Holistic 3D Scene Parsing and Reconstruction from a Single RGB Image Supplemental Material

Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, Song-Chun Zhu

Center for Vision, Cognition, Learning and Autonomy
University of California, Los Angeles

1 Learning of Prior Knowledge

The learning process of our method contains two major steps: i) collecting the statistics of scene categories, object categories, object sizes and supporting relations from SUN RGB-D dataset [1]; ii) collecting the statistics of grouping occurrences and the geometric relations between objects and human from Watch-n-Patch [2].

Using SUN RGB-D, we model the prior of scene types, object categories and support relations by multinoulli distributions. For example, a lamp is supported by the floor with a probability of 0.4 and by a desk with a probability of 0.2. The branching probability is simply counting the frequency of each alternative choice. The distribution of the object sizes is learned via non-parametric kernel density estimation.

The human-centric grouping occurrence and human-object interactions in 3D space are learned from the Watch-n-Patch. This dataset collects the RGB-D videos of human activities in offices and kitchens. Since some activities are irrelevant with objects, we learn the activities of 'reading', 'play-computer', 'take-item' and 'put-down-item' in all the office videos. For each activity, we first extract key frames from each sequence with group activity labels. Then we compute the occurrence frequency of the objects around human within a distance threshold, and model the prior of object category using a multinomial distribution. The geometric relations between the objects and humans are similarly learned by fitting normal distributions of relative distance, height, and orientation between each joint of a human pose and the object center.

2 2D Room Layout Estimation

Similar to [3], we use a keypoint-based room layout representation to train our network. Figure 1 shows the regular room types defined in [4] with their respective keypoints.

Our model is able to predict both keypoint and room type from a full image using a single model. To achieve this goal, we increase the number of channels in the output layer to match the total number of keypoints (in total 48) of all 11

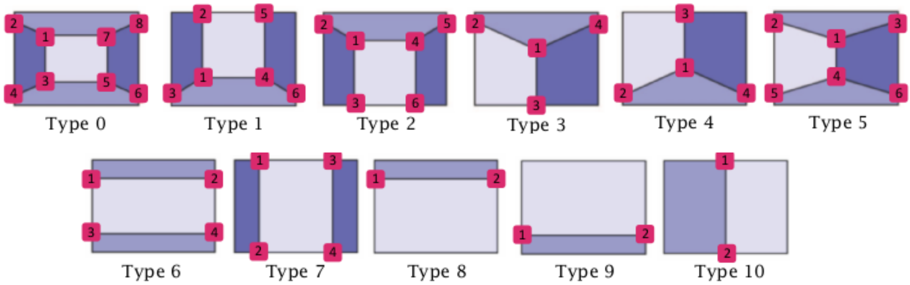

Fig. 1: Types of room layout. The room types are defined in [4]. These 11 room types cover most of possible situations of indoor scenes under Manhattan world assumption [5].

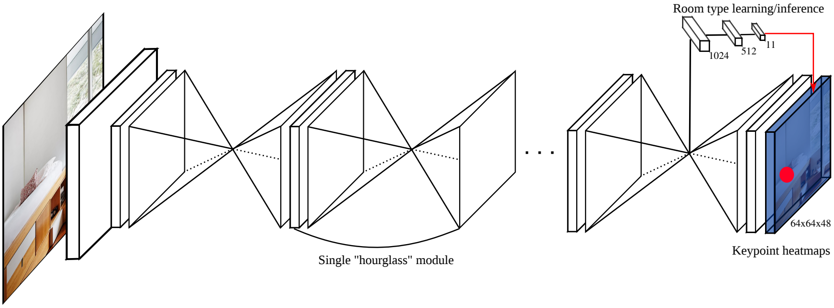

Fig. 2: Network Architecture. The "hourglass" modules work as encoder-decoders which allow for repeated bottom-up, top-down inference.

room types. The cost function is the same as described in [3], which incorporates the Euclidean loss for layout heatmap regression and the cross-entropy loss for room type estimation.

Figure 2 shows our network architecture. Compared with [3], we use the "stacked hourglass" network [6] as our basic network architecture rather than SegNet [7]. Our network consists of multiple stacked hourglass modules which allow for repeated bottom-up, top-down inference.

