Deep Multi Depth Panoramas for View Synthesis Supplementary Material

1 Network Architecture

Our novel learning based framework (see Fig. 3 in the main paper) uses a 3D CNN to predict per-view MPIs as described in Sec. 4.1 in the main paper. We show the detailed layers of our 3D CNN in Tab. 1.

Layer	kernel size	stride	dilation	in	out	activation	input
conv1_1	3	1	1	15	8	ReLU	PSVs
conv1_2	3	2	1	8	16	ReLU	conv1_1
conv2_1	3	1	1	16	16	ReLU	conv1_2
conv2_2	3	2	1	16	32	ReLU	conv2_1
conv3_1	3	1	1	32	32	ReLU	conv2_2
conv3_2	3	1	1	32	32	ReLU	conv3_1
conv3_3	3	2	1	32	64	ReLU	conv3_2
conv4_1	3	1	2	64	64	ReLU	conv3_3
conv4_2	3	1	2	64	64	ReLU	conv4_1
conv4_3	3	1	2	64	64	ReLU	conv4_2
up5		2		128	128		conv3_3 + conv4_3
conv5_1	3	1	1	128	32	ReLU	nnup5
conv5_2	3	1	1	32	32	ReLU	conv5_1
conv5_3	3	1	1	32	32	ReLU	conv5_2
up6		2		64	64		conv2_2 + conv5_3
conv6_1	3	1	1	64	16	ReLU	nnup6
conv6_2	3	1	1	16	16	ReLU	conv6_1
up7		2		32	32		conv1_1 + conv6_2
conv7_1	3	1	1	32	8	ReLU	nnup7
conv7_2	3	1	1	8	8	ReLU	conv7_1
conv7_3	3	1	1	8	5	ReLU	conv7_2
weights	1	1	1	5	5	Softmax	conv7_3
alpha	1	1	1	1	1	Sigmoid	conv7_3

Table 1. Our 3D CNN structure. **in** and **out** denote the input and output channel counts, respectively. **input** denotes the input of each layer and + means concatenation in channel dimension. Layers starting with "up" means 2× nearest neighbor upsampling. We compute the blending weights for all views in the PSVs with a softmax layer and the alpha values with a sigmoid layer.

2 Additional Results

We now demonstrate additional results of our method. Tab. 2 shows the ablation study on the effects of end-to-end training. It shows the difference between naively converting the MPIs to MDPs in post-process versus training the method end-to-end. Simply binning as a post-process operation is not adequate to resolve depth conflicts, which the end-to-end network learns to handle via subtle changes in alpha values to better match the overall 3D geometry of the scene. Tab. 2 demonstrates that training the end-to-end network leads to a 1.32 dB improvement over naively binning into MDPs as a post-processing technique. Fig. 1 shows an example of a scene with the 16 input images from the 360° device and our panorama rendering results using five-layer MDPs at two different locations. Similar to Fig. 1 in the main paper, our novel MDPs enable realistic novel view synthesis with translational motions for 360° panorama rendering. To illustrate the details of our MDPs, in Fig. 2, we show the reconstructed five-layer RGBDα images in the MDPs used in rendering the scene of Fig. 1. Note that the MDPs effectively model the scene content at multiple depth ranges. Similar to MPIs [2], our MDPs accumulate information at subsequent layers. Specifically, a farther layer corresponds to a more complete RGBDα panorama image, and the farthest layer covers the entire 360° range. In this way, our MDPs effectively model the background content in a scene and enable realistic 360° rendering with challenging occlusions. Similarly, we also show panorama rendering results of a synthetic scene at two different translated locations in Fig. 4 and the corresponding MDP layers in Fig. 3. In Fig. 5, we show the difference among different number of MDP layers. A single layer of MDP is not enough to cover the disoccluded regions, causing holes to appear. Adding an extra layer helps, but is still not enough to handle large translations. The ideal number of layers depends on scene complexity, but for typical cases, we found out that five layers is generally enough to generate visually pleasing results. We also show the comparison in Fig. 6 between our method and a modified weighted average of Mildenhall et al. [1]. Instead of choosing the weights according to camera position, we choose the cosine similarity between source and target camera orientations as our weighting to blend the MPIs. Insets in fig. 6 shows that depth conflicts might become apparent at image boundaries, whereas our MDP method alleviates the artifacts by projecting onto a canonical panorama representation and resolving depth conflicts.

