Supplementary of NeuRIS

Jiepeng Wang[1], Peng Wang[1], Xiaoxiao Long[1], Christian Theobalt[2], Taku Komura[1], Lingjie Liu[2], and Wenping Wang[3]

[1] The University of Hong Kong {jpwang,pwang3,xxlong,taku}@cs.hku.hk
[2] Max Planck Institute for Informatics {lliu,theobalt}@mpi-inf.mpg.de
[3] Texas A&M University wenping@tamu.edu

1 Implementation details

Evaluation. As shown in Fig. 1, the output mesh by some TSDF-based (truncated signed distance function) methods [7, 13] is double-layered while the GT mesh is single-layered. Thus, it is not fair if using the original double-layered mesh to perform evaluations directly. Thus, for the double-layered mesh, we remove the back layer according to the visibility of triangle faces from cameras. In other words, if a triangle face is not visible to all cameras, this face will be removed. Another situation is that some areas in the predicted mesh by some methods, such as [7, 15], are out of the scope of the ground truth (GT) mesh as illustrated in the main paper. These areas in the predicted mesh should also be removed to guarantee fair comparisons.

To address these issues, we remove faces in the areas in a (predicted) mesh that are not observed in its corresponding GT mesh or not visible to all cameras. Specifically, we clean the mesh following the steps below: (1) we first remove the faces that cannot be observed by any input view; (2) then we filter faces at empty regions of GT mesh by 2D masks, which are rendered from GT mesh for each input view. For each mask pixel, if there is no projection from the GT mesh for its incompleteness, the corresponding value is false, otherwise the value is true.

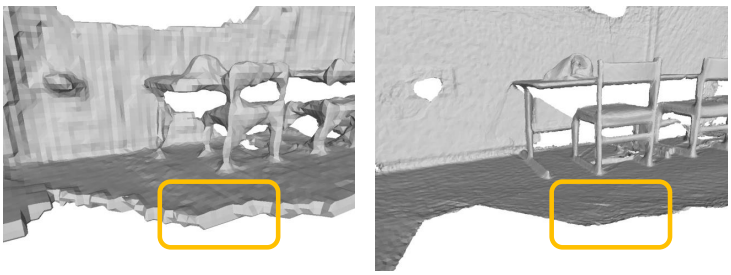

Fig. 1. Visualization of double-layered mesh and single-layered mesh.

Training. Sphere initialization [2] is applied to the SDF (signed distance function) network. But different from [15] where the SDF values of the outer and inner region of the sphere are initialized with positive and negative values respectively, we reverse the signs, i.e., the SDF values inside the sphere are positive. This is because for cameras capturing indoor scenes [4] they usually look from inside out, while for those capturing small objects [1] they usually look from outside in. Besides, the weights of color loss λ_c, prior loss λ_p and Eikonal loss λ_{eik} used in the training process are 1.0, 1.0, and 0.1 respectively. And the robust threshold ϵ in patch match is 0.66, following OpenMVS [8].

Patch match. We follow OpenMVS [8] using a patch size of 11×11 with a step size 2 for patch-match in all our experiments. Because captured scenes in ScanNet are usually texture-less with small baselines and large image noises, it is hard to use conventional stereo pair selection strategies [11, 10, 6], to select neighbor views for all input images. Thus, for simplicity, we instead choose 6 adjacent views as neighbor referencing views. For the data of indoor scenes with sparse images instead of videos, such as Hypersim [9], we follow [8] to choose neighbor views. Besides, in our pre-experiments, we find patch-match is not robust to image noises at texture-less regions, where the numerator and denominator of Eq. 5 in the main paper can be both very small. To this end, we add a small value at the numerator, which can guarantee a more robust calculation of NCC at texture-less areas therefore the normal priors at such areas will not be wrongly removed.

Normal network. For fair evaluations, we first divide the Scannet dataset into training split with 1180 scenes and testing split with 433 scenes. Then, we retrain the normal network [5] on our training split instead of using the officially pre-trained model provided by [5]. The two splits have no overlapping scenes and all scenes we used are in the testing split. Note that a real physical scene often has multiple corresponding video sequences in the ScanNet dataset which were captured in different camera trajectories. Thus, in our experiments, all the sequences related to the scenes we used are also divided into the testing split.

2 Limitations

NeuRIS may fail in the scenes with low lighting conditions, where the structural information cannot be well reflected in the captured images. Besides, we also notice that the normal estimations may be not correct when the camera observation angle is tilted greatly or there is a picture on the wall. If many input views are captured under such conditions, our method may not produce satisfactory geometry or contain artifacts. Leveraging multi-view information of input images instead of monocular input may help to get better normal estimations [3] and further help to produce better geometry in our optimization framework, which is out of the scope of this paper and can be explored as a future direction.

