Supplementary Material for Learning Signed Distance Field for Multi-view Surface Reconstruction

1. Implementation Details

We first provide additional implementation details that are not discussed in the main paper due to the space limit.

1.1. Baselines

Vis-MVSNet We use the official Vis-MVSNet implementation [5] in our experiments. In MVS step, the source image number is set to 2 and the depth sample number is set to 256. Output depth map size is set to 512×384 for *DTU*, 600×400 for *EPFL* and 640×360 for *Tanks and Temples*. In depth filtering step, the probability thresholds are set to $0.9, 0.7, 0.3$ for *DTU* and *EPFL* and $0.9, 0.9, 0.8$ for *Tanks and Temples*. The number of geometric consistency is 4 for *DTU*, 3 for *EPFL* and 5 for *Tanks and Temples*.

In order to use sPSR for mesh reconstruction, normal maps are additionally computed from filtered depth maps of Vis-MVSNet. Fused point clouds of both Vis-MVSNet and Colmap are further clipped by the bounding box used in MVSDF, and meshes are reconstructed by sPSR with octree depth as 9 and trim parameter as 5. The implementation of sPSR is provided by Open3D [6].

IDR In order to test IDR on the *EPFL* dataset, we manually create the image masks and use the same bounding box as in MVSDF. We use the official IDR implementation in our experiments [4]. The network is trained by 10000 epochs. The scheduling of learning rate and alpha value is also scaled accordingly.

1.2. MVSDF

MVS Module In our MVS Module, parameters of source image number, depth sample number, output depth map size and probability thresholds are all set to the same as Vis-MVSNet baseline. For the feature loss, we use deep image feature maps from the last scale, whose size is the same as the final depth map.

Loss Weights During training, weights of Eikonal loss, indicator loss and render loss are all fixed to $w_E, w_P, w_R = [0.1, 0.01, 0.5]$. Weights of the distance loss and feature loss will be changed in different training stages: 1) in the first $1/6$ of the training, $w_D = 1.0$ and $w_F = 0$; 2) in the next $1/3$, $w_D = w_F = 0.1$; 3) in the last $1/2$, $w_D = w_F = 0.01$.

We observe that optimization by feature loss and render loss may diverge in the second and third stage of our training process (see the **Training** part in Sec. 4.1 in the main paper). To ensure that the surface can only be locally refined within certain range, we only decrease the weight of distance loss for sample points whose absolute value of calculated distance $|l(\mathbf{x})|$ is less than 5% of the bounding box size.

1.3. Evaluations

Chamfer Distance For *DTU* dataset, the Chamfer distance is calculated by the provided MATLAB code [1]. For *EPFL* dataset, we use our own evaluation script. First, we crop the ground truth mesh by the manual image masks. Then both input mesh and ground truth mesh are sampled to point clouds by a predefined sample number. The reported value is the average of Chamfer distance from the input to the ground truth and also from the ground truth to the input. For both directions, we excluded points with distance larger than 0.8.

PSNR We only evaluate PSNR using pixels located in predefined masks. For *DTU*, we use the perfect mask provided by IDR [4]. For the other two datasets, we gather the render mask from all methods and take the intersection of them as the predefined mask.

2. Ablation Study on MVS module

We additionally study on how would the MVS depth map quality affect the geometry accuracy of MVSDF. The following two settings are tested on *DTU* dataset: 1) *lowres*: input depth maps are down sampled to 256×192, which represents a MVS module of lower quality and 2) *filtered*: input depth maps are precomputed and geometrically filtered as in [3], which represents a MVS module of higher accuracy. Quantitative results are shown in Tab. 1. We find that both *lowres* and *filtered* settings generates similar results to the proposed setting *full*. In fact, the MVS module is mainly used for recovering the correct initial surface topology, and it could be switched to other MVS algorithms if necessary.

