Supplementary Material for "Generating Holistic 3D Scene Abstractions for Text-based Image Retrieval"

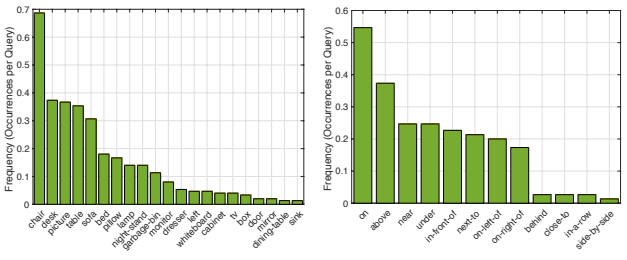
Ang Li Jin Sun Joe Yue-Hei Ng Ruichi Yu Vlad I. Morariu Larry S. Davis
Institution for Advanced Computer Studies
Univesrity of Maryland, College Park, MD 20742
{angli,jinsun,yhng,richyu,morariu,lsd}@umiacs.umd.edu

1. Introduction

The supplementary material for the submission "Generating Holistic 3D Scene Abstractions for Text-based Image Retrieval" is presented below, which provides additional experimental analysis and qualitative results.

2. Additional details about datasets

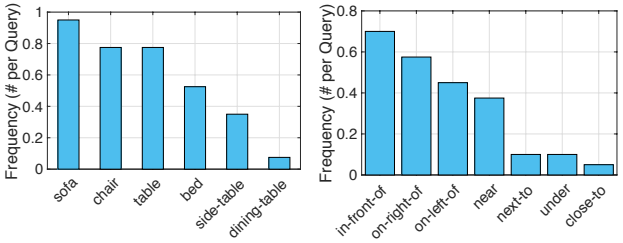
We have 150 of the images annotated in the SUN RGB-D test dataset. We show the statistics about queries used in SUN RGB-D evaluation. In average, SUN RGB-D annotation has 4.26 objects, 2.65 relations, 19.85 words and 2.69 sentences per query. Fig. 1(a) shows the averaged occurrences per query of each object category and Fig. 1(b) shows the averaged occurrences per query of each spatial relation category. The object and relation categories are sorted in the descending order w.r.t. the frequency.



(a) (b)

Figure 1. SUN RGB-D query statistics: (a) frequency (occurrences per query) of objects and (b) frequency (occurrences per query) of spatial relations.

In average, 3DGP annotation has 3.06 objects, 1.94 relations, 17.06 words and 1.94 sentences per query. Similarly, we show the object category and spatial relation frequencies in Fig. 2. Different from the statistics of SUN RGB-D where spatial relation `on` has the highest frequency, spatial relations in 3DGP are mostly horizontal. This is because, for 3DGP, we only have DPM detectors for 6 furniture categories and all of them are on the floor.



(a) (b)

Figure 2. Query statistics in 3DGP evaluation: (a) Frequency (occurrences per query) of objects, and (b) frequency (occurrences per query) of spatial relations.

3. Additional qualitative retrieval results

We provide additional results in SUN RGB-D for top-3 retrievals in Fig. 3. The ground truth image is shown with a blue bar on its top. Although it happens rare in this evaluation, there are cases when there are images other than the ground truth that meet the descriptions of the query (e.g., the last example in Fig. 3). Qualitative results with matched 3D layout are shown in Fig. 4. The figure shows the 3D layouts with camera location corresponding to the best matched 2D spatial layouts (from 5 layout samples).

4. Learned 2D spatial relationships in baseline

The learned distributions of 2D spatial relationships in the nearest neighbor baseline algorithm are shown in Fig. 5. The figure shows the relationship between the subject and the object (*subject-relation-object*) w.r.t. all eight atomic spatial relations (other relations are built upon these atomic relations). For each relationship, the annotated bounding boxes of each pair of subjects and objects are normalized (rescaled in both $x-$ and $y-$ coordinates) so that the subject bounds to a 1×1 square with bottom left $(0, 0)$ and top right $(1, 1)$. All of the normalized relation annotations are visualized in the figure. The nearest neighbor classifier is based on the IOU scores of normalized bounding boxes.

Figure 3. Top 3 retrieved images in SUN RGB-D. Ground truth images appear with **blue** bars on top. Green bounding boxes are detection outputs matching the generated 2D layouts. Red boxes are missing objects (not detected) w.r.t. the expectation of generated 2D layouts.

