Supplementary Materials for S-VolSDF: Sparse Multi-View Stereo Regularization of Neural Implicit Surfaces

Haoyu Wu Alexandros Graikos Dimitris Samaras
Stony Brook University
{haoyuwu,agraikos,samaras}@cs.stonybrook.edu

In Sec. 1 we report additional results on 3D reconstructions, novel view synthesis, the implicit surface optimization process, scalability, and limitations of our method. In Sec. 2 we describe in further detail our experiment settings. We also include a supplementary video that compares the results of our method against various baselines.

1. Additional Results

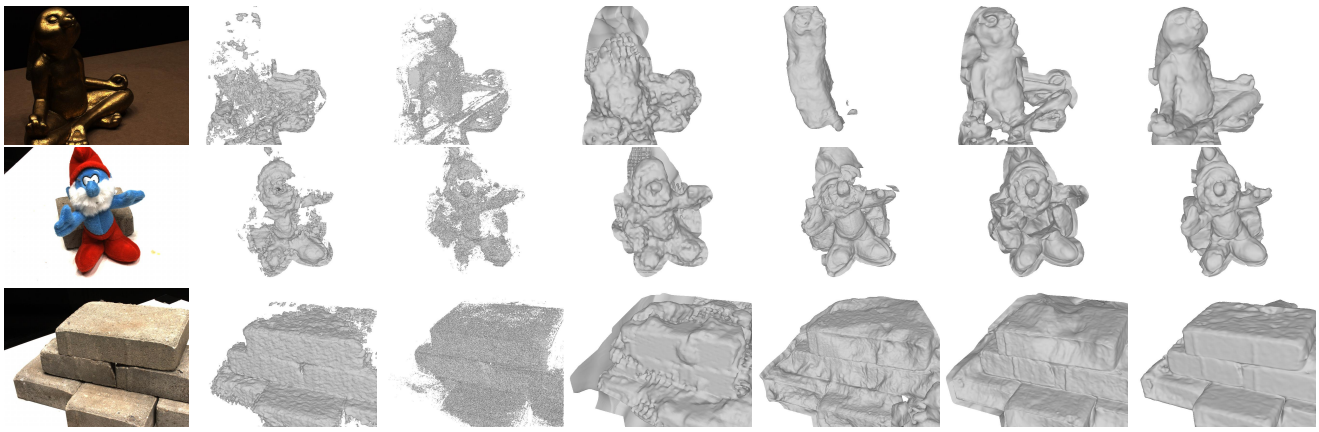


Figure 1. Additional 3D reconstruction results of neural rendering methods on DTU. Our results appear more complete and accurate.

Additional Results on 3D Reconstructions. We showcase additional meshes extracted from neural rendering methods on three-view 3D reconstruction for the DTU [1] and BlendedMVS [18] datasets (Fig. 1 and Fig. 2). We provide more point cloud visualizations of the results when combining our method with different MVS models in Fig. 3 and Fig. 4.

Additional Results on Novel View Synthesis. In Fig. 5 and Fig. 2 we showcase additional qualitative comparisons between our method and the baselines on novel view synthesis for the DTU and BlendedMVS datasets.

Optimization Process. In Fig. 6, we show an example of how the implicit surface evolves during the optimization process. Our output surface reconstruction after 10-15 minutes of training (on an NVIDIA A5000 GPU) is already more accurate than the reconstruction of a fully trained VolSDF [19] (typically 4-10 hours).

Scalability. We conduct an ablation study on the scalability of our method. Fig. 7 and Tab. 1 show that as the input views become denser, the performance of our method, measured by surface reconstruction and novel view synthesis quality, improves and is consistently better than CasMVSNet [7] and VolSDF [19]. Tab. 2 shows that, for three given views, the reconstruction quality of our method remains the same when varying the input image resolution. CasMVSNet [7] and VolSDF [19] perform worse when lowering the image resolution.

Ablation Study on Different MVS Models. In Tab. 3, we provide an extended ablation study across all three MVS mod-

	PSNR ↑			SSIM ↑			LPIPS ↓		
	3-views	6-views	9-views	3-views	6-views	9-views	3-views	6-views	9-views
VolSDF [19]	16.99	20.19	**23.04**	0.786	0.823	0.836	0.332	0.317	0.310
Ours	20.21	20.80	22.98	0.820	0.824	0.832	0.321	0.318	0.309
Ours$_{IR}$	**20.58**	**21.48**	23.01	**0.855**	**0.872**	**0.895**	**0.157**	**0.145**	**0.128**

Table 1. Quantitative results on novel view synthesis with 3-9 input views on DTU.

