Semantic Clustering for Robust Fine-Grained Scene Recognition Supplementary Material

Marian George[1], Mandar Dixit[2], Gábor Zogg[1], and Nuno Vasconcelos[2]

[1] Department of Computer Science, ETH Zurich, Switzerland
[2] Statistical and Visual Computing Lab, UCSD, CA, United States

0.1 Qualitative Analysis of Discovered Clusters

In Figure 1, we show sample images from each discovered cluster when using $k = 5$ clusters. Our discovered clusters are semantically meaningful, where each cluster represents scene classes that share common objects, e.g. cluster 1 contains images of flowers and vegitables shared between florist, grocery store, and restaurant classes. In a similar manner, cluster 2 contains images of shelves shared between bookstore, clothes shop and pharmacy classes. Also, cluster 4 show images of seating areas in clothing store, coffee shop, restaurant, shoe shop and sports store. This emphasizes the effectiveness of our semantic clustering approach, and that the proposed represntation successfully exploits the underlying semantic structure in the different scene classes.

0.2 Combining our proposed object-based representation with holistic scene representation, namely Places CNN

We studied the complementarity of object-based and holistic representations for scene classification. Table 1 shows the accuracy of fusing the proposed object based representations with the holistic features derived from layer fc7 of the Places CNN. Combining the two representations produced the best results on both datasets, enabling gains of 3% on SnapStore and around 10% on MIT 67 datasets. This shows that the two representations indeed contain complementary information.

0.3 Posterior class probabilities on MIT67 dataset

In Figure 2, we show the matrix of posterior class probabilities learned by the OOM, for soft detections on MIT67. The figure shows the pobabilities $p(c|o_i;\theta)$ at the confidence level $\theta = 0.1$. The OOM captures the informative objects for each scene class, e.g, desk for office and stretcher for operating room (Fig. 2a). Posterior class probabilities for non-discriminative objects are almost uniform for all scene class as they are hardly detected in any of the scenes (Figure 2b).

Table 1: Classification accuracy for the combination of object-based and holistic classification (Places fc7 features)

Dataset/Method	Places fc7	OOM[RCNN] (Ours)	OOM[CNN] (Ours)	Combined
SnapStore	44.2	47.9	45.4	**51.0**
MIT67	68.2	49.4	68.6	**79.1**

Table 2: Training/Testing configuration for experiment in Sec. 6.3 in main text

Dataset	Places	SUN	SnapStore
number of training images	5363	2548	3590
number of test images	350	300	338

0.4 Qualitative results on MIT67 dataset

In Figure 5b in the main text, we show the top four correctly-classified scene classes in MIT67 sorted from top to bottom by decreasing classification accuracy. For each scene class, we show the most popular object classes (most popular object on the left). The localization accuracy of bounding boxes is less than that of the RCNN (hard detections) method but still informative of the presence and approximate location of a certain object.

Failure cases for MIT67 include prisoncell, elevator, and casino classes. Such classes are characterized by a distinctive global structure with very few or no objects (e.g., elevator).

0.5 Training/Testing configuration of cross-recognition on multiple datasets (Section 6.3)

We ran experiments on the **9** common store categories between the 3 datasets. Namely, the following classes were considered: bookstore, coffee shop, clothing store, florist, restaurant, pharmacy, shoe shop, supermarket, and toystore. In Table 2, we show the number of training images and testing images for each of SUN, Places, and SnapStore datasets. For SnapStore phone test set, we used 264 test images that cover the 9 classes.

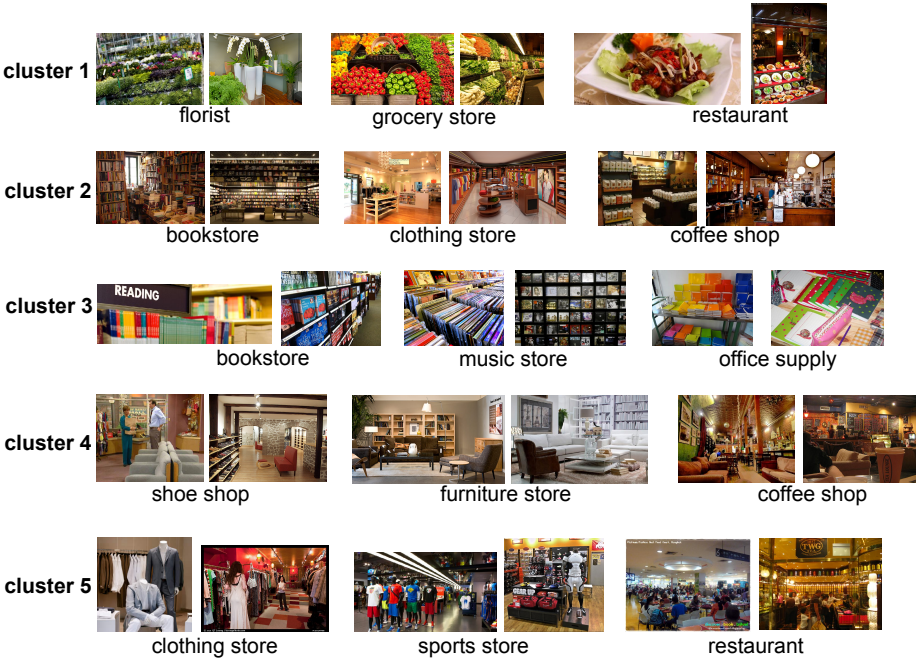
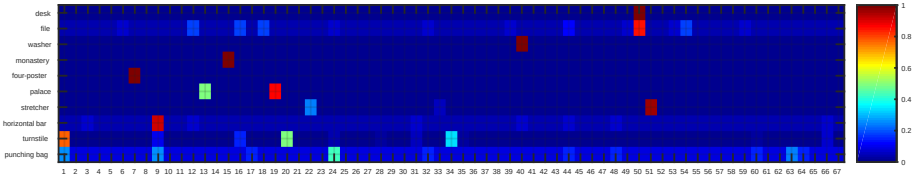
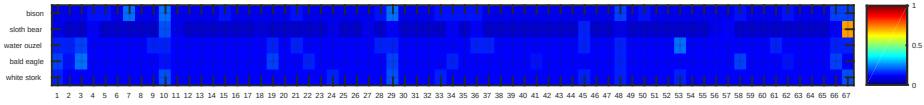


Fig. 1: Sample images from each discovered cluster when using $k = 5$ clusters. Each row shows images from one cluster, specifically 2 images from 3 classes of the cluster. Each cluster represents semantically related classes, e.g. **cluster** 1 contains images of flowers and vegitables shared between florist, grocery store, and restaurant classes. In a similar manner, **cluster** 2 contains images of shelves shared between bookstore, clothes shop, coffee shop and pharmacy classes. **Cluster 3** contains close-up images of books, notebooks, and CDs in bookstore, office supplies and music store. Also, **cluster** 4 show images of seating areas in furniture store, clothing store, coffee shop, restaurant, shoe shop and sports store. Finally, **cluster 5** represents images where people are salient in the scene.



(a) *



(b)

Fig. 2: Scene likelihoods for all scene classes for (a) the top 10 discriminative objects and (b) the least discriminative objects using soft detections (CNN) on MIT67 dataset.
∗-scene names corresponding to relevant IDs:
1: airport inside,
7: bedroom,
9: bowling,
13: church inside,
15: cloister,
19: concert hall,
20: corridor,
22: dentaloffice,
24: elevator,
34: inside bus,
40: laundromat,
50: office,
51: operating room.