PVP-Recon: Progressive View Planning via Warping Consistency for Sparse-View Surface Reconstruction - Appendix

A1 ADDITIONAL EXPERIMENT DETAILS

We conduct experiments on three datasets. To evaluate the reconstructed surfaces, we use the DTU [Jensen et al. 2014] and BlendedMVS [Yao et al. 2020] datasets. To evaluate the rendering quality, we use the Blender [Mildenhall et al. 2020] dataset. For a scene in these datasets, there are typically 50-150 views. We directly set the camera poses of all dense views as candidate viewpoints of *PVP-Recon*, but assume that images under these viewpoints are currently not acquired to simulate our problem setting. Specially, since generalization-based baselines (SparseNeuS [Long et al. 2022], VolRecon [Ren et al. 2023]) require a long pretraining process on multiple scenes, while other methods only optimize on a single scene, we do not compare the time consumption of generalization-based methods. Also, we do not compare generalization-based baselines on the Blender dataset, as these methods are pretrained on DTU scenes and cannot generalize to Blender scenes.

For the DTU dataset, the resolution of our input images is 800×600. For the BlendedMVS dataset, the resolution of our input images is 768×576. For the Blender dataset, the resolution of our input images is 800 × 800. The total iterations of surface optimization are 10,000. To accelerate training, we also employ NerfAcc [Li et al. 2023a] for efficient sampling in the volume rendering pipeline.

Although *PVP-Recon* outperforms other baselines on most scenes of the DTU dataset, we observe that **scan106** is an exception. The ground-truth 3D model of **scan106** has a hole, while our optimized SDF tends to fill this hole to generate a watertight mesh surface, leading to a drop in the reconstruction accuracy. Nevertheless, *PVP-Recon* produces comparable or better results on other scenes and achieves the lowest Chamfer distance on average.

A2 ABLATION OF NORMAL PRIOR

In our loss term, we add a normal loss that constrains the normal vectors rendered by the reconstruction module to be consistent with the pseudo ground-truth normal vectors predicted by Omnidata [Eftekhar et al. 2021]. Omnidata is a state-of-the-art monocular surface normal estimation model trained on 14.5 million images. We assume that Omnidata provides valuable normal prior information that facilitates the optimization of mesh surfaces.

In Figure A1, we show the ablation results of removing normal prior regularization. Note that the normal prior serves as a necessary constraint, especially in textureless regions where color does not provide sufficient supervision. The reconstruction accuracy decreases when normal constraints are removed.

A3 THE DIFFERENCE BETWEEN NEURALANGELO

Both our reconstruction module and Neuralangelo [Li et al. 2023b] adopt multi-resolution hash features to represent SDF, as hash features excel at capturing fine-grained details. Additionally, Neuralangelo uses a coarse-to-fine training scheme that gradually reduces the step size of its proposed numerical gradients and increases the resolution of hash features. The entire optimization process is lengthy (as also mentioned in their paper). Different from theirs, our scheme focuses on solving the severe overfitting problem under sparse-view SDF optimization. We design a progressive training scheme that linearly activates hash features according to training iterations, and a directional Hessian loss for further regularization. Furthermore, we first use down-sampled images to enlarge the sampling receptive field, and switch to full-resolution images after the training process stabilizes. By leveraging the above techniques, our model converges much faster. Experiment results show that our *PVP-Recon* can generate better results in ten minutes, while Neuralangelo requires several hours of optimization.

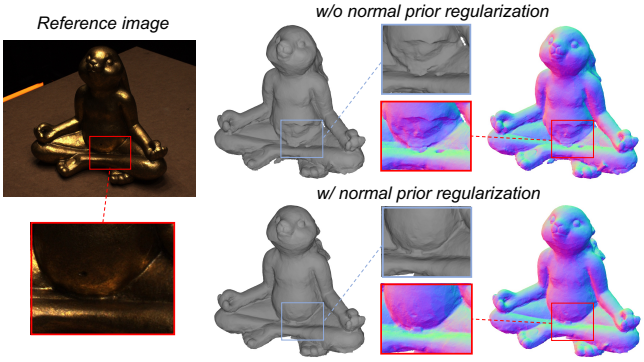

Fig. A1. Ablation results of the normal prior (DTU scan110). Without the normal constraints, artifacts will appear in textureless regions.

A4 CHOICE OF TARGET WARPED VIEWS

Our view planning module utilizes a warp-based scoring strategy. To assess the potential contribution of a candidate camera pose, we render the image and depth under this pose using the reconstruction module, and warp the rendered image to one of the existing training views. The problem is how to decide which training image to warp to. In practice, we warp to the training image whose camera pose is closest to this candidate pose. This strategy is simple and effective, and ensures that the source and target images have considerable overlap to avoid meaningless warping results. We also extend this strategy by warping the rendered image to K-nearest training images and averaging the scores. Table A1 reports the reconstruction quality by setting K to different values ($K = 1, 2, 3$). We observe that directly warping the rendered image to its closest training image ($K = 1$) yields the best results. Therefore, we use $K = 1$ as the default setting in our warp-based scoring strategy.

REFERENCES

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. 2021. Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets From 3D Scans. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal,*

Table A1. The reconstruction results using different target warped views on the DTU dataset. We report the Chamfer distance (mm) ↓.

Configs	scan55	scan65	scan69	scan83	scan105	scan106	scan110	scan114	scan118	scan122	mean
K=1	0.51	1.15	0.75	1.28	0.84	0.91	0.95	0.46	0.72	0.50	0.81
K=2	0.59	1.24	0.80	1.31	0.85	1.03	1.10	0.49	0.71	0.54	0.87
K=3	0.56	1.23	0.81	1.29	0.89	0.92	1.09	0.53	0.73	0.53	0.86

QC, Canada, October 10-17, 2021. IEEE, US, 10786–10796.

Rasmus Ramsbøl Jensen, Anders Lindbjerg Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. 2014. Large Scale Multi-view Stereopsis Evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, Washington, DC, 406–413.

Ruilong Li, Hang Gao, Matthew Tancik, and Angjoo Kanazawa. 2023a. Nerfacc: Efficient sampling accelerates nerfs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, US, 18537–18546.

Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H. Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. 2023b. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, US, 8456–8465.

Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. 2022. SparseNeuS: Fast Generalizable Neural Surface Reconstruction from Sparse Views. In *ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII*, Vol. 13692. Springer, Heidelberg, Germany, 210–227.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, Vol. 12346. Springer, Heidelberg, Germany, 405–421.

Yufan Ren, Fangjinhua Wang, Tong Zhang, Marc Pollefeys, and Sabine Süsstrunk. 2023. VolRecon: Volume Rendering of Signed Ray Distance Functions for Generalizable Multi-View Reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*. IEEE, US, 16685–16695.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. 2020. BlendedMVS: A Large-Scale Dataset for Generalized Multi-View Stereo Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, New York, NY, 1787–1796.