	Chamfer ↓			PSNR ↑			SSIM ↑			LPIPS ↓		
Resolution	Low	Mid	High	Low	Mid	High	Low	Mid	High	Low	Mid	High
CasMVSNet [7]	1.92	1.86	1.87		—			—			—	
VolSDF [19]	2.56	2.80	2.70	16.99	15.52	15.75	0.786	0.771	0.790	0.332	0.352	0.346
Ours	**1.32**	**1.33**	**1.33**	20.21	19.63	19.97	0.820	0.822	0.833	0.321	0.330	0.330
Ours$_{IR}$		—		**20.58**	**19.98**	**20.30**	**0.855**	**0.853**	**0.858**	**0.157**	**0.178**	**0.186**

Table 2. Quantitative results with different image resolutions: low (576×768), mid (864×1152), and high (1152×1536), on DTU.

els: TransMVSNet [6], UCSNet [4], and CasMVSNet [7]. It validates the importance of using probability volumes, soft consistency check, and generalized cross-entropy loss, consistent with our main text's ablation study findings.

Chamfer (mm)↓	TransMVSNet [6]	UCSNet [4]	CasMVSNet [7]
MVS Model	2.915	2.201	1.920
MVS + **Ours**	1.798	1.519	**1.320**
only soft consistency	2.627	1.901	1.711
MSE loss	2.233	2.019	1.792
w/o prob. volume	2.692	1.791	1.543
w/o GCE loss	2.525	1.702	1.534

Table 3. Ablation study on different MVS models, on DTU.

Additional Comparison with Related Work. In Tab. 4, we provide additional comparisons with regularization based approach including DS-NeRF [5], which utilizes estimated depth from structure-from-motion [12], and MonoSDF [22], which [14] utilizes monocular depth estimation. Because their depth priors are either sparse or often not accurate enough, providing only approximated structures or shapes, their results are worse than ours.

	MonoSDF [22]	DS-NeRF [5]	**Ours**
Chamfer (mm)↓	2.141	1.792	**1.32**

Table 4. Additional comparison with related work, on DTU.

Limitations. While our method is also capable of refining the probability volumes of the finer stages of MVS, we notice that the benefits diminsh since there is not as much uncertainty in later stages. Our method applied to stages 1, 1,2, and 1,2,3 of MVS resulted in chamfer distances of 1.320, 1.312, and 1.309, respectively.

Evaluation on Objects with Glossy Material. Although our method may not work well for texture-less or glossy surfaces due to the introduction of MVS. Surprisingly, as shown in Fig. 8 and Tab. 5, our method still surpasses VolSDF in reconstructing complex glossy surfaces. We suspect that our noise-tolerant optimization and MVS models operating on features instead of pixels make our pipeline more robust to specular reflections that violate multi-view consistency. Further research on this problem would be quite interesting.

2. Experimental Settings

Hyperparameters. We observe a strong over-fitting tendency for VolSDF [19] with sparse input views. This over-fitting is due to the usage of the view direction to explain object color in different views, and therefore we set the positional encoding level of view direction to 1 for VolSDF and our method. We use the same loss functions as VolSDF [19], along with our

	PSNR ↑	SSIM ↑	LPIPS ↓	MAE° ↓
VolSDF [19]	20.71	0.943	0.126	32.96
Ours	20.97	0.944	0.124	**29.26**
Ours$_{IR}$	**21.50**	**0.944**	**0.081**	

Table 5. Results on Shiny Dataset (6 scenes, from Ref-NeRF [13]). Mean angular error (MAE) is used in evaluating normal vectors.

weight loss $\mathcal{L}_{weight}$ and a sparsity regularization $\mathcal{L}_{sparse}$. Both $\mathcal{L}_{weight}$ and $\mathcal{L}_{sparse}$ are weighted with a value of 1.0. The ϵ in $\mathcal{L}_{sparse}$ is 0.001. Moreover, we do not apply weight loss for rays with weak MVS supervision (i.e. the sum of consistency-weighted probability along the ray is less than 0.001). We found that our weight loss is highly tolerant to parameter choices. We used grid search to find the best q but determined that all q in $[0.2, 0.8]$ yield satisfactory results (overall error: 1.32-1.44). We set $q = 0.5$ in all our experiments.

