NeRSP: Neural 3D Reconstruction for Reflective Objects with Sparse Polarized Images
Supplementary Material

Yufei Han[1†] Heng Guo[1†*] Koki Fukai[2†] Hiroaki Santo[2] Boxin Shi[3,4] Fumio Okura[2]
Zhanyu Ma[1] Yunpeng Jia[1]

[1]Beijing University of Posts and Telecommunications
[2]Graduate School of Information Science and Technology, Osaka University
[3]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[4]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

{hanyufei, guoheng, mazhanyu}@bupt.edu.cn shiboxin@pku.edu.cn
{santo.hiroaki, okura, fukai.koki}@ist.osaka-u.ac.jp xibei156@163.com

A. Photometric and geometric cues of NeRSP

A.1. Derivation of geometric cue

As shown in Fig. S1, given a scene point observed by different views, its surface normal at the target view can be represented by the azimuth and elevation angles ϕ and θ respectively, *i.e.*,

$$\mathbf{n} = \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} = \begin{bmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{bmatrix}. \tag{1}$$

The relationship between the azimuth angle and the element of the surface normal can be formulated as

$$n_y \cos\phi - n_x \sin\phi = 0. \tag{2}$$

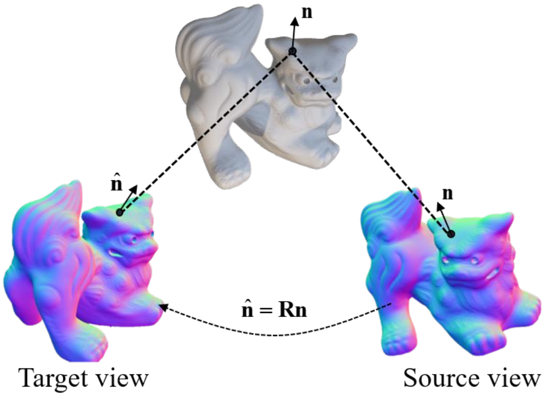

Figure S1. A scene point observed by the target view and the source view.

The surface normal at the target view can be calculated by rotating the normal at the source view, *i.e.* $\hat{\mathbf{n}} = \mathbf{R}\mathbf{n}$. Given the rotation matrix from the calibrated camera poses as $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3]^\top$, Eq. (2) based on $\hat{\mathbf{n}}$ can be formulated as

$$\mathbf{r}_1^\top \mathbf{n} \cos\phi - \mathbf{r}_2^\top \mathbf{n} \sin\phi = 0. \tag{3}$$

Following MVAS [2], we can rearrange Eq. (3) to get the orthogonal relationship between the surface normal and the projected tangent vector $\mathbf{t}(\phi)$ as defined below,

$$\mathbf{n}^\top \underbrace{(\cos\phi\mathbf{r}_1 - \sin\phi\mathbf{r}_2)}_{\mathbf{t}(\phi)} = 0. \tag{4}$$

This conclusion on azimuth angle can be extend to the an-

[†] Equal contribution. [*] Corresponding author.

gle of polarization (AoP). The π ambiguity can be naturally resolved as Eq. (4) stands if we add ϕ by π. The $\pi/2$ ambiguity can be addressed by using a pseudo-projected tangent vector $\hat{\mathbf{t}}(\phi)$ such that

$$\mathbf{n}^\top \underbrace{(\sin\phi\mathbf{r}_1 + \cos\phi\mathbf{r}_2)}_{\hat{\mathbf{t}}(\phi)} = 0. \tag{5}$$

If one scene point $\mathbf{x}$ is observed by f views, we can stack Eq. (4) and Eq. (5) based on different rotations and observed AoPs, leading to a linear system

$$\mathbf{T}(\mathbf{x})\mathbf{n}(\mathbf{x}) = \mathbf{0}. \tag{6}$$

We treat this linear system as our geometric cue for multi-view polarized 3D reconstruction.

