Supplemental Material:
Understanding Indoor Scenes using 3D Geometric Phrases

Wongun Choi[1], Yu-Wei Chao[1], Caroline Pantofaru[2], and Silvio Savarese[1]

[1]University of Michigan, Ann Arbor, MI, USA
[2]Google, Mountain View, CA, USA*
{wgchoi, ywchao, silvio}@umich.edu, cpantofaru@google.com

1. Complete Set of Learned GPs Π

Fig. 1 shows the 10 GPs Π learned by the proposed training method. As shown in the figure, the training method learns GPs that appear frequently in realistic indoor scenes. Notice that training method can learn GPs with arbitrary numbers of constituent objects and that the cardinality of a GP is not predefined.

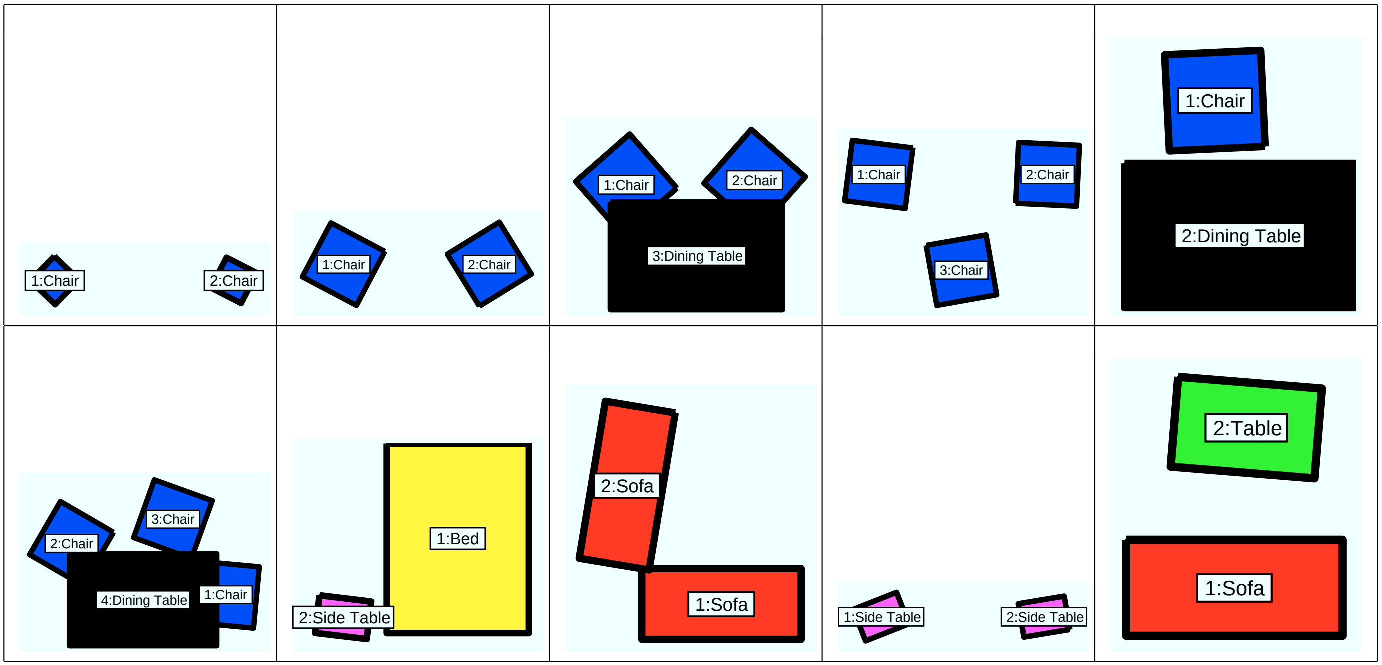

Figure 1. The complete set of learned GP models Π generated by our learning algorithm. Notice that all the learned GPs embed spatially meaningful configurations of objects. A GP hypothesis can have arbitrary orientation.

*This work was done while C. Pantofaru was at Willow Garage, Inc.

2. Example results

In Fig. 2, we present additional examples of results. The left columns show the output of the baseline layout estimator [2]. The middle columns show the result of our system projected into the 2D image. The right columns show the results of our system in 3D, from a top-down view. These example results suggest that our method is capable of producing spatially consistent interpretation of indoor scenes, in which the configurations of objects, layout and scene type are compatible.