Rendering Pipeline. In testing, our method utilizes image-based rendering. We merge source pixels from multiple source images for a target pixel. More specifically, we first render depth maps for all source views. Then, for a target view, we render its depth map and project its pixels back to the source views and we apply consistency check on the back-projected depths with the source depth to determine its visibility on source views and retrieve the interpolated source pixel colors. The blending weights for pixel colors from different source views are based on the cosine between the target and source pixels' view directions, computed using *softmax* with a temperature of 20. In areas where there are no valid pixels to blend (i.e., the geometric consistency check fails for all source views), we use the rendered colors. Finally, a 4-level Laplacian pyramid [2] is used to smoothly blend source pixels.

MVS Models. In our experiments, we compare our proposed method against TransMVSNet [6], CasMVSNet [7], and UCSNet [4]. We employ the official implementation of each method provided by the authors and use their published pre-trained models. To ensure a fair comparison, the weights for all three models we used were pre-trained exclusively on the DTU dataset [1] with ground-truth depth as supervision.

Denser Plane Sweep. The main difference in our training scheme, compared to MVS models, is the usage of a denser plane sweep, which we also implemented for all baseline MVS models, reducing their overall error by 33% on average.

The Choice of CasMVSNet and VolSDF. In our method, we select CasMVSNet [7] as the MVS model and VolSDF [19] as the neural rendering model. We opt for CasMVSNet as it is the representative coarse-to-fine MVS model, and we find no substantial improvement in other recent MVS models when compared to CasMVSNet for sparse-input scenarios, as demonstrated in the main text. We use VolSDF, which is a state-of-the-art implicit surface reconstruction method, as demonstrated in [9, 19]. Nevertheless, other neural rendering models like NeRF [10] and NeuS [15] can also be used in our method but the differences in the overall performance are a subject for future work.

Metrics. The Chamfer distance is the average of the Accuracy (the distance from the reconstructed point cloud to reference) and Completeness (the distance from reference to reconstruction). The use of stronger geometric/photometric filtering can lead to better accuracy, but at the expense of completeness, and vice-versa. Given this trade-off between accuracy and completeness in point cloud filtering, we choose to employ the Chamfer distance metric as our primary measure in the main text, following [20, 19]. We present the Accuracy-Completeness trade-off in Fig. 9. The results reveal that we consistently attain roughly 30% higher completeness than the baseline across all accuracy levels.

Datasets. For the DTU dataset [1], we combine the scans used in [20, 19, 21] with the ones used in conventional MVS settings [6, 17], and remove the training scans of common MVS models. Specifically, we use scans 21, 24, 34, 37, 38, 40, 82, 106, 110, 114, and 118 for our evaluation. For evaluation on DTU, we adhere to the standard protocol in [1, 19, 11].

The BlendedMVS dataset [18] lacks a standard evaluation protocol for sparse-view scenarios. Therefore, we adopted a similar evaluation protocol to DTU; select three sparse input views with a relatively wide baseline and evaluate using object masks. Similar to DTU, only scene objects are used in the evaluation. This is simply performed by removing the plane from the ground truth point cloud. The sparse view indexes we adopt are: Doll: 9, 10, 55; Egg: 9, 52, 59; Head: 22, 26, 27; Angel: 11, 39, 53; Bull: 32, 42, 47; Robot: 28, 34, 57; Dog: 2, 5, 25; Bread: 16, 21, 33; Camera: 10, 16, 60. For reference, we offer quantitative comparisons without using object masks or removing the plane in Tab. 6.