3 Evaluation metrics

Table 1 defines the the 3D geometry metrics and 2D depth metrics used for evaluations. In general, F-score is considered as the most proper metric to evaluate the quality of 3D geometry [13], which contains the information of both accuracy and completeness.

2D		3D	
Metric	Definition	Metric	Definition
Abs Rel	$\frac{1}{n}\sum \lvert d_o - d_g \rvert / d_g$	Accuracy	$\mathbf{mean}_{p_o \in P_o}(\mathbf{min}_{p_g \in P_g} \lvert\lvert p_o - p_g \rvert\rvert)$
Sq Rel	$\frac{1}{n}\sum \lvert d_o - d_g \rvert^2 / d_g$	Comp.	$\mathbf{mean}_{p_g \in P_g}(\mathbf{min}_{p_o \in P_o} \lvert\lvert p_o - p_g \rvert\rvert)$
RMSE	$\sqrt{\frac{1}{n}\sum \lvert d_o - d_g \rvert^2}$	Prec.	$\mathbf{mean}_{p_o \in P_o}(\mathbf{min}_{p_g \in P_g} \lvert\lvert p_o - p_g \rvert\rvert < 0.05)$
RMSE log	$\sqrt{\frac{1}{n}\sum \lvert \log(d_o) - \log(d_g) \rvert^2}$	Recall	$\mathbf{mean}_{p_g \in P_g}(\mathbf{min}_{p_o \in P_o} \lvert\lvert p_o - p_g \rvert\rvert < 0.05)$
$\delta < 1.25$	$\frac{1}{n}\sum(\max(\frac{d_o}{d_g}, \frac{d_g}{d_o}) < 1.25)$	F-score	$\frac{2 \times Prec. \times Recall}{Prec. + Recall}$

Table 1. Evaluation metrics used in our paper. n is the number of pixels with valid depth in ground truth (GT) depth map. d_o and d_g are the predicted and GT depths, respectively. p_o and p_g are the vertices in predicted mesh P_o and GT mesh P_g respectively.

4 More evaluation results

Evaluation baseline NeRF. For the baseline NeRF, we use the implementation in NeuS. And we use the density level set 20 to extract surfaces, which can produce best geometries with smallest reconstruction errors in our pre-experiments as shown in Table 2. Fig. 2 shows the extracted geometries of one scene with different threshold.

Table 2. Quantitative evaluation results between the GT mesh and the extracted surfaces from NeRF with different thresholds over 8 scenes.

Threshold	Accu. ↓	Comp.↓	Prec.↑	Recall↑	F-score↑
0	0.337	0.418	0.041	0.038	0.038
10	**0.121**	0.082	0.347	0.457	0.390
20	0.127	**0.080**	0.404	**0.512**	**0.436**
30	0.147	0.093	0.412	0.493	0.424
40	0.169	0.114	0.409	0.440	0.387
50	0.196	0.139	0.409	0.381	0.348
75	0.255	0.236	0.422	0.253	0.259
100	0.280	0.349	**0.448**	0.162	0.187

2D depth evaluation. We additionally report the evaluation results using 2D metrics on depth maps here. Note that some methods [7, 13, 15, 17] as well as

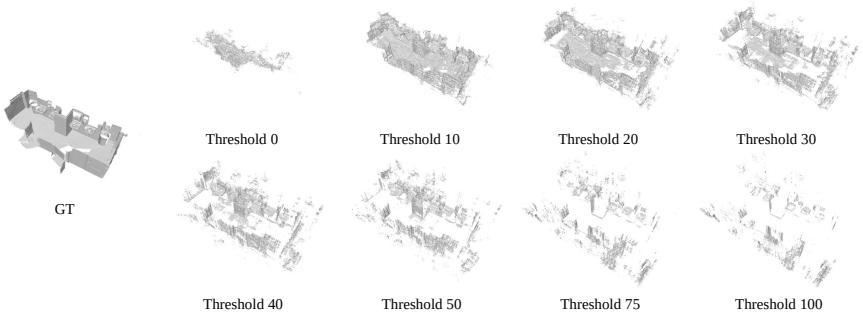

Fig. 2. Visualization of extracted surfaces from NeRF results of one scene with different threshold values.

ours do not generate depth maps explicitly. For these methods we render depth maps of each input view from their output meshes. As shown in Table 3, our method can surpass almost all existing methods, except for DeepV2D whose depth maps are re-scaled according to GT depth maps. If no scaling strategy is adopted (i.e., 'DeepV2D-no scale' in Table 3), our method is significantly better. For NerfingMVS [16], it failed on most (5/8) room-scale scenes in our pre-experiments and it only showed the results of reconstructing local room regions in its original paper. Thus, we only report the averaged scores on succeeded scenes here. (Note that our results here are for room-scale reconstructions while the results in the original paper of NerfingMVS [16] is for a local region in rooms, which cannot be compared directly.)

