- Supplementary Material -
Boosting Monocular Depth Estimation with Lightweight 3D Point Fusion

Lam Huynh[1] Phong Nguyen[1] Jiri Matas[2] Esa Rahtu[3] Janne Heikkilä[1]
[1]University of Oulu [2]Czech Technical University in Prague [3]Tampere University

The supplementary material is organized as follows: Section 1 presents additional qualitative results for the KITTI [5] dataset; Section 2 shows more results on the NYU-Depth-v2; Section 3 experiments with an in-house dataset namely OuKi; and Section 4 defines the evaluation metrics. We also include a video that demonstrates the depth estimation performance on the KITTI and OuKi dataset using extremely sparse sets of 3D points.

1. Additional qualitative results on KITTI

This section provides further results for different sparsity on KITTI. Figure 1 and 2 shows the results for our method and for [3, 1] using a varying number of input 3D points. The proposed method clearly produces depth maps with less errors than state-of-the-art approaches especially with a small number of input 3D points. These results suggest that high-quality depth maps can be obtained by using only a few LiDAR points enabling more cost efficient solutions.

2. Additional results on NYU-Depth-v2

Figure 3 presents additional results for our method and NLSPN [3] on the NYU-Depth-v2 dataset [4] using a varying number of randomly selected input points. The proposed method preserves both coarse and fine structures in all tested cases.

3. Experiment with OuKi dataset

We utilized a Kinect-v2 or an Android phone to record a set of videos, namely OuKi, to further assess the generalization properties of the proposed method. This dataset will be made publicly available upon the publication of the paper.

Dense depth prediction using COLMAP points. The OuKi test set consists of 597 RGB frames with ground truth depth maps from indoor environments. The frames are pre-processed with COLMAP to obtain the camera poses and the sparse 3D point cloud. Table 1 contains the performance metrics for our method, NLSPN [3] and MVSNet [6] using the collected dataset. Compared to the NYU-v2 results, we obtain similar performance, while NLSPN [3] and MVS-Net [6] perform worse. These results indicate that the proposed method can generalise well to environments unseen at the training time.

Table 1. Evaluation results on OuKi dataset. Metrics with ↓ mean lower is better and ↑ mean higher is better.

Method	#3D pts	#params	REL↓	RMSE↓	δ_1↑
NLSPN [3]	32	25.8M	0.340	0.915	0.635
NLSPN [3]	128	25.8M	0.232	0.534	0.811
MVSNet [6]	-	124.5M	0.062	0.307	0.933
Ours	32	**8.7M**	0.096	0.313	0.907
Ours	128	**8.7M**	**0.059**	**0.271**	**0.988**

Figure 4 show the qualitative examples from the OuKi test set. The proposed approach clearly preserves the scene structure and details compared to baseline methods.

Dense depth prediction using ARCore points. The recent AR frameworks provide 3D points of the environment, which can be utilised for dense depth estimation. To this end, we collected video sequences using an Android phone and used ARCore [2] to produce a sparse 3D point cloud of the scene. Figure 6 presents a challenging sample from OuKi dataset using the ARCore points. The results with different number of input points confirm that 1) coarse details are well-preserved for all sparsity cases, and 2) depth estimates are consistently better when using more points.

Dense depth prediction from two images. We provide examples where we reconstructed a very sparse set of 3D points (32 points) from two images and utilized those as input to our model. The dense depth maps obtained with this setting using our method, NLSPN [3] and MVSNet [6] are illustrated in Figure 5. The proposed method produces high quality depth maps with significantly less distortions than baseline approaches.

4. Definitions of the evaluation metrics

For NYU-Depth-v2 [4] the evaluation results are calculated for pixels with depth values in the range [0.0, 10.0] while for KITTI [5] the valid range is [0.0, 90.0]. We evaluate the performance for our model and for baselines using the following standard metrics:

- Mean absolute relative error (REL):

$$\frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{d}_i - d|}{d} \tag{1}$$

- Root mean square error (RMSE):

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{d}_i - d_i)^2} \tag{2}$$

- Thresholded accuracy (δ_i):

$$max(\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}) = \delta^i < 1.25^i \quad (i = 1, 2, 3) \tag{3}$$

