Supplementary Material for Joint 3D Layout and Depth Prediction from a Single Indoor Panorama Image

Wei Zeng[1], Sezer Karaoglu[1,2], and Theo Gevers[1,2]

[1] Computer Vision Laboratory, University of Amsterdam, The Netherlands
{w.zeng,th.gevers}@uva.nl
[2] 3DUniversum, Science Park 400, The Netherlands
{s.karaoglu,theo.gevers}@3duniversum.com

1 Network Architecture Details

The detailed network architectures for the coarse depth estimation and semantic segmentation are shown in Table 1. The encoder is based on the ResNet-18 architecture with modified input blocks [4]. The rectangle filters in the input blocks are connected in parallel and resolutions vary to account for different distortion levels. The encoder is shared for both the coarse depth estimation and semantic segmentation. The decoders restore the original input resolution by means of up-sampling operators followed by 3×3 convolutions. Skip connections are also added to link to the corresponding resolution in the encoder. The two decoders do not share weights. The network architecture for the layout depth estimation is similar to this architecture however without input blocks and only one decoder with respect to the layout depth estimation.

The detailed network architecture for the depth refinement is shown in Table 2. The first part is the semantic-guided depth fusion network. The input is the concatenation of the coarse depth prediction, estimated layout depth and semantic segmentation and the output is the semantic-guided attention map. This attention map maximizes the exploitation of the coarse depth and layout depth. Then, the depth refinement module takes the fused depth as input to predict the final refined depth. The encoder-decoder architecture is similar to the previous depth estimation network.

2 Layout Depth Generation

Here we describe how to generate the layout depth map from the original corner labeling for supervised learning. The 3D layout can be recovered by fitting planar surfaces to the corner positions. As shown in Fig. 1, we assume the dimensions of the panorama image is $W \times H$. Since the panorama image covers 360 degree field of view horizontally and 180 degree field of view vertically, so $W = 2H$, and the focal length is $W/2\pi$, which is also the radius of the cylinder: $R = W/2\pi$. From Fig. 1, it can be derived that for a 3D point (x, y, z):

Table 1. Details of the network architecture for the coarse depth estimation and semantic segmentation. The encoder uses modified input blocks in front of the ResNet-18 architecture to reduce the distortion effect. The network for the layout depth estimation is similar to this architecture with slight modifications

name	**layer**	**kernel size**	**output feature map size**
Input_Block_1	conv	5x11x8	512x1024x8
	conv	3x9x8	512x1024x8
	conv	5x7x8	512x1024x8
	conv	7x7x8	512x1024x8
	concat		512x1024x32
Input_Block_2	conv	3x9x16	512x1024x16
	conv	3x7x16	512x1024x16
	conv	3x5x16	512x1024x16
	conv	5x5x16	512x1024x16
	concat		512x1024x64
Conv1	conv	3x3x64 3x3x64 }x2	256x512x64
Conv2	conv	3x3x128 3x3x128 }x2	128x256x128
Conv3	conv	3x3x256 3x3x256 }x2	64x128x256
Conv4	conv	3x3x512 3x3x512 }x2	32x64x512
	up-sampling		64x128x512
De-conv4_Sem	conv	3x3x256	64x128x256
	up-sampling		128x256x256
De-conv3_Sem	conv	3x3x128	128x256x128
	up-sampling		256x512x128
De-conv2_Sem	conv	3x3x64	256x512x64
	up-sampling		512x1024x64
De-conv1_Sem	conv	3x3x64	512x1024x64
De-conv0_Sem	conv	3x3x13	512x1024x13
De-conv4_Dep	conv	3x3x256	64x128x256
	up-sampling		128x256x256
De-conv3_Dep	conv	3x3x128	128x256x128
	up-sampling		256x512x128
De-conv2_Dep	conv	3x3x64	256x512x64
	up-sampling		512x1024x64
De-conv1_Dep	conv	3x3x64	512x1024x64
De-conv0_Dep	conv	3x3x1	512x1024x1