The input to the network is 256x256. The output of the network is the room type keypoint heatmaps in a resolution 64x64 within a respect room type category label. We use the Adam optimizer [8] with batch size 16, initial learning rate 0.0001. We train 150 epochs, which takes about 2 days on a 12GB NVIDIA TitanX GPU. We also degrade the gradient of background pixels by multiplying them with a factor of 0.2 to prevent the output converges to zero due to the imbalance between foreground and background distribution.

3 Implementation Details

For 2D object detection, we fine-tune the object detector on SUN RGB-D with 30 object categories. Since [4] and [9] have no ground-truth of the camera parameter, we train the 2D layout estimation module using [4] as the initial model,

Table 1: Quantitative comparisons of 2D layout estimation on LSUN [4] and Hedau dataset [9]

Method	LSUN Keypoint Error (%)	LSUN Pixel Error (%)	Hedau Pixel Error(%)
Hedau *et al.* (2009) [9]	15.48	24.23	21.20
Zhao *et al.* (2013) [15]	-	-	14.50
Mallya *et al.* (2015) [16]	11.02	16.71	12.83
Dasgupta *et al.* (2016) [17]	8.20	10.63	9.73
Ren *et al.* (2016) [18]	7.57	5.23	8.67
Izadinia *et al.* (2017) [19]	-	10.04	10.15
Lee *et al.* (2017) [3]	6.30	9.86	8.34
Zhao *et al.* (2017) [20]	5.29	**3.84**	**6.60**
Ours (init.)	**5.22**	4.53	7.03

followed by using the feature of the heatmap to further train camera parameter and scene category on SUN RGB-D. During the initialization and joint inference process, we use the depth estimation model as described in [10], surface normal estimation in [11], and semantic segmentation in [12]. These models are trained on the training set of the SUN RGB-D or NYU v2 dataset [13] (included in the SUN RGB-D). Here, we further incorporate human context inference on the subset of offices and skip it on other scenes. During joint inference, we fix the scene category, object categories and support relations to reduce the computational complexity. We used OpenGL [14] to render the depth, surface normal and segmentation map. Rendering each map takes about 1 second. On average, our joint inference process takes about one hour for each image using a single CPU core.

4 Supplemental Experimental Results

4.1 Evaluation of 2D Layout Estimation

We evaluate the 2D layout estimation without joint inference on LSUN dataset [4] and Hedau dataset [9]. The LSUN dataset consists of 4000 training, 394 validation and 1000 test images. The Hedau dataset contains 209 training, 56 validation and 105 test images. We follow the standard evaluation procedure [17] and use pixel errors and keypoint errors as two evaluation metrics. Pixel errors compute the pixel-wise error between the ground truth and estimations of the surface label, and the keypoint errors only considers the average Euclidean distance between the annotated and estimated keypoints. As reported in Table 1, our approach achieves 5.22% keypoint error which outperforms all existing methods and comparable pixel error with the previous best results [20] on both LSUN and Hedau dataset.

4.2 Evaluation of Camera Parameter Estimation

We compute the mean absolute error between our estimation and the groundtruth on testing set of SUN RGB-D. As shown in Table 2, comparing with the traditional geometry-based method [9], the proposed method gains a significant improvement. Quantitative results of the comparison over all the scene categories are shown in Figure 3. Empirically, geometry-based methods perform poorly in cluttered scenes (*e.g.*, storage rooms) and perform well in clear scenes with clear orthogonal lines (*e.g.*, receptions). Our method provides a good estimation which applies to most of the indoor scenes, improve the generalization ability of the monocular reconstruction algorithms.

Figure 3 shows the comparison in detail over all categories. We can see that the geometry-based method results in large errors over cluttered scenes like storage rooms and performs well over the scenes with clear lines in three orthogonal directions like receptions. Our method provides a good estimation which applies to most of the indoor scenes, improve the generalization ability for the *single view reconstruction* algorithms.

