Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval Supplementary Material

1. Network Architecture of the Proposed Model

Here we offer a more detailed description of the proposed model to facilitate re-implementation. A schematic illustration of the network architecture of the proposed model can be found in Fig. 2 of the main paper. It shows that the model is a Siamese triplet network with three branches of identical architectures and shared parameters. In this section, we further describe the detailed network architecture for each branch. Table 1 shows that each network branch has 7 convolutional layers and 2 fully connected layers. The first 5 convolutional layers are the same as the Sketch-a-Net [5] and the last 2 convolutional layers are part of the proposed attention module. Note that the output of the attention module is a 7×7 attention mask which is used to re-weight the feature map of pool5. The attention module shortcut connection (see Sec. 3.2 of the main paper) takes place before Layer No. 12 and the coarse-fine fusion shortcut connection occurs before Layer No. 14 (as detailed in Sec. 3.3 of the main paper). The final output of each branch is a 512D feature vector which is then subjected to our HOLEF loss.

Layer No.	Input Layer(s)	Layer Type	Kernel Size	Stride	Pad	Output
0	-	Input	-	-	-	$225 \times 225 \times 1$
1	0	Conv1	15×15	3	0	$71 \times 71 \times 64$
2	1	Pool1	3×3	2	0	$35 \times 35 \times 64$
3	2	Conv2	5×5	1	0	$31 \times 31 \times 128$
4	3	Pool2	3×3	2	0	$15 \times 15 \times 128$
5	4	Conv3	3×3	1	1	$15 \times 15 \times 256$
6	5	Conv4	3×3	1	1	$15 \times 15 \times 256$
7	6	Conv5	3×3	1	1	$15 \times 15 \times 256$
8	7	Pool5	3×3	2	0	$7 \times 7 \times 256$
9	8	attention-conv1	1×1	1	0	$7 \times 7 \times 256$
10	9	attention-conv2	1×1	1	0	$7 \times 7 \times 1$
11	$8 \oplus (10 \otimes 8)$	GAP*	-	-	-	$1 \times 1 \times 256$
12	$8 \oplus (10 \otimes 8)$	FC6	1×1	1	0	$1 \times 1 \times 512$
13	12	FC7	1×1	1	0	$1 \times 1 \times 256$
14	$11 \odot 13$	feature for HOLEF loss	-	-	-	$1 \times 1 \times 512$

Table 1. The detailed configuration of each branch of the proposed model. '*' When concatenating the output of the FC7 layer with the attended feature map, we applied global average pooling (GAP) on the attended feature map to reduce the dimensionality of the feature map first. "$\oplus$" is element-wise sum, "$\otimes$" is element-wise product and "$\odot$" is concatenation.

2. Experiments on the Sketchy database

In this section, we present further experimental results on the recently released Sketchy database [3]. In the original experiments presented in the main paper, three datasets are chosen, namely QMUL-Shoe, QMUL-Chair and the new Handbag dataset. Since we focus on the task of retrieving visually similar object instances from the same category, these three datasets are the most appropriate because each contains only one object category and many object instances are visually highly similar (see Fig. 1(a)). Importantly, they have the highest numbers of instances per category among all existing FG-SBIR datasets – there are 419, 297, and 568 instances of shoes, chairs and handbags in the three datasets respectively. Among them, 115,

Figure 1. (a) Examples of the QMUL-Shoe dataset and the Handbag dataset. (b) Examples of the Sketchy database.

97, and 168 instances are used for testing which represent challenging FG-SBIR tasks. Other FG-SBIR datasets do exist, however they have much less instances per category compared to the three datasets above. Among them, the largest one is the Sketchy database [3].

2.1. Dataset and Settings

The Sketchy database [3] includes 74,425 sketches and 12,500 gallery photos spanning 125 categories. In each category, there are 100 object instances. Each instance has one photo and at least 5 corresponding sketches. Following the dataset partitions in [3], we used 90% object instances for training and the rest for testing. Specifically, 11,250 of 12,500 photos and 68,543 of 74,425 sketches are used for training. In the provided test data partition, the gallery set contains 1250 photos (10 photos per category) and the probe set contains 6312 query sketches.