A.2. Derivation of photometric cue

Following the polarized BRDF model [1], the output stokes vector can be decomposed into the diffuse and specular parts modeled via $\mathbf{H}_d$ and $\mathbf{H}_s$ correspondingly, *i.e.*,

$$\mathbf{s}_o(\mathbf{v}) = \int_\Omega \mathbf{H}_d\mathbf{s}_i(\boldsymbol{\omega})\ d\boldsymbol{\omega} + \int_\Omega \mathbf{H}_s\mathbf{s}_i(\boldsymbol{\omega})\ d\boldsymbol{\omega}. \tag{7}$$

The diffuse stokes component under a single light can be formulated as

$$\mathbf{H}_d\mathbf{s}_i(\boldsymbol{\omega}) = \rho_d L(\boldsymbol{\omega})\boldsymbol{\omega}^\top\mathbf{n}T_i^+T_i^- \begin{bmatrix} T_o^+ \\ T_o^- \cos(2\phi_n) \\ -T_o^- \sin(2\phi_n) \\ 0 \end{bmatrix}, \tag{8}$$

where ρ_d denotes the diffuse albedo, ϕ_n is the azimuth angle of incident light onto the plane perpendicular to the surface normal, $T_{i,o}^+$ and $T_{i,o}^-$ denote the calculations of Fresnel transmission coefficients [1] that are related to the angle between view direction and surface normal. Following the notions in PANDORA [3], we rewrite the diffuse stokes vector under environment light as

$$\int_\Omega \mathbf{H}_d\mathbf{s}_i(\boldsymbol{\omega})\ d\boldsymbol{\omega} = L_d \begin{bmatrix} T_o^+ \\ T_o^- \cos(2\phi_n) \\ -T_o^- \sin(2\phi_n) \\ 0 \end{bmatrix}, \tag{9}$$

where $L_d = \int_\Omega \rho L(\boldsymbol{\omega})\boldsymbol{\omega}^\top\mathbf{n}T_i^+T_i^-\ d\boldsymbol{\omega}$ is denoted as diffuse radiance. Instead of calculating from the equation, the diffuse radiance as a spatially-varying variable is mapped directly from a neural point feature extracted by a coordinate-based MLP.

On the other hand, the specular stokes vector under a single light direction $\boldsymbol{\omega}$ in the polarimetric BRDF model can be defined as

$$\mathbf{H}_s\mathbf{s}_i(\boldsymbol{\omega}) = \rho_s L(\boldsymbol{\omega})\frac{DG}{4\mathbf{n}^\top\mathbf{v}} \begin{bmatrix} R^+ \\ R^- \cos(2\phi_h) \\ -R^- \sin(2\phi_h) \\ 0 \end{bmatrix}, \tag{10}$$

where ρ_s denotes the specular albedo; D and G denote the normal distribution and shadowing term in the Microfacet model [8], which can be controlled by surface roughness; R^+ and R^- denote the calculations of the Fresnel reflection coefficients [1], which are related to the angle between surface normal and incident light direction; ϕ_h is the incident azimuth angle w.r.t. the half vector $\mathbf{h} = \frac{\boldsymbol{\omega}+\mathbf{v}}{\|\boldsymbol{\omega}+\mathbf{v}\|_2^2}$. Following the notions in PANDORA [3], we rewrite the specular stokes vector under environment light as

$$\int_\Omega \mathbf{H}_s\mathbf{s}_i(\boldsymbol{\omega})\ d\boldsymbol{\omega} = L_s \begin{bmatrix} R^+ \\ R^- \cos(2\phi_h) \\ -R^- \sin(2\phi_h) \\ 0 \end{bmatrix}, \tag{11}$$

where $L_s = \rho_s \int_\Omega L(\boldsymbol{\omega})\frac{DG}{4\mathbf{n}^\top\mathbf{v}}\ d\boldsymbol{\omega}$ denotes specular radiance. With the spilt-sum approximation [5], we can further approximate $L_s \approx \frac{\rho_s DG}{4\mathbf{n}^\top\mathbf{v}}\int_\Omega L(\boldsymbol{\omega})\ d\boldsymbol{\omega}$. Combining with the diffuse stokes vector shown in Eq. (9), we build the photometric cue based on the following polarimetric image formation model

$$\mathbf{s}_o(\mathbf{v}) = L_d \begin{bmatrix} T_o^+ \\ T_o^- \cos(2\phi_n) \\ -T_o^- \sin(2\phi_n) \\ 0 \end{bmatrix} + L_s \begin{bmatrix} R^+ \\ R^- \cos(2\phi_h) \\ -R^- \sin(2\phi_h) \\ 0 \end{bmatrix}. \tag{12}$$