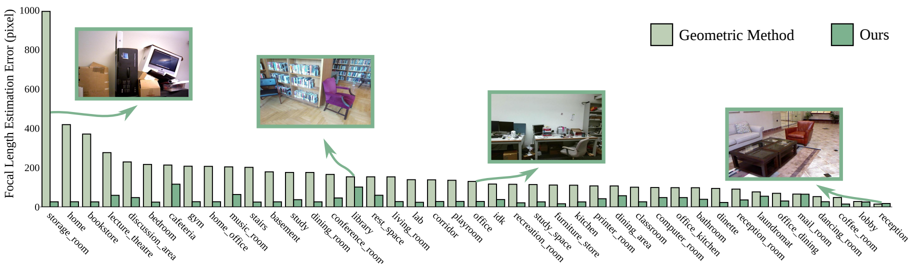

Fig. 3: Estimation error of focal length.

4.3 Evaluation of 3D Layout Estimation

Figure 4 shows the comparison with 3DGP over all categories; we can also observe that 3DGP fails in some scene categories such as dinette and cafeteria, which further reflects the drawbacks of the geometry-based methods.

4.4 Evaluation of 3D Object Detection

Table 3 shows the evaluation of 3D object detection over 30 categories of objects.

Table 2: Camera parameter estimation.

Method	Mean Absolute Error		
	focal length	pitch	roll
Hedau et al. [9]	141.78	3.45	33.85
Ours	**35.87**	**3.12**	**7.60**

Table 3: Comparisons of 3D object detection on SUN RGB-D dataset.

Method	bed	chair	sofa	table	desk	toilet	fridge	sink	bathtub	bookshelf	counter	door	dresser	lamp	tv
[21]	5.62	2.31	3.24	1.23	-	-	-	-	-	-	-	-	-	-	-
Ours (init.)	45.55	5.91	23.64	4.20	2.50	1.91	14.00	2.12	0.55	2.16	0.34	0.01	5.69	1.12	0.62
Ours (joint.)	**58.29**	**13.56**	**28.37**	**12.12**	**4.79**	**16.50**	**15.18**	**2.18**	**2.84**	**7.04**	**1.60**	**1.56**	**13.71**	**2.41**	**1.04**
nightstand	books	tvstand	sofachair	cabinet	endtable	dressermirror	person	recyclebin	curtain	whiteboard	mirror	picture	paper	computer	
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
5.83	0.00	3.04	8.87	0.00	0.65	17.16	1.31	0.00	0.27	0.00	0.00	0.00	0.00	0.00	
8.80	**0.02**	**6.69**	**16.99**	**0.48**	**3.15**	**19.43**	**4.04**	**0.63**	**0.40**	**0.20**	**0.00**	**0.00**	**0.00**	**0.00**	

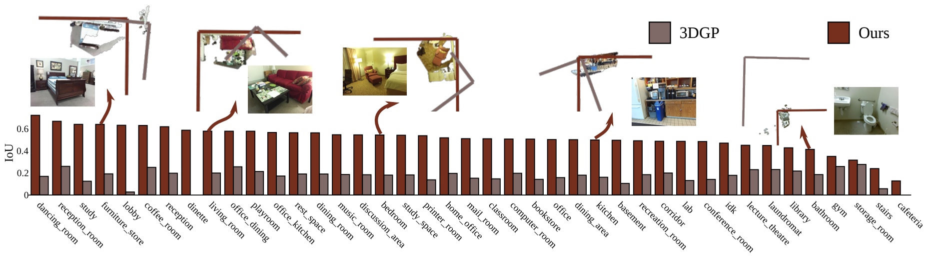

Fig. 4: Quantitative comparisons of 3D layout estimation.