Some examples of the Sketchy database can be seen in Fig. 1(b). Compared with the QMUL-Shoe, QMUL-Chair and Handbag datasets (Fig. 1(a)), there are a number of key differences: (i) The Sketchy database has much more sketch-photo pairs, but within each category it has much less (10 photos per category in the gallery set). (ii) The photos in Sketchy were selected from ImageNet, i.e., from Google Image Search, whilst the other three datasets are collected from online shopping websites consisting of real product catalogue photos. As a result the Sketchy photos have cluttered background and variable object poses as well as partial occlusions. In contrast, those catalogue photos have uniformed (clean) background and poses.

With these differences, the FG-SBIR tasks are also quite different. In particular, a model for the Sketchy database needs to first distinguish different categories and then differentiate different same-category object instances. This needs to be achieved with the presence of cluttered background and variable object poses. From the outset, this may seem to be a more difficult task than the same-category FG-SBIR tasks for the other three datasets. However, since the same-category FG-SBIR task is much more challenging than the category recognition task, especially when the Sketchy database contain only 10 photos per category as opposed to around 100 in QMUL-Shoe, QMUL-Chair and Handbag, FG-SBIR on Sketchy is actually a much easier task – this is partially reflected by the Top 1 matching accuracy normalised by the gallery size obtained in our experiments[1].

2.2. Model Implementation

Model architecture The model presented in the main paper needs to be modified in order to cope with the additional challenges of (1) cluttered background, (2) variable pose, and (3) object categorisation. In particular, we can no longer extract edge maps from photos and feed them together with sketches into a Siamese network. This is because the edge maps will remove too much information (colour and texture) that could be useful for object categorisation and background-

[1]The top-k ranking accuracy needs to be interpreted together with the gallery size: a top-1 accuracy of 50% with a gallery size of 10 is roughly equivalent of an accuracy of 5% with a gallery size of 100.

Layer No.	Input Layer(s)	Layer Type	Kernel Size	Stride	Pad	Output
0	-	Input	-	-	-	$227 \times 227 \times 3$
1	0	Conv1	11×11	4	0	$57 \times 57 \times 96$
2	1	Pool1	3×3	2	0	$28 \times 28 \times 96$
3	2	Conv2	5×5	1	2	$28 \times 28 \times 256$
4	3	Pool2	3×3	2	0	$13 \times 13 \times 256$
5	4	Conv3	3×3	1	1	$13 \times 13 \times 384$
6	5	Conv4	3×3	1	1	$13 \times 13 \times 384$
7	6	Conv5	3×3	1	1	$13 \times 13 \times 384$
8	7	Pool5	3×3	2	0	$6 \times 6 \times 256$
9	8	attention-conv1	1×1	1	0	$6 \times 6 \times 256$
10	9	attention-conv2	1×1	1	0	$6 \times 6 \times 256$
11	10	attention-conv3	1×1	1	0	$6 \times 6 \times 1$
12	$8 \oplus (11 \otimes 8)$	GAP*	-	-	-	$1 \times 1 \times 256$
13	8	FC6	1×1	1	0	$1 \times 1 \times 4096$
14	13	FC7	1×1	1	0	$1 \times 1 \times 4096$
15	14	FC8	1×1	1	0	$1 \times 1 \times 256$
16-1	$15 \odot 12$	feature for HOLEF loss	-	-	-	$1 \times 1 \times 512$
16-2	15	feature for classification loss	-	-	-	$1 \times 1 \times 256$

Table 2. The detailed configuration of each branch of the hybrid Siamese-Heterogeneous triplet ranking network. '*' When concatenating the output of the FC8 layer with the attended feature map, we applied global average pooling (GAP) on the attended feature map to reduce the dimensionality of the feature map first. "$\oplus$" is element-wise sum, "$\otimes$" is element-wise product and "$\odot$" is concatenation.

foreground separation. Therefore, the first change is that the photo branch needs to now take the photo image raw pixel values directly as input. This change means that the Siamese network architecture is unsuitable because the domain gap between photo and sketch is too great for the parameter sharing between the corresponding branches to be carried out end-to-end.

Instead, a hybrid Siamese-Heterogeneous triplet ranking network is formulated for the Sketchy database. More specifically, each branch has a similar architecture as AlexNet [2], consisting of 5 convolutional+pooling layers and 3 fully connected layers (one more FC layer than AlexNet to reduce the feature dimension from 4096D to 256D). Importantly, all the convolutional+pooling layers, as well as the attention modules for the photo and sketch domains/branches are now learned independently, i.e., they are heterogeneous branches. After those layers, the three FC layers are tied, i.e., Siamese. This architecture allows the domain specific features at the earlier layers to be learned independently to accommodate the large domain gap before the domains are aligned in the tied FC layers for cross-domain matching. The detailed configurations of each network branch are summarised in Table 2.