B. Implementation details

This section presents the rendering details of our Synthetic Multi-view Polarized image dataset SMVP3D, and the training details of NeRSP.

B.1. Dataset

We provide SMVP3D, which contains images of five synthetic reflective objects under natural illumination. For each object, we render 48 views and record the corresponding ground truth (GT) surface normal maps. We use Mitsuba3 [4] as the rendering engine, with the BRDF type set to polarized plastic material in our rendering. For the diffuse albedo ρ_d, we utilize a spatially varying albedo texture to enhance the realism of our rendering results. At the same time, we keep the specular albedo ρ_s at a constant value of 1.0 and set the surface roughness to 0.05. This approach ensures uniform reflectivity across the surfaces of the objects. The resulting polarized images are rendered at a resolution of 512 × 512 pixels.

B.2. Training

The hyperparameters λ_g, λ_m, and λ_e in our loss function are set to 1, 1, and 0.1, respectively. During the training process, we employ a warm-up strategy following PANDORA [3], where for the first $1,000$ epochs, we consider only unpolarized information in the photometric cue and

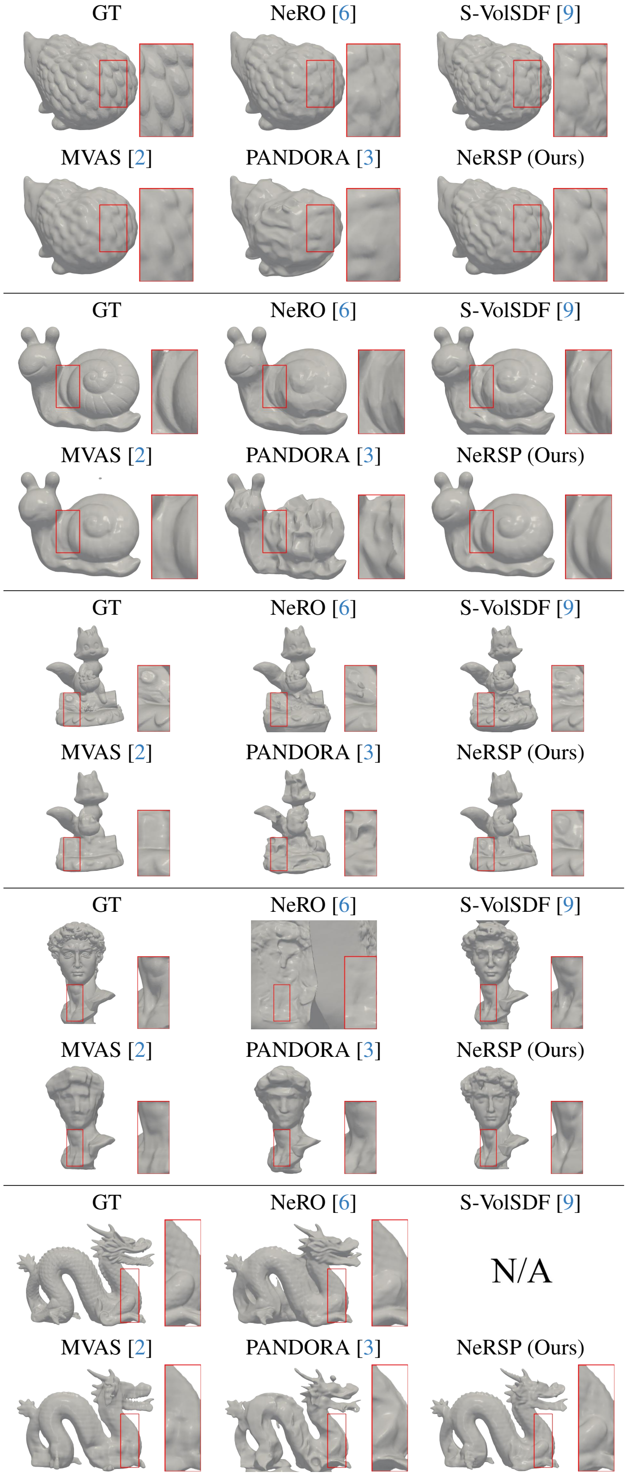

Figure S2. Qualitative evaluation of shape reconstruction on SMVP3D.

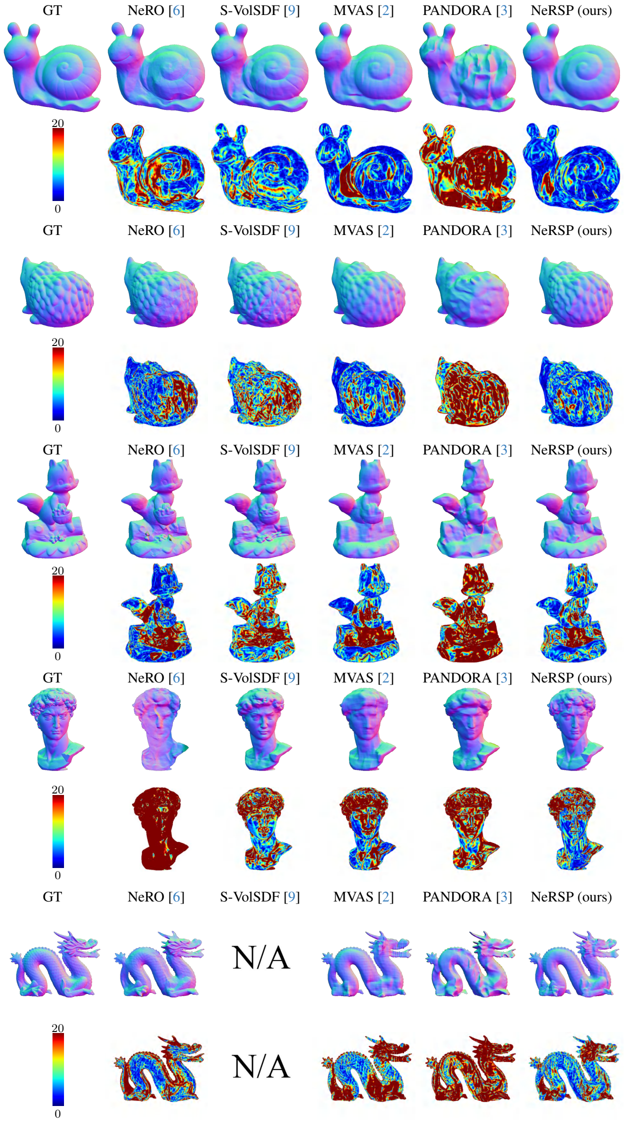

Figure S3. Qualitative evaluation of surface normal estimation on our SMVP3D. Even and odd rows show the surface normal estimates and the corresponding angular error distributions.

assume that the object's specular component is 0. In all experiments, we use a resolution of 512×512 for training and testing on SMVP3D, and 512×612 for real-world datasets. Our method generally converges around $100,000$ epochs, which takes about 6 hours on an Nvidia RTX 3090 GPU, with the memory consuming around $8,000$ MB.

C. BRDF estimation and re-rendering results

Figure S4 (top) presents our estimation of roughness, diffuse, and specular components. The estimates are a bit noisy due to only 6 views. Similar to Ref-NeRF [7] where illumination is implicitly controlled via IDE, we cannot conduct relighting experiments. Therefore, we show the novel view synthesis results instead, as visualized in Fig. S4 (bottom). Compared with existing methods, our re-rendering images are closer to the corresponding real-world observations.