	lowres	filtered	distance only	no render	no feature	full
24	0.83	**0.79**	3.48	1.53	1.02	0.83
37	**1.35**	1.65	5.67	4.38	1.80	1.76
40	1.11	**0.85**	3.73	1.59	1.09	0.88
55	0.46	0.45	2.85	0.68	0.58	**0.44**
63	1.22	**1.05**	3.64	1.99	1.65	1.11
65	1.13	1.06	4.27	2.09	1.18	**0.90**
69	0.84	0.80	2.76	1.15	0.80	**0.75**
83	1.30	1.30	3.95	2.72	1.71	**1.26**
97	1.06	**0.98**	3.14	1.33	1.23	1.02
105	**1.11**	1.15	4.64	3.01	1.31	1.35
106	**0.77**	1.01	3.49	1.25	0.95	0.87
110	0.81	**0.71**	3.40	1.74	0.95	0.84
114	0.35	0.35	1.81	0.53	0.37	**0.34**
118	0.48	0.52	3.14	1.07	0.62	**0.47**
122	0.49	0.53	3.37	1.22	0.65	**0.46**
Mean	0.89	**0.88**	3.56	1.75	1.06	**0.88**

Table 1. Quantitative results of ablation study on DTU dataset. The proposed method could consistently generate high quality reconstructions in spite of the depth map quality of the MVS module.

3. Tanks and Temples Dataset

We additionally conduct experiments on *Francis*, *M60* and *Panther* of the *Tanks and Temples* dataset and report the PSNR scores of Colmap, Vis-MVSNet and the proposed method in Tab. 2. According to the result, our method consistently outperforms other methods on the rendered image quality.

	Colmap	Vis-MVSNet	MVSDF (Ours)
Family	21.50	22.09	**26.11**
Francis	18.25	20.07	**25.58**
Horse	18.62	18.25	**26.43**
M60	17.46	17.41	**20.64**
Panther	19.73	19.99	**23.93**
Mean	19.11	19.56	**24.54**

Table 2. Quantitative results on Tanks and Temples dataset.

4. Additional Qualitative Results

Here we show additional qualitative results on *DTU* (Fig. 1,2), *EPFL* (Fig. 3) and *Tanks and Temples* (Fig. 4) datasets.