Apart from the model architecture adaptation, the learning objective also needs to be modified for this dataset. In particular, in order to learn features that can help the object categorisation task, an object classification loss is added to the final feature output in addition to the HOLEF loss. This loss is given a weight of 0.1 in our implementation to balance its importance against that of the HOLEF loss.

Implementation details Our model is implemented on TensorFlow. Each branch is pretrained by ImageNet only. This is different from the model in the main paper which is pretrained with two more stages on the TU-Berlin sketch recognition dataset [1] and a category-level SBIR dataset [4]. It is also different from the models in [3], which has an additional stage of pretraining using TU-Berlin and a privately collected Flickr photo dataset. The initial learning rate is 0.0001. The batch size is 160 and we exhaustively generate triplet pairs within the same category for HOLEF loss. We do random crop to each input image with a crop size of 227×227 for data augmentation.

2.3. Results

Comparison against the state-of-the-art The only work that published results on the Sketchy dataset is [3] which evaluated a number of baselines. We compare with the strongest baselines from [3] including: **GN Triplet**: This is the best model to date. It uses GoogLeNet as the base network in a heterogeneous architecture (i.e. all branches are untied from end-to-end). **GN Siamese**: same GoogLeNet as the base network but with a Siamese architecture. **AN Siamese**: The base network of this model is similar to ours and based on AlexNet but the architecture is Siamese. All three deep triplet ranking baselines use triplet ranking and classification losses. **Humans**: This is obtained by asking human participants to browse through the

Method	acc.@1
GN Triplet [3]	37.10%
GN Siamese [3]	27.36%
AN Siamese [3]	21.36%
Our model	**43.03**%
Human [3]	54.27%

Table 3. Comparison against published results on the Sketchy database

Method	acc.@1	acc.@10
Base	40.85%	88.83%
Base + CFF	42.68%	88.93%
Base + HOLEF	41.25%	89.08%
Full: Base + CFF + HOLEF	43.03%	88.89%

Table 4. Contributions of the different model components

1,250 gallery photos to find the best match for a given query sketch. The results in Table 3 shows that our model is about 6% higher than the best reported result on this dataset using the top-1 ranking accuracy, although there is still a fairly big gap between our model and the human performance. As we shall see next, this good performance is partly due to the hybrid Siamese-Heterogeneous architecture, whilst the proposed attention module with CFF and HOLEF loss also play an important role. Note this higher result is obtained without using the privately collected Flickr dataset for additional model pretraining which is shown to be useful in [3].

Ablation study To investigate the contribution of the novel model component, namely attention modelling with coarse-fine fusion (CFF) and the HOLEF loss, we compare our full model (**Full: Base+CFF+HOLEF**) with three stripped-down versions: baseline model with coarse-fine fusion (**Base+CFF**), baseline model with HOLEF loss (**Base+HOLEF**) and baseline without either (**Base**). Table 4 shows that each component has a positive effect on the SBIR performance and the best performance is obtained when both components are combined. Note that comparing the Base model (40.85%) with the best reported result (37.10%) obtained by GN Triplet [3], we can see that the hybrid Siamese-Heterogeneous architecture is also important for this cross-category FG-SBIR task.

Qualitative results Fig. 2 shows some qualitative results obtained using our model. In each example (row), the query sketch and the top-10 ranked photos are shown. These results suggest that: (1) The model does a very good job in distinguishing different object categories – most of the 10 photos of the same category as the query sketch are ranked in the top 10. (2) The model is fairly robust against cluttered background and pose variation (see the trumpet example). The results also reveal some of the limitations of this database: (a) The pose now plays a critical role for the SBIR task and the boundary between instance recognition and pose recognition is somewhat blurred. For example, in the elephant example, the 10 elephants in the gallery all have different poses and partial occlusions. The model is not really recognising the identity of each instance (it is after all an extremely difficult task even for humans to tell the difference between two elephants, especially given the abstract sketch as input) – it is more about recognising the pose. This is perhaps contradictory to the original purpose of a typical FG-SBIR application of searching online shopping catalogues, whereby the pose issue does not exist. (b) For some general object categories such as pear and starfish, without colour and texture information, using a sketch to retrieve the correctly matched photo is an ill-conditional problem. For example, the query pear sketch has barely enough information to indicate it is a pear that is being searched, and there is very little cues on exactly which pear is the correct match.