5 More Qualitative Results

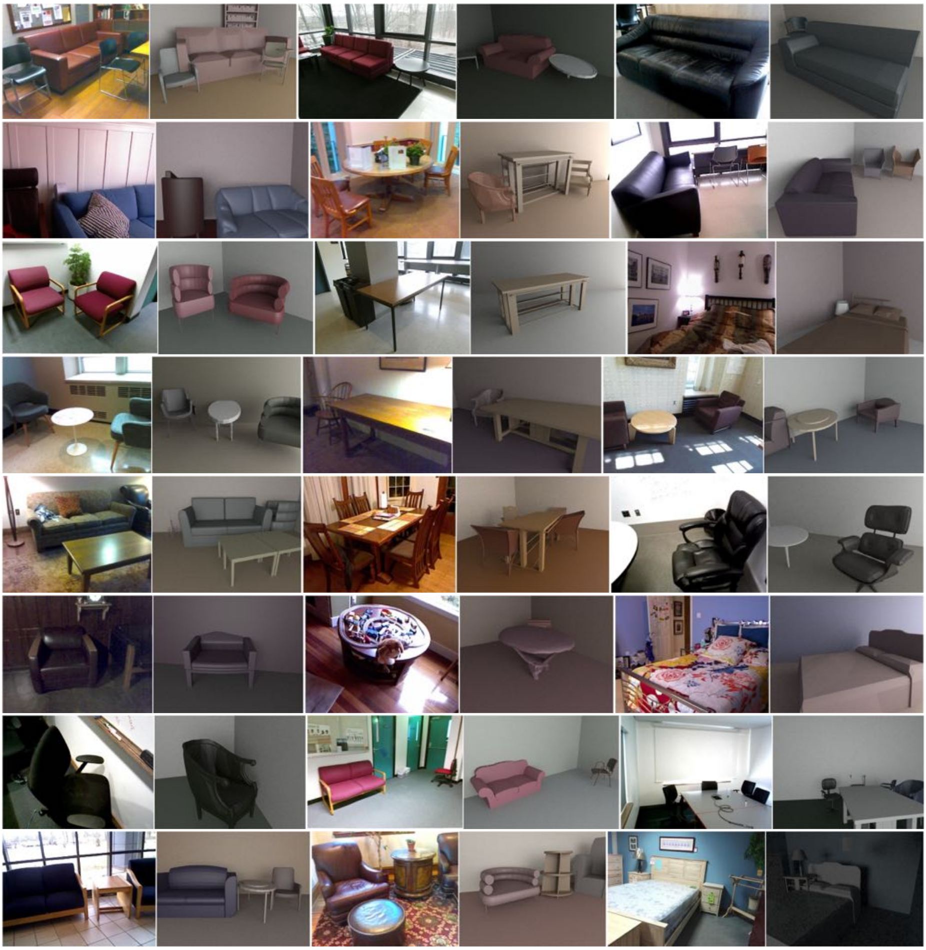

Fig. 5: More qualitative results (cont.)