Table 2. Details of the network architecture for the depth refinement. The first part is the semantic-guided depth fusion network which outputs the attention map to fuse the previous depth maps. The following encoder-decoder architecture is similar to the previous depth estimation network

name	**layer**	**kernel size**	**output feature map size**
	concat		512x1024x15
Conv1_Sem_guided	conv	3x3x32	256x512x32
Conv2_Sem_guided	conv	3x3x64	128x256x64
Conv3_Sem_guided	conv	3x3x128	64x128x128
Conv4_Sem_guided	conv	3x3x256	32x64x256
	up-sampling		64x128x256
De-conv3_Sem_guided	conv	3x3x128	64x128x128
	up-sampling		128x256x128
De-conv2_Sem_guided	conv	3x3x64	256x512x64
	up-sampling		512x1024x64
De-conv1_Sem_guided	conv	3x3x1	512x1024x1
	fusion		512x1024x1
Conv1	conv	3x3x64 3x3x64 }x2	256x512x64
Conv2	conv	3x3x128 3x3x128 }x2	128x256x128
Conv3	conv	3x3x256 3x3x256 }x2	64x128x256
Conv4	conv	3x3x512 3x3x512 }x2	32x64x512
	up-sampling		64x128x512
De-conv4_Dep_refine	conv	3x3x256	64x128x256
	up-sampling		128x256x256
De-conv3_Dep_refine	conv	3x3x128	128x256x128
	up-sampling		256x512x128
De-conv2_Dep_refine	conv	3x3x64	256x512x64
	up-sampling		512x1024x64
De-conv1_Dep_refine	conv	3x3x64	512x1024x64
De-conv0_Dep_refine	conv	3x3x1	512x1024x1

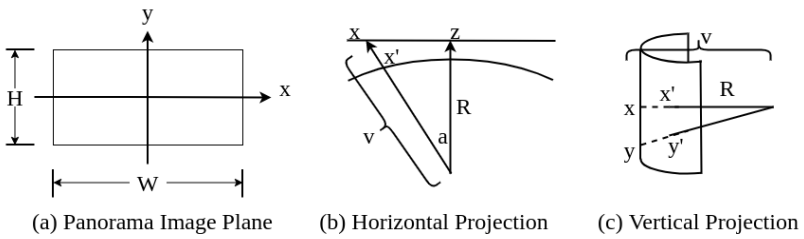
$$\alpha = \arctan(\frac{x}{z}) \tag{1}$$

$$v = \sqrt{x^2 + z^2} \tag{2}$$

$$x' = R\alpha = R \cdot \arctan(\frac{x}{z}) \tag{3}$$

$$y' = R \cdot \arctan(\frac{y}{\sqrt{x^2 + z^2}}) \tag{4}$$

$$d' = \sqrt{x^2 + y^2 + z^2} \tag{5}$$



(a) Panorama Image Plane (b) Horizontal Projection (c) Vertical Projection

Fig. 1. Geometry derivation for equirectangular reprojecting layout depth map from 3D layout points

where x' and y' are the reprojected coordinates on the panorama. d' is the corresponding depth value. Fig. 2 shows additional results for the layout depth map.

3 Additional Qualitative Results for Layout Prediction

Additional qualitative results for layout prediction are shown in Fig. 3. The first four rows demonstrate the results of the baseline LayoutNet [5] and our proposed method on the Stanford 2D-3D dataset. The last four rows are computed for the PanoContext dataset. For each example, we show the predicted layout (LayoutNet: blue, our proposed method: green) and the ground truth (orange) under equirectangular view. By explicitly incorporating the layout depth map, the proposed method can locate the corners more precisely (avoiding locations in the middle of the wall which has continuous depth, e.g. the third, fourth and eighth examples for the Stanford 2D-3D and the third, sixth and eighth examples for the PanoContext). Constrained by the layout depth map, the proposed method is also able to handle occluded corners (e.g. the second, fourth and sixth examples for the Stanford 2D-3D and the fifth example for the PanoContext). Thus, it can be derived that for both the Stanford 2D-3D and PanoContext dataset, the proposed method obtains better performance for layout prediction.

Additional qualitative results for non-cuboid room layout prediction are shown in Fig. 4 and Fig. 5. To verify the generalization ability of our proposed method to non-cuboid layout, we fine-tune our model on the non-cuboid rooms labeled by [2]. It can be shown that the proposed method is able to handle non-cuboid layout rooms.