End-to-end Training	**PSNR↑**	**SSIM↑**
Without	25.07	0.8282
With	**26.39**	**0.8664**

Table 2. Ablation study on end-to-end training. We show that using end-to-end training in our proposed pipeline improves the rendered image quality.

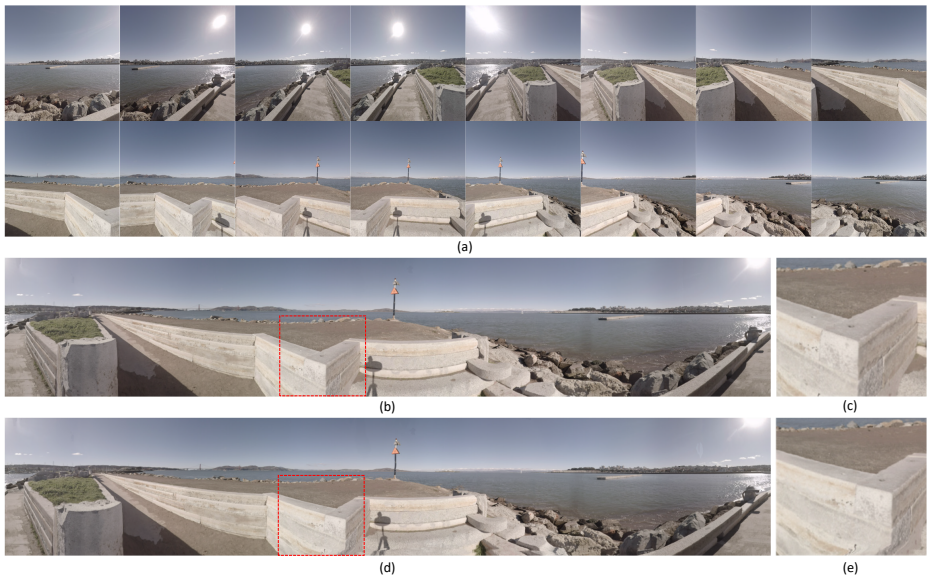

Fig. 1. Per-view input images and the rendered panoramas. We show the input images from all 16 cameras in (a), a synthesized panorama result at the center of the device in (b), and a panorama result at a translated position (to the left 15cm and to the back 20cm) in (d). We also show insets (c)(e) of the panoramas with the same crop (red rectangle), which illustrate the difference between the results caused by the translation.