References

[1] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? In *TOG*, 2012. 3

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3

[3] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: Learning to retrieve badly drawn bunnies. In *SIGGRAPH*, 2016. 1, 2, 3, 4

[4] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy. Sketch me that shoe. In *CVPR*, 2016. 3

[5] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. Hospedales. Sketch-a-net that beats humans. *BMVC*, 2015. 1

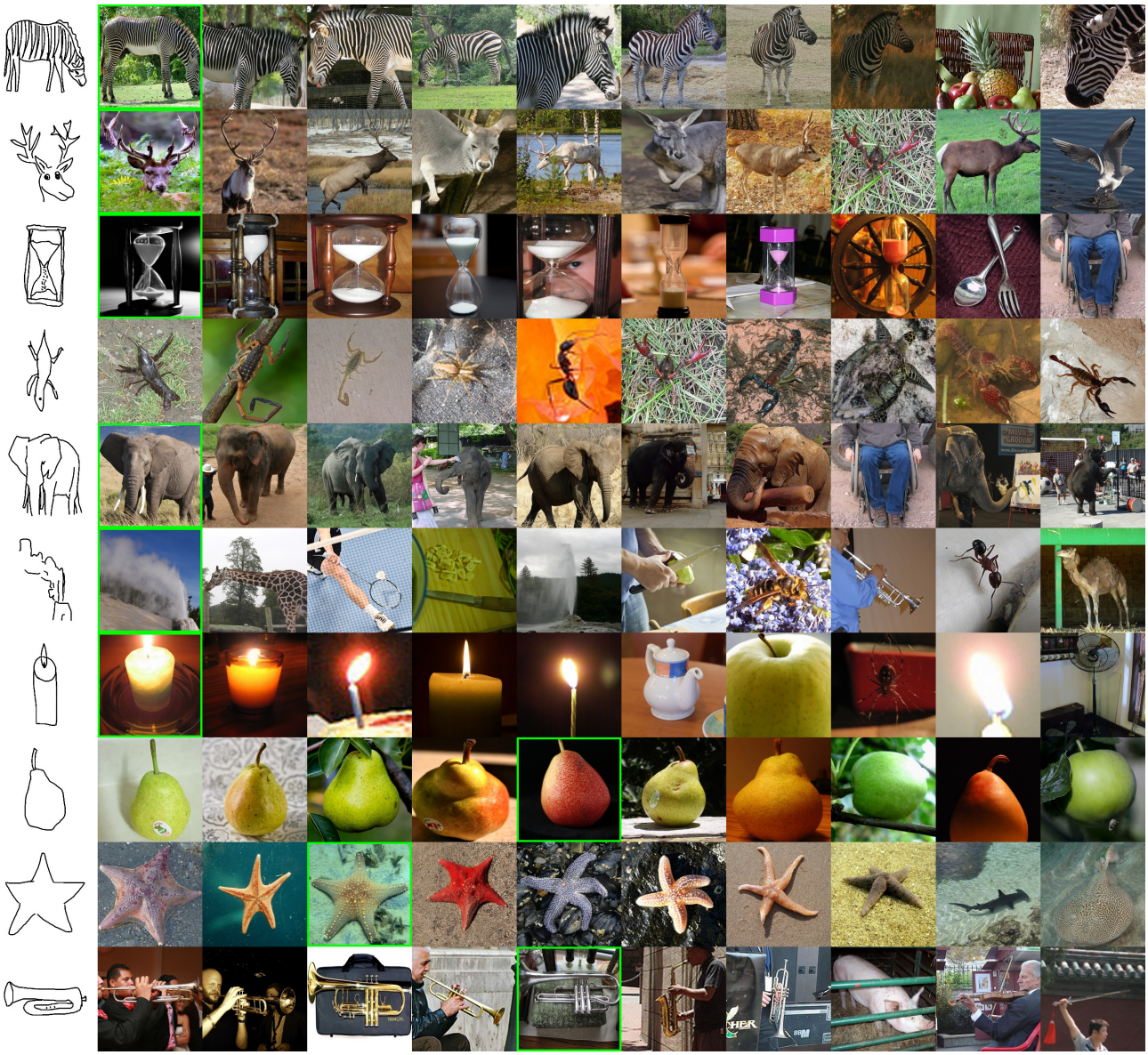

Figure 2. Qualitative performance on the sketchy dataset. In each row, the query sketch is shown on the left followed by the top-10 photos retrieved. The correctly matched photos are indicated by green boxes.