4 Additional Qualitative Results for Depth Estimation

Additional qualitative results for depth estimation are shown in Fig. 6. The baseline RectNet [4], state-of-the-art Plane-aware network [1] and our proposed method are compared. Additional 3D reconstruction comparison of the depth estimation are shown in Fig. 7 and Fig. 8. In Fig. 7, the panorama input splits

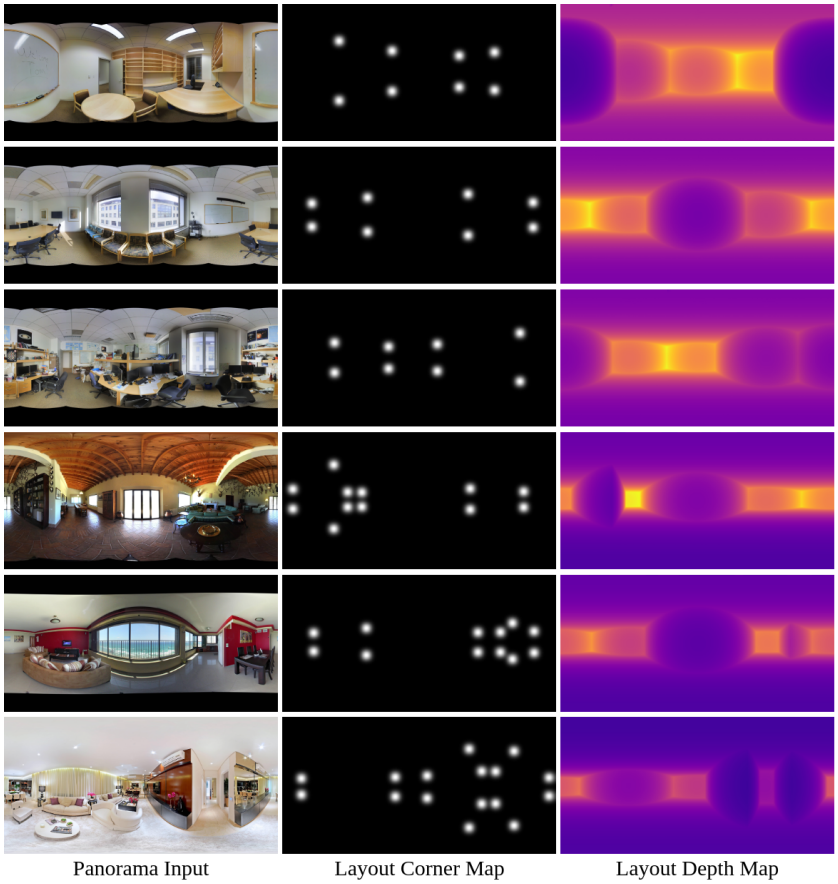


Fig. 2. Additional results of the layout depth map. From left to right: the panorama input image, the original layout corner map and the corresponding layout depth map. The first three rows are images from the Stanford 2D-3D dataset and the last three rows are from the non-cuboid rooms of the PanoContext dataset labeled by [2]

the window into two parts, the leftmost and rightmost part of the panorama. Without any constraint, the RectNet [4] estimates the discontinuous depth for the window, resulting to disjointed 3D reconstruction, as circled by the red dash ellipses. Constrained by the layout depth map, our proposed method correctly estimates the continuous depth for the window. Explicitly inter-positioning the layout depth, the 3D reconstruction of the proposed method also obtains more planar ceiling and walls. In Fig. 8, similar disjointed 3D reconstruction is shown for the whiteboard of RectNet, but our 3D reconstruction can overcome this issue and preserve more planarity. Additional internal qualitative comparison between the coarse depth estimation and the final refined depth are shown in Fig. 9, which provides more insights about the depth refinement. The depth map below the

coarse depth map is the estimated layout depth map which we use to refine the depth estimation. For the first example, the depth of the window region is incorrect for the coarse depth estimation. Combined with the layout depth map, the refined depth map correctly estimates the depth for the ambiguous window region. For the second example, the depth of the right bookcases are too difficult to estimate for the coarse depth estimation. Constrained by the layout depth map, the proposed method obtains proper depth estimation.

