PlaneRCNN: 3D Plane Detection and Reconstruction from a Single Image Supplementary Material

Anonymous CVPR submission

Paper ID 704

This supplementary material provides additional details and more qualitative results. In Sec 1, we first provide more details of the depth estimation network, the refinement network, and the warping operation. In Sec 2, we show more results images and comparison in addition to the results shown in Fig.5 and Fig.6 of the original submission. In Sec 3, we discuss more details about the occlusion reasoning experiment.

Finally, we visualize the reconstructed 3D planar regions with varying views in the attached **supplementary video**.

1. Architectural details

1.1. Depth estimation network

The depth estimation network is built upon the feature pyramid network (FPN) [4] in Mask R-CNN, whose outputs are feature maps with resolutions $2 \times 2, 4 \times 4, 8 \times 8, 16 \times 16, 32 \times 32$. Starting from the smallest feature map, we use a block which consists of a 3×3 convolution with stride 1, an upsampling with factor 2, and another 3×3 convolution with stride 1, to process and get a new feature map of higher resolution. We concatenate the upsampled feature map with the next FPN feature map and use another block to process and upsample. After all the feature maps are processed, a 3×3 convolution regresses an one-channel depthmap, which is upsampled to the image solution with bilinear interpolation.

1.2. Refinement network

The detailed architecture of the refinement network is illustrated in Fig. 1.

1.3. Warping operation

The warping operation consists of an unprojection, a coordinate frame transformation, and a projection. Given the camera intrinsics, we first unproject the pixel in the nearby view (u^n, v^n) using the camera intrinsics K, as $X^n = K^{-1}h(u^n, v^n)\hat{D}^n(u^n, v^n)$ where $\hat{D}^n$ is the ground truth depthmap for the nearby view and h converts (u^n, v^n) to the homogeneous representation. We then transform X^n to the current view using rotation R and translation t as $X^c = RX^n + t$. Finally, we get the warped pixel coordinate (u^w, v^w) by projection $(u^w, v^w) = h^{-1}(KX^c)$, where h^{-1} converts the homogeneous coordinate back to the 2D representation.

2. More qualitative results

We show more qualitative results of our method on test scenes from ScanNet in Fig. 2 and Fig. 3 and more comparison against PlaneNet [5] and PlaneRecover [10] on unseen datasets in Fig. 4 and Fig. 5.

3. Occlusion reasoning details

The key challenge for training the network with occlusion reasoning is to generate ground truth complete mask for supervision. Besides projecting 3D planes to the image domain without depth checking, we further complete the mask for layout structures based on the fact that layout planes are behind other geometries. First, we collect all planes which have layout labels (e.g., *wall* and *floor*), and compute the convexity and concavity between two planes in 3D space. Then for each combination of these planes, we compute the corresponding complete depthmap by using the greater depth value for two convex planes and using the smaller value for two concave ones. A complete depthmap is valid if 90% of the complete depthmap is behind the visible depthmap (with $0.2m$ tolerance to handle noise). We pick the valid complete depthmap which has the most support from visible regions of layout planes.

References

[1] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 5

[2] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. In *ACM transactions on graphics (TOG)*, volume 24, pages 577–584. ACM, 2005. 6