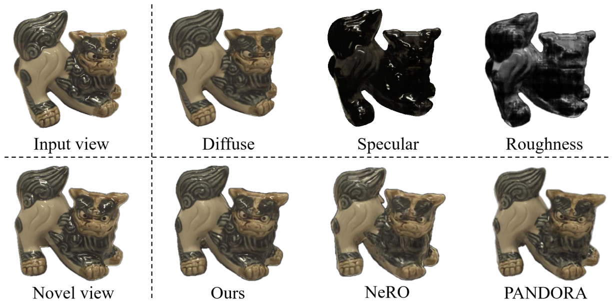

Figure S4. (Top) Estimated BRDF from our method. (Bottom) Comparison of novel view synthesis.

D. Additional results on our datasets

In this section, we present additional results of shape reconstruction on SMVP3D and Real-world Multi-view Polarized image dataset RMVP3D.

D.1. Evaluation on SMVP3D

We present the qualitative reconstruction results of baseline methods and our approach in Fig. S2. The results from MVAS [2] lack detail, as the photometric cue is not taken into account. While NeRO [6] offers improved shape reconstructions, it fails to provide a reliable surface for textureless objects, such as DAVID. S-VolSDF [9] uses a coarse-to-fine Multi-View Stereo (MVS) approach and shows increased sensitivity to texture information on object surfaces, which sometimes leads to misinterpreting texture details as structural features. PANDORA [3] has difficulty in effectively separating albedo and specular information, leading to unreliable reconstruction results. Our method, NeRSP, effectively utilizes both photometric and geometric cues, resulting in reconstructions that more accurately reflect the GT structure.

We also display the surface normal estimates and the corresponding angular error distributions in Fig. S3, which consistently show that NeRSP achieves better shape reconstruction results for reflective surfaces with sparse input views.

D.2. Evaluation on RMVP3D

In this section, we present another object reconstruction result on RMVP3D. Figure S5 shows that NeRO [6], MVAS [2] and NeRSP are able to accurately reconstruct a simple spherical object with a reflective surface. In contrast, S-VolSDF [9] and PANDORA [3] can not decomposing the albedo and specular component of the surface, resulting in distortion in the shape reconstruction process.

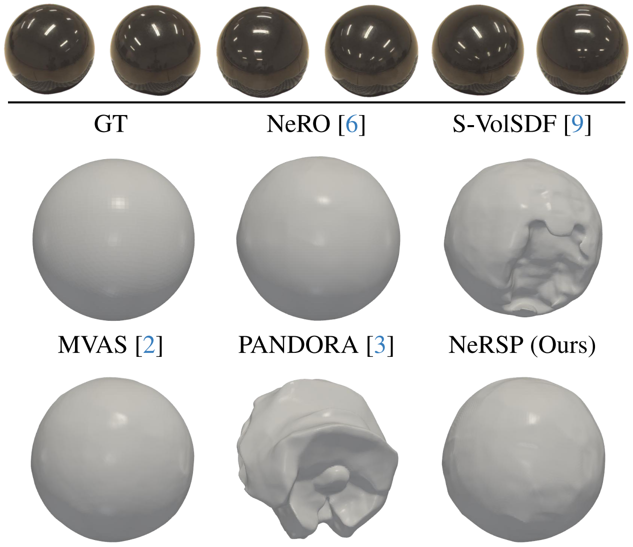

Figure S5. Qualitative evaluation of shape reconstruction BALL.

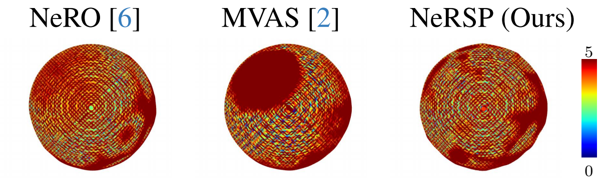

Figure S6. The Chamfer Distance maps clipped between 0 and $5\,mm$ for the estimated shapes of BALL from NeRO [6], MVAS [2], and NeRSP.