Table 3. Quantitative evaluation on ScanNet using 2D depth metrics over 8 room-scale scenes. For NerfingMVS, the scores are averaged on 3 scenes because it failed on other 5 scenes.

Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25 \uparrow$
COLMAP[10]	0.155	0.168	0.515	0.576	0.796
NeuralRecon[13]	0.108	0.146	0.469	0.620	0.901
Atlas[7]	0.079	0.090	0.297	0.293	0.927
DeepV2D[14]	0.065	**0.021**	0.173	**0.094**	0.959
DeepV2D-no scale	0.160	0.071	0.299	0.176	0.774
NerfingMVS[16]	0.079	0.050	0.266	0.192	0.935
NeuS[15]	0.154	0.124	0.419	0.372	0.784
Ours	**0.045**	0.022	**0.169**	0.102	**0.962**

Generalizability. We evaluate NeuRIS on two other indoor datasets to test the generalizability on large-scale indoor scenes: Hypersim [9] and Replica [12]. As shown in Fig. 3, our method generalizes well to unseen datasets and outperforms the baseline NeuS[15].

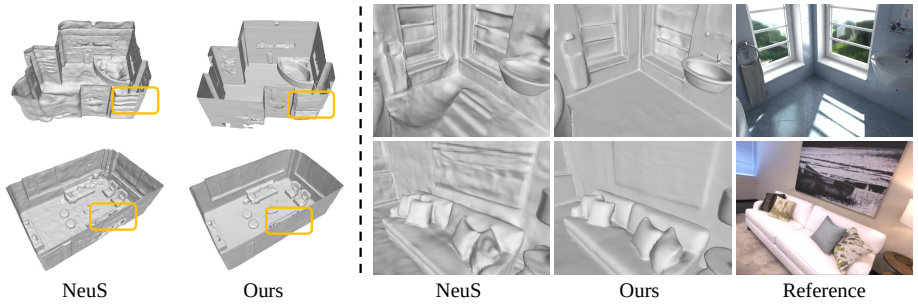

Fig. 3. Evaluation on Hypersim (the first row) and Replica (the second row) datasets. The left part shows the top view of the whole rooms while the right part shows the zoom-in view of the marked areas.

More qualitative results on ScanNet. Fig. 4, Fig. 5 and Fig. 6 show more qualitative results of geometry, novel view synthesis and normal predictions, respectively. Our method can produce much more complete and precise geometry, images with higher quality and better normal predictions.

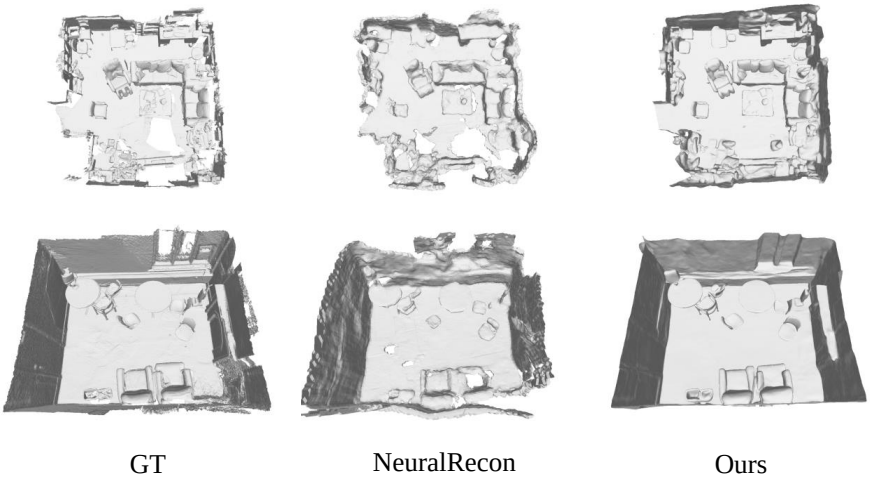

Fig. 4. More qualitative geometry comparisons. The three columns show the GT mesh, reconstructed mesh by NeuralRecon[13], and reconstructed mesh by our method, respectively.

