Indoor Scene Structure Analysis for Single Image Depth Estimation

Wei Zhuo, Mathieu Salzmann, Xuming He, Miaomiao Liu
Australian National University & NICTA

Single image depth estimation is a very challenging and ambiguous task. In recent years, much progress has been made by exploiting training image-depth pairs [1, 2, 3, 4, 5, 6, 7]. These approaches, however, typically model depth only at local scale. For instance, [3] predict the depth of each pixel individually. While, in contrast, [2, 4, 5, 6, 7] encode some higher-level information by modeling the relationship of neighboring superixels, the resulting methods still lack reasoning about the global structure of the scene.

In this paper, we propose to exploit high-level scene structure for detailed single image depth estimation. To this end, we introduce an approach that relies on a hierarchical representation of the scene depth encoding local, mid-level and global information. This lets us model the detailed depth of a scene while still benefiting from information about its global structure.

More specifically, our hierarchical representation of the scene depth consists of three layers: superpixels, regions and layout. The superpixels allow us to model the local depth variations in the scene. In contrast, the regions and layout let us account for mid- and large-scale scene structures. We model the depth estimation problem with a Conditional Markov Random Field (CRF) with variables for each layer in our hierarchy. This CRF allows us to encode interactions within and across layers, and thus to effectively exploit local and global information jointly. To this end, let us denote by Y, R and L the variables that represent local depth, mid-level and global structures, respectively. Inference in our CRF is achieved by minimizing the energy

$$E(Y,R,L) = E_l(Y) + E_m(Y,R) + E_g(Y,L) \ , \tag{1}$$

where the first term englobes a unary and a pairwise potential for the local variables only, the second term contains a unary potential for the region variables and a pairwise potential modeling the relationships between superpixels and regions, and the last term encodes the relationship between superpixels and global scene layout. As illustrated by Fig. 1, inference in our model therefore yields depth estimation ranging from coarse (i.e., layout) to fine (i.e., superpixels) levels of details.

We demonstrate the effectiveness of our model on two standard indoor datasets. Here, we compare our results with several baselines using five error metrics: the average relative error (**rel**), the average log_{10} error (**log10**), the root mean square error (**rms**), and two percentages of pixels whose depth error is below a threshold [3]. The results on the NYUv2 depth dataset [8] are shown in Table 1. Our model outperforms DepthTransfer and DC-Depth on all metrics, and SemanticDepth on one out of two thresholds, despite the fact that this baseline relies on additional semantic information. In Table 2, we analyze the different parts of our model. This analysis evidences the fact that each layer contributes to improving the final accuracy. It also reveals that the mid-level structures yield the major contribution. More experiments are provided in the paper, which, altogether, evidence the benefits of exploiting higher-level scene structure for single image depth estimation.

Method	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$
DepthTransfer [2]	0.374	0.134	1.12	49.81%	79.46%
DC-Depth [5]	0.335	0.127	1.06	51.55%	82.32%
SemanticDepth [3]	-	-	-	**54.22%**	82.90%
Ours	**0.305**	**0.122**	**1.04**	52.50%	**83.77%**

Table 1: **NYUv2: Comparison of our approach with the baselines.** Note that we outperform the two baselines (DepthTransfer and DC-Depth) working under the same settings as us. Note also that we outperform the SemanticDepth approach on one out of two thresholds, despite the fact that we do not make use of any pixel label information.

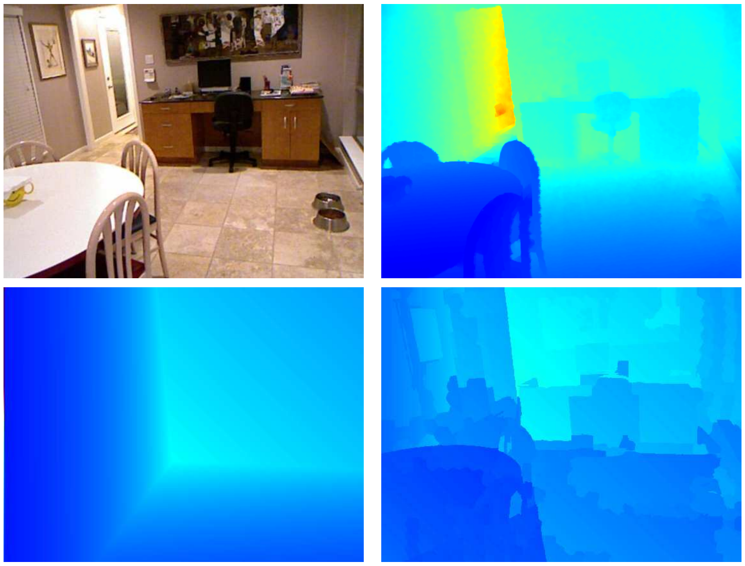

Figure 1: **Depth estimation from a single image:** (Top) Image and ground-truth depth map. (Bottom) Estimated layout and detailed depth map. Color indicates depth (red is far, blue is close).

Method	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$
Ours-local	0.334	0.128	1.05	50.35%	82.31%
Ours-mid	0.312	0.123	**1.03**	52.08%	**83.92%**
Ours-global-only	0.325	0.128	1.07	50.38%	82.06%
Ours	**0.305**	**0.122**	1.04	**52.50%**	83.77%

Table 2: **NYUv2: Ablation study.** We evaluate the influence of the different components of our model. Note that each part of our model contributes to the final results, with a strong influence of the mid-level structures.

[1] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*, 2014.

[2] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth extraction from video using non-parametric sampling. In *Computer Vision–ECCV 2012*, pages 775–788. Springer, 2012.

[3] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 89–96. IEEE, 2014.

[4] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1253–1260. IEEE, 2010.

[5] Miaomiao Liu, Mathieu Salzmann, and Xuming He. Discrete-continuous depth estimation from a single image. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 716–723. IEEE, 2014.

[6] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. 3-d depth reconstruction from a single still image. *IJCV*, 2007.

[7] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 2009.

[8] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.