- Mean absolute error (MAE):

$$\frac{1}{N}\sum_{i=1}^{N}|\hat{d}_i - d| \tag{4}$$

- Mean absolute error of the inverse depth (iMAE):

$$\frac{1}{N}\sum_{i=1}^{N}|\hat{p}_i - p_i| \tag{5}$$

- Root mean square error of the inverse depth (iRMSE):

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - p_i)^2} \tag{6}$$

where N is the number of valid pixel; $\hat{d}_i$ and d_i are the predicted and ground truth depth value at pixel i; $\hat{p}_i$ and p_i are the inverse value of the predicted and ground truth depth at pixel i. Higher thresholded accuracies δ_1, δ_2 and δ_3 figures mean better results, while lower REL, MAE, RMSE, iMAE, iRMSE values are better.

References

[1] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Towards precise and efficient image guided depth completion. *ICRA*, 2021. 1

[2] Google Inc. *ARCore Resources*, Released 01 Mar 2018. 1

[3] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Proc. of European Conference on Computer Vision (ECCV)*, 2020. 1, 6

[4] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. 1, 2

[5] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 1, 2

[6] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. *European Conference on Computer Vision (ECCV)*, 2018. 1, 6

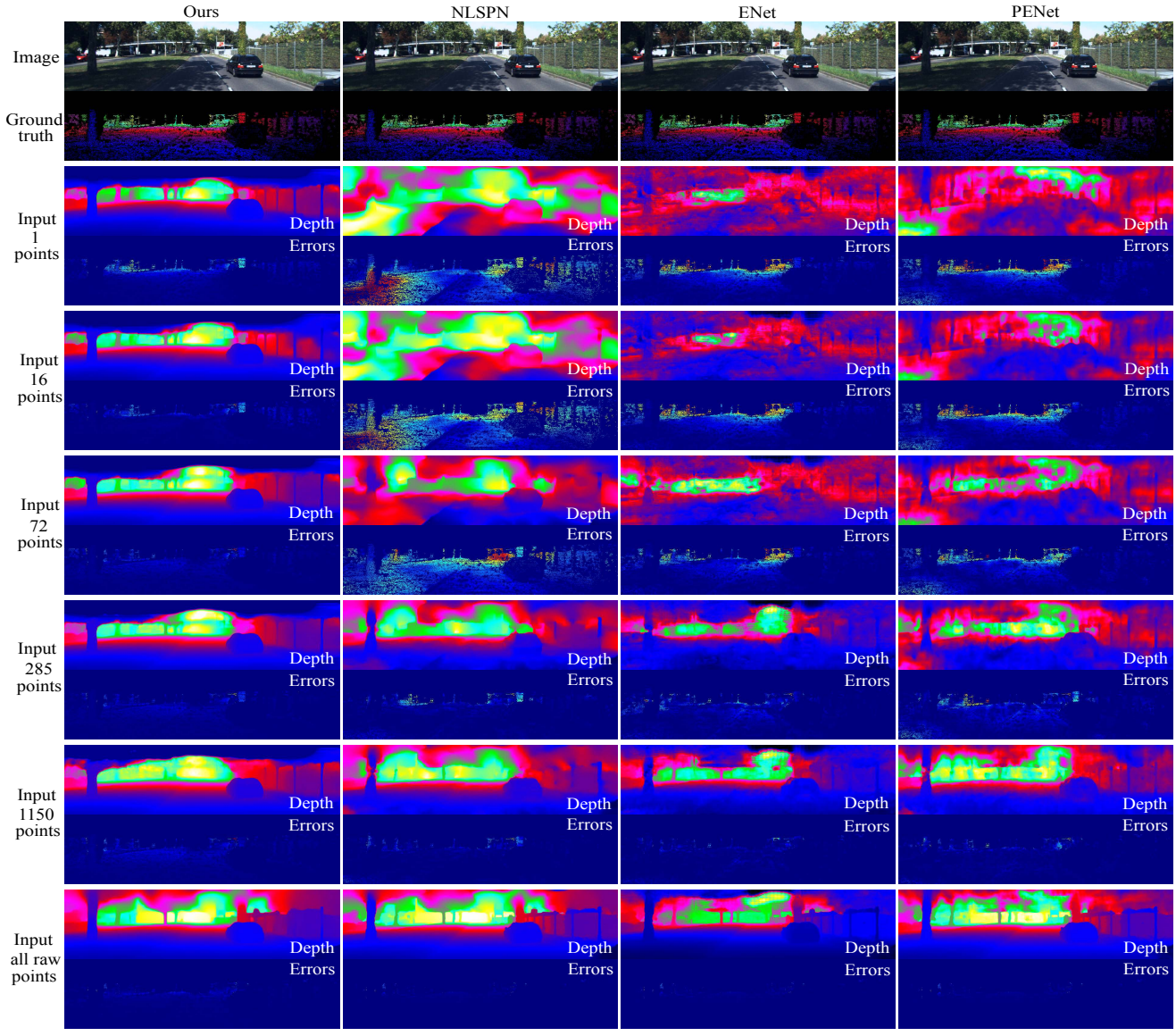

Figure 1. An example from the KITTI validation set using 1, 16, 32, 72, 285, 1150 randomly sampled points. (Ground truth, depth and error maps are at the same scale for visualization)