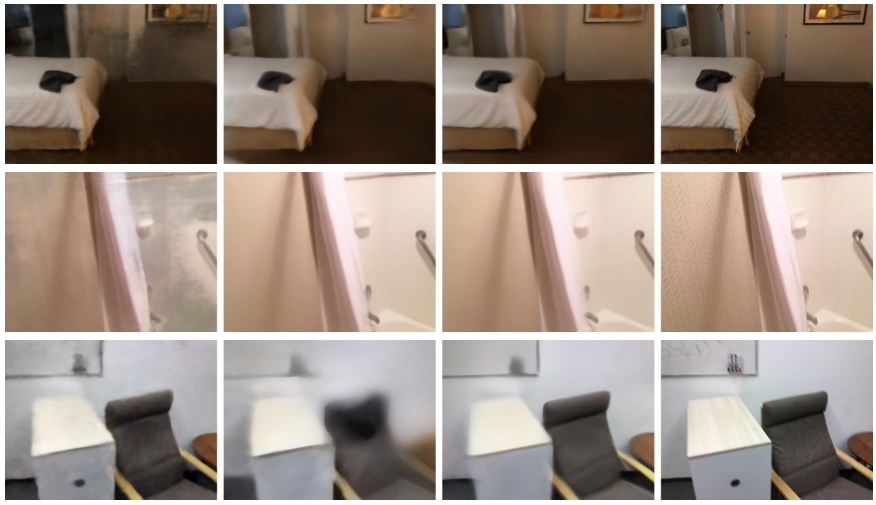

Fig. 5. More novel view synthesis results.

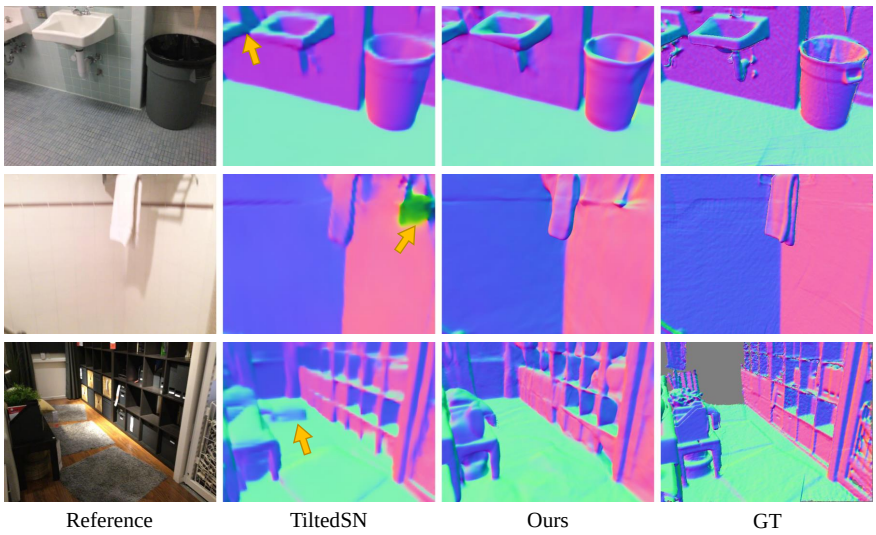

Fig. 6. More qualitative normal comparisons.

References

1. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. International Journal of Computer Vision pp. 1–16 (2016)
2. Atzmon, M., Lipman, Y.: Sal: Sign agnostic learning of shapes from raw data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2565–2574 (2020)
3. Bae, G., Budvytis, I., Cipolla, R.: Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In: International Conference on Computer Vision (ICCV) (2021)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2017)
5. Do, T., Vuong, K., Roumeliotis, S.I., Park, H.S.: Surface normal estimation of tilted images via spatial rectifier. In: Proc. of the European Conference on Computer Vision. Virtual Conference (August 23–28 2020)
6. Li, J., Li, E., Chen, Y., Xu, L., Zhang, Y.: Bundled depth-map merging for multi-view stereo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2769–2776. IEEE (2010)
7. Murez, Z., van As, T., Bartolozzi, J., Sinha, A., Badrinarayanan, V., Rabinovich, A.: Atlas: End-to-end 3d scene reconstruction from posed images. In: ECCV (2020), https://arxiv.org/abs/2003.10432
8. openMVS: https://github.com/cdcseacave/openMVS
9. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: International Conference on Computer Vision (ICCV) 2021 (2021)
10. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)
11. Shen, S.: Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes. IEEE transactions on image processing **22**(5), 1901–1914 (2013)
12. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H.M., Nardi, R.D., Goesele, M., Lovegrove, S., Newcombe, R.: The Replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
13. Sun, J., Xie, Y., Chen, L., Zhou, X., Bao, H.: NeuralRecon: Real-time coherent 3D reconstruction from monocular video. CVPR (2021)
14. Teed, Z., Deng, J.: Deepv2d: Video to depth with differentiable structure from motion. arXiv preprint arXiv:1812.04605 (2018)
15. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2021)
16. Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In: ICCV (2021)
17. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. Advances in Neural Information Processing Systems **34** (2021)