SceneSketcher: Fine-Grained Image Retrieval with Scene Sketches

Fang Liu[1,2⋆], Changqing Zou[3⋆], Xiaoming Deng[1†], Ran Zuo[1,2], Yu-Kun Lai[4], Cuixia Ma[1,2†], Yong-Jin Liu[5,†], Hongan Wang[1,2]

[1]State Key Laboratory of Computer Science and Beijing Key Lab of Human-Computer Interaction, Institute of Software, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences [3] HMI Lab, Huawei Technologies
[4] Cardiff University [5] Tsinghua University

1 Network Details

The node feature vector of our scene graph is extracted with Inception-V3, and its dimension is 2048. Category label c_i of each node is encoded to a 300-d vector by Word2Vec [1]. We denote the spatial information of the node by a 4-d vector p_i indicating the top left and bottom right coordinates of the node bounding box. We adopt two graph convolutional layers in our graph encoder. Table 1 shows the details of our graph encoder. Moreover, we use the *leaky_relu* activation function to replace the conventional *relu* function, which enables the GCN layers to be better trained with triplet loss.

Table 1. Details of our graph encoder. GCN1 and GCN2 are denotes as the first and the second graph convolutional layers, respectively. 'Batch' means batch size.

Layer name	Input	Adjacency matrix	Output dimension
GCN1	(Batch, # of Node, 2352)	(# of Node,# of Node)	(Batch, # of Node, 128)
GCN2	(Batch, # of Node, 128)	(# of Node,# of Node)	(Batch, # of Node, 32)

2 Details of Our Extended Database

In order to investigate the performance of our method using a larger image gallery, we extend our scene sketch database with natural images from Coco-stuff [2], named our extended scene sketch database. We select 21,379 natural images, the objects of which are within the 14 categories in our scene sketch database. These natural images do not have corresponding sketches in our scene sketch database. Then we split these natural images into a test dataset with 5,000 images and a training dataset with 16,379 images, and combine them with the images of the training and test dataset in our scene sketch database.

Table 2 shows the number of sketches and images in our extended database.

⋆ indicates equal contributions. † indicates corresponding author.

Table 2. Summary of our extended scene sketch database.

	# of Sketches	# of Images
Training dataset	1,015	1,015+16,379
Test dataset	210	210+5,000

3 Fine-grained Retrieval

To analyze the performance of our fine-grained scene-level SBIR, we made some images that are extremely similar in overall layout of sketches, category of objects, and their position and shape.

Firstly, we pick up 10 extremely similar images of airplanes (or elephants) from our Extended database mentioned in Section 4.5 in our paper. Corresponding scene sketches are drawn to conduct the SBIR task. We aim at exploring the sorting results of the 10 images with different sketches as inputs. We can infer from the visual results shown in Fig. 1 that the number of instances is the primary factor determining the search results, and then goes for the IoU. We also compare our method with *Sketch me that shoe* [3] on these images, results are shown in Fig. 2.

Secondly, we manually made some similar images using Poisson blending algorithm and Photoshop image editing software (see Fig. 3 and Fig. 4). These images all have the same background, so the SBIR task becomes even more challenging.

These experiments demonstrate that our method can effectively capture the details of the object class, position of objects and their relationships, and can retrieve the desired fine-grained images.

Fig. 1. Retrieval results of our fine-grained scene-level SBIR framework. The ground true matches are highlighted with red rectangles.

