Surface-Centric Modeling for High-Fidelity Generalizable Neural Surface Reconstruction (Supplementary Materials)

Rui Peng[1,2], Shihe Shen[1], Kaiqiang Xiong[1], Huachen Gao[1],
Jianbo Jiao[3], Xiaodong Gu[4], and Ronggang Wang✉[1,2]

[1]School of Electronic and Computer Engineering, Peking University
[2]Peng Cheng Laboratory [3]University of Birmingham [4]Alibaba
ruipeng@stu.pku.edu.cn rgwang@pkusz.edu.cn

A Results of Fine-tuning

As the qualitative and quantitative comparisons shown in our main paper, the reconstructions of our model without fine-tuning exhabit finest geometric details even compared with some fine-tuned models like SparseNeuS-ft [8]. Here, we illustrate some results of our model after fast fine-tuning. Different with methods [5,15] that require reconstructing separate cost volume for each view, our model only builds the global volume, which makes our model easily fine-tuned (only 2.5k iterations, about 10 minutes). The quantitative results in Tab. A show that our model still ranks the first in most scenes and has the best mean chamfer distance. Meanwhile, it is worth noting that our volume is sparse and more memory and computationally efficient. And the qualitative results of some scenes are visulized in Fig. A. Note that there are only three input views during fine-tuning.

Table A: Quantitative results of the fine-tuned model on DTU dataset. Best results in each category are in **bold** and the second best are in underline.

Method	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
IBRNet-ft [13]	1.67	2.97	2.26	1.56	2.52	2.30	1.50	2.05	2.02	1.73	1.66	1.63	1.17	1.84	1.61	1.90
SparseNeuS-ft [8]	1.29	2.27	1.57	0.88	1.61	1.86	1.06	1.27	1.42	1.07	0.99	0.87	0.54	1.15	1.18	1.27
GenS-ft [9]	0.91	2.33	1.46	**0.75**	**1.02**	1.58	0.74	1.16	1.05	**0.77**	0.88	0.56	**0.49**	**0.78**	**0.93**	1.03
SuRF-ft (Ours)	**0.73**	**2.11**	**1.39**	0.83	1.05	**1.53**	**0.68**	**1.03**	**1.02**	0.84	**0.85**	**0.46**	**0.49**	0.84	1.00	**0.99**

Table B: Reproduced results of VolRecon [10] and ReTR [7] in two image sets.

Method	Image Set	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
VolRecon [10]	Set0	1.27	2.66	1.54	1.04	1.41	1.94	1.10	1.53	1.36	1.08	1.18	1.37	0.74	1.22	1.26	1.38
	Set1	1.80	3.46	2.14	1.12	1.92	1.74	1.17	1.72	1.63	1.31	0.94	1.46	0.78	1.23	1.30	1.58
	Average	1.54	3.05	1.84	1.08	1.67	1.84	1.13	1.63	1.49	1.19	1.06	1.42	0.76	1.22	1.29	1.48
ReTR [7]	Set0	1.05	2.32	1.47	0.97	1.22	1.52	0.88	1.30	1.29	0.87	1.07	0.76	0.58	1.11	1.12	1.17
	Set1	1.42	2.95	1.76	0.99	1.55	1.59	0.92	1.49	1.50	1.19	0.79	0.89	0.60	1.09	1.21	1.33
	Average	1.23	2.63	1.62	0.98	1.38	1.56	0.90	1.39	1.39	1.02	0.93	0.83	0.59	1.10	1.16	1.25

