Supplementary Material for "NerfingMVS: Guided Optimization of Neural Radiance Fields for Indoor Multi-view Stereo"

Yi Wei[1,2], Shaohui Liu[3], Yongming Rao[1,2], Wang Zhao[4], Jiwen Lu[1,2]*, Jie Zhou[1,2]
[1]Department of Automation, Tsinghua University, China
[2]Beijing National Research Center for Information Science and Technology, China
[3]ETH Zurich [4]Department of Computer Science and Technology, Tsinghua University, China
y-wei19@mails.tsinghua.edu.cn; b1ueber2y@gmail.com; raoyongming95@gmail.com;
zhao-w19@mails.tsinghua.edu.cn; {lujiwen, jzhou}@tsinghua.edu.cn

K	α_l	α_h	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
2	0.05	0.15	0.055	0.006	0.083	0.075	0.977	0.998	**1.000**
8	0.05	0.15	0.054	0.006	0.084	0.074	0.979	**0.999**	**1.000**
4	0.01	0.3	0.054	0.007	0.087	0.080	0.971	0.997	**1.000**
4	0.05	0.3	0.055	0.007	0.087	0.079	0.976	0.998	**1.000**
4	0.01	0.15	0.053	0.006	0.083	0.075	0.980	0.998	**1.000**
4	0.05	0.15	**0.051**	**0.005**	**0.076**	**0.069**	**0.987**	0.998	**1.000**

Table 1: Hyperparameter analysis. The experiment was conducted on scene0521.

A. Implementation Details

To train the proposed system, we mostly followed NeRF [5]. Specifically, we sampled 64 points in each ray and used a batch of 1024 rays. Since we did not adopt coarse-to-fine strategy in the sampling process, we only need one network (the architecture is same with [5]) to optimize the neural radiance fields. We added random Gaussian noise with zero mean and unit variance to the density σ to regularize the network. In addition, following [5], positional encoding was also employed. Adam was adopted as our optimizer with the initial learning rate as 5×10^{-4} and decayed exponentially to 5×10^{-5}. We utilized PyTorch [7] in our implementation. Each scene was trained with 200K iterations on a single RTX 2080 Ti.

Error metrics. We follow the metrics in [2, 4, 6, 10, 11, 13] to evaluate depth estimation results:

- Abs Rel: $\frac{1}{|T|}\sum_{y\in T}|y-y^*|/y^*$
- Sq Rel: $\frac{1}{|T|}\sum_{y\in T}||y-y^*||^2/y^*$
- RMSE: $\sqrt{\frac{1}{|T|}\sum_{y\in T}||y-y^*||^2}$
- RMSE log: $\sqrt{\frac{1}{|T|}\sum_{y\in T}||\log y-\log y^*||^2}$
- $\delta < t$: % of y s.t. $\max(\frac{y}{y^*}, \frac{y^*}{y}) = \delta < t$

where y and y^* indicate predicted and groundtruth depths respectively, and T indicates all pixels on the depth image.

*Corresponding author.

B. Baseline Method Details

We compared our results with several state-of-the-art depth estimation method, which can be roughly classified as four categories:

Conventional multi-view stereo: COLMAP [8, 9], ACMP [12]. COLMAP is a non-learning MVS method for 3D reconstruction building upon PatchMatch stereo [1]. Based on COLMAP, ACMP introduces planar models to solve low-textured areas in complex indoor environments.

Learning-based multi-view stereo: DELTAS [10], Atlas [6]. These two methods are trained on ScanNet with groundtruth depth supervision. For DELTAS, we used two neighboring frames as the reference frames.

Monocular depth estimation: Mannequin Challenge [3]. Mannequin Challenge is a state-of-the-art monocular depth estimation method. We directly used their pretrained weight for evaluation.

Video-based depth estimation: CVD [4], DeepV2D [11]. For video-based methods, we sorted images in a scene according to the timeline. DeepV2D is trained on ScanNet with groundtruth depth supervision.

C. Hyperparameter Analysis

To further demonstrate the effectiveness of our method, we did hyperparameter analysis for the number of used minimum errors K, and the bounds α_l, α_h used in the guided sampling process. The experiments were conducted on

scene0521. Table 1 shows experimental results. We find that using a K that is too small or too large will degrade the performance. On the one hand, it is possible to satisfy the multi-view consistency check although the depths are not correct. Small K will increase the probability of this phenomenon. On the other hand, there are pixels that do not overlap across some view pairs. Thus, the projection errors on some views are invalid and a large K may cover these invalid views. In addition, a large upper bound α_h or a small lower bound α_l for sampling range will lead to worse results, which indicates the necessity to set bounds in sampling process.

D. Additional Qualitative Results

In addition, we attach a video demo in the submitted supplementary material to show qualitative comparisons of multi-view depth estimation between our method and state-of-the-art methods [4–6].

References

[1] Michael Bleyer, Christoph Rhemann, and Carsten Rother. PatchMatch Stereo-Stereo Matching with Slanted Support Windows., 2011. 1

[2] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *CVPR*, pages 2189–2199, 2020. 1

[3] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, pages 4521–4530, 2019. 1

[4] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *TOG*, 39(4):71–1, 2020. 1, 2

[5] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*, pages 405–421. Springer, 2020. 1, 2

[6] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *ECCV*, 2020. 1, 2

[7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 1

[8] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 1

[9] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, pages 501–518. Springer, 2016. 1

[10] Ayan Sinha, Zak Murez, James Bartolozzi, Vijay Badrinarayanan, and Andrew Rabinovich. Deltas: Depth estimation by learning triangulation and densification of sparse points. In *ECCV*, 2020. 1

[11] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. In *ICLR*, 2020. 1

[12] Qingshan Xu and Wenbing Tao. Planar prior assisted patchmatch multi-view stereo. In *AAAI*, volume 34, pages 12516–12523, 2020. 1

[13] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 1851–1858, 2017. 1