BGDNet: Background-guided indoor panorama depth estimation Supplementary Material

Jiajing Chen*†[1], Zhiqiang Wan*[2], Manjunath Narayana[2], Yuguang Li[2], Will Hutchcroft[2], Senem Velipasalar[1], and Sing Bing Kang[2]

[1]Syracuse University
[2]Zillow Group

We show results of more analysis and experimental results to further demonstrate the effectiveness of our proposed method.

1. Comparison of D_{Bg} and D_S

In this section, we perform an analysis of the D_S and D_{Bg}. As we explained in section 4.1.5 in the main paper, the background depth map D_S, obtained by SAM, is typically inaccurate in the ceiling and ceiling-adjacent wall regions. To address this problem, we combine D_S and D_H to obtain the D_{Bg}, for the final depth estimation task. Fig 1 shows the heatmap of depth MAE of D_S and D_{Bg}

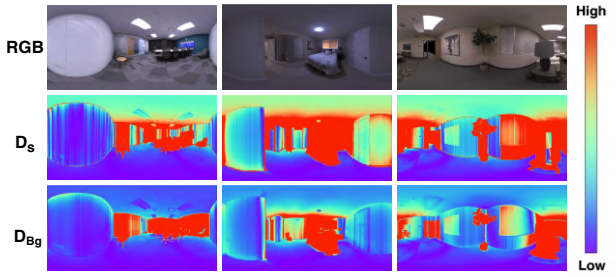

Figure 1. RGB image and the corresponding depth error map of D_S and D_{Bg}. Note that D_{Bg} has less error in the ceiling and some wall regions.

Although SAM and HorizonNet were not trained on the Replica dataset, they exhibit excellent depth estimation capabilities on the background portion of the panoramic image. Additionally, it should be noted that D_S has a higher error compared to D_{Bg} specifically in the ceiling area and some walls. This is because there exists some error in the ceiling-wall connection boundary generated by SAM. This error produces inaccurate estimated distance from the camera to the ceiling, which adversely affects depth estimation on the ceiling and ceiling-connected wall.

Model	MAE↓	RMSE↓	RMSE (log)↓	δ^1↑	δ^2↑	δ^3↑
BGDNet w/ D_S	0.1709	0.3553	0.1331	0.8517	0.9339	0.9616
BGDNet w/ D_{Bg}	**0.1678**	**0.3456**	**0.1334**	**0.8554**	**0.9365**	**0.9624**

Table 1. BGDNet performance with D_S and D_{Bg} as background depth map respectively.

To further show the effectiveness of combining D_S and D_{Bg}, we train and evaluate BGDNet with D_S and D_{Bg} respectively on the Replica dataset. The results are shown on Table 1. Compared with D_S, D_{Bg} helps BGDNet achieve a better performance.

2. Effect of Depth Replacing Threshold

In our proposed Background Depth Replacement Module, as described in section 4.3 of the main paper, we replace a portion of the depth map predicted by the network with the depth from D_C under the condition that the difference between D_S and D_H is less than a threshold value α for that portion. In this section, we conducted an experiment to investigate the impact of different α values on the accuracy of the final depth estimation. The results of this experiment are presented in Table 2.

The table clearly illustrates the impact of selecting a small threshold value (close to 0) in the Background Depth Replacement Module, as depicted in Figure 3 of the main paper. In such instances, there is minimal replacement occurring, resulting in the final output P being similar to P_N, which is directly generated by the Depth Prediction Module. Consequently, this scenario leads to an increase in error.

*Equal contribution.
†This work was done when Jiajing Chen was an intern at Zillow.
[1]{jchen152,svelipas}@syr.edu
[2]{zhiqiangw,manjun,yuguangl,willhu,singbingk}@zillowgroup.com

