PMNI: Pose-free Multi-view Normal Integration for Reflective and Textureless Surface Reconstruction
Supplementary Material

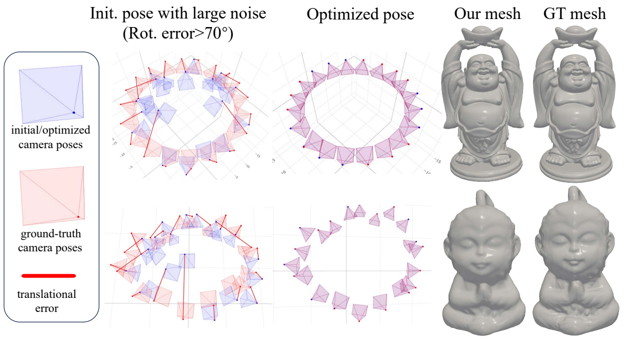
Init. pose with large noise
(Rot. error>70°)
Optimized pose
Our mesh
GT mesh
initial/optimized camera poses
ground-truth camera poses
translational error

Figure 1. Our method is robust to large errors in the initial pose.

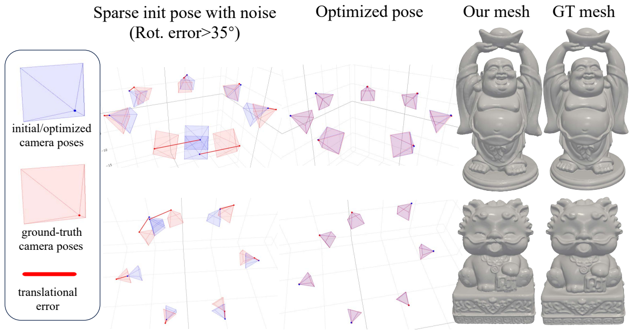
Sparse init pose with noise
(Rot. error>35°)
Optimized pose
Our mesh
GT mesh
initial/optimized camera poses
ground-truth camera poses
translational error

Figure 2. Our method is robust against sparse input with only 7 normal maps.

1. Additional results

This part introduces addtional results of our experiments. The contents can also be found in our supplementary video.

1.1. Robustness against initial pose

In contrast to the conservative pose initialization in the experiments of the main text, this section evaluates the ability of our method under large initialization rotation errors. As illustrated in Fig. 1, the results indicate that even with rotation errors exceeding 70° at initialization, our method successfully recovers the pose and achieves high-quality 3D shape reconstruction. These results underscore the robustness and effectiveness of our method.

1.2. Robustness against sparse inputs

In this part, we test the robustness of our method for pose recovery and shape reconstruction under sparse viewpoints. The results show that even in challenging scenarios with only 7 normal maps as input and initial pose rotation errors exceeding 35°, our method can still accurately recover both the pose and the mesh shape. The result is shown as Fig. 2.

1.3. Robustness against view-independent surface normals

As shown in Fig. 3, we test two special case for our method: ball and cylinder. We render their normal in Blender Software. The initial pose for optimization is set as shown in Fig. 3. The recovered shape is accurate, but the recovered camera pose differs from the ground truth due to view-dependent surface normals as the geometric is symmetric.

1.4. Additional results on our RT3D dataset

As shown in Table 1, We quantitatively evaluate the results of our method on our RT3D dataset. Experimental results show that our method achieves significant advantages in shape reconstruction compared to SuperNormal [1] with slight noise in pose. The results of our pose-free method remains comparable with SuperNormal [1] with ground truth pose.

2. Initialization of pose

In the experimental section of our main text, we determine the radius as described and initialize using a circle as shown in Fig. 4. This pose initial setting is used to test our method on DiLiGenT-MV [2] and our RT3D dataset.

Supplementary video available at: https://youtu.be/hLIZG24m1Wo

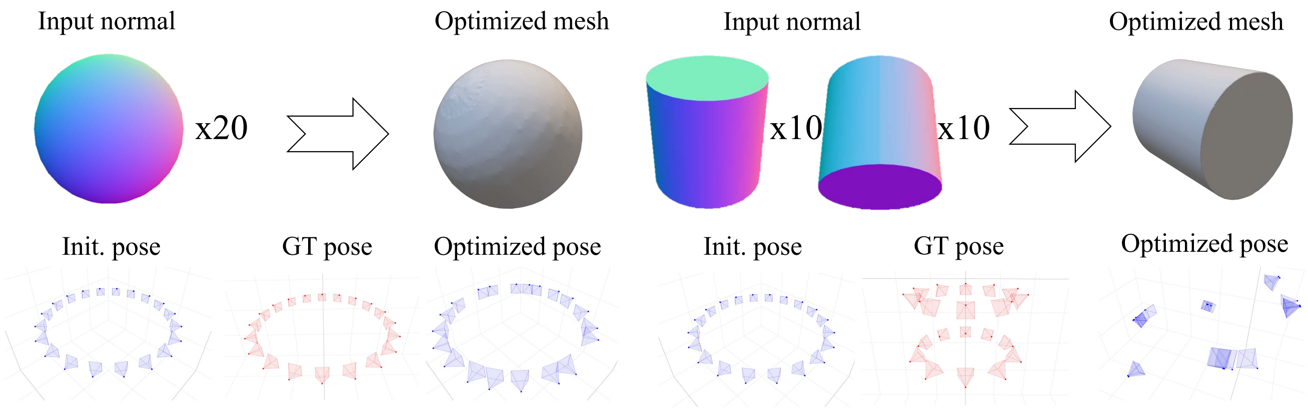

Figure 3. For view-independent surface normals, our method still recovers correct shape.