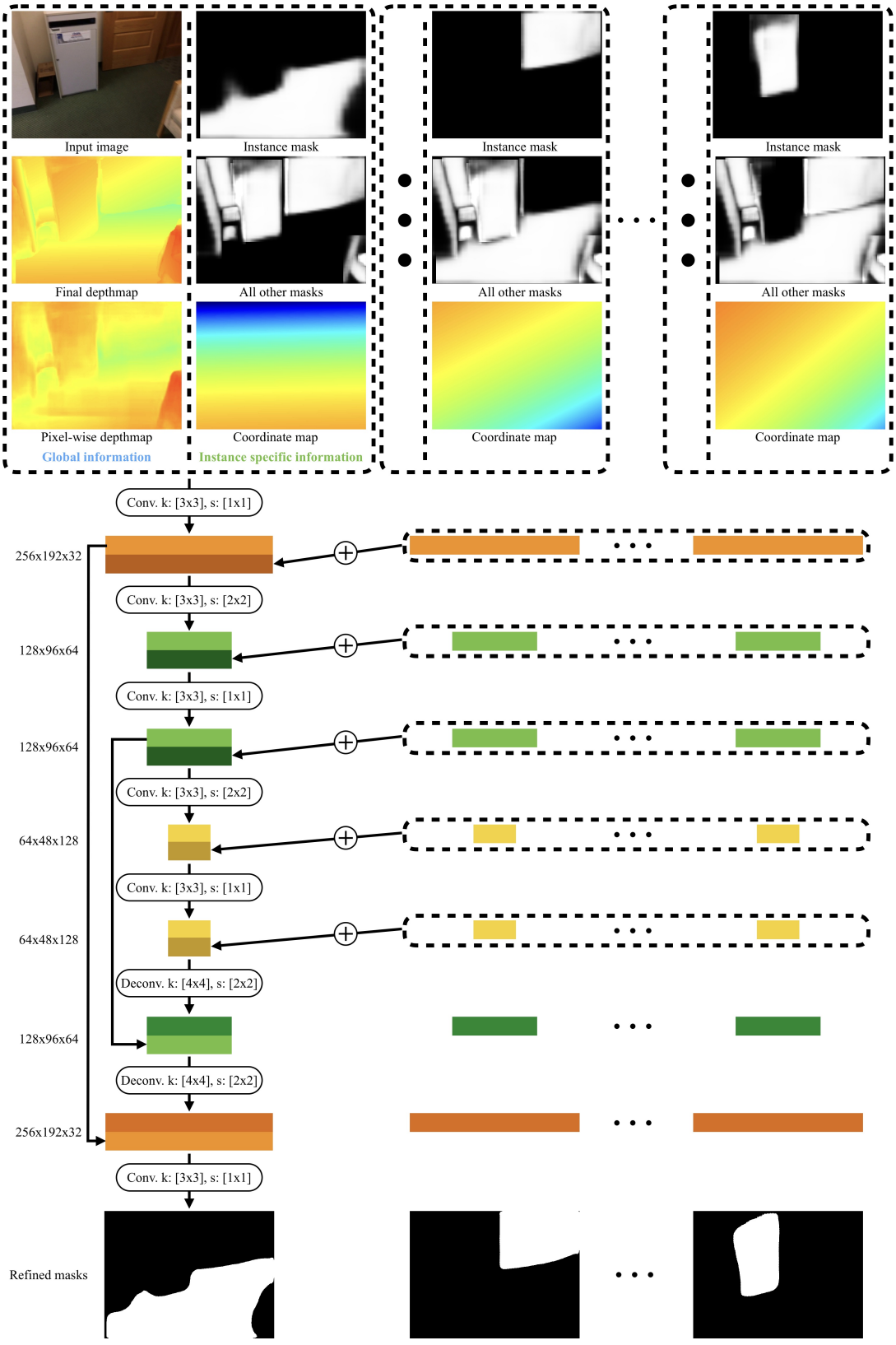


Figure 1. Refinement network architecture. The network takes both global information (i.e., the input image, the reconstructed depthmap and the pixel-wise depthmap) and instance-specific information (i.e., the instance mask, the union of other masks, and the coordinate map of the instance) as input and refines instance mask with a U-Net architecture [6]. Each convolution in the encoder is replaced by a ConvAccu module to accumulate features from other masks.