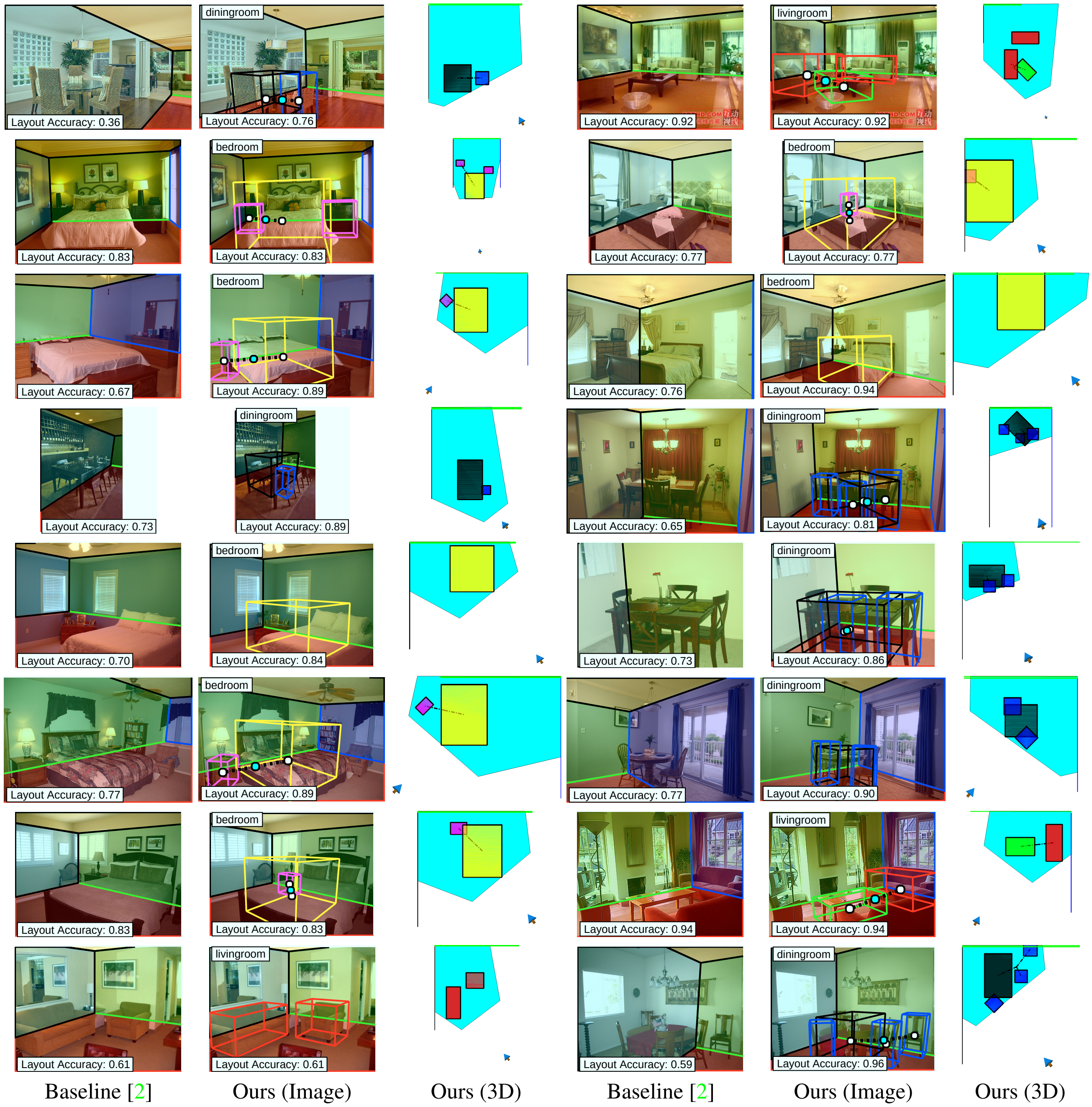

Figure 2. Example results obtained by the baseline layout estimator [2], ours overlaid on the image and ours shown in 3D space (top-view). The camera viewpoint is shown as blue arrows.