To distinguish among the reconstruction results of NeRO [6], MVAS [2], and NeRSP, we visualize the Chamfer Distance for the meshes reconstructed by each method. As shown in Fig. S6, the color of each point indicates its Chamfer Distance, which is clipped between 0 and $5\,mm$. These illustrations show that the reconstruction error associated with NeRSP is smaller compared to that of the other two methods.

E. Ablation study on surface reflectance

Our method aims at reflective surface reconstruction, and it can also be applied to recovering the shape with rough surfaces. As an example, we re-render the SNAIL object with its specular albedo ρ_s reducing from 1.0 to 0.1. The mean angular error (MAE) of the estimated surface normal at 6 input views from different methods are shown in Table S1. The qualitative evaluation of the surface normal estimation and the corresponding angular error distribution of different methods under the same input view are shown in Fig. S7. These experiments indicate that most methods improve reconstruction quality on rough surfaces compared to reflective surfaces. In particular, our method consistently delivers the most reliable surface reconstruction of the object.

F. Ablation study on #views

Our NeRSP aims at the reconstruction of reflective surfaces under sparse input views. The experiments shown in the

Table S1. Comparison on surface normal estimation on Snail evaluated by mean angular error (MAE) (↓).

Reflectance type	NeRO [6]	S-VolSDF [9]	MVAS [2]	PANDORA [3]	NeRSP
Reflective	11.45	7.59	6.19	16.54	**4.82**
Rough	5.94	8.12	5.75	8.63	**4.18**

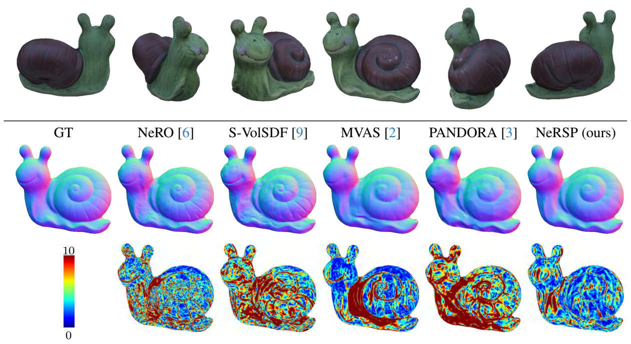

Figure S7. Qualitative evaluation of surface normal estimation on Snail with less reflective reflectance. Top row shows the 6 input views. Second and third rows show the surface normal estimates and the corresponding angular error distributions.

Table S2. Qualitative evaluation on Lion measured by Chamfer Distance (↓) under different input views.

#Views	NeRO [6]	S-VolSDF [9]	MVAS [2]	PANDORA [3]	NeRSP
3	34.48	31.50	**23.96**	24.44	24.01
6	10.74	7.39	7.51	15.04	**5.18**
12	5.50	6.80	5.31	12.1	**4.29**
24	4.96	6.14	5.32	12.5	**4.11**

main paper take 6 sparse views as input. To evaluate our method under the different numbers of input views (*i.e.*, #views), we conduct experiments on the real-world object Lion under the setting of 3, 6, 12, and 24 views. Figure S8 visualizes the recovered shapes, while the qualitative evaluation with Chamfer Distance is presented in Table S2.

Under sparse input views, such as 3, existing methods struggle to recover plausible results. This is mainly because they focus either on photometric cue or geometric cue. Taking S-VolSDF [9] as an example, the estimated shape, as observed in close-up views, is heavily influenced by the corresponding texture. This leads to incorrect shapes due to the shape-radiance ambiguity under sparse views. By addressing both the geometric and the photometric cues, our NeRSP reduces the ambiguity under sparse inputs. As a result, we achieve more reasonable shape reconstruction.

This observation remains valid when the number of input views exceeds 12. As shown in Table S2, our NeRSP consistently achieves the smallest Chamfer Distance with an increasing number of input views. This shows the effectiveness of our method on reflective surfaces over a wide range of views.