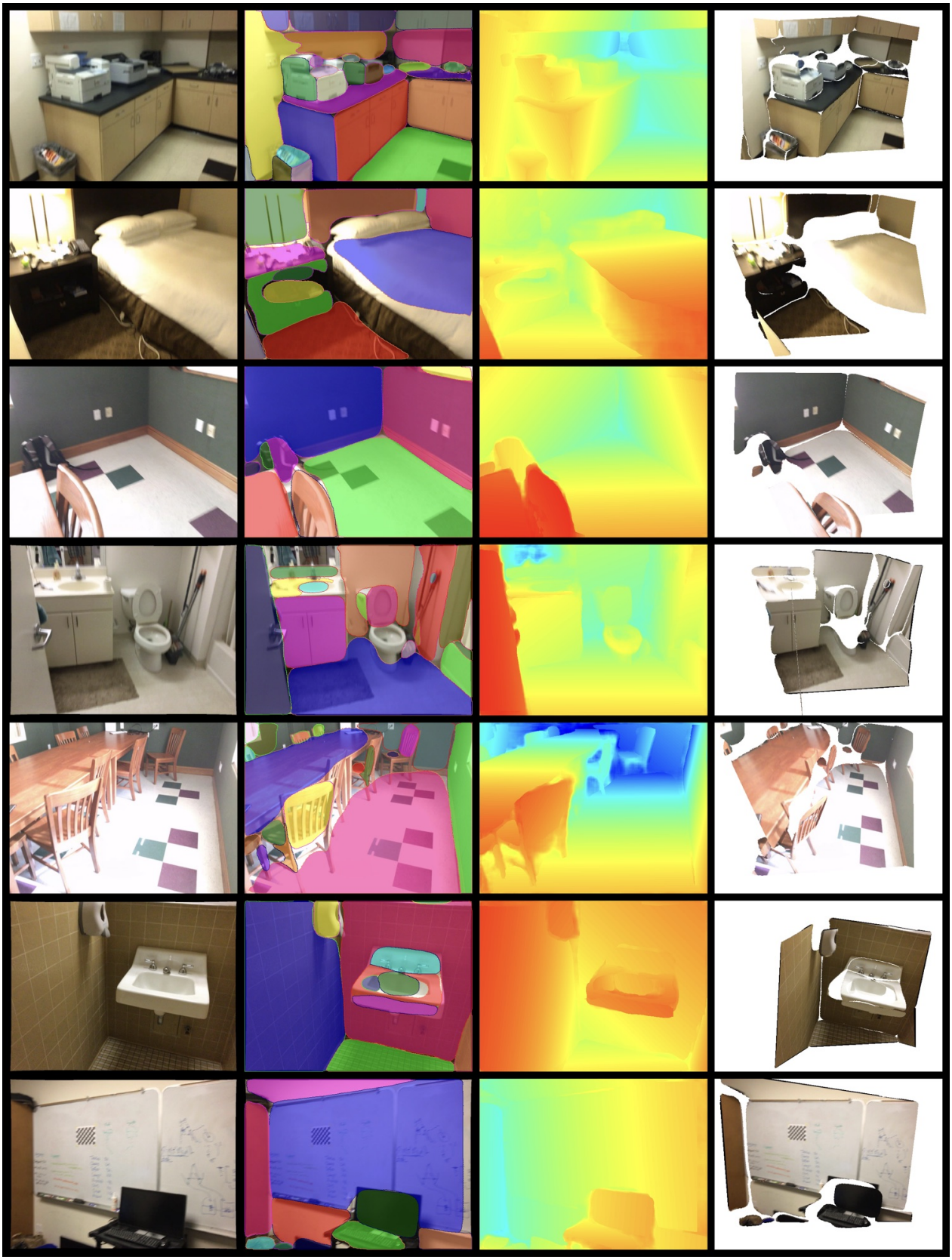


Figure 2. More qualitative results on test scenes from the ScanNet dataset.

