. In Fig. 12,

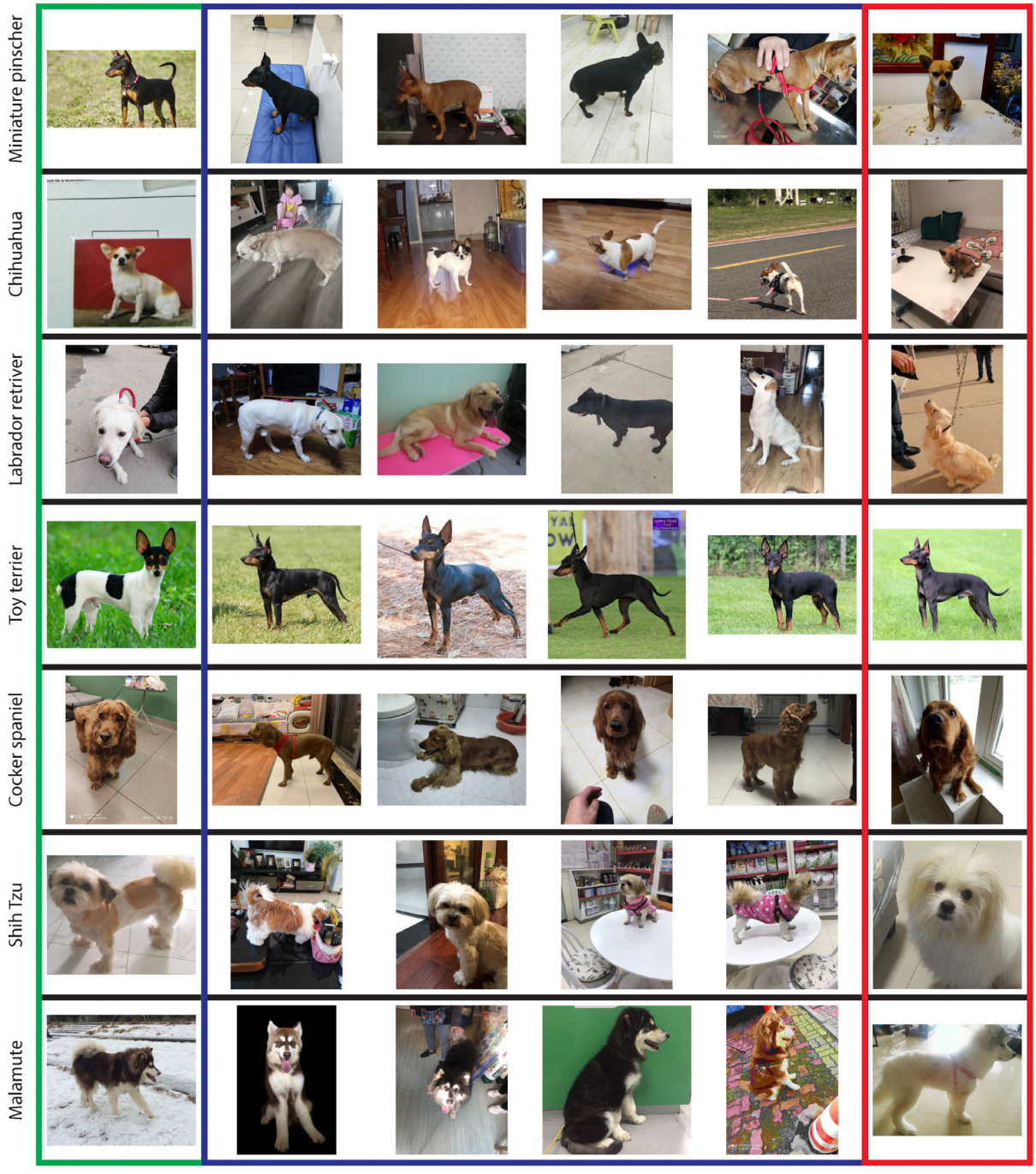

Fig. 12 Qualitative comparison of WS_DAN models trained on Stanford Dogs and Tsinghua Dogs. Dogs in each row belong to the same breed. WS_DAN trained on Tsinghua Dogs classifies the dogs correctly except for the last column, while the one trained on Stanford Dogs gives a correct classification only for the first column.

we qualitatively show classification results for seven breeds. For dogs in columns 2–5, the model trained on Tsinghua Dogs succeeded in finding the right breed, while the model trained on Stanford Dogs failed. Real-world dogs are captured from various directions, giving a wide variation in appearance even for the same breed. Our Tsinghua Dogs incorporates more diversity and is thus more suitable for developing generalized deep neural models for fine-grained dog classification.