3. Robust 3D Object Localization via a Common Ground Plane Assumption

We estimate the 3D extent of the room space by using the method introduced in [7, 3]. Given an image, one can find three (mutually orthogonal) vanishing points by identifying the points where many parallel lines are intersecting. From this, we can estimate the intrinsic camera parameters and camera rotation with respect to the room space (K, R) using the vanishing points [7]. Given the pair of camera parameters K, R and a layout hypothesis l_i, we obtain the corresponding 3D cubic room representation by finding a 3D cuboid that is consistent with the hypothesis l_i (Fig. 3). Such a 3D cuboid can be estimated upto scale due to scale ambiguity as shown in the Fig. 3.

Given a 2D bounding detection o and associated pose p, we localize the object in 3D space as a 3D cuboid O by the following optimization,

$$\hat{O} = \underset{O}{\operatorname{argmin}} \, ||o - P(O, p, K, R)||_2^2 \tag{1}$$

where $P(\cdot)$ is a camera projection function that projects the 3D cuboid O and generates a fitted bounding box in the image plane. We measure the fitness of a projected bounding box by evaluating the euclidean distance between o and $P(O)$. The above optimization is quickly solved with a simplex search method [5].

In practice, an individually estimated 3D cuboid model $\hat{O}$ may be inaccurate due to noisy detection output and intra-class variation in the size of objects as shown in Fig. 4. Also, in order to estimate the absolute scale of the 3D room space, we need to first find the camera height h_c. In order to tackle these issues, we introduce a flexible model that allows each object to have a small variation in its size and assumes a shared ground plane [4]. Each object class is assumed to have a cuboid model with given mean dimensions and one degree of variance in the scale α. Given a set of object hypotheses with an associated 3D cuboid O_i, we can obtain the scale α_i of all of the objects and the height of the camera h_c using the following optimization:

$$\underset{h_c, \alpha_i}{\operatorname{argmin}} \sum_i (\alpha_i min_y(O_i) - h_c)^2 + C \sum_i log(\alpha_i)^2, \; s.t. \; \forall \alpha_i > 0 \tag{2}$$

where $min_y(\cdot)$ gives the minimum y value of a 3D cuboid (bottom of the cuboid). The objective function penalizes i) having any objects floating or submerged into the floor and ii) objects deformed too much from the mean 3D model. For any configuration of positive object hypotheses, we run this optimization to obtain the 3D configuration of the image as shown in the Fig. 4 right bottom. We use $C = 0.1$ in practice.

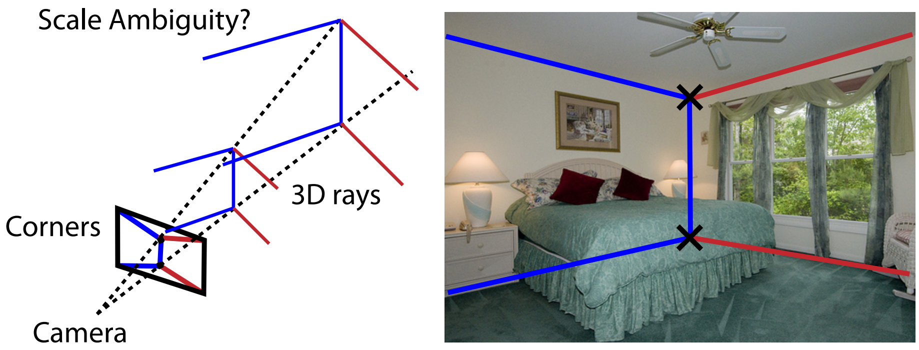

Figure 3. Given the camera parameters K, R and a layout hypothesis l_i (shown as a lines on the image), the 3D room representation can be obtained by finding the cubes that are intersecting with the 3D rays at its corners (corner rays). Corner rays can be obtained by identifying the rays that intersect both the camera aperture and the layout corners (shown as black crosses) in the image plane. Due to scale ambiguity, there exist infinitely many cubes that are consistent with a layout hypothesis. We identify the unique cube by applying the common ground plane assumption (see text).