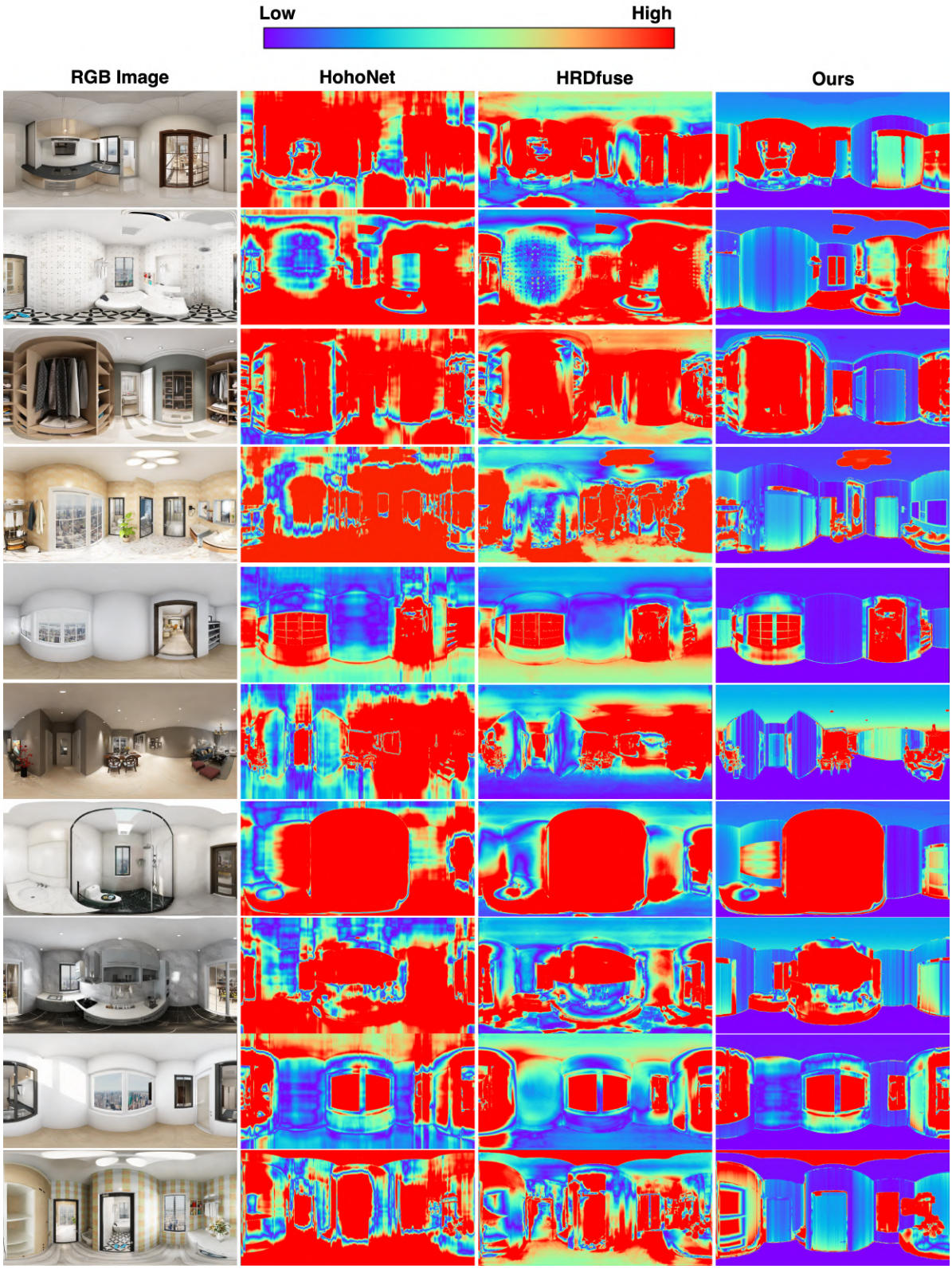

Figure 2. Qualitative result on images from Structured3D. In the visual representation, the purple area indicates a low error, whereas the red area indicates a high error. From this observation, it is evident that our proposed method outperforms alternative approaches and yields superior performance.