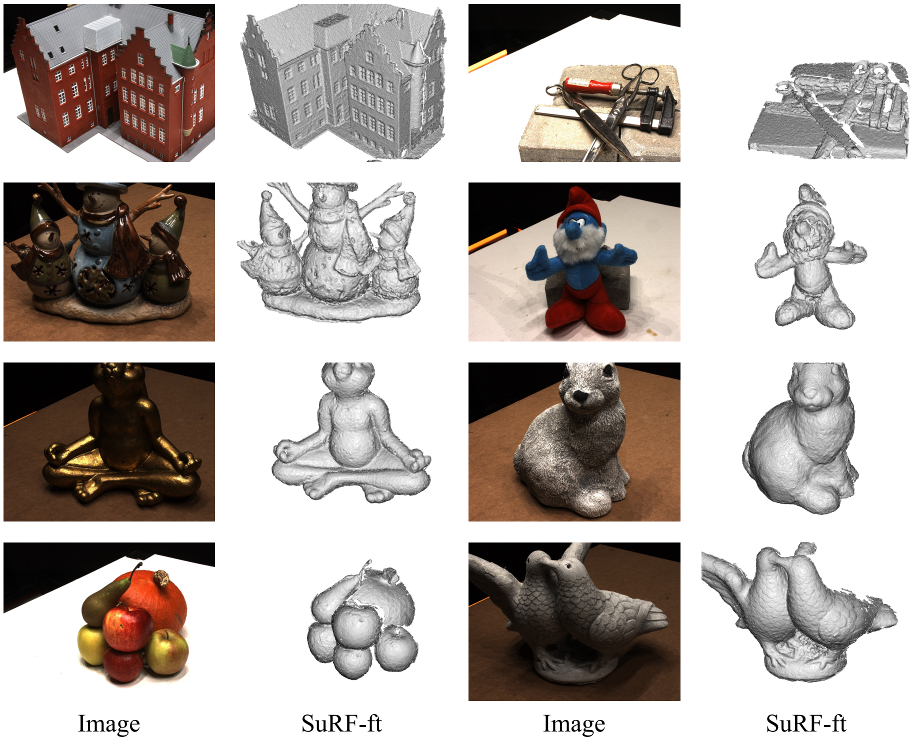

Fig. A: Visualization of fine-tuning results.

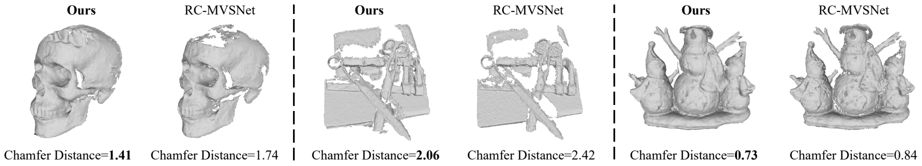

Fig. B: Visual and metric comparisons with RC-MVSNet.

B More Comparisons with RC-MVSNet

We show some visual and metric comparisons with the TSDF fusion result of RC-MVSNet [1] in Fig. B. The results show that the reconstruction of our model is smoother and more complete, especially in low-texture regions, leading to better results in the chamfer distance metric. To further verify the effectiveness of our surface-centric modeling, we compare with two baselines which directly use the surface point of RC-MVSNet to prune voxels: Baseline1 directly replaces the surface region of our trained model with that of RC-MVSNet; Baseline2 uses the surface region of RC-MVSNet to train a new model. Results in Tab. C show that even simply using the surface region of RC-MVSNet can achieve superior results. And our full model, trained together with the surface location module (Our matching field), achieves the best performance. This is reasonable because

Table C: More ablation studies about directly using the surface region of RC-MVSNet.

	RC-MVSNet	Baseline1	Baseline2	Ours
Chamfer Distance	1.22	1.13	1.16	**1.05**

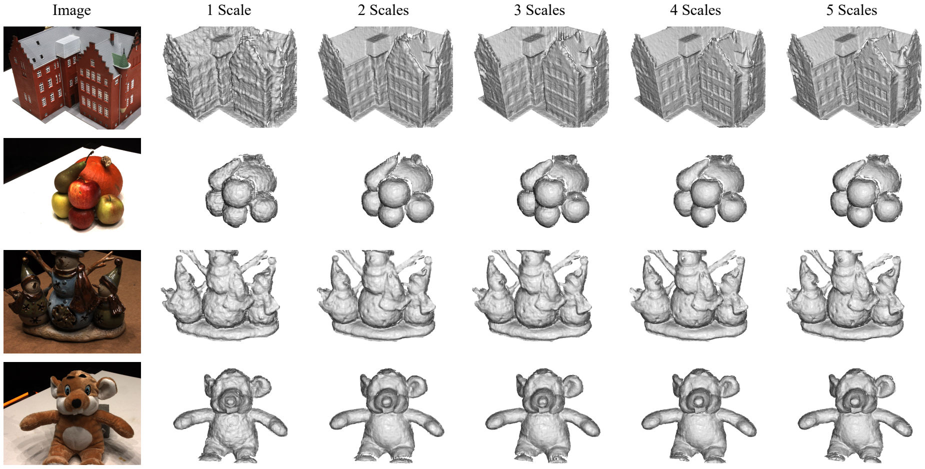

Fig. C: Visual comparison with different number of scales on DTU dataset.

the surface region of these two baselines was not optimized or corrected with the model when directly using the results of RC-MVSNet.