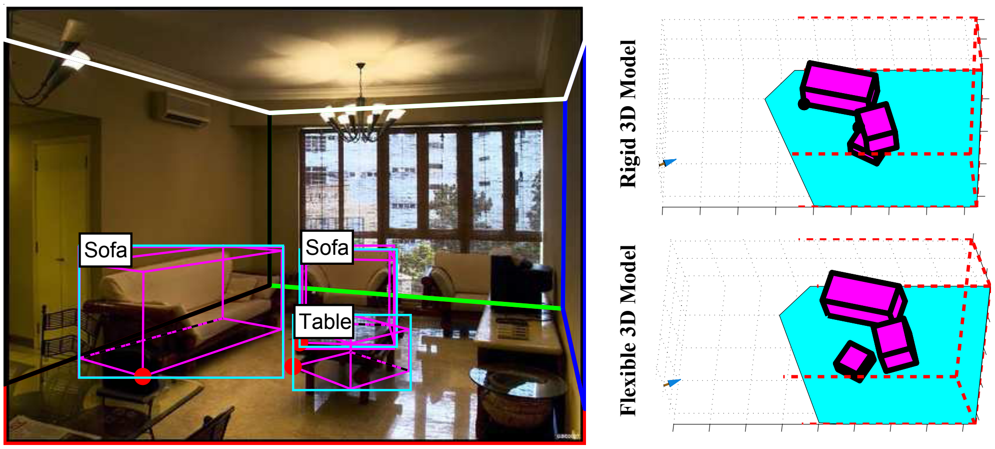

Figure 4. 3D interpretation of the room and objects given the layout and object hypotheses. The left image shows an example of an image with wall face layouts and object hypotheses (bounding box and reprojected polygon) in the image plane. The right two images show the estimated room space (blue arrow for the camera, red-dotted lines for edges and cyan plane for the ground floor) and object cuboids (magenta colored boxes) in 3D space (top: with rigid 3D model and bottom: with flexible 3D - common ground model). As shown in the figure, the rigid 3D model assumption introduces huge error in 3D localization (table is located below and at the simlar distance as a sofa), yet the common grounded model enables the system to obtain a better 3D estimation.

4. Loss Definition

Given a ground truth label $y_i = (C, H, V_T)$ and an estimated label y of an image i, we define the loss function $\delta(y, y_i)$ as a combination of three components: i) object detection loss $\delta_d(V_T, V_{Ti})$, ii) scene classification loss $\delta_s(C, C_i)$ and iii) layout estimation loss $\delta_l(H, H_i)$. Here, V_T represents all positive sets of detection hypotheses in y, H is the selected layout hypothesis in y and C is the scene type of y.

The detection loss is represented as a sum of individual detection losses. Considering the whole set of detection hypotheses $\mathbb{V}_T$, the detection loss is defined as follows:

$$\begin{aligned}\delta_d(V_T, V_{Ti}) = \sum_{V \in \mathbb{V}_T} & \mathbb{I}(V \in V_T) l_{fp}(V, V_{Ti}) \\ & + \mathbb{I}(V \notin V_T) l_{fn}(V, V_{Ti})\end{aligned} \tag{3}$$

The false positive loss $l_{fp}(V, V_{Ti})$ is set to 1 if V does not overlap with any ground truth object with an overlap ratio [1] larger than 0.5. On the other hand, the false negative loss $l_{fn}(V, V_{Ti})$ is set to 1 if V does overlap with any ground truth object in V_{Ti} with an overlap ratio larger than 0.5.

We incorporate the hinge loss [6] as a classification loss.

$$\delta_s(C, C_i) = \mathbb{I}(C \neq C_i) \tag{4}$$

Finally, the layout loss is defined similarly to the one proposed by [2].

$$\delta_l(H, H_i) = \delta_{l1}(H, H_i) + \delta_{l2}(H, H_i) \tag{5}$$

$$\delta_{l1}(H, H_i) = \sum_{k \in [1,5]} d(H_k, H_{ki}) \tag{6}$$

$$\delta_{l2}(H, H_i) = \sum_{k \in [1,5]} 1 - \frac{Area(H_k \cap H_{ki})}{Area(H_k \cup H_{ki})} \tag{7}$$

where H_k is the k^{th} face of a layout hypothesis H, e.g. floor, left wall, or ceiling. $d(H_k, H_{ki}) = 1$ if one of the two is visible and the other is not. $d(H_k, H_{ki}) = 0$ if both are visible or not visible.

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) challenge. *IJCV*, 2010. 4

[2] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered room. In *ICCV*, 2009. 2, 4

[3] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *ECCV*, 2010. 3

[4] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 2008. 3

[5] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM J. on Optimization*, 1998. 3

[6] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. MIT Press, 2006. 4

[7] C. Rother. A new approach for vanishing point detection in architectural environments. *Journal Image and Vision Computing (IVC)*, 2002. 3