References

[1] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

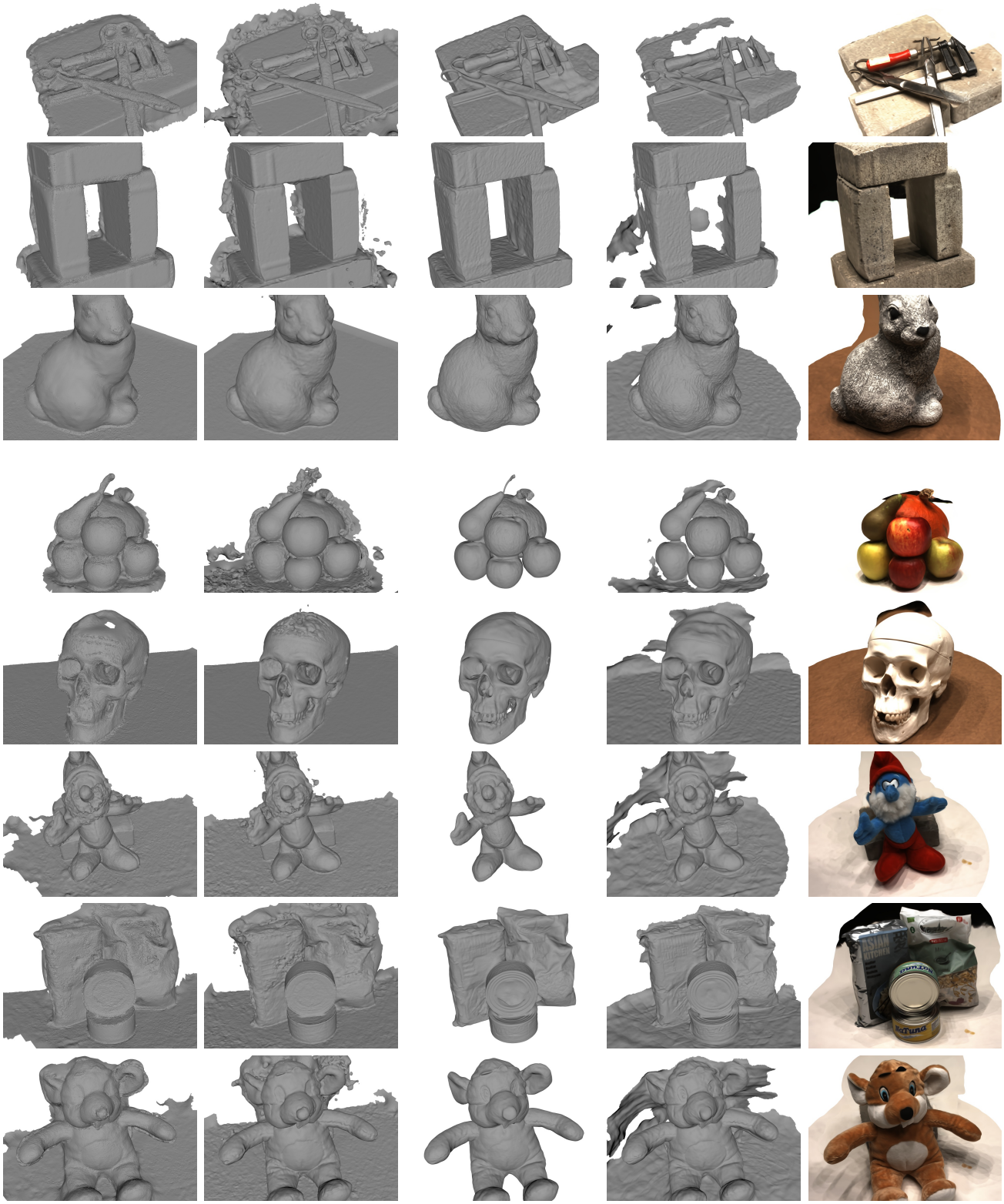
[2] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[3] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2018. 1

[4] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Neural Information Processing Systems (NeurIPS)*, 2020. 1

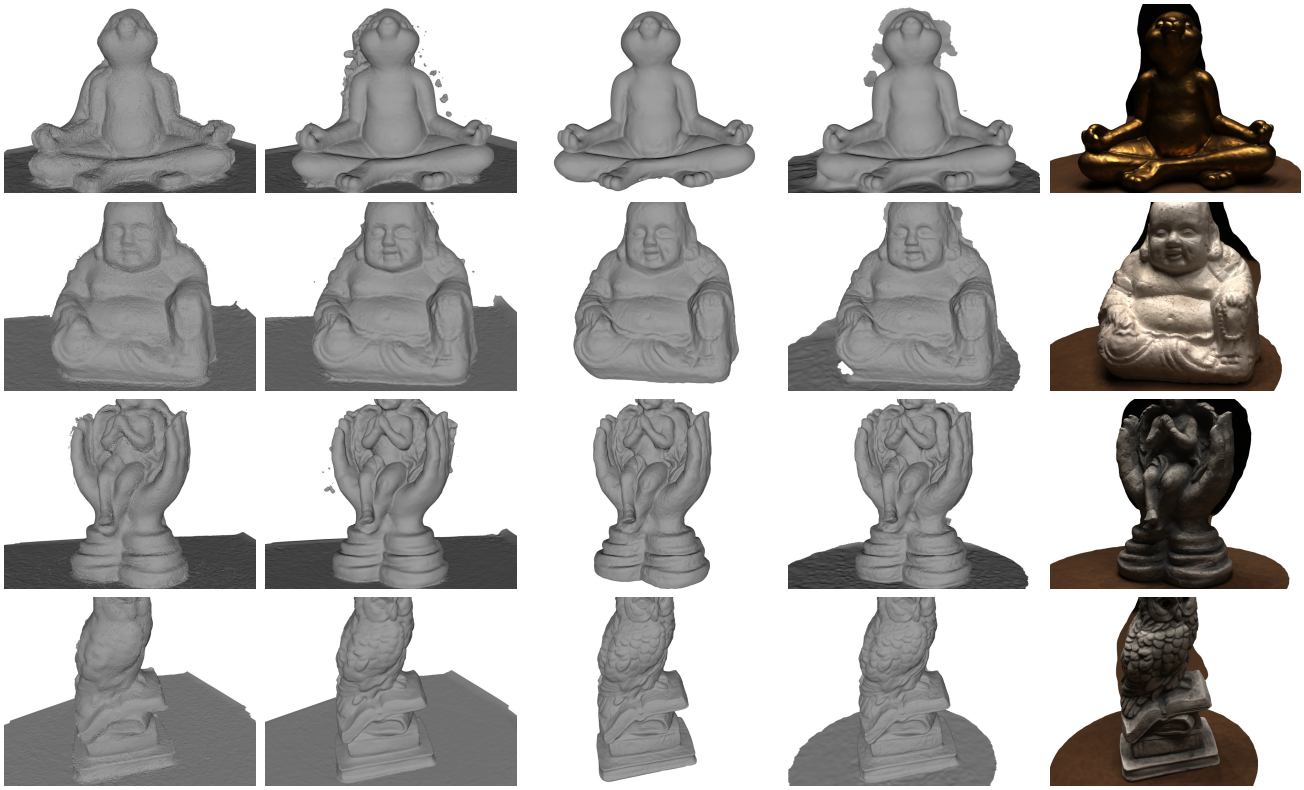
[5] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. In *British Machine Vision Conference (BMVC)*, 2020. 1