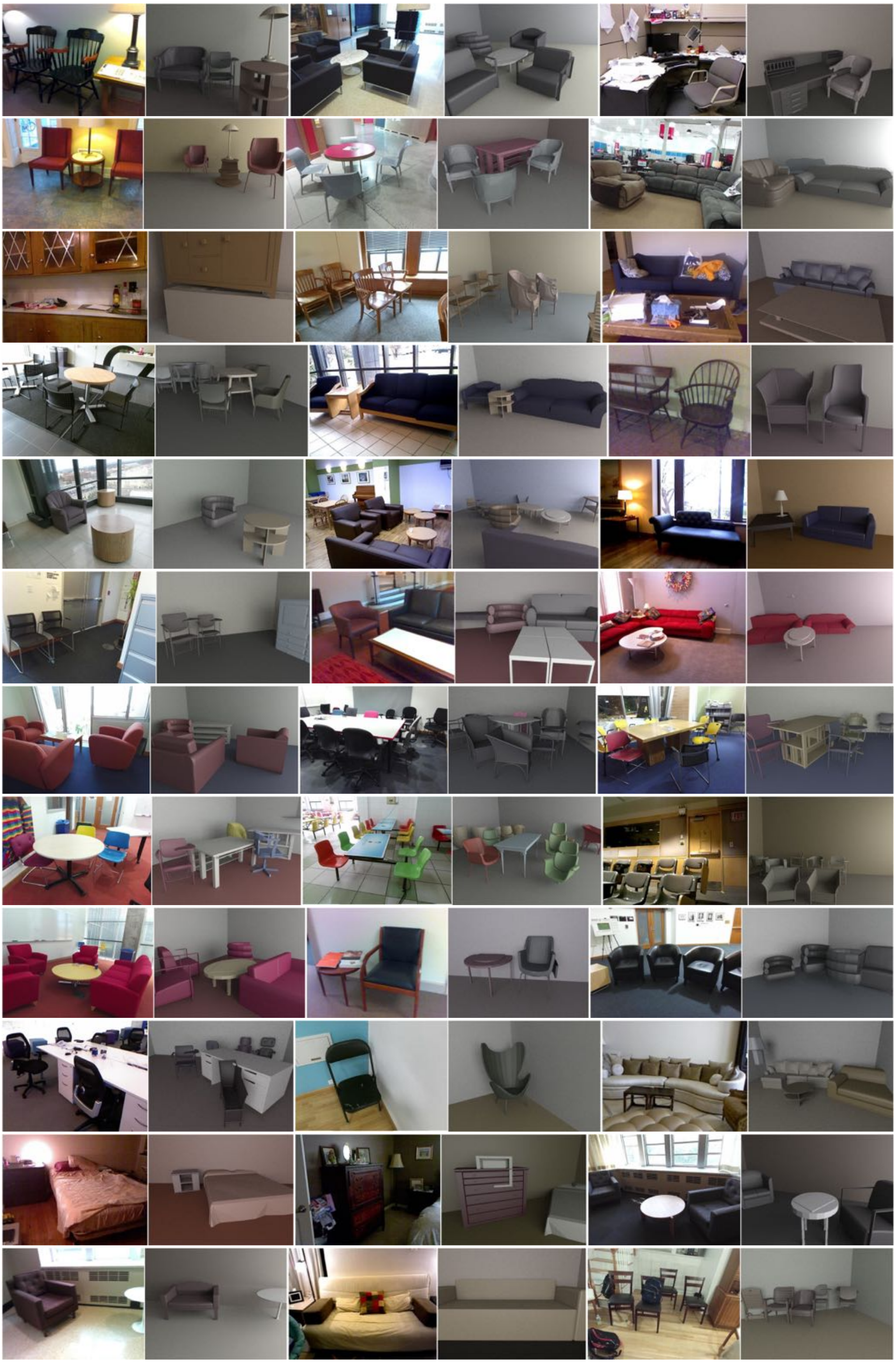

Fig. 6: More qualitative results