5 Timing Statistics

Table 3 summarizes the time comparison for a single forward pass of the network and the post optimization step between LayoutNet [5], DuLa-Net [3], HorizonNet [2] and our proposed method. Note that the computation time of our proposed method is with the depth refinement module. Optimization directly in 3D space makes our proposed method more efficient. Our proposed method is the fastest for both the network prediction and the optimization step.

Table 3. Time consumption comparison for a single forward pass of the neural network and the optimization step between different methods

Method	Optimization avg. CPU Time(ms)	Network avg. GPU Time(ms)
LayoutNet [5]	1583	39
DuLa-Net [3]	22	35
HorizonNet [2]	18	58
Ours	**15**	**32**

References

1. Eder, M., Moulon, P., Guan, L.: Pano popups: Indoor 3d reconstruction with a plane-aware network. In: 2019 International Conference on 3D Vision (3DV). pp. 76–84. IEEE (2019)
2. Sun, C., Hsiao, C.W., Sun, M., Chen, H.T.: Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1047–1056 (2019)
3. Yang, S.T., Wang, F.E., Peng, C.H., Wonka, P., Sun, M., Chu, H.K.: Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3363–3372 (2019)
4. Zioulis, N., Karakottas, A., Zarpalas, D., Daras, P.: Omnidepth: Dense depth estimation for indoors spherical panoramas. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 448–465 (2018)
5. Zou, C., Colburn, A., Shan, Q., Hoiem, D.: Layoutnet: Reconstructing the 3d room layout from a single rgb image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2051–2059 (2018)

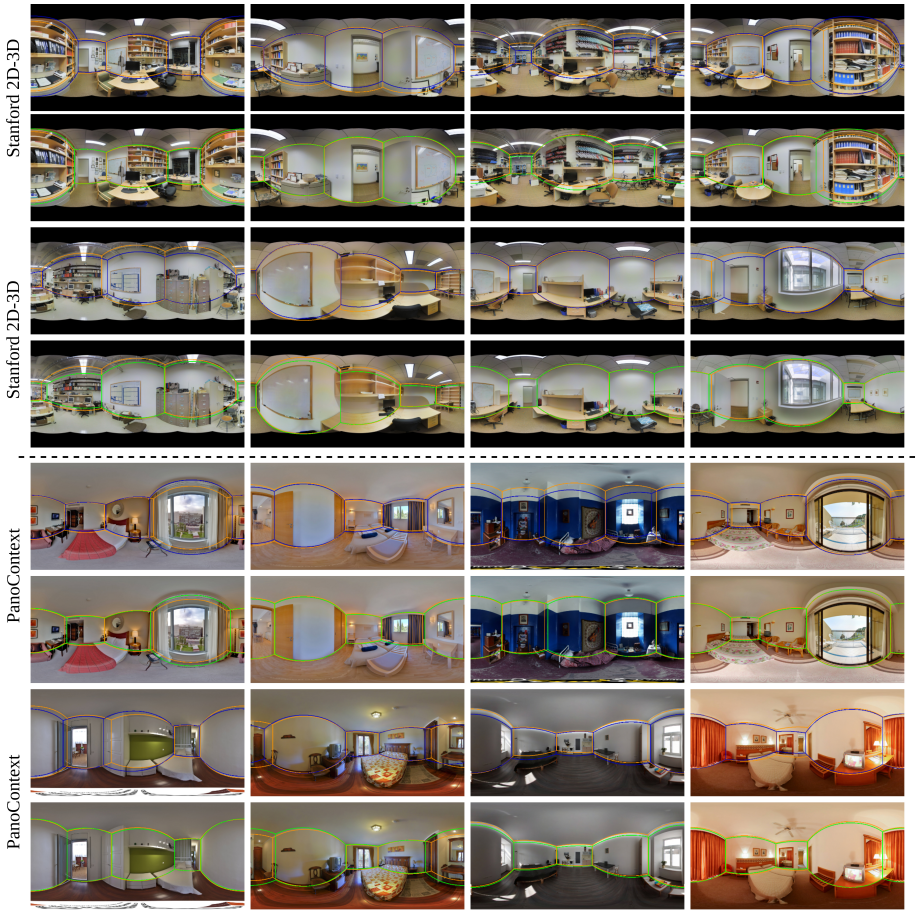


Fig. 3. Additional qualitative results on layout prediction. Results are shown of testing the baseline LayoutNet [5] and our proposed method on the Stanford 2D-3D dataset (top four rows) and PanoContext dataset (bottom four rows). For each example, we show the predicted layout (LayoutNet: blue, the proposed method: green) and the ground truth (orange) under equirectangular view

