PlaneNet: Piece-wise Planar Reconstruction from a Single RGB Image
Supplementary materials

Chen Liu[1] Jimei Yang[2] Duygu Ceylan[2] Ersin Yumer[3] Yasutaka Furukawa[4]
[1]Washington University in St. Louis [2]Adobe Research [3]Argo AI [4]Simon Fraser University
chenliu@wustl.edu {jimyang,ceylan}@adobe.com meyumer@gmail.com furukawa@sfu.ca

The supplementary material explains our modifications to the three baseline methods to enable their executions on our problem setting (i.e., a single image input).

NYU Toolbox is the first baseline [2]. We follow Wang *et al*. [3] and add the same extension to the toolbox as follows: If a plane occupies more than 90% pixels of a semantic mask, the entire semantic mask will be assigned to that plane.

Manhattan World Stereo (MWS) requires a 3D point cloud as the input [1]. Instead of using multi-view stereo, we un-project the input depthmap to obtain a point cloud. The data term in their MRF formulation requires the visibility information associated with multiple views. The information does not exist for our case, and we instead use the data term in NYU Toolbox [2]. This term utilizes pixel-wise surface normals, which are calculated from the depthmap.

Piece-wise Planar Stereo (PPS) extracts vanishing directions (instead of the three Manhattan axes) and generates plane hypotheses from every pair of the extracted directions. PPS also reconstructs 3D lines by multi-view stereo techniques and fits additional planes, whose step is impossible and excluded in our experiments. However, we do not expect much performance degradation, as our input is a dense depthmap and most of the line information are covered through vanishing line extraction. The input point-cloud is generated by un-projecting the input depthmap as in the case of MWS. The data term in the MRF formulation is again replaced by the one in NYU Toolbox.

Applying the model trained on ScanNet to NYUv2. We found that as shown in Figure 1 the network before fine-tuning is able to predict reasonable planes and their masks while the network after fine-tuning predicts less but more confident planes; 2) the predicted depth map before fine-tuning has a global bias towards shallow scenes and the fine-tuning seems to correct this bias. We hypothesize that without plane supervision during fine-tuning, the network chooses an easier path to learn the non-plane depth values instead of planes.

Figures 2, 3 and 4 show more results of our method on the ScanNet and NYUv2 datasets.

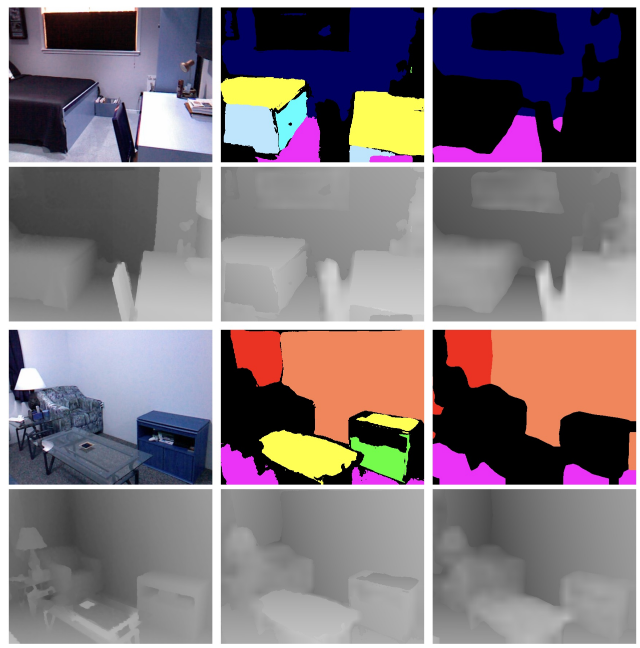

Figure 1: The first column shows the input image and the ground-truth depth map. The second column shows the results (plane segmentation and final depth map) generated by the network before fine-tuning. The third column shows the results after fine-tuning.

References

[1] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Manhattan-world stereo. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1422–1429. IEEE, 2009. 1

[2] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. *Computer Vision–ECCV 2012*, pages 746–760, 2012. 1

[3] P. Wang, X. Shen, B. Russell, S. Cohen, B. Price, and A. L. Yuille. Surge: Surface regularized geometry estimation from a single image. In *Advances in Neural Information Processing Systems*, pages 172–180, 2016. 1

Figure 2: More experimental results on the ScanNet dataset. From left to right: an input image, plane segmentation, depthmap reconstruction, and 3D rendering of our depthmap.