[6] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 1



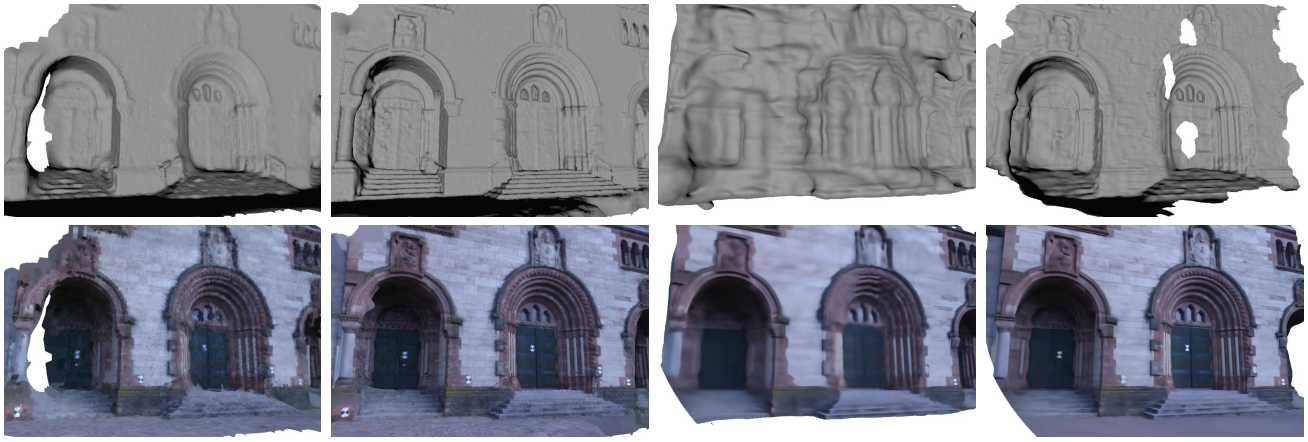
Colmap | Vis-MVSNet | IDR (perfect mask) | MVSDF (Ours) | MVSDF (Ours) Render

Figure 1. Qualitative Results on DTU dataset.