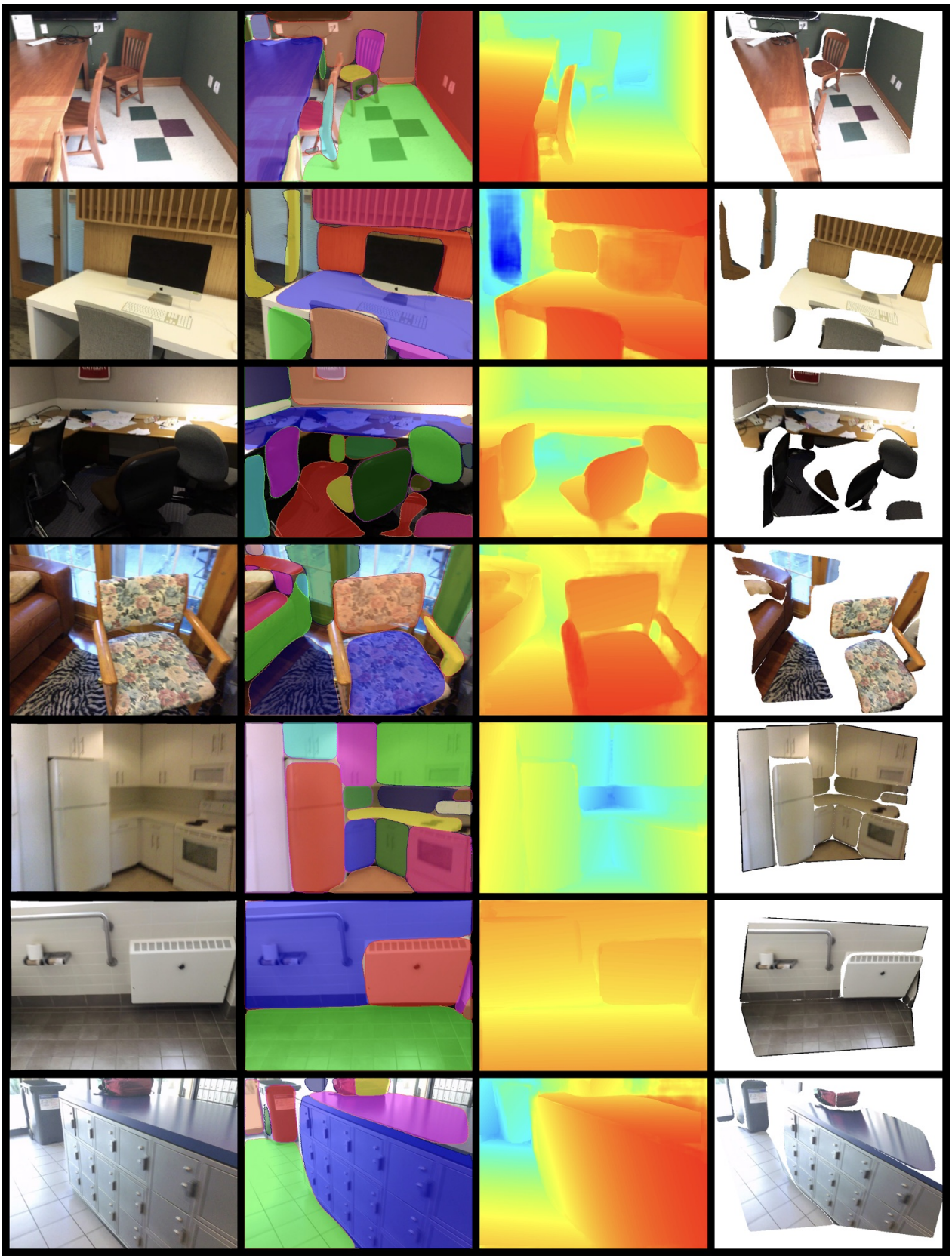