Fig. 4. Additional qualitative results for non-cuboid room layout prediction. It can be derived that the proposed method can also handle non-cuboid layout rooms

Fig. 5. Additional qualitative results for non-cuboid room layout prediction. It can be derived that the proposed method can also handle non-cuboid layout rooms

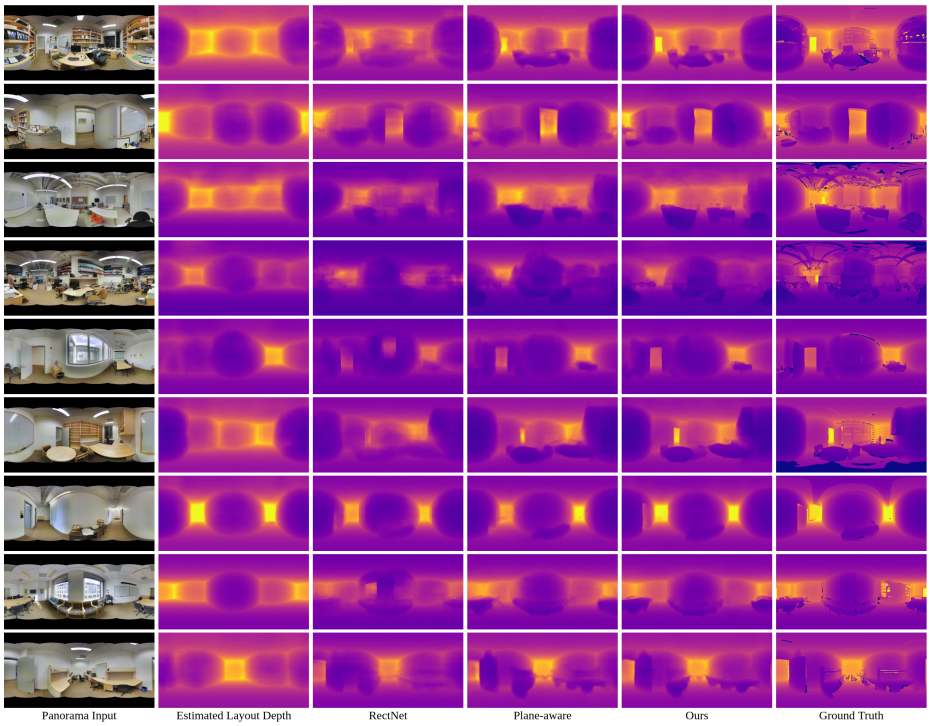


Fig. 6. Additional qualitative results on depth estimation. Results are shown for experiments of testing the estimated layout depth map, the baseline RectNet [4], Plane-aware network [1] and our proposed method on the Stanford 2D-3D dataset