Table 1. Quantitative evaluation of shape and camera pose recovery on our RT3D dataset. SuperNormal [1] with noisy camera poses is indicated with * marker. The best and second-best results are labeled in **bold** and underlined.

Method	Metric	MONKEY	CAT	PINEAPPLE	DOG	DRAGON	TIGER	Average
SuperNormal [1]	CD ↓	**0.228**	**0.209**	0.189	**0.232**	**0.160**	**0.219**	**0.206**
SuperNormal [1] *		0.671	0.745	0.555	0.739	0.434	0.490	0.606
PMNI (Ours)		0.251	0.318	**0.160**	0.271	0.204	0.245	0.241
SuperNormal [1]	F1-score ↑	**0.974**	**0.956**	0.986	**0.954**	**0.994**	**0.965**	**0.972**
SuperNormal [1] *		0.464	0.473	0.571	0.365	0.690	0.645	0.663
PMNI (Ours)		0.973	0.906	**0.995**	0.922	0.964	0.941	0.950
PMNI	RPEr(°) ↓	0.230	0.356	0.258	0.258	0.439	0.582	0.354
	RPEt ↓	0.011	0.017	0.008	0.010	0.011	0.026	0.014

While this setup may seem special, the results in Subsection 1.1 show that our method remains effective even with large pose rotation errors exceeding 70°, demonstrating the robustness of our method in causal capture.

In fact, our method does not rely on fixed camera tilt angles, heights, or radii, nor does it require uniformly sampled camera positions. Instead, capturing in a single direction (clockwise or counterclockwise) is sufficient. Our method is able to reconstruct high quality 3D model without camera pose in causal capture.

3. Implementation Details

3.1. Network architecture

Fig. 5 shows our neural SDF architecture. $\mathbf{x}$ represents the point in the space, ϕ represents the learnable hash encoding parameters. In a word, our SDF network is a MLP with point $\mathbf{x}$ as input and SDF function value $f(\mathbf{x})$ as output.

3.2. Training details

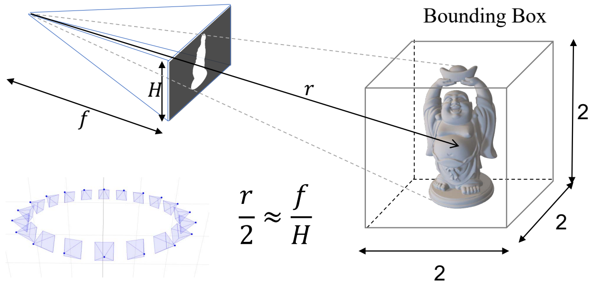

Figure 4. Camera pose initialization.

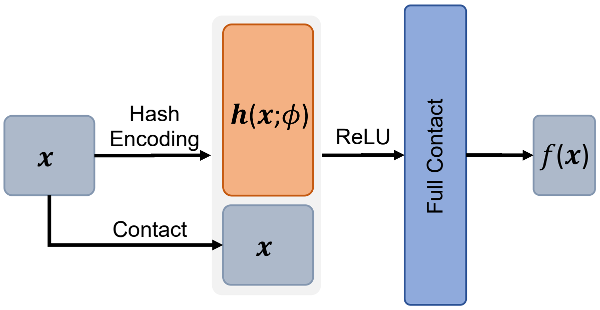

Figure 5. Neural SDF architecture.

We conduct our training on a single NVIDIA 4090 GPU, consisting of a total of 30,000 epochs. In each epoch, 4,096 pixels are sampled from each image. Two Adam optimizers are used, one for the SDF network and the other for the pose parameters. The learning rate for the SDF network is fixed at 5×10^{-4} throughout the training. For the pose parameters, the learning rate is set to 1×10^{-3} for epochs less than 10,000, and 5×10^{-4} for epochs between 10,000 and 30,000. For the weights, during epochs less than 10,000, the weights for $\mathcal{L}_{mask}$ and $\mathcal{L}_{eikonal}$ are set to 1, the weight for $\mathcal{L}_{ni}$ is set to 0.3. As initial pose is totally wrong, the weight for $\mathcal{L}_{normal}$ is 0 in this period. Between epochs 10,000 and 20,000, the weights for $\mathcal{L}_{normal}$, $\mathcal{L}_{mask}$, and $\mathcal{L}_{eikonal}$ terms are set to 1, and the weight for $\mathcal{L}_{ni}$ remains 0.3. Between epochs 20,000 and 30,000, the weights for $\mathcal{L}_{normal}$, $\mathcal{L}_{mask}$, $\mathcal{L}_{eikonal}$ and $\mathcal{L}_c$ are set to 1, while the weight for $\mathcal{L}_{ni}$ is reduced to 0 as it is only a prior not fully reliable. It should be noted that different weight strategies were employed on public dataset and our real-world captured dataset to achieve better performance, with specific implementation details available in our code repository.

References

[1] Xu Cao and Takafumi Taketomi. Supernormal: Neural surface reconstruction via multi-view normal integration. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20581–20590, 2024. 1, 2

[2] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. 29:4159–4173, 2020. 1