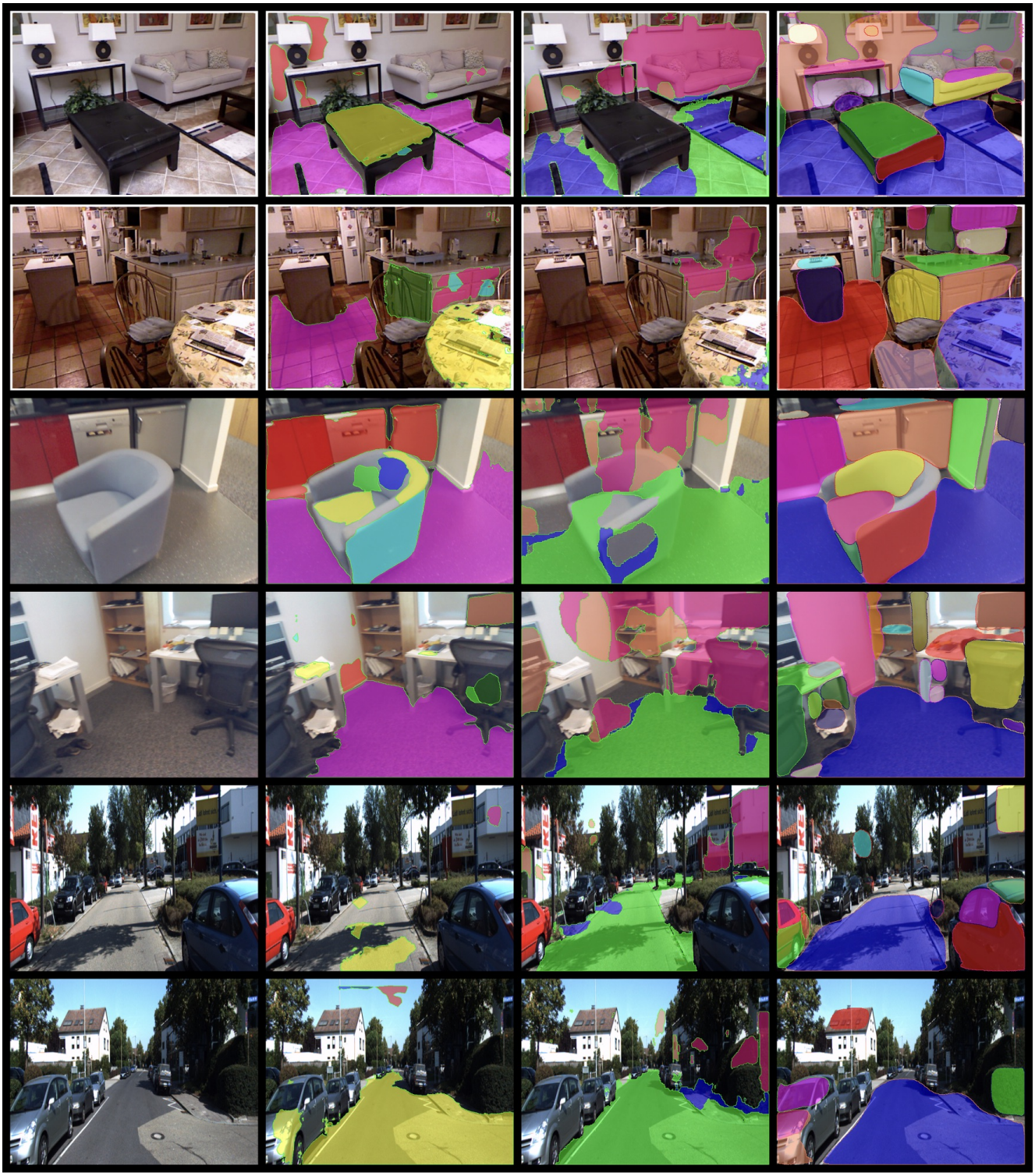