C Detailed Results of VolRecon and ReTR

As mentioned in our main paper, we report the reproduced results of VolRecon [10] and ReTR [7] on two image sets using their official repositories and released model checkpoints. The detailed reproduction results of all scenes at two image sets are illustrated in Tab. B, which are slightly different from the results reported in their papers. We speculate that there are something inconsistent in the experimental configurations, but this inconsistency doesn't affect the valuable of their contributions.

D More Ablation Results

Here, we report more ablation results of our model, and we set the training time to a quarter of the overall process (different from our main paper to save time) and only test on the first image set for convenience.

Number of scales. We conduct some ablations to evaluate the effect of the number of scales. We set the resolution of the finest stage of each model to be similar. The results in Tab. D show that the overall quality first remarkably increases and then slightly decreses, reaching the optimum in 4 scales. We illustrate some visual results of the model with differnet scales in Fig. C. The single-scale model performs the worst, with reconstructions that are noisy and lack geometric detail, while the four-scales model can reconstruct smooth geometry and restore more geometric details.

Table D: Ablation results on DTU dataset.

Number of scales	Surface sampling	Cross-scale fusion	Mean
1 scales	✓	✓	1.38
2 scales	✓	✓	1.22
3 scales	✓	✓	1.15
4 scales	✓	✓	**1.11**
5 scales	✓	✓	1.13
4 scales	✓	✗	1.15
4 scales	✗	✓	1.13

Ablations on loss weight. We further conduct some experiments to verify the effect of the weight of each loss term on model performance. Concretely, we change the weight of the pseudo loss β and the weight combination of different stages of matching field loss μ^j. The ablation model is based on the 4-scales model and the results are shown in Tab. E. Through the comparison between model A and model B, we can see that the pseudo point clouds generated from the unsupervised multi-view stereo method [1] can guide the model towards better convergence. To avoid the influence of erroneous pseudo points, we apply a very strict filtering strategy, *i.e.*, Only point clouds whose projection distance from at least 3 viewpoints does not exceed 0.2 pixels and whose relative depth error does not exceed 0.001 can be left. From the results of the model (B, C, D) adopt different weight combinations of the matching field loss, we can see that model B which has $\mu^1 : \mu^2 : \mu^3 : \mu^4 = 0.25 : 0.50 : 0.75 : 1.00$ performs the best and model D performs the worst. This indicates that applying greater weight to the high-resolution scale is beneficial to model convergence. Because there is no need to obtain very accurate predictions in the low-resolution scale, and the gradient of the high-resolution scale will be transmitted back to the low-resolution scale, it is reasonable to have a lower weight in the low-resolution scale. And we show some visual comparisons of these models in Fig. D.

Table E: Ablation results of loss weight on DTU dataset.

Method	β	$\mu^1 : \mu^2 : \mu^3 : \mu^4$	Mean
A	0.0	0.25 : 0.50 : 0.75 : 1.00	1.17
B	1.0	0.25 : 0.50 : 0.75 : 1.00	**1.11**
C	1.0	1.00 : 1.00 : 1.00 : 1.00	1.16
D	1.0	1.00 : 0.75 : 0.50 : 0.25	1.25

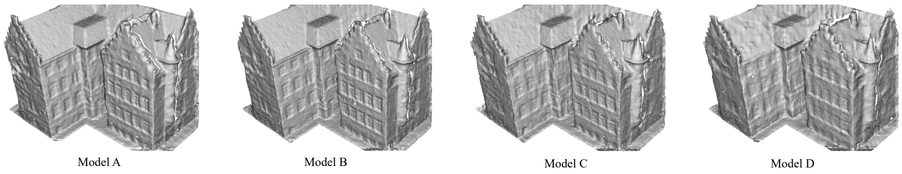

Fig. D: Visual comparison of reconstructions with different loss weights.

