Supplementary Material for Human-centric Indoor Scene Synthesis Using Stochastic Grammar

Siyuan Qi[1] Yixin Zhu[1] Siyuan Huang[1] Chenfanfu Jiang[2] Song-Chun Zhu[1]

[1] UCLA Center for Vision, Cognition, Learning and Autonomy
[2] UPenn Computer Graphics Group

1. Simulated Annealing

The simulated annealing schedule is important for synthesizing realistic scenes. In our experiments, we set the total sampling iterations to 20000, and it takes around 20 minutes to sample an interior layout. We use the following simulated schedule for sampling:

$$T(t) = \frac{T_0}{\ln(1+t)} \tag{1}$$

where $T(t)$ is the temperature at iteration t. Geman *et al.* [5] proved that $T(t) \geq \frac{T_0}{\ln(1+t)}$ is a necessary and sufficient condition to ensure convergence to the global minimum with probability one.

2. Data Effectiveness

We further demonstrate that our data can be utilized to improve performance on two scene understanding tasks: depth estimation and surface normal estimation from single RGB images. We show that the performance of state-of-art methods can be improved when trained with our synthesized data along with natural images.

Depth estimation Single-image depth estimation is a fundamental problem in computer vision, which has found broad applications in scene understanding, 3D modeling, and robotics. The problem is challenging since no reliable depth cues are available. In this task, the algorithms output a depth image based on a single RGB input image.

To demonstrate the efficacy of our synthetic data, we compare the depth estimation results provided by models trained following protocols similar to those we used in normal prediction with the network in [6]. To perform a quantitative evaluation, we used the metrics applied in previous work [3]:

- Abs relative error: $\frac{1}{N}\sum_p \frac{\left|d_p - d_p^{gt}\right|}{d_p^{gt}}$,
- Square relative difference: $\frac{1}{N}\sum_p \frac{\left|d_p - d_p^{gt}\right|^2}{d_p^{gt}}$,
- Average $\log_{10}$ error: $\frac{1}{N}\sum_x \left|\log_{10}(d_p) - \log_{10}(d_p^{gt})\right|$,
- RMSE :$\sqrt{\frac{1}{N}\sum_x \left|d_p - d_p^{gt}\right|^2}$,
- Log RMSE:$\sqrt{\frac{1}{N}\sum_x \left|\log(d_p) - \log(d_p^{gt})\right|^2}$,
- Threshold: $\%$ of d_p s.t. $\max\left(\frac{d_p}{d_p^{gt}}, \frac{d_p^{gt}}{d_p}\right) <$ threshold,

where d_p and d_p^{gt} are the predicted depths and the ground truth depths at the pixel indexed by p, respectively, and N is the number of pixels in all the evaluated images. The first five metrics capture the error calculated over all the pixels; lower values are better. The threshold criteria capture the estimation accuracy; higher values are better.

Table 1 summarizes the results. We can see that the model pretrained on our dataset and fine-tuned on the NYU-Depth V2 dataset achieves the best performance, both in error and accuracy. This demonstrates the usefulness of our dataset in improving algorithm performance in scene understanding tasks.

Surface normal estimation Predicting surface normals from a single RGB image is an essential task in scene understanding since it provides important information in recovering the 3D structure of the scenes. We train a neural network with our synthetic data to demonstrate that the perfect per-pixel ground truth generated using our pipeline could be utilized to improve upon the state-of-the-art performance on a specific scene understanding task. Using the fully convolutional network model described by Zhang *et al.* [7], we compare the normal estimation results given by models trained under two different protocols: (i) the network is directly trained and tested on the NYU-Depth V2 dataset, and (ii) the network is first pre-trained using our synthetic data, then fine-tuned and tested on NYU-Depth V2.

Following the standard evaluation protocol [4, 1], we evaluate a per-pixel error over the entire dataset. To evaluate the prediction error, we computed the mean, median, and RMSE of angular error between the predicted normals and ground truth normals. Prediction accuracy is given by

Table 1: Depth estimation with different training protocols.

		Error					Accuracy		
pre-Train	fine-Tune	Abs Rel	Sqr Rel	Log10	RMSE(linear)	RMSE(log)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
NYUv2	-	0.233	0.158	0.098	0.831	0.117	0.605	0.879	0.965
Ours	-	0.241	0.173	0.108	0.842	0.125	0.612	0.882	0.966
Ours	NYUv2	**0.226**	**0.152**	**0.090**	**0.820**	**0.108**	**0.616**	**0.887**	**0.972**

Table 2: Normal estimation with different training protocols.

pre-train	fine-tune	mean↓	median↓	$11.25°\uparrow$	$22.5°\uparrow$	$30°\uparrow$
NYUv2		27.30	21.12	27.21	52.61	64.72
Eigen [2]		22.2	15.3	38.6	64.0	73.9
[7]	NYUv2	21.74	14.75	39.37	66.25	76.06
ours+[7]	NYUv2	**21.47**	**14.45**	**39.84**	**67.05**	**76.72**

calculating the fraction of pixels that are correct within a threshold t, where $t = 11.25°, 22.5°, 30°$. Our experimental results are summarized in Table 2. By utilizing our synthetic data, the model achieves better performance. The error mainly accrues in the area where the ground truth normal map is noisy. We argue that part of the reason is due to the sensor's noise or sensing distance limit. Such results in turn imply the importance to have perfect per-pixel ground truth for training and evaluation.

3. More Qualitative Results

See page 3-9.

References

[1] A. Bansal, B. Russell, and A. Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. *arXiv preprint arXiv:1604.01347*, 2016.

[2] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015.

[3] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014.

[4] D. F. Fouhey, A. Gupta, and M. Hebert. Data-driven 3d primitives for single image understanding. In *ICCV*, 2013.

[5] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *PAMI*, 1984.

[6] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, 2015.

[7] Y. Zhang, S. Song, E. Yumer, M. Savva, J.-Y. Lee, H. Jin, and T. Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *CVPR*, 2017.