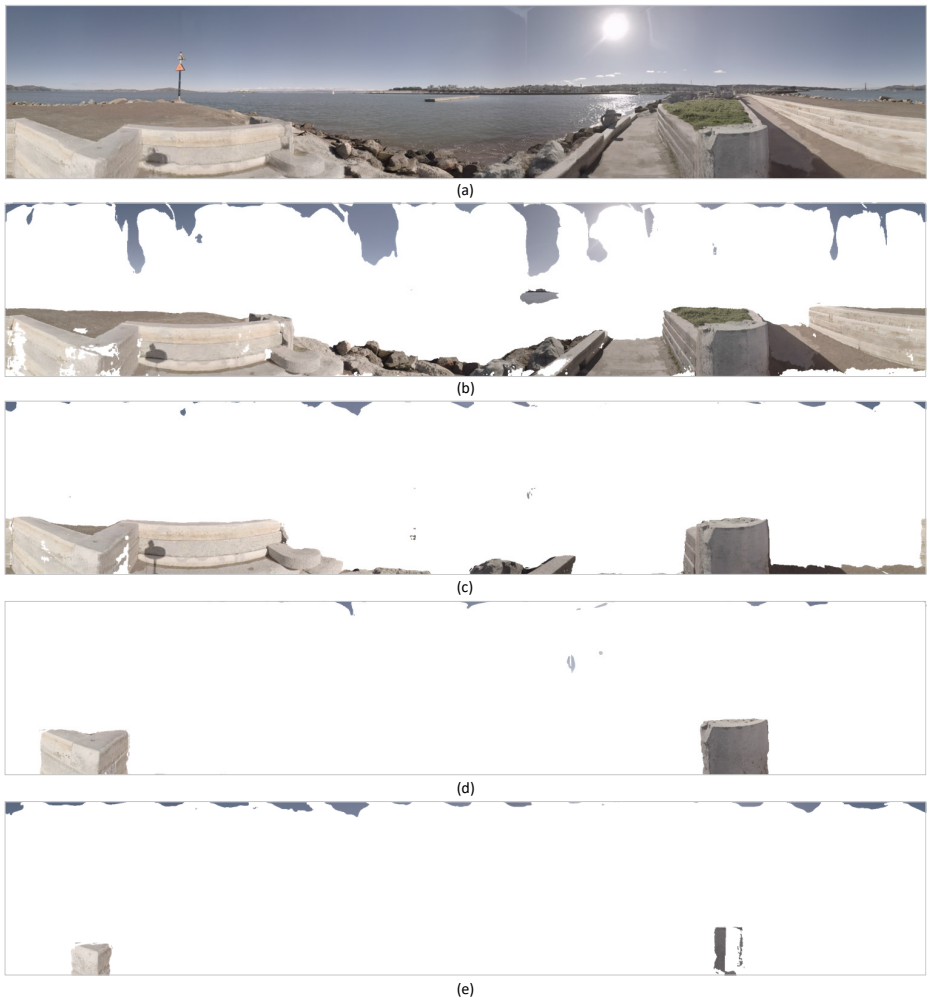

Fig. 2. Individual layers of the MDPs from of the real scene in Fig. 1. (a)-(e) farthest layer to closest.

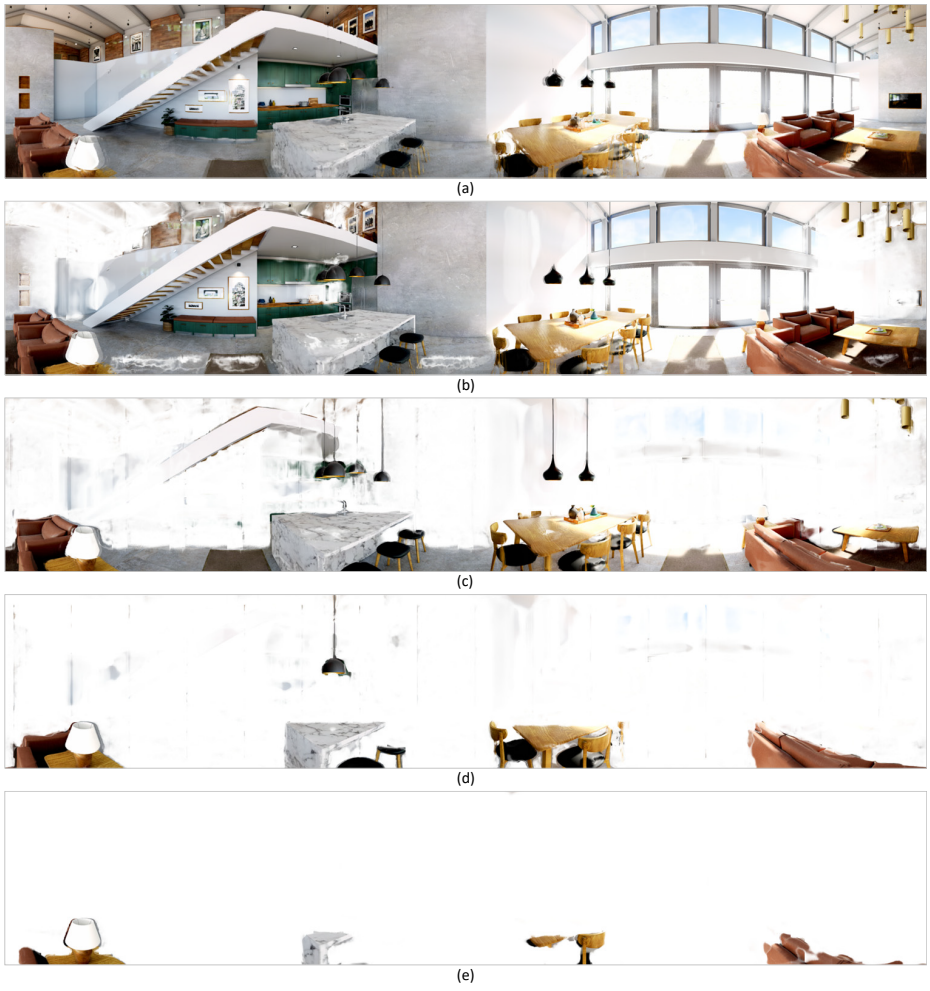

Fig. 3. Individual layers of the MDPs of a synthetic scene. (a)-(e) farthest layer to closest.