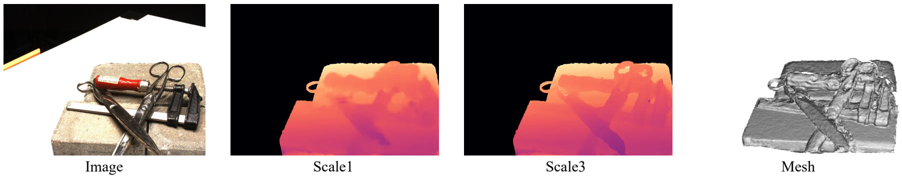

Fig. E: Visualization of the located surface region from the matching field at different scales. We visualize the depth of the middle surface for convenience.

Ablations on the range of surface region. Here, we employ an additional ablation experiment to study the sensitivity of the range of surface regions. ϵ^0 is the range of the first scale, and its value is fixed at 1, which represents covering the entire near and far area. And the value of later scales means the percentage of coverage. As the results shown in Tab. F, the differences of these three groups of experiments are not large, as long as the surface region is gradually tightened, and model F which has a range combination of $\epsilon^1 : \epsilon^2 : \epsilon^3 : \epsilon^4 = 1.00 : 0.30 : 0.10 : 0.01$ performs the best.

Table F: Ablation results of the range of surface regions.

Method	$\epsilon^1 : \epsilon^2 : \epsilon^3 : \epsilon^4$	Mean
B	1.00 : 0.40 : 0.10 : 0.01	1.11
E	1.00 : 0.30 : 0.10 : 0.01	**1.10**
F	1.00 : 0.30 : 0.05 : 0.01	1.12

E Visualization of the Surface Region

To understand how the surface region changes as scale increases, we show some visualization results of the surface region at different scales in Fig. E. For convenience, we show the depth of the middle position of the surface region. We can see that the located surface region at the higher resolution scale is indeed sharper, which proves the effectiveness of our design.

F Limitations and Future Work

Despite exhibiting efficiency over existing methods, our model still struggled to extract the surface in real time due to the inherent drawback of MLP-based implicit methods. In the future, we will be focusing on addressing this deficiency issue, and we have constructed a lite-version model, which will be released later. Furthermore, we plane to train our model on more large-scale dataset like Objaverse [2] and expand the scale of the model like [4, 6].

G More Results

Because C2F2NeuS [15] doesn't release the code, the memory of C2F2NeuS in Tab. 2 of our main paper is refer to the implementation of CasMVSNet [3]. Fig.

F shows additional comparisons with COLMAP [11], NeuS [12], SparseNeuS [8], SparseNeuS-ft [8], VolRecon [10] and ReTR [7] on DTU dataset. We can see that our method can stably achieve superior results and exhibit finer geometry details. We further show some visual comparisons with the fast per-scene overfitting method Voxurf [14] in Fig. G. While Voxurf still requires more than 30 minutes of training time per scene, it struggles to reconstruct smooth and accurate surface from sparse inputs.

References

1. Chang, D., Božič, A., Zhang, T., Yan, Q., Chen, Y., Süsstrunk, S., Nießner, M.: Rc-mvsnet: unsupervised multi-view stereo with neural rendering. In: ECCV (2022)
2. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: CVPR. pp. 13142–13153 (2023)
3. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: CVPR. pp. 2495–2504 (2020)
4. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. In: ICLR (2024)
5. Johari, M.M., Lepoittevin, Y., Fleuret, F.: Geonerf: Generalizing nerf with geometry priors. In: CVPR. pp. 18365–18375 (2022)
6. Li, J., Tan, H., Zhang, K., Xu, Z., Luan, F., Xu, Y., Hong, Y., Sunkavalli, K., Shakhnarovich, G., Bi, S.: Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In: ICLR (2024)
7. Liang, Y., He, H., Chen, Y.c.: Rethinking rendering in generalizable neural surface reconstruction: A learning-based solution. In: NeurIPS (2023)
8. Long, X., Lin, C., Wang, P., Komura, T., Wang, W.: Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In: ECCV. pp. 210–227 (2022)
9. Peng, R., Gu, X., Tang, L., Shen, S., Yu, F., Wang, R.: Gens: Generalizable neural surface reconstruction from multi-view images. In: NeurIPS (2023)
10. Ren, Y., Zhang, T., Pollefeys, M., Süsstrunk, S., Wang, F.: Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In: CVPR. pp. 16685–16695 (2023)
11. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR. pp. 4104–4113 (2016)
12. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In: NeurIPS (2021)
13. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: CVPR. pp. 4690–4699 (2021)
14. Wu, T., Wang, J., Pan, X., Xu, X., Theobalt, C., Liu, Z., Lin, D.: Voxurf: Voxel-based efficient and accurate neural surface reconstruction. In: ICLR (2022)
15. Xu, L., Guan, T., Wang, Y., Liu, W., Zeng, Z., Wang, J., Yang, W.: C2f2neus: Cascade cost frustum fusion for high fidelity and generalizable neural surface reconstruction. In: ICCV (2023)