In the context of novel view synthesis, it is noteworthy that while the BlendedMVS dataset has 360-degree views of an object, the sparse inputs partially cover the frontal area. Consequently, conducting novel view synthesis on all images, including the back views, is unreasonable. Therefore, we choose to evaluate the closest 12 views in each scene. The indexes for evaluation are: Doll: 0, 13, 19, 20, 22, 31, 33, 35, 36, 37, 58, 61; Egg: 1, 8, 12, 14, 23, 27, 37, 39, 49, 65, 68, 71; Head:

Scene	Doll	Egg	Head	Angel	Bull	Robot	Dog	Bread	Camera	**Mean**
MVSNeRF [3]	22.3	-9.7	-30.8	38.1	4.1	24.8	-2.7	2.7	8.6	6.4
GeoNeRF [8]	**48.8**	37.9	3.6	37.6	-7.8	30.3	29.4	19.1	9.2	23.1
CasMVSNet [7]	46.2	47.7	-0.2	45.8	-6.6	41.5	41.3	8.9	31.8	28.5
Ours	47.8	**62.0**	**23.3**	**54.7**	**20.6**	**49.7**	**48.0**	**59.9**	**49.3**	**46.1**

Table 6. BlendedMVS 3D reconstruction results without applying object masks on the reconstruction results. Since there are no units in BlendedMVS, we report relative improvement (in %) over VolSDF [19] in terms of Chamfer distance.

0, 1, 6, 7, 11, 13, 15, 16, 25, 28, 31, 33; Angel: 0, 2, 9, 23, 29, 30, 46, 48, 50, 59, 68, 71; Bull: 0, 13, 16, 17, 20, 24, 26, 41, 44, 55, 57, 58; Robot: 1, 2, 10, 13, 22, 25, 40, 55, 73, 75, 80, 88; Dog: 0, 6, 7, 8, 10, 13, 14, 17, 22, 23, 27, 29; Bread: 8, 10, 17, 18, 24, 25, 26, 27, 28, 30, 43, 47; Camera: 18, 25, 59, 65, 68, 83, 89, 92, 94, 118, 133, 136.