G. Evaluation on polarimetric MVIR dataset

Besides the real-world experiments on PANDORA dataset [3] and our RMVP3D, we also provide the evaluation on a multi-view polarized images dataset present in PMVIR [10]. As shown in Fig. S9, we visualize the shape recovery results from PANDORA [3] and ours, taking 6 sparse views as input. Since there is no GT shape in this dataset, we use the results from PMVIR [10] as a reference, which takes 31 and 56 views as input for the camera and the car scene, respectively. We observe that our results are more reasonable compared to those using PANDORA [3], demonstrating the effectiveness of our method on sparse 3D reconstruction.

References

[1] Seung-Hwan Baek, Daniel S Jeon, Xin Tong, and Min H Kim. Simultaneous acquisition of polarimetric SVBRDF and normals. *ACM TOG*, 37(6):268–1, 2018. 2

[2] Xu Cao, Hiroaki Santo, Fumio Okura, and Yasuyuki Matsushita. Multi-View Azimuth Stereo via Tangent Space Consistency. In *CVPR*, pages 825–834, 2023. 1, 3, 4, 5, 6

[3] Akshat Dave, Yongyi Zhao, and Ashok Veeraraghavan. Pandora: Polarization-aided neural decomposition of radiance. In *ECCV*, pages 538–556, 2022. 2, 3, 4, 5, 6

[4] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org. 2

[5] Brian Karis and Epic Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3):1, 2013. 2

[6] Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. NeRO: Neural Geometry and BRDF Reconstruction of Reflective Objects from Multiview Images. *arXiv preprint arXiv:2305.17398*, 2023. 3, 4, 5, 6

[7] 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 *CVPR*, pages 5481–5490, 2022. 3

[8] Bruce Walter, Stephen R Marschner, Hongsong Li, and Kenneth E Torrance. Microfacet models for refraction through rough surfaces. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 195–206, 2007. 2

[9] Haoyu Wu, Alexandros Graikos, and Dimitris Samaras. S-VolSDF: Sparse Multi-View Stereo Regularization of Neural Implicit Surfaces. *arXiv preprint arXiv:2303.17712*, 2023. 3, 4, 5, 6

[10] Jinyu Zhao, Yusuke Monno, and Masatoshi Okutomi. Polarimetric multi-view inverse rendering. *IEEE TPAMI*, 2022. 5, 6