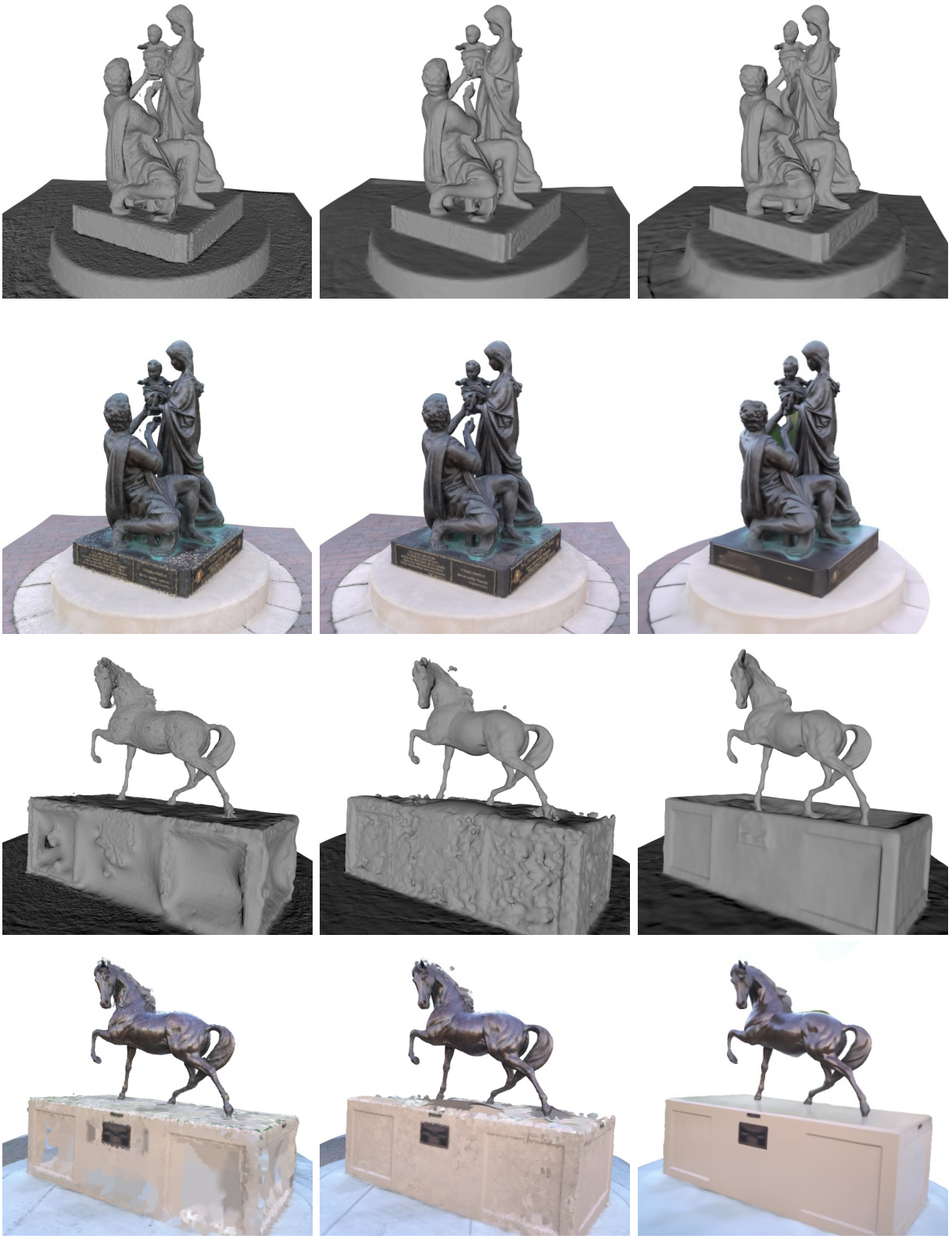
Colmap Vis-MVSNet IDR (perfect mask) MVSDF (Ours) MVSDF (Ours) Render

Figure 2. Qualitative Results on DTU dataset.



Colmap Vis-MVSNet IDR MVSDF (Ours)

Figure 3. Qualitative results on EPFL dataset.



Colmap Vis-MVSNet MVSDF (Ours)

Figure 4. Qualitative results on Tanks and Temples dataset.