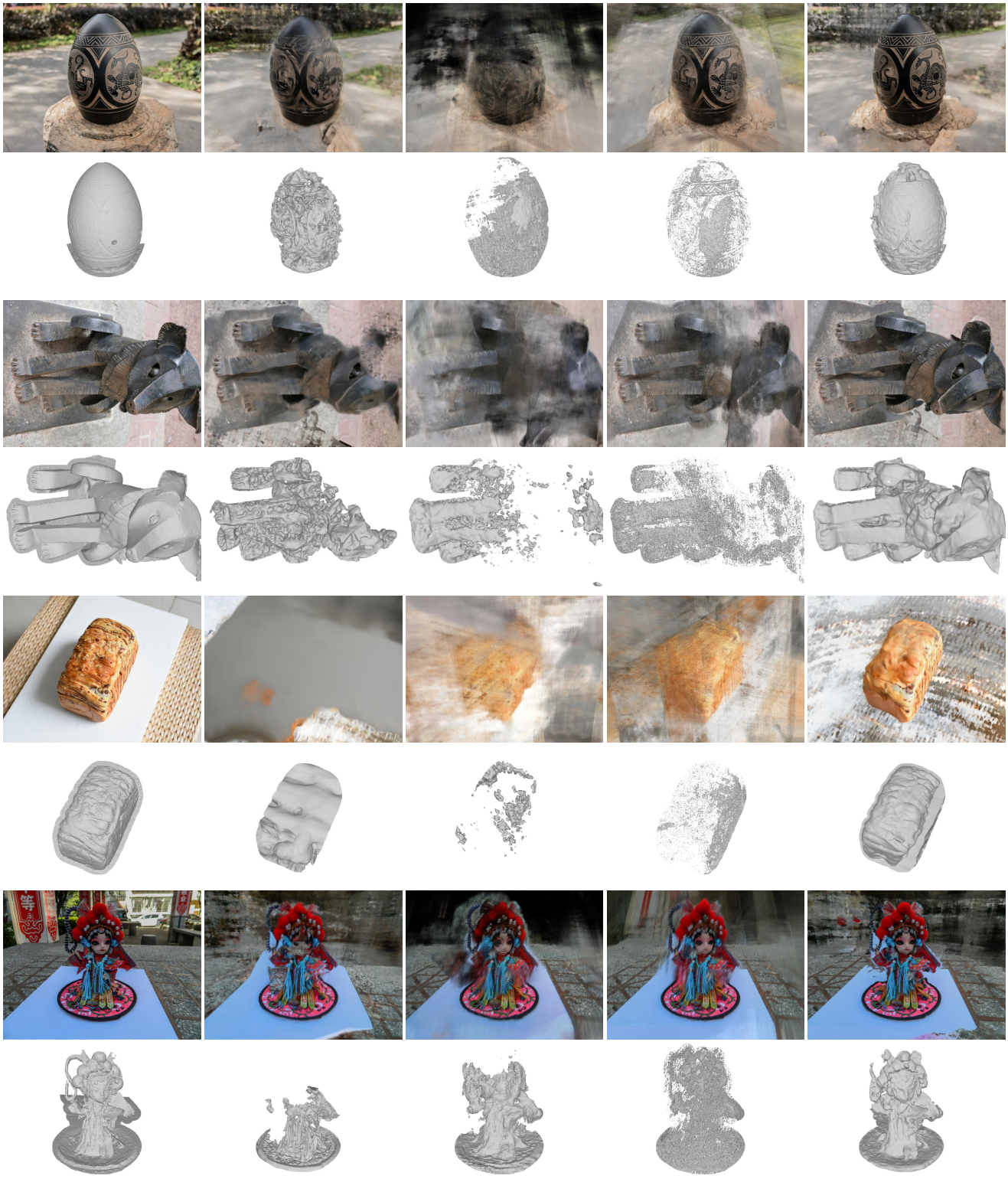


Figure 2. Additional 3D reconstruction and novel view synthesis comparisons on BlendedMVS. Our results appear more complete and accurate.

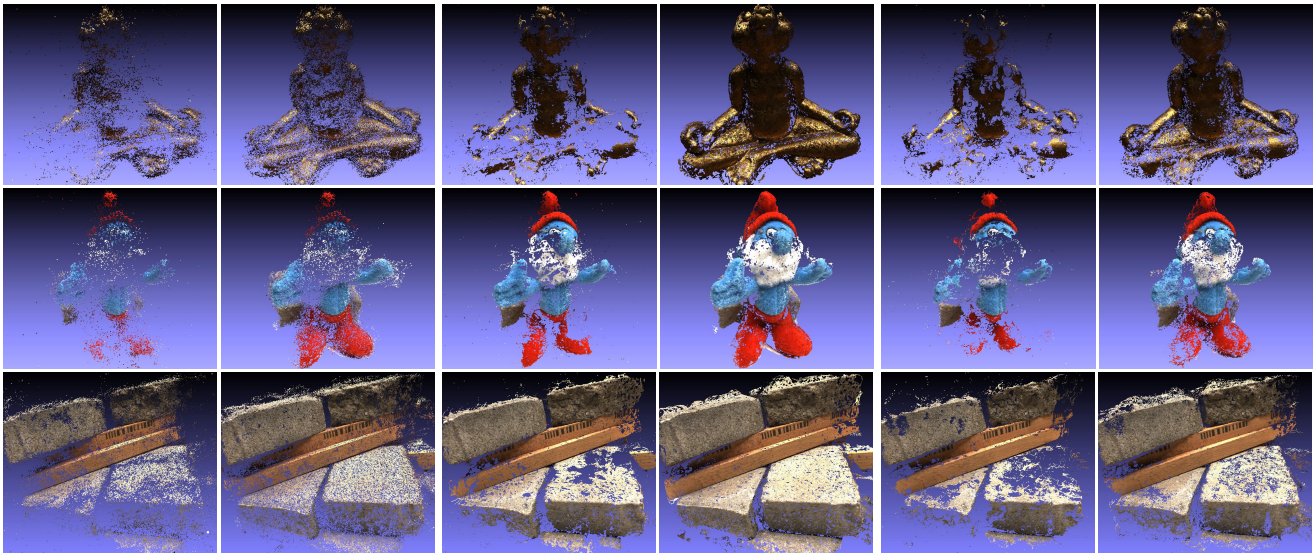


Figure 3. Additional point cloud visualization on DTU. Results improve in all combinations of our method with different MVS models.