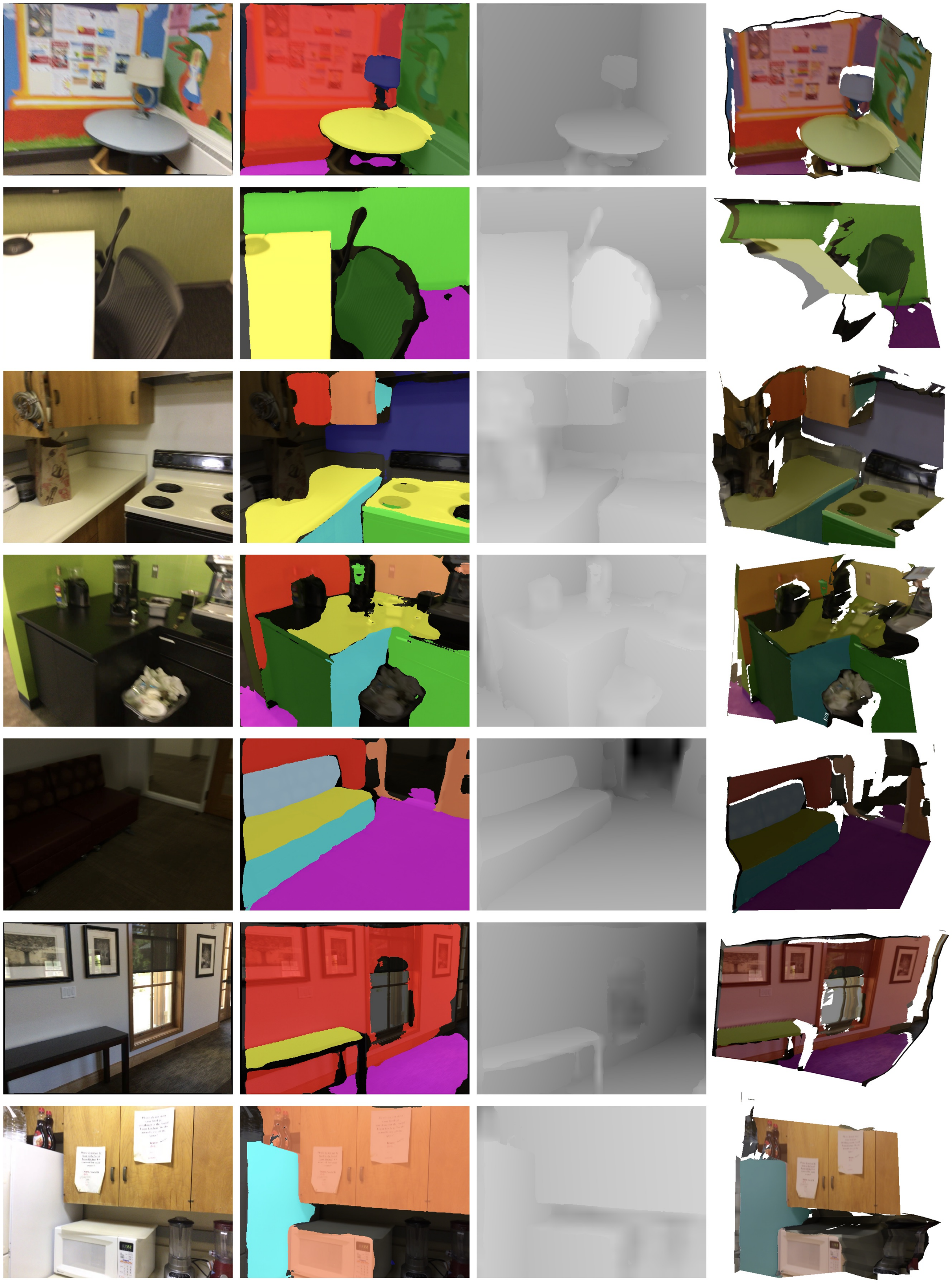

Figure 3: More experimental results on the ScanNet dataset. From left to right: an input image, plane segmentation, depthmap reconstruction, and 3D rendering of our depthmap.

