Supplementary Material for
Robust Geometry-Preserving Depth Estimation Using Differentiable Rendering

Chi Zhang[1*], Wei Yin[2*], Gang Yu[1†], Zhibin Wang[1], Tao Chen[3], Bin Fu[1], Joey Tianyi Zhou[5], Chunhua Shen[4]

[1] Tencent [2] DJI Technology [3] Fudan University [4] Zhejiang University
[5] Centre for Frontier AI Research, A*STAR [5] Institute of High Performance Computing , A*STAR

[1]{johnczhang, skicyyu, billzbwang, brianfu}@tencent.com
[2]yvanwy@outlook.com [3]eetchen@fudan.edu.cn [4]chunhua@me.com [5]joey.tianyi.zhou@gmail.com

1. Introduction

In our supplementary materials, we offer supplementary information regarding our research, which includes details on the dataset specifications, additional qualitative findings, and specifications on the algorithm used.

2. Method Detail

In this section, we present the derivation of Eq. 9 and Eq. 10 in the main body of our paper. The transformation of the camera amounts to the transformation of the 3D points. Given a rotation matrix R, transition matrix T, source point vector X, target point vector X', and rotation center $T_{\texttt{center}}$, the forward mapping function can be written as:

$$X' = R(X - T_{\texttt{center}}) + T + T_{\texttt{center}}, \tag{1}$$

and the backward mapping function is written as :

$$X = R_{\texttt{inv}} X\prime + T_{\texttt{inv}}. \tag{2}$$

The derivation is as follows:

$$\begin{aligned} X' &= R(X - T_{\texttt{center}}) + T + T_{\texttt{center}}, && (3)\\ R^T X' &= R^T R(X - T_{\texttt{center}}) + R^T(T + T_{\texttt{center}}), && (4)\\ R^T R(X - T_{\texttt{center}}) &= R^T X' - R^T(T + T_{\texttt{center}}), && (5)\\ X - T_{\texttt{center}} &= R^T X' - R^T(T + T_{\texttt{center}}), && (6)\\ X &= R^T X' - R^T(T + T_{\texttt{center}}) + T_{\texttt{center}}. && (7) \end{aligned}$$

Therefore, $R_{\texttt{inv}} = R^{\mathrm{T}}$, and $T_{\texttt{inv}} = -R^{\mathrm{T}}(T + T_{\texttt{center}}) + T_{\texttt{center}}$.

3. Dataset Information

Table 1 and Table 2 present detailed information on the training sets and evaluation datasets used in our study, respectively.

*Equal contributions.
†Corresponding author.

Dataset	Scene	Data Num.	Annotations
Taskonomy [11]	Indoor	114k	Metric depth
DIML [5]	Outdoor	121K	Disparity
Holopix50K [4]	Indoor & Outdoor	48K	Disparity
HRWSI [9]	Indoor & Outdoor	20K	Disparity

Table 1: Information of training datasets. We use the same mixed datasets from Leres [10].

Dataset	Scene	Data Num.	metric	Source
NYU [8]	Indoor	654	AbsRel, δ_1, RMSE	Kinect
Scannet [2]	Indoor	700	AbsRel δ_1	Kinect
KITTI [3]	Outdorr	652	AbsRel, δ_1, RMSE	LiDAR
ETH3D [6]	Indoor & Outdoor	431	AbsRel, δ_1	LiDAR
2D3D-S [1]	Indoor	1000	RMSE	LiDAR

Table 2: Information of evaluation datasets. The raw 2D3D dataset contains 64,235 images, and we sample 1K images for evaluation.

4. More Qualitative Results

We include additional qualitative results in our supplementary materials. Our attached files contain animated GIFs that provide visualizations of point clouds from multiple perspectives, displaying the structures in greater detail. Additionally, we create demo videos of 3D photo applications [7] based on our depth estimators, showcasing the robustness of our depth estimation approach.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv: Comp. Res. Repository*, page 1702.01105, 2017. 2

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5828–5839, 2017. 2

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3354–3361. IEEE, 2012. 2

[4] Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A large-scale in-the-wild stereo image dataset. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, June 2020. 2

[5] Youngjung Kim, Hyungjoo Jung, Dongbo Min, and Kwanghoon Sohn. Deep monocular depth estimation via integration of global and local predictions. *IEEE Trans. Image Process.*, 27(8):4131–4144, 2018. 2

[6] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3260–3269, 2017. 2

[7] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 8028–8038, 2020. 2

[8] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. Eur. Conf. Comp. Vis.*, pages 746–760. Springer, 2012. 2

[9] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 611–620, 2020. 2

[10] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021. 2

[11] Amir Zamir, Alexander Sax, , William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* IEEE, 2018. 2