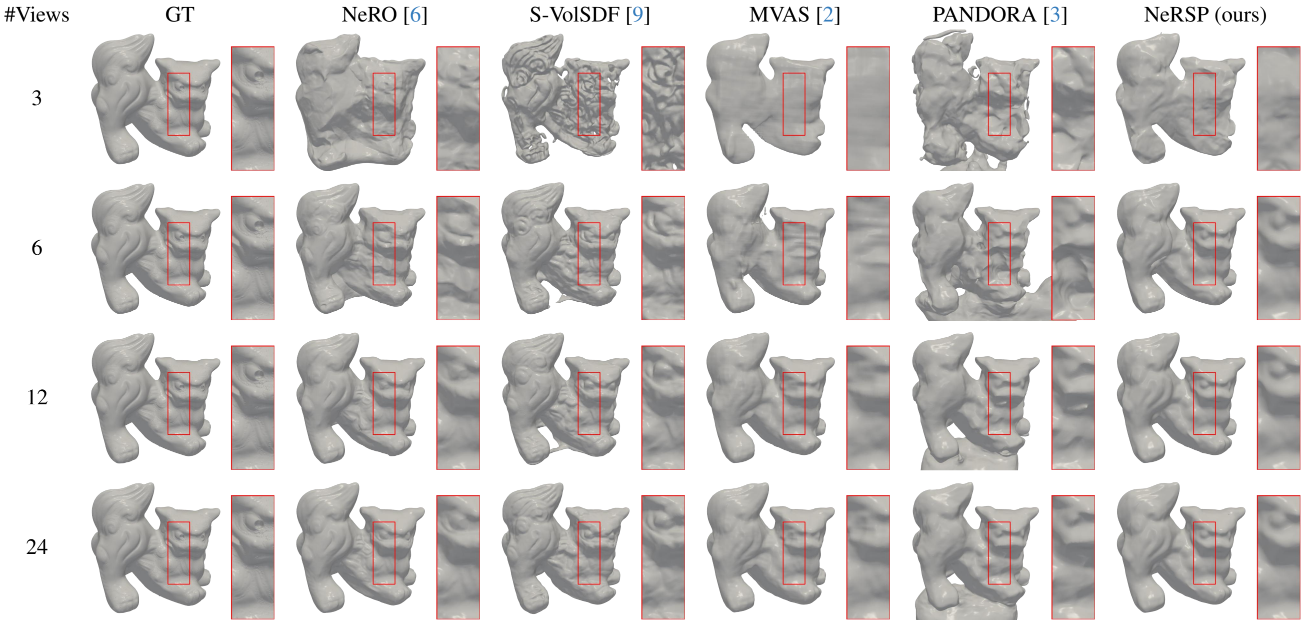

Figure S8. Qualitative results of shape reconstruction on LION with different input #views.

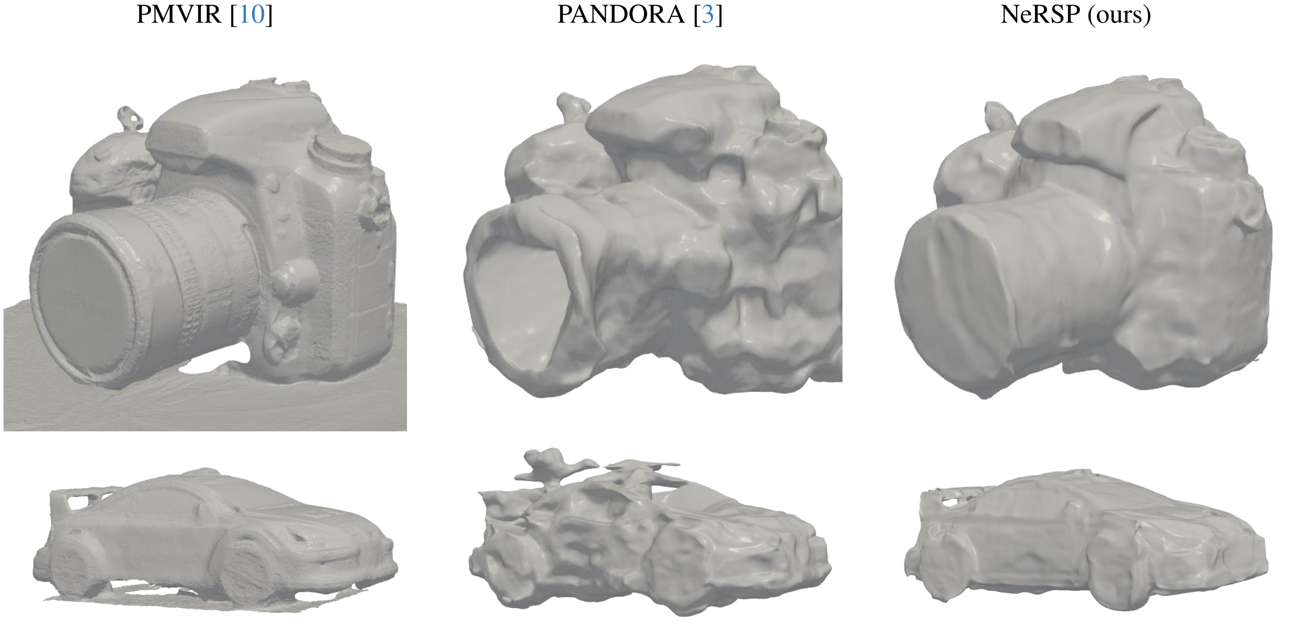

Figure S9. Shape estimation results on the Polarimetric MVIR dataset [10]. PANDORA [3] and our NeRSP use polarized images with 6 sparse views as input. As a reference, PMVIR [10] uses 31 and 56 input views on the camera and the car cases, respectively.