Figure 3. More qualitative results on test scenes from the ScanNet dataset.



Input image PlaneNet PlaneRecover Ours

Figure 4. More plane segmentation results on unseen datasets without fine-tuning. From left to right: input image, PlaneNet [5] results, PlaneRecover [10] results, and ours. From top to the bottom, we show two examples from each dataset in the order of NYUv2 [9], 7-scenes [8], and KITTI [1].

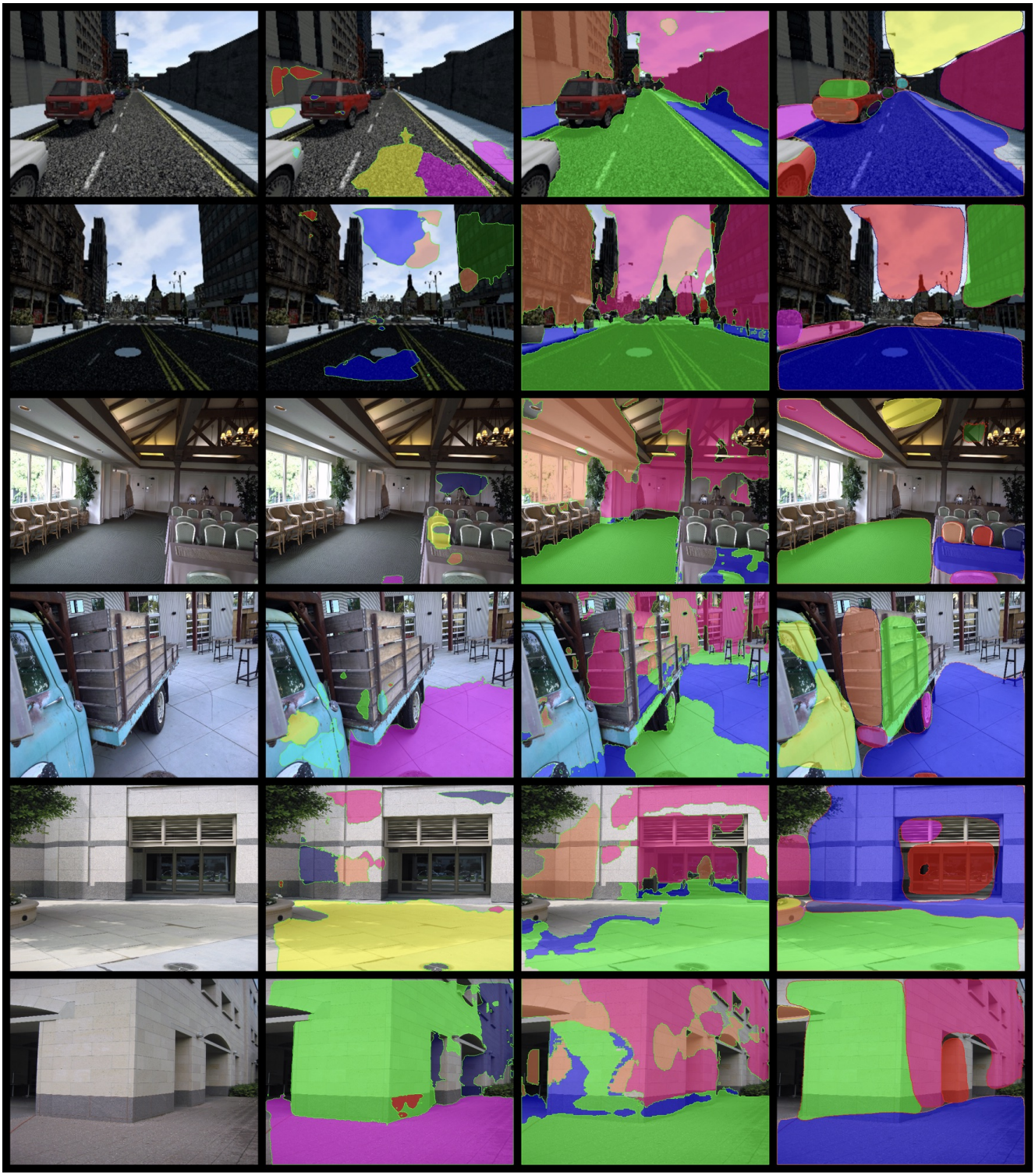


Figure 5. More plane segmentation results on unseen datasets without fine-tuning. From left to right: input image, PlaneNet [5] results, PlaneRecover [10] results, and ours. From top to the bottom, we show two examples from each dataset in the order of SYNTHIA [7], Tank and Temple [3], and PhotoPopup [2].

[3] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):78, 2017. 6

[4] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017. 1

[5] C. Liu, J. Yang, D. Ceylan, E. Yumer, and Y. Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2579–2588, 2018. 1, 5, 6

[6] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[7] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 6

[8] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2937, 2013. 5

[9] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. 5

[10] F. Yang and Z. Zhou. Recovering 3d planes from a single image via convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 1, 5, 6