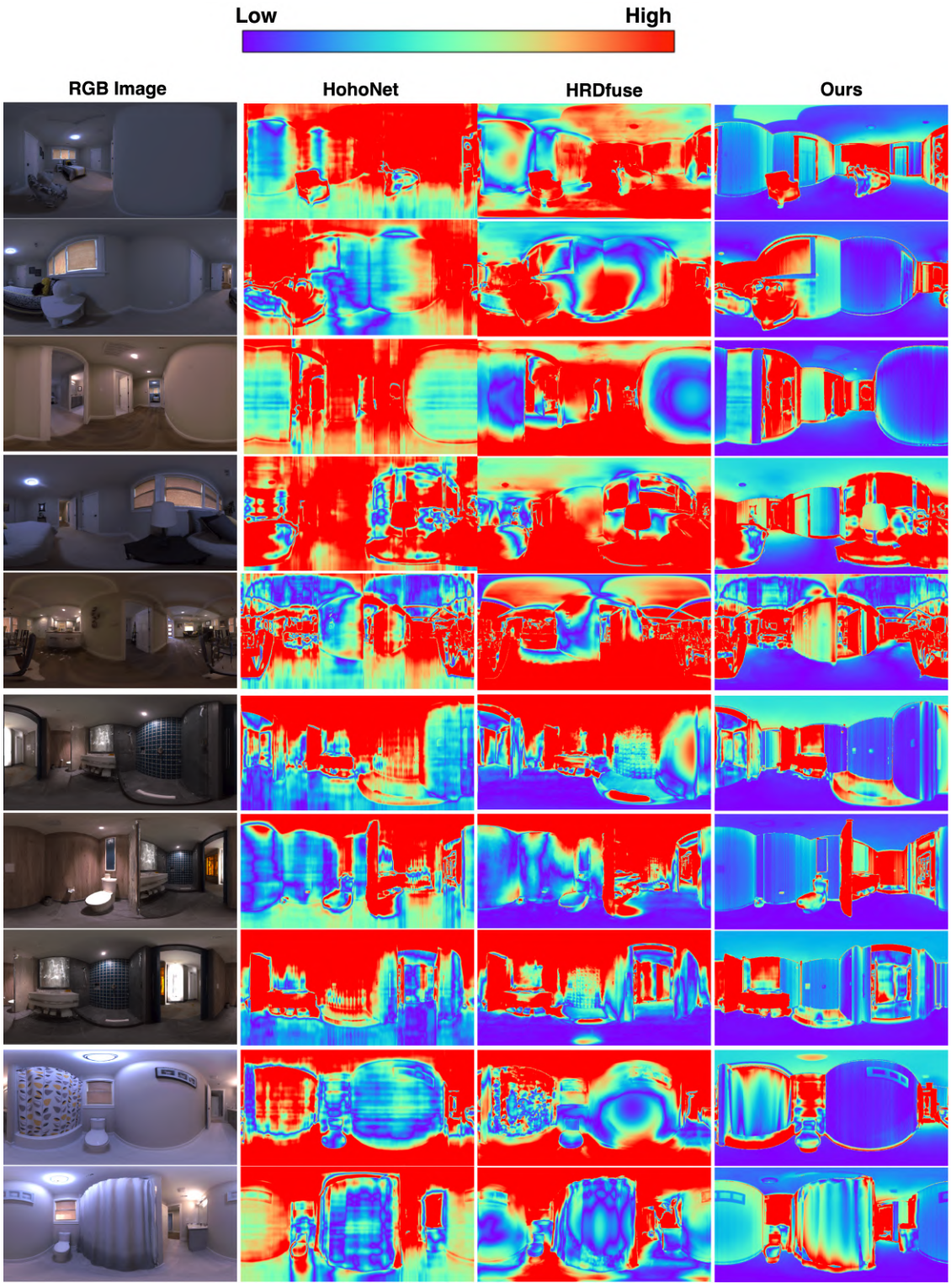

Figure 3. Qualitative result on images from Replica. Our method shows an overall better accuracy than the SOTA method.

Furthermore, with an increase in the value of α, there is a notable decrease in error. Additionally, within a wide range of α values (0.22-0.62), our BGDNet consistently demonstrates stable and exceptional performance across all rele-

α	MAE↓	RMSE↓	RMSE (log)↓	δ^1↑	δ^2↑	δ^3↑
0.02	0.1919	0.3567	0.1431	0.828	0.9266	0.9537
0.22	0.1687	0.3463	0.135	0.8552	0.9351	0.9608
0.42	**0.1678**	**0.3456**	**0.1334**	**0.8554**	**0.9365**	**0.9642**
0.62	0.1692	0.3467	0.1345	0.8548	0.9352	0.9613

Table 2. Performance of our method with different threshold values.

vant metrics.

3. More Qualitative Result

In figure 2, we present additional qualitative results comparing our proposed BGDNet with state-of-the-art (SOTA) methods. Although all networks are trained on the Replica dataset, the testing images are from the Structured3D dataset. It is evident in the comparison that predictions from HohoNet and HRDfuse exhibit significant areas of high error, indicated by the red regions. In contrast, our predictions demonstrate superior performance in most areas of the image, indicated by the purple or blue regions. Additionally, Figure 3 shows the qualitative results obtained when the network is trained and tested on the Replica dataset. In this comparison, our proposed BGDNet exhibits overall better performance than the SOTA method.

Although our proposed method performs well in predicting most background areas, it still exhibits a higher error in areas that are not flat. For instance, if there is a light on the ceiling or a concave feature such as a window or door on the wall, the error will be substantial in these regions. We will address this problem in future works.