5 Conclusions

This paper has introduced a new challenging fine-grained classification dog dataset, Tsinghua Dogs. Our dataset contains 130 dog breeds and 70,428 images, with bounding boxes annotated for locations of the dog and its head. The diversity of the dataset and its additional annotation allow the construction of more robust and accurate deep neural fine-grained models needed for real-world applications.

Acknowledgements

The authors would like to thank Wei-Yu Xie for his assistance on paper writing, and also thank Qiu Xin and Zhi-Ping Zhang for much help on image processing and labeling. This work was supported by the National Natural Science Foundation of China (Project Nos. 61521002 and 61772298), a Research Grant of Beijing Higher Institution Engineering Research Center, and Tsinghua–Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] Cai, S.; Zuo, W.; Zhang, L. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In: Proceedings of the IEEE International Conference on Computer Vision, 511–520, 2017.

[2] Cui, Y.; Song, Y.; Sun, C.; Howard, A.; Belongie, S. J. Large scale fine-grained categorization and domain-specific transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4109–4118, 2018.

[3] Wang, Y.; Morariu, V. I.; Davis, L. S. Learning a discriminative filter bank within a CNN for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4148–4157, 2018.

[4] Yang, Z.; Luo, T. G.; Wang, D.; Hu, Z. Q.; Gao, J.; Wang, L. W. Learning to navigate for fine-grained classification. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science Vol. 11218.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 438–454, 2018.

[5] Khosla, A.; Jayadevaprakash, N.; Yao, B.; Li, F.-F. Novel dataset for fine-grained image categorization. In: Proceedings of the 1st Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[6] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, Vol. 1, 1097–1105, 2012.

[7] Chen, L.; Yang, M. Semi-supervised dictionary learning with label propagation for image classification. *Computational Visual Media* Vol. 3, No. 1, 83–94, 2017.

[8] Chen, K. X.; Wu, X. J. Component SPD matrices: A low-dimensional discriminative data descriptor for image set classification. *Computational Visual Media* Vol. 4, No. 3, 245–252, 2018.

[9] Ren, J. Y.; Wu, X. J. Vectorial approximations of infinite-dimensional covariance descriptors for image classification. *Computational Visual Media* Vol. 3, No. 4, 379–385, 2017.

[10] Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. Computation & Neural Systems Technical Report, CNS-TR-2011-001. California Institute of Technology, 2011.

[11] Liu, J.; Kanazawa, A.; Jacobs, D.; Belhumeur, P. Dog breed classification using part localization. In: Proceedings of the 12th European Conference on Computer Vision, Vol. Part I, 172–185, 2012.

[12] Berg, T.; Belhumeur, P. N. POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 955–962, 2013.

[13] Branson, S.; Horn, G. V.; Belongie, S.; Perona, P. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint* arXiv:1406.2952, 2014.

[14] Zhang, N.; Donahue, J.; Girshick, R.; Darrell, T. Part-based R-CNNs for fine-grained category detection. In: *Computer Vision–ECCV 2014. Lecture Notes in Computer Science Vol. 8689.* Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer Cham, 834–849, 2014.

[15] Lin, D.; Shen, X.; Lu, C.; Jia, J. Deep LAC: Deep localization, alignment and classification for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1666–1674, 2015.

[16] Lam, M.; Mahasseni, B.; Todorovic, S. Fine-grained recognition as HSnet search for informative image parts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6497–6506, 2017.

[17] Chen, Y.; Bai, Y.; Zhang, W.; Mei, T. Destruction and construction learning for finegrained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5157–5166, 2019.

[18] Ge, W. F.; Lin, X. R.; Yu, Y. Z. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. *arXiv preprint* arXiv:1903.02827, 2019.

[19] Du, R. Y.; Chang, D. L.; Bhunia, A. K.; Xie, J. Y.; Ma, Z. Y.; Song, Y. Z.; Guo, J. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. *arXiv preprint* arXiv:2003.03836, 2020.

[20] Zheng, H.; Fu, J.; Mei, T.; Luo, J. Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, 5219–5227, 2017.

[21] Fu, J.; Zheng, H.; Mei, T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4476–4484, 2017.

[22] Zheng, H.; Fu, J.; Zha, Z.; Luo, J.; Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5012–5021, 2019.

[23] Zhang, F.; Li, M.; Zhai, G.; Liu, Y. Three-branch and multi-scale learning for fine-grained image recognition (TBMSL-Net). *arXiv preprint* arXiv:2003.09150, 2020.

[24] Sun, G. L.; Cholakkal, H.; Khan, S.; Khan, F. S.; Shao, L. Fine-grained recognition: Accounting for subtle differences between similar classes. *arXiv preprint* arXiv:1912.06842, 2019.

[25] Lin, T.-Y.; RoyChowdhury, A.; Maji, S. Bilinear CNN models for fine-grained visual recognition. In: Proceedings of the IEEE international conference on computer vision, 1449–1457, 2015.