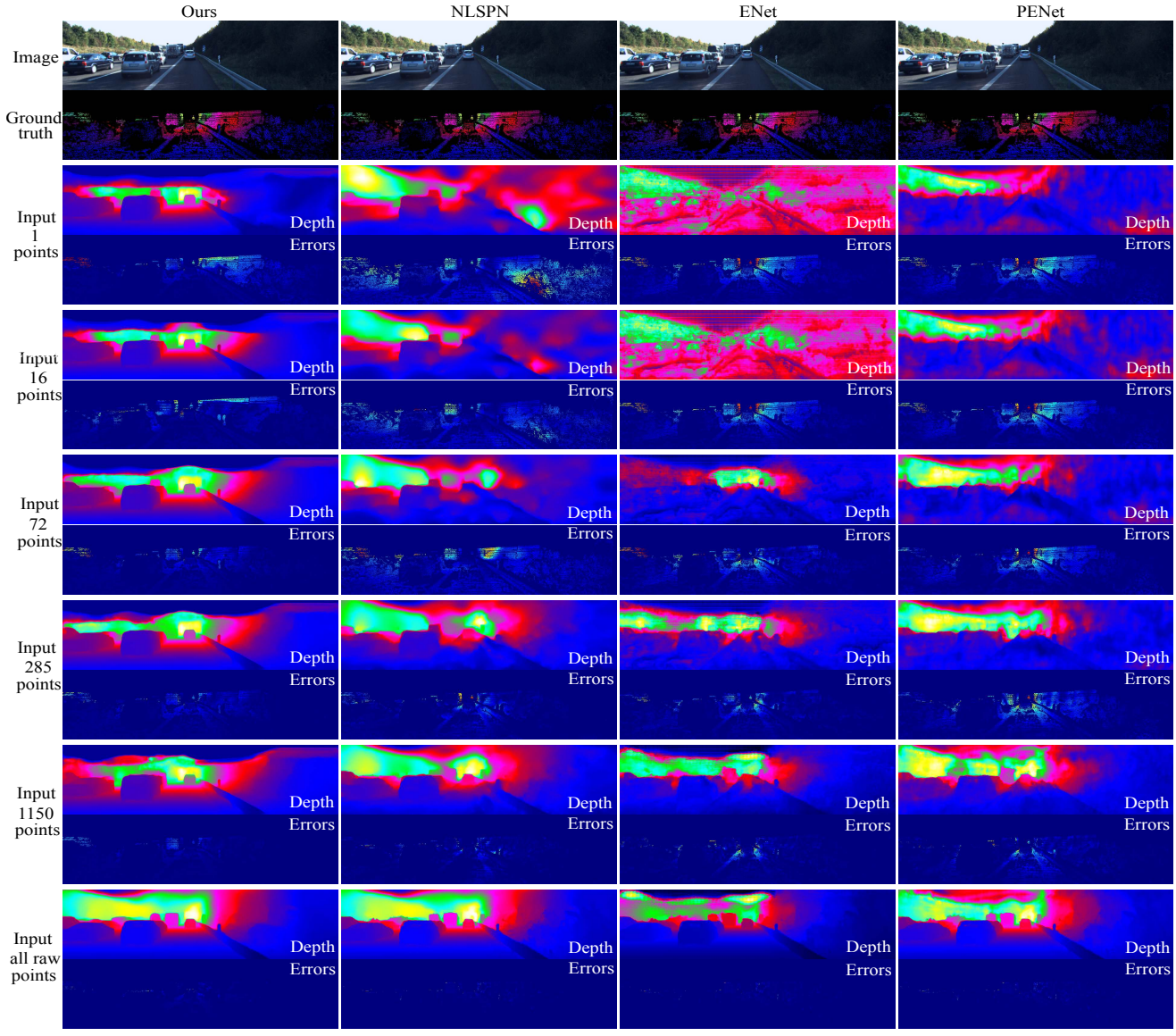

Figure 2. An example from the KITTI validation set using 1, 16, 32, 72, 285, 1150 randomly sampled points. (Ground truth, depth and error maps are at the same scale for visualization)