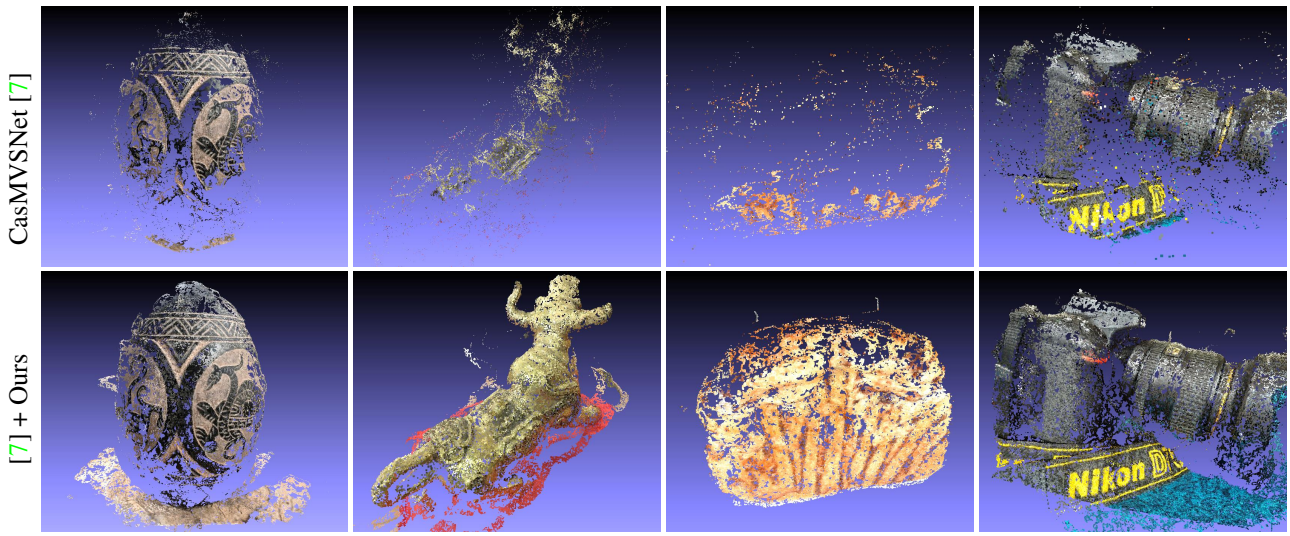


Figure 4. Point cloud visualization on BlendedMVS when combining our method with CasMVSNet [7].

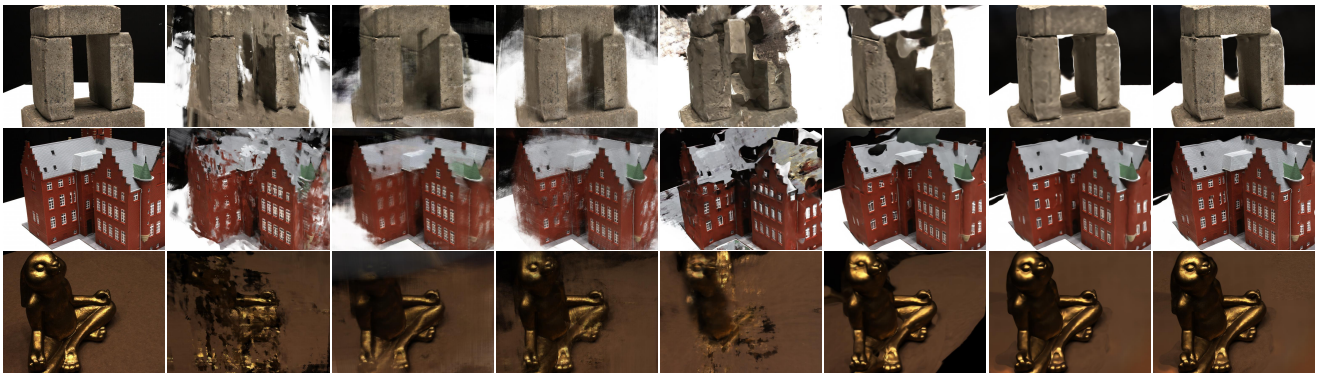


Figure 5. Additional novel view synthesis comparison on DTU. Our method leads to more accurate novel views.