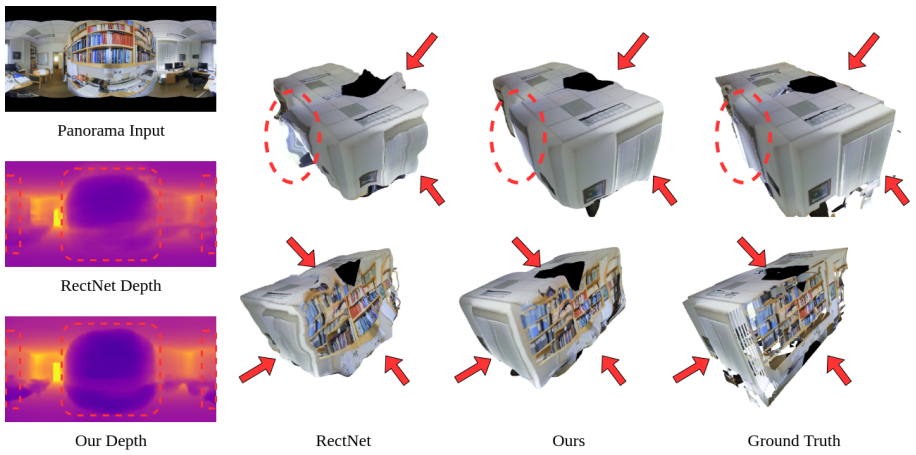


Fig. 7. Additional 3D reconstruction comparison. Due to the explicitly inter-positioning of the layout depth, the proposed method predicts a relatively good depth map for the distant regions. So the 3D reconstruction of the proposed method is more proper and provides more planar ceiling and walls

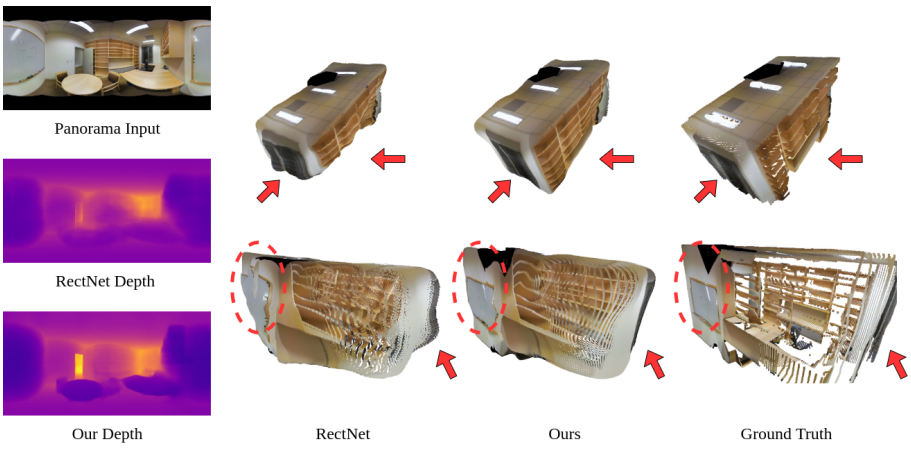


Fig. 8. Additional 3D reconstruction comparison. The proposed method preserves more accurate scale of the room and the wall planes are more consistent

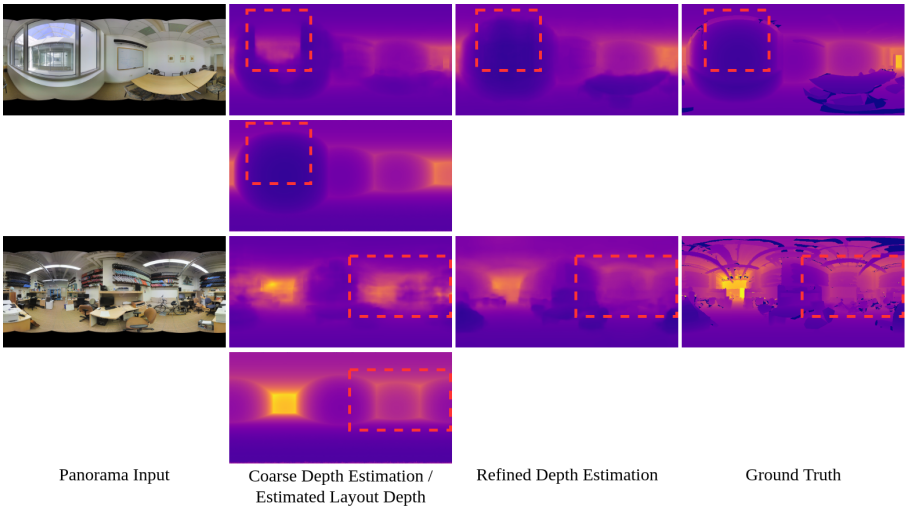


Fig. 9. Additional internal qualitative comparison between the coarse depth estimation and the final refined depth. Constrained by the layout depth map, the proposed method refines better depth estimation based on the coarse depth estimation