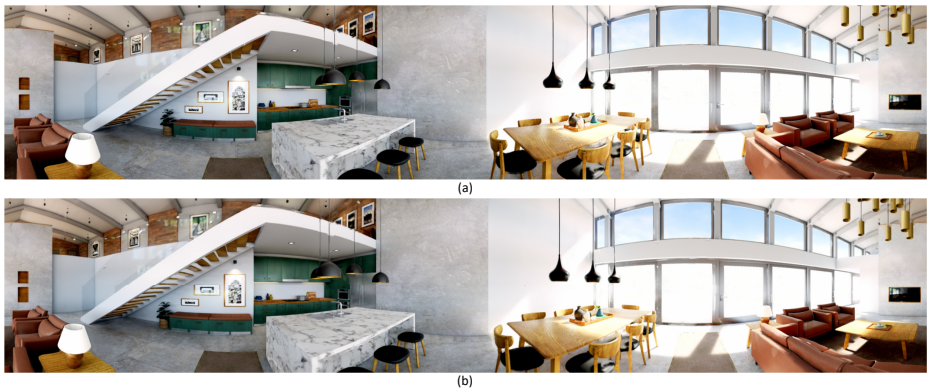

Fig. 4. Rendered panoramas with different camera poses of a synthetic scene. Note how that lamp on the left moves away from the camera.

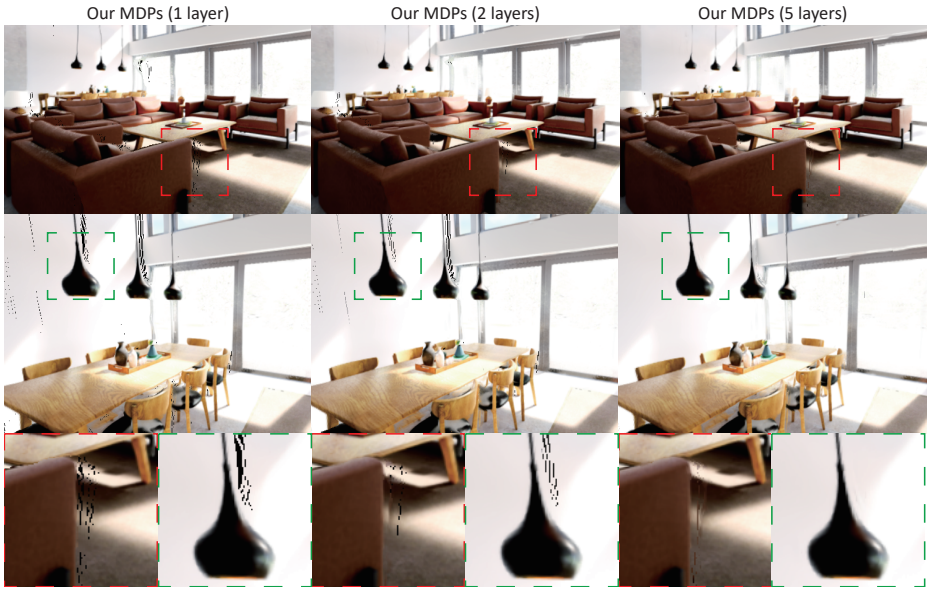

Fig. 5. Demonstration of the visual quality with different numbers of MDP layers. The user can determine the number of MDP layers as a trade-off between quality and memory usage.

Fig. 6. The comparison between our method and the modified weighted average of multiple MPIs, similar to that in Mildenhall et al. [1]. The table in the orange insets shows clear improvement of our method by resolving the depth conflicts.

References

1. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019)
2. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. In: SIGGRAPH (2018)