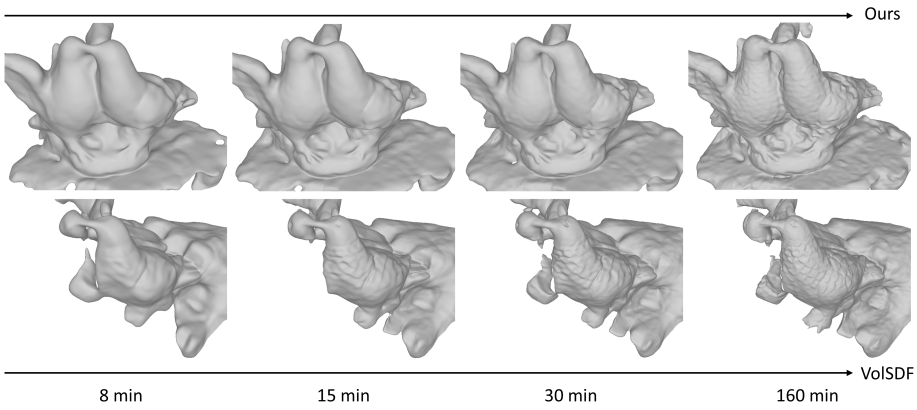


Figure 6. An example of the implicit surface during the optimization process. We show that, with only 10-15 minutes of training, our output surface reconstruction is already reasonably good to guide finer stage of MVS, compared to the sub-optimal results of VolSDF [19].

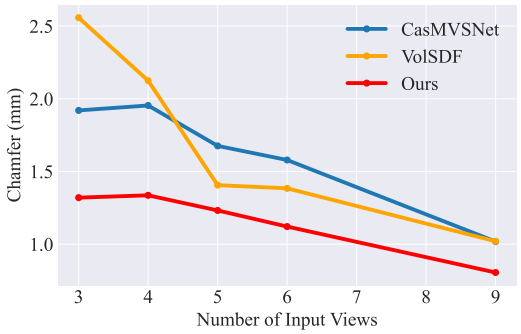
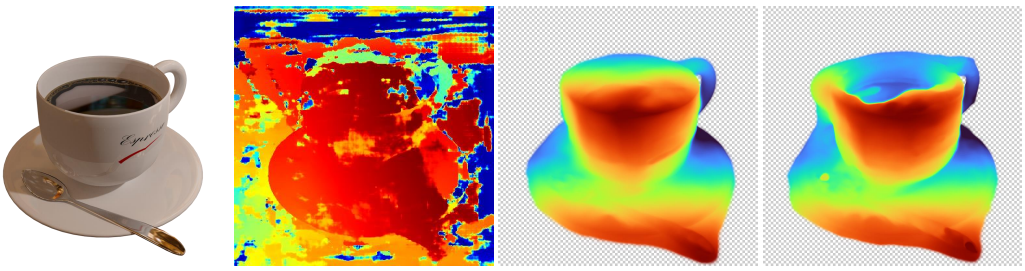


Figure 7. Quantitative results on 3D reconstruction with 3-9 input views on DTU.



GT CasMVSNet [7] VolSDF [19] Ours

Figure 8. Depth map predictions on Shiny Dataset.

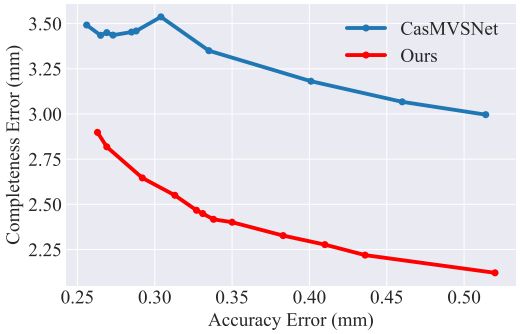


Figure 9. Completeness error and Accuracy error trade-off.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 1, 3

[2] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pages 671–679. Elsevier, 1987. 3

[3] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 1, 4, 5, 7

[4] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020. 2, 3, 6

[5] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 2

[6] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8585–8594, 2022. 2, 3, 6

[7] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020. 1, 2, 3, 4, 6, 8

[8] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022. 1, 4, 5, 7

[9] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. *arXiv preprint arXiv:2206.05737*, 2022. 1, 3

[10] 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 *European conference on computer vision*, pages 405–421. Springer, 2020. 3

[11] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 3

[12] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[13] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 3

[14] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. *arXiv preprint arXiv:2303.16196*, 2023. 2

[15] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1, 3, 7

[16] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 7

[17] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 3

[18] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020. 1, 3

[19] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 1, 2, 3, 4, 5, 7, 8

[20] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 3

[21] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 3

[22] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665*, 2022. 2