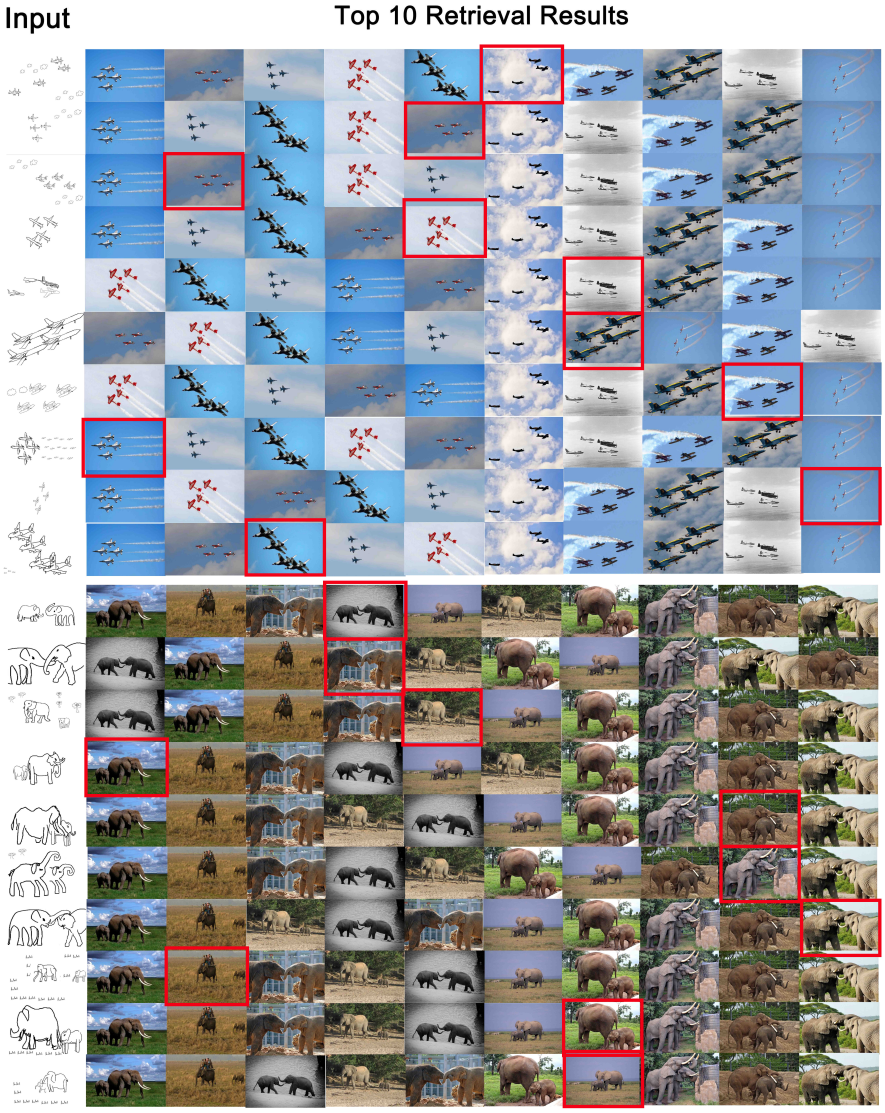

Fig. 2. Retrieval results of *Sketch me that shoe* [3]. The ground true matches are highlighted with red rectangles.

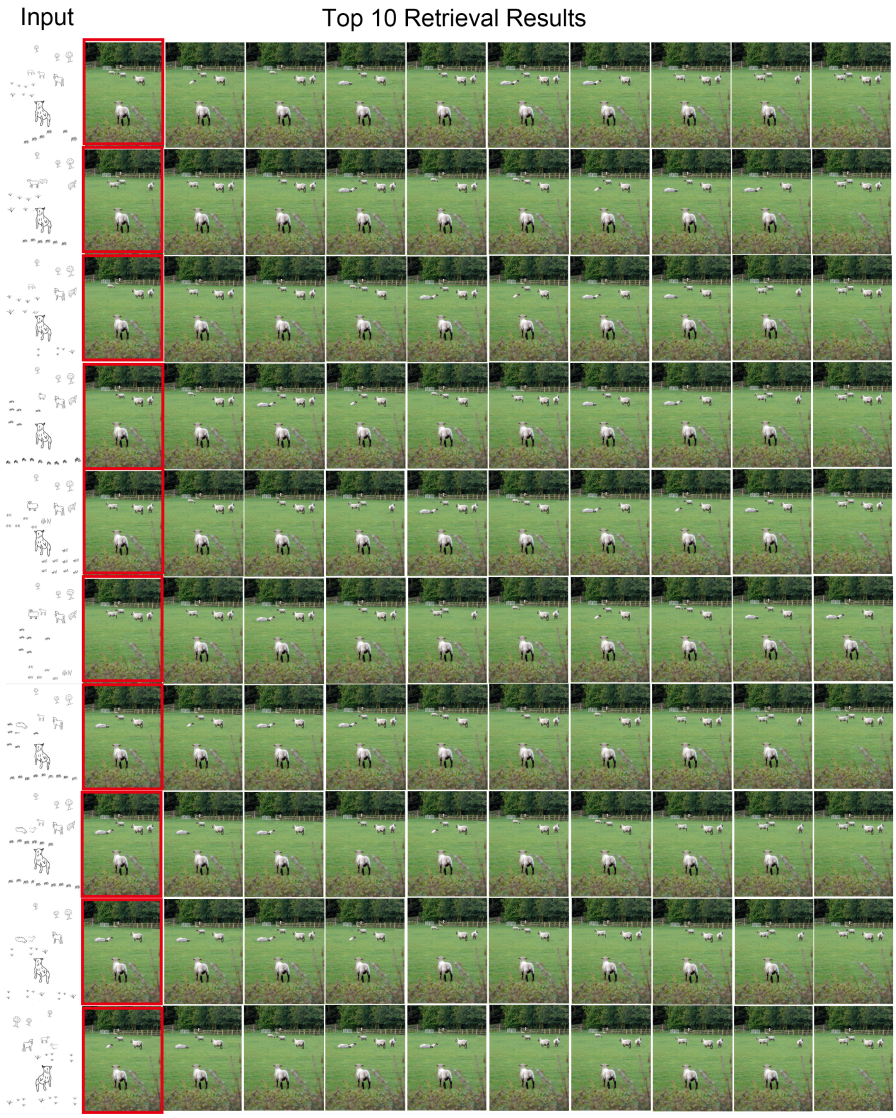

Fig. 3. Retrieval results of our fine-grained scene-level SBIR framework. The ground true matches are highlighted with red rectangles.

Fig. 4. Retrieval results of our fine-grained scene-level SBIR framework. The ground true matches are highlighted with red rectangles.

References

1. https://code.google.com/archive/p/word2vec/
2. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1209–1218 (2018)
3. Yu, Q., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., Loy, C.C.: Sketch me that shoe. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 799–807 (2016)