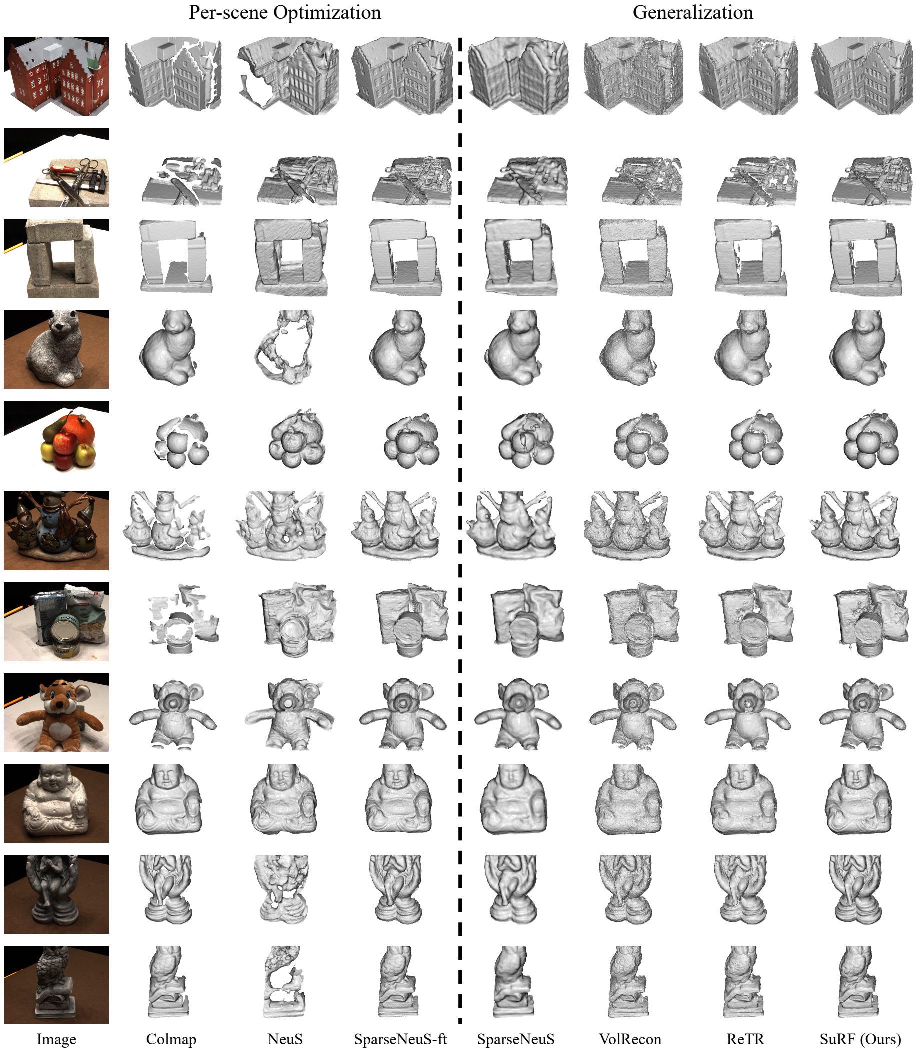

Fig. F: More qualitative comparisons on DTU dataset.

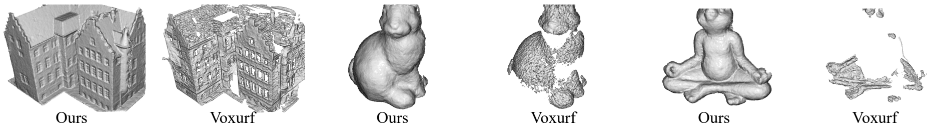

Fig. G: Comparison with Voxurf on DTU dataset with 3 inputs.