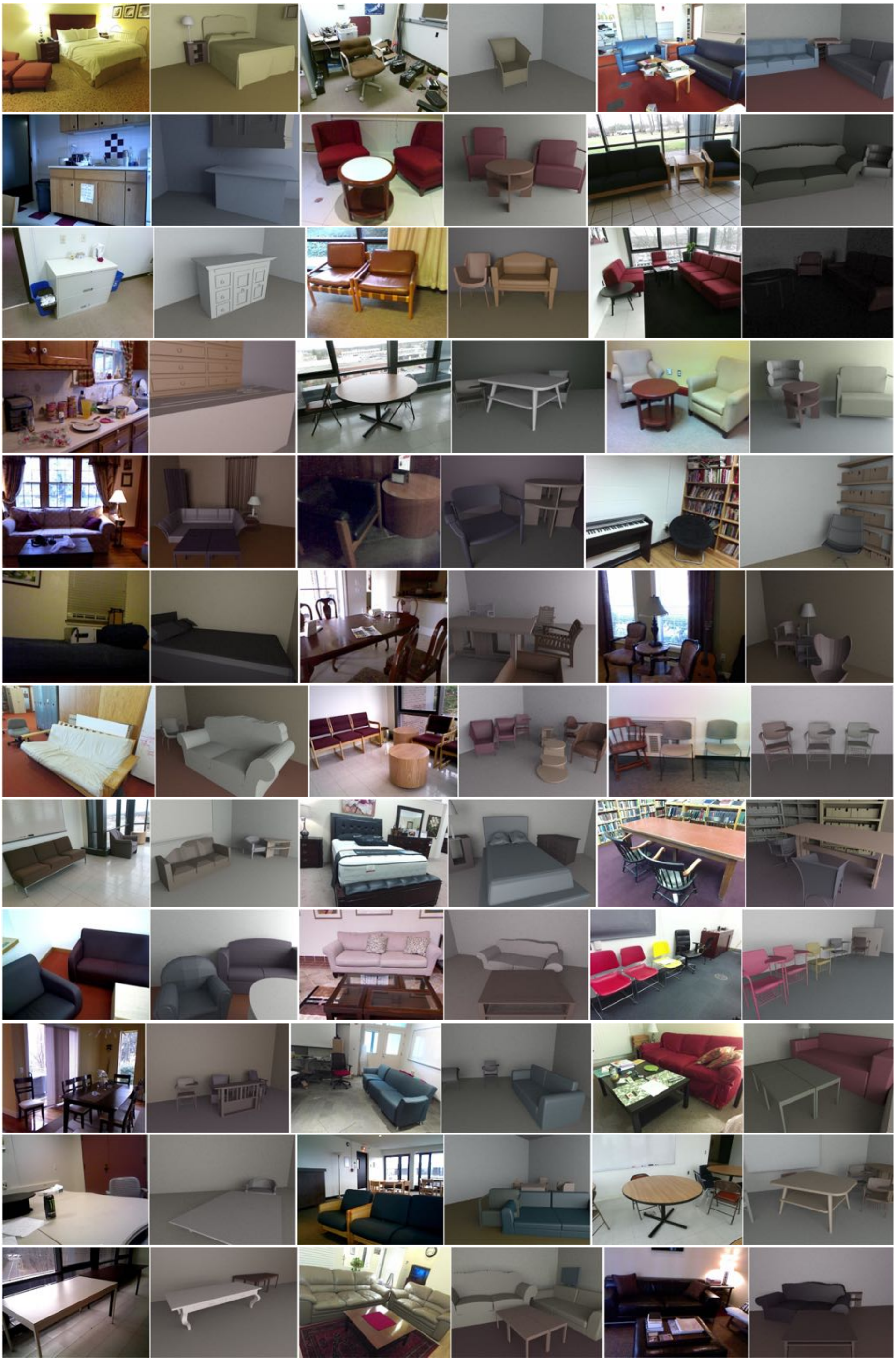

Fig. 7: More qualitative results (cont.)