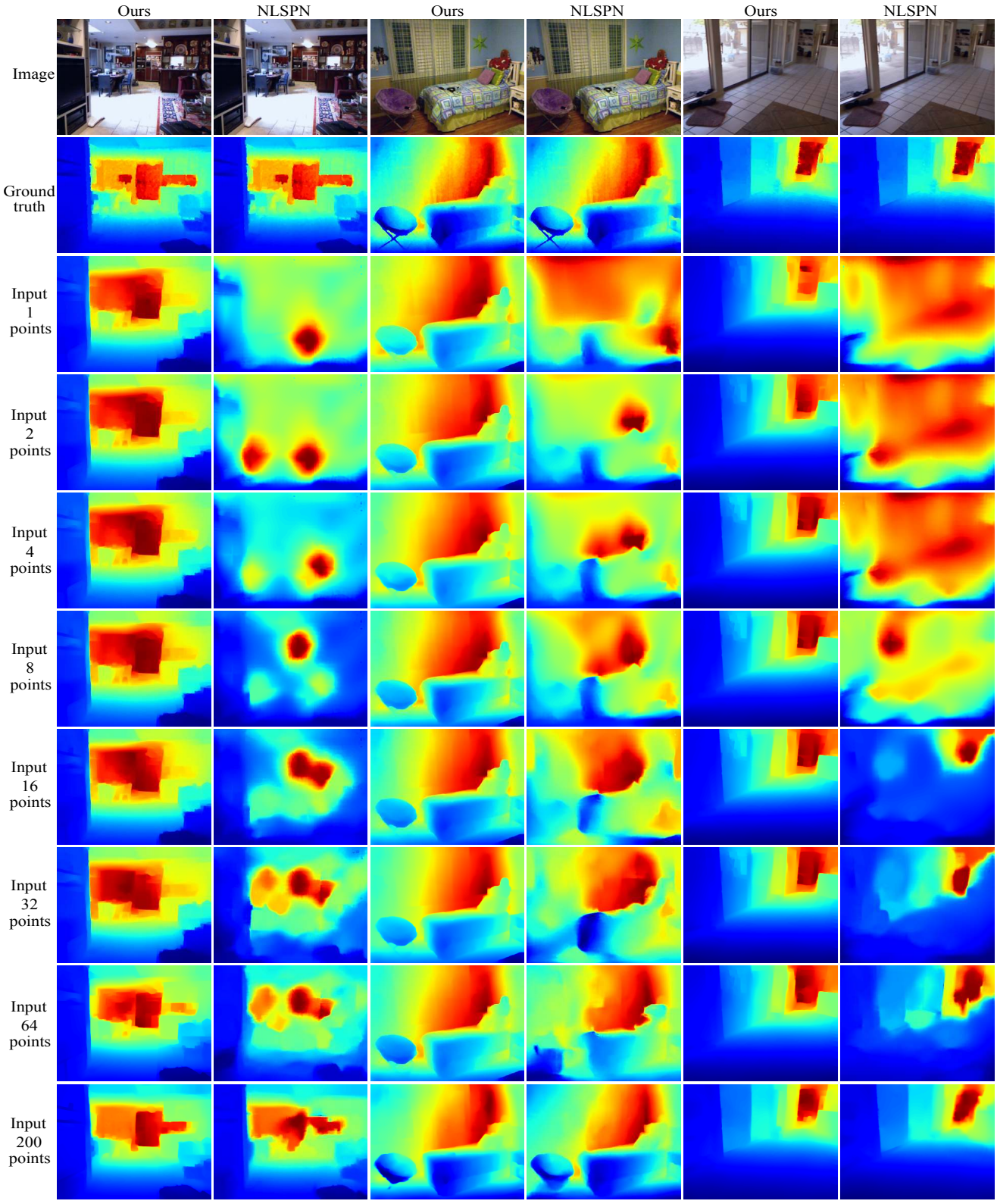

Figure 3. Examples from NYU-v2 with different number of input 3D points.

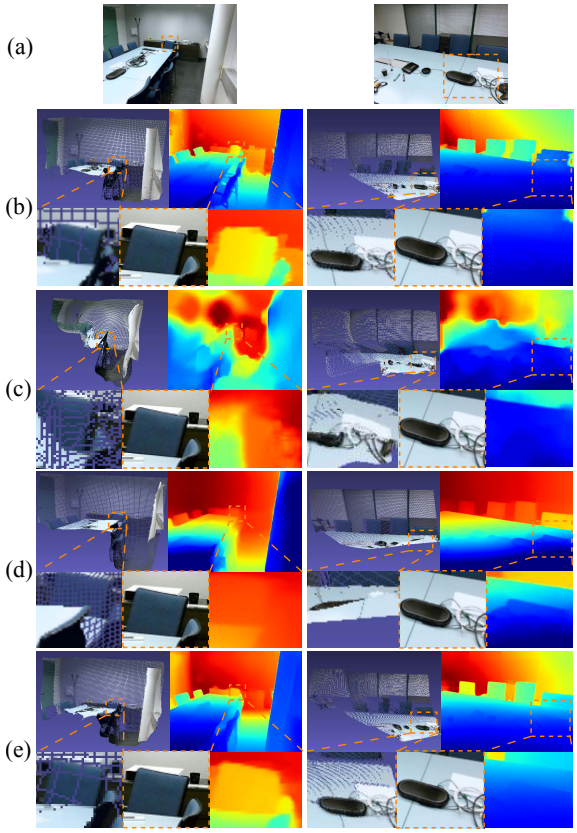

Figure 4. OuKi examples are captured using a Kinect-v2 (a). Dense depth maps and reconstructed point cloud: (b) ground truth, (c) NLSPN [3], (d) MVSNet [6], and (e) the proposed method.