[26] Gao, Y.; Beijbom, O.; Zhang, N.; Darrell, T. Compact bilinear pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 317–326, 2016.

[27] Yu, C.; Zhao, X.; Zheng, Q.; Zhang, P.; You, X. Hierarchical bilinear pooling for fine-grained visual recognition. In: *Computer Vision–ECCV 2018. Lecture Notes in Computer Science Vol. 11220.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 595–610, 2018.

[28] Wang, Y.; Choi, J.; Morariu, V. I.; Davis, L. S. Mining discriminative triplets of patches for fine-grained classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1163–1172, 2016.

[29] Zhang, X.; Zhou, F.; Lin, Y.; Zhang, S. Embedding label structures for finegrained feature representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1114–1123, 2016.

[30] Dubey, A.; Gupta, O.; Raskar, R.; Naik, N. Maximum-entropy fine grained classification. *arXiv preprint* arXiv:1809.05934, 2018.

[31] Qian, Q.; Jin, R.; Zhu, S.; Lin, Y. Fine-grained visual categorization via multi-stage metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3716–3724, 2015.

[32] Sun, M.; Yuan, Y.; Zhou, F.; Ding, E. Multi-attention multi-class constraint for fine-grained image recognition. In: *Computer Vision–ECCV 2018. Lecture Notes in Computer Science Vol. 11220.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 834–850, 2018.

[33] Dubey, A.; Gupta, O.; Guo, P.; Raskar, R.; Farrell, R.; Naik, N. Pairwise confusion for fine-grained visual classification. In: *Computer Vision–ECCV 2018. Lecture Notes in Computer Science Vol. 11216.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 71–88, 2018.

[34] Zhuang, P.; Wang, Y.; Qiao, Y. Learning attentive pairwise interaction for fine-grained classification. *arXiv preprint* arXiv:2002.10191, 2020.

[35] Xu, Z.; Huang, S.; Zhang, Y.; Tao, D. Augmenting strong supervision using web data for finegrained categorization. In: Proceedings of the IEEE International Conference on Computer Vision, 2524–2532, 2015.

[36] Niu, L.; Veeraraghavan, A.; Sabharwal, A. Fine-grained classification using heterogeneous web data and auxiliary categories. *arXiv preprint* arXiv:1811.07567, 2018.

[37] Torralba, A.; Efros, A. A. Unbiased look at dataset bias. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1521–1528, 2011.

[38] Hu, T.; Qi, H. G.; Huang, Q. M.; Lu, Y. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv preprint* arXiv:1901.09891, 2019.

[39] Krause, J.; Stark, M.; Deng, J.; L. Fei-Fei. 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, 554–561, 2013.

[40] Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv preprint* arXiv:1306.5151, 2013.

[41] Nilsback, M.; Zisserman, A. Automated flower classification over a large number of classes. In: Proceedings of the 6th Indian Conference on Computer Vision, Graphics & Image Processing, 722–729, 2008.

[42] Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 248–255, 2009.

[43] Everingham, M.; van Gool, L.; Williams, C. K. I.; Winn, J.; Zisserman, A. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision* Vol. 88, No. 2, 303–338, 2010.

[44] Lin, T.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; Dollár, P. Microsoft COCO: Common objects in context. *arXiv preprint* arXiv:1405.0312, 2014.

[45] Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* Vol. 13, No. 4, 600–612, 2004.

[46] Russell, B. C.; Torralba, A.; Murphy, K. P.; Freeman, W. T. LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision* Vol. 77, Nos. 1–3, 157–173, 2008.

[47] Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K. Q. Densely connected convolutional networks. *arXiv preprint* arXiv:1608.06993, 2016.

Ding-Nan Zou is a master candidate in the Department of Computer Science and Technology at Tsinghua University, Beijing. His research interests include computer graphics and computer vision, especially dog face and iris recognition.

Song-Hai Zhang received his Ph.D. degree in computer science and technology from Tsinghua University, Beijing, in 2007. He is currently an associate professor in the Department of Computer Science and Technology at Tsinghua University. His research interests include image and video analysis and processing as well as geometric computing.

Tai-Jiang Mu is currently an assistant researcher in the Department of Computer Science and Technology, Tsinghua University, Beijing, where he received his bachelor and doctor degrees in computer science and technology in 2011 and 2016, respectively. His research interests include visual media learning, SLAM, and human robot interaction.

Min Zhang is a researcher in Harvard Medical School, Brigham and Women's Hospital. She received her Ph.D. degree in computer science from Stony Brook University and the other Ph.D. degree in mathematics from Zhejiang University. She is an expert in the fields of geometric modeling, medical imaging, graphics, visualization, machine learning, 3D technologies, etc.