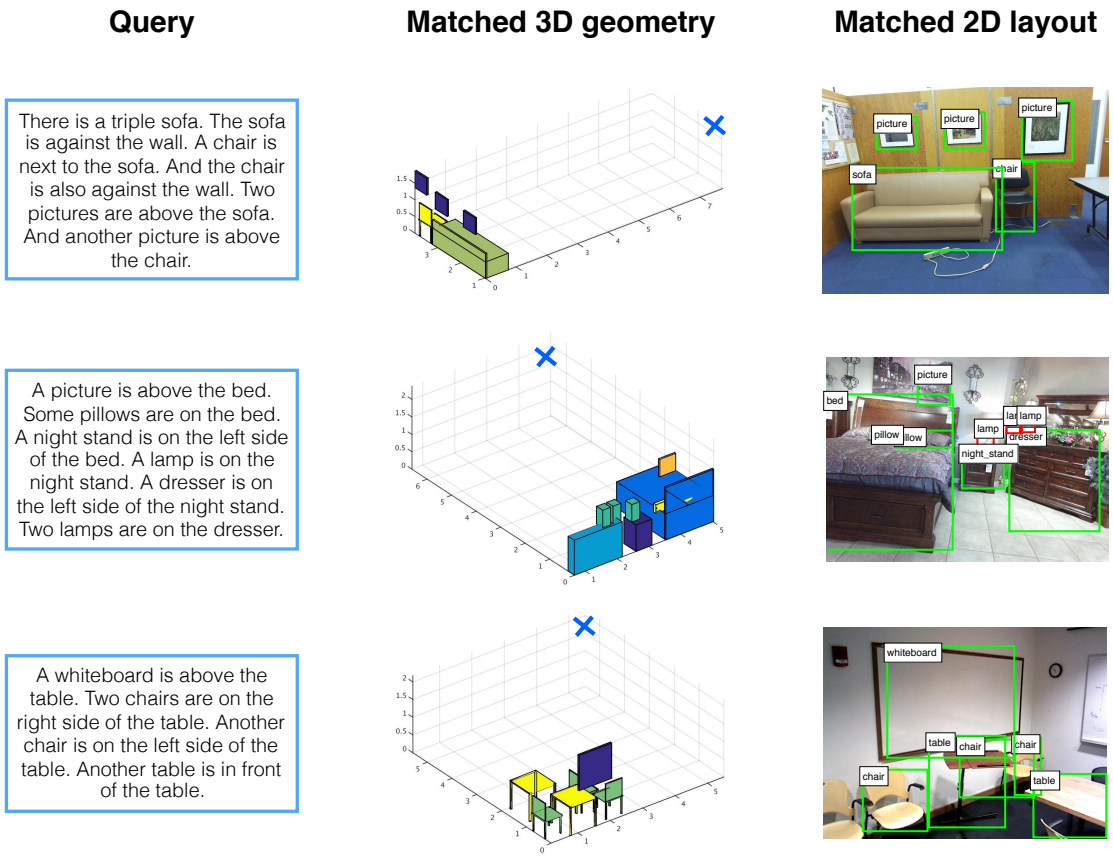
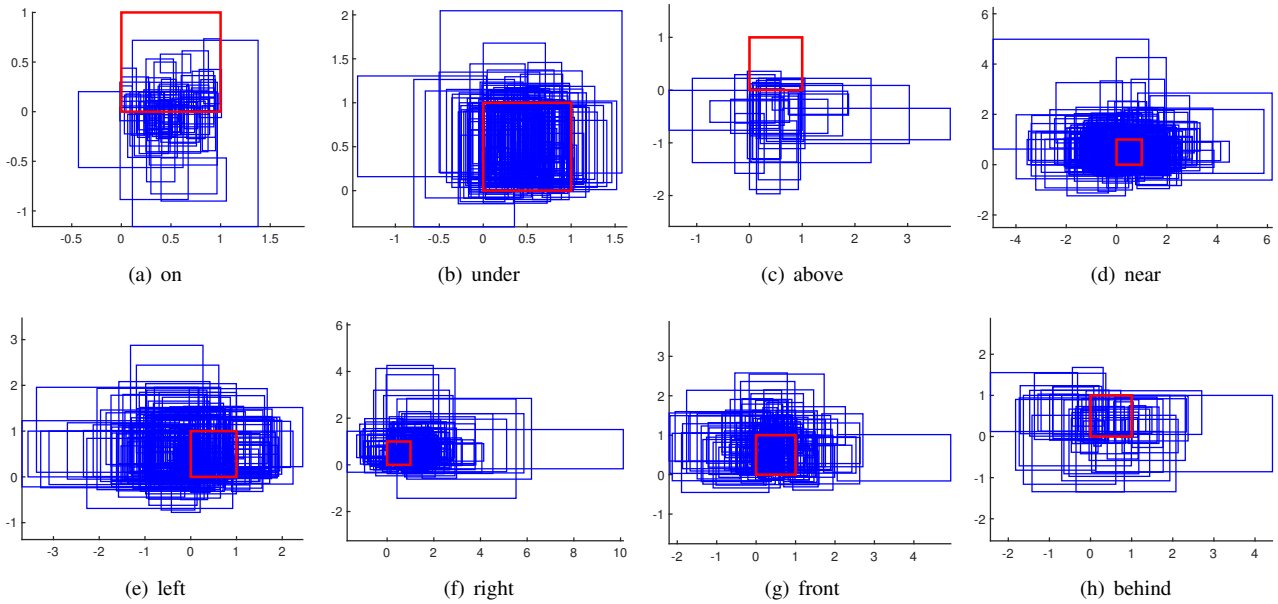


Figure 4. Matched 3D and 2D layouts based on our greedy 2D layout matching for three ground truth images in SUN RGB-D. Blue crosses represent camera locations. Green bounding boxes are object detection outputs that match the 2D layouts generated from the text queries. Red bounding boxes represent a missing object (not detected by the object detector) within the expected region proposed by 2D layouts.



(a) on (b) under (c) above (d) near

(e) left (f) right (g) front (h) behind

Figure 5. Learned distribution of 2D spatial relations in **subject-relation-object** relationships. Red bounding boxes represent the subject and blue bounding boxes represent the sampled objects in the annotations corresponding to each relation. The subject is normalized to 1×1 squares (with bottom-left $(0, 0)$ and top-right $(1, 1)$) and all objects are rescaled with the same normalization factors in x-y coordinates.