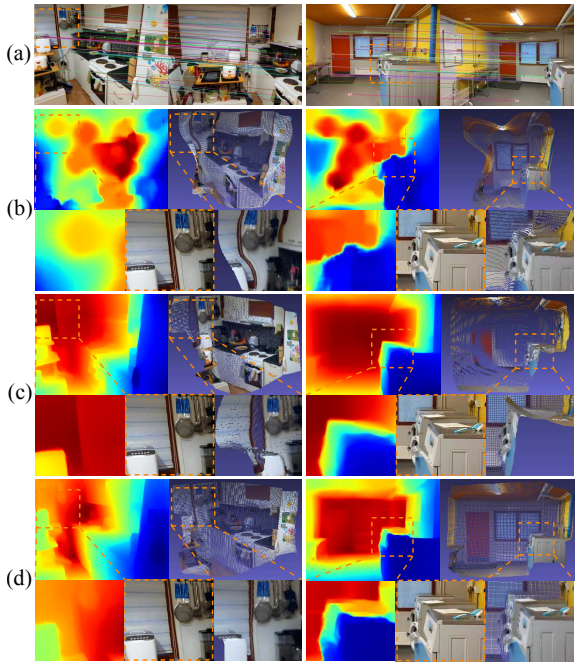

Figure 5. OuKi examples are captured using an Android phone (a). Dense depth maps and reconstructed point cloud from two images: (b) NLSPN [3], (c) MVSNet [6], and (d) the proposed method.

Figure 6. An example from OuKi dataset with different sparsity. Input 3D points are enhanced for visualization.