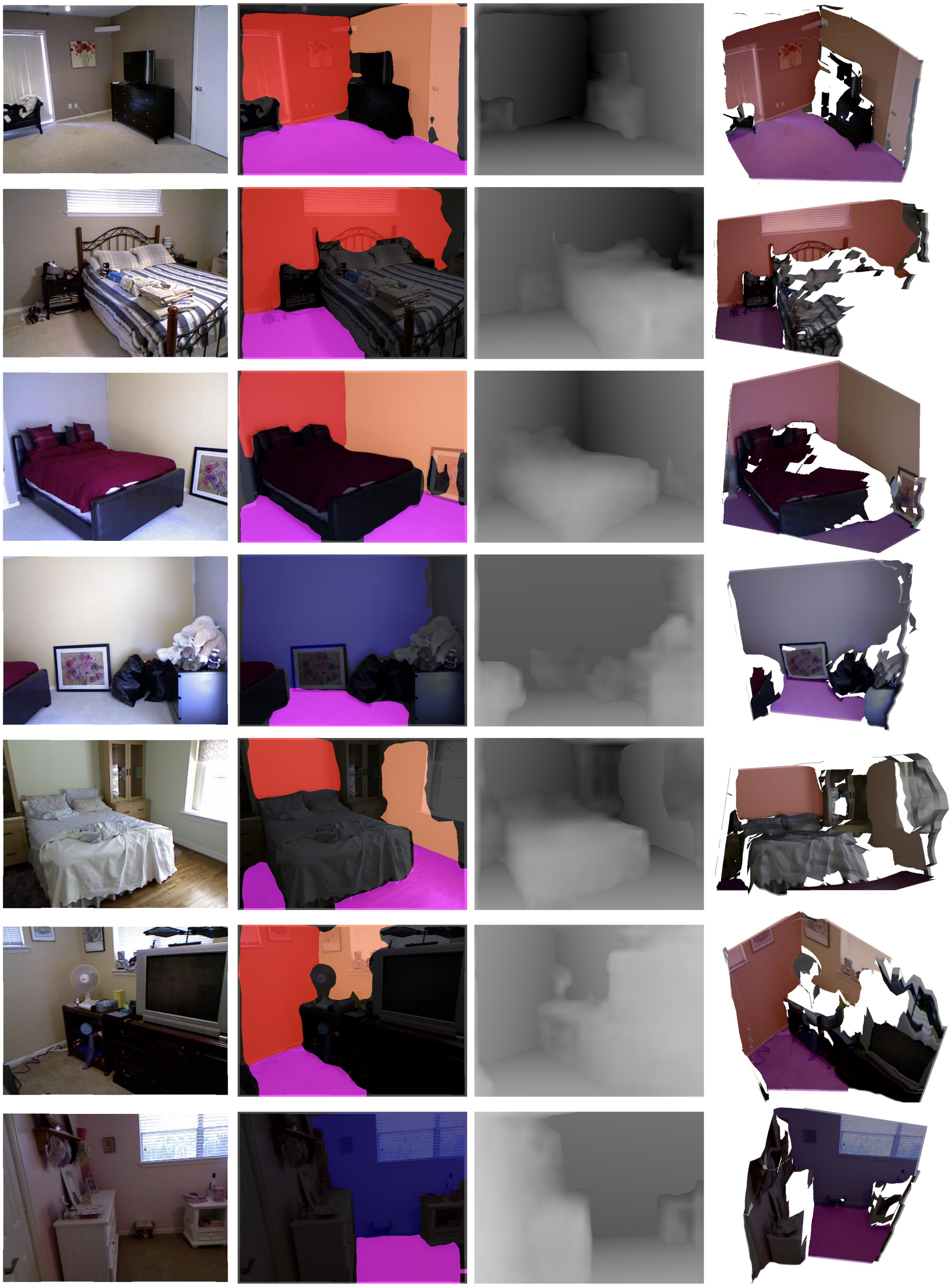

Figure 4: More experimental results on the NYUv2 dataset. From left to right: an input image, plane segmentation, depthmap reconstruction, and 3D rendering of our depthmap. The network is fine-tuned on the dataset without plane supervision, and our algorithm tends to miss small planes at cluttered regions.