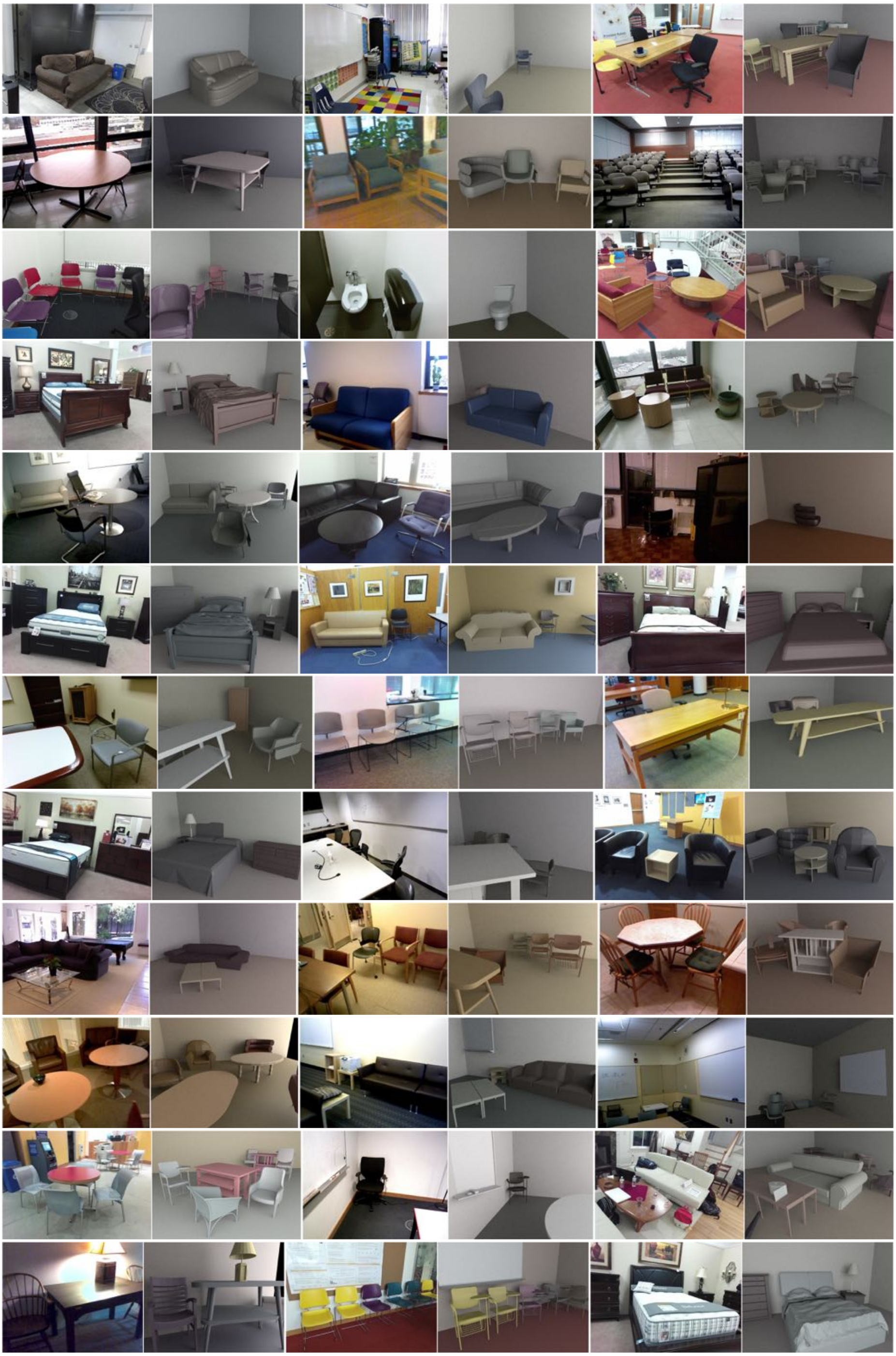

Fig. 8: More qualitative results (cont.)

References

1. Song, S., Lichtenberg, S.P., Xiao, J.: Sun RGB-D: A RGB-D scene understanding benchmark suite. In: CVPR. (2015)
2. Wu, C., Zhang, J., Savarese, S., Saxena, A.: Watch-n-patch: Unsupervised understanding of actions and relations. In: CVPR. (2015)
3. Lee, C.Y., Badrinarayanan, V., Malisiewicz, T., Rabinovich, A.: Roomnet: End-to-end room layout estimation. In: ICCV. (2017)
4. Zhang, Y., Yu, F., Song, S., Xu, P., Seff, A., Xiao, J.: Large-scale scene understanding challenge: Room layout estimation. In: CVPR Workshop. (2015)
5. Coughlan, J.M., Yuille, A.L.: Manhattan world: Compass direction from a single image by bayesian inference. In: CVPR. (1999)
6. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV. (2016)
7. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for scene segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **39**(12) (2017) 2481–2495
8. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: CVPR. (2009)
10. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: International Conference on 3D Vision (3DV). (2016)
11. Zhang, Y., Song, S., Yumer, E., Savva, M., Lee, J.Y., Jin, H., Funkhouser, T.: Physically-based rendering for indoor scene understanding using convolutional neural networks. In: CVPR. (2017)
12. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: CVPR. (2017)
13. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. (2012)
14. Shreiner, D., Group, B.T.K.O.A.W., et al.: OpenGL programming guide: the official guide to learning OpenGL, versions 3.0 and 3.1. Pearson Education (2009)
15. Zhao, Y., Zhu, S.C.: Scene parsing by integrating function, geometry and appearance models. In: CVPR. (2013)
16. Mallya, A., Lazebnik, S.: Learning informative edge maps for indoor scene layout prediction. In: ICCV. (2015)
17. Dasgupta, S., Fang, K., Chen, K., Savarese, S.: Delay: Robust spatial layout estimation for cluttered indoor scenes. In: CVPR. (2016)
18. Ren, Y., Li, S., Chen, C., Kuo, C.C.J.: A coarse-to-fine indoor layout estimation (cfile) method. In: Asian Conference on Computer Vision (ACCV). (2016)
19. Izadinia, H., Shan, Q., Seitz, S.M.: Im2cad. In: CVPR. (2017)
20. Zhao, H., Lu, M., Yao, A., Guo, Y., Chen, Y., Zhang, L.: Physics inspired optimization on semantic transfer features: An alternative method for room layout estimation. In: CVPR. (2017)
21. Choi, W., Chao, Y.W., Pantofaru, C., Savarese, S.: Understanding indoor scenes using 